Convolutional neural network-based approach for segmentation of left ventricle myocardial scar from 3D late gadolinium enhancement MR images

Fatemeh Zabihollahy[a)]
Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada

James A. White
Stephenson Cardiac Imaging Centre, Libin Cardiovascular Institute of Alberta University of Calgary, Calgary, AB, USA

Eranga Ukwatta
Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada

Purpose: Accurate three-dimensional (3D) segmentation of myocardial replacement fibrosis (i.e., scar) is emerging as a potentially valuable tool for risk stratification and procedural planning in patients with ischemic cardiomyopathy. The main purpose of this study was to develop a semiauto-mated method using a 3D convolutional neural network (CNN)-based for the segmentation of left ventricle (LV) myocardial scar from 3D late gadolinium enhancement magnetic resonance (LGE-MR) images.

Methods: Our proposed CNN is built upon several convolutional and pooling layers aimed at choosing appropriate features from LGE-MR images to distinguish between myocardial scar and healthy tissues of the left ventricle. In contrast to previous methods that consider image intensity as the sole feature, CNN-based algorithms have the potential to improve the accuracy of scar segmentation through the creation of unconventional features that separate scar from normal myocardium in the feature space. The first step of our pipeline was to manually delineate the left ventricular myocar-dium, which was used as the region of interest for scar segmentation. Our developed algorithm was trained using 265,220 volume patches extracted from ten 3D LGE-MR images, then was validated on 450,454 patches from a testing dataset of 24 3D LGE-MR images, all obtained from patients with chronic myocardial infarction. We evaluated our method in the context of several alternative methods by comparing algorithm-generated segmentations to manual delineations performed by experts.

Results: Our CNN-based method reported an average Dice similarity coefficient (DSC) and Jaccard Index (JI) of 93.63% ± 2.6% and 88.13% ± 4.70%. In comparison to several previous methods, including K-nearest neighbor (KNN), hierarchical max flow (HMF), full width at half maximum (FWHM), and signal threshold to reference mean (STRM), the developed algorithm reported significantly higher accuracy for DSC with a *P*-value less than 0.0001.

Conclusions: Our experimental results demonstrated that our CNN-based proposed method yielded the highest accuracy of all contemporary LV myocardial scar segmentation methodologies, inclusive of the most widely used signal intensity-based methods, such as FWHM and STRM. To our knowl-edge, this is the first description of LV myocardial scar tissue segmentation from 3D LGE-MR images using a CNN-based method. © *2019 American Association of Physicists in Medicine* [https://doi.org/10.1002/mp.13436]

1. INTRODUCTION

Replacement myocardial fibrosis or "scar" is a recognized substrate for the development of malignant ventricular arrhythmias and is a dominant cause of sudden cardiac death in the population.[1] The most common mechanism of scar for-mation in the heart is from myocardial infarction (MI), an irreversible death of muscle cells related to blockage of a coronary artery. Following this injury, patients are at elevated subsequent risk for re-entrant-type ventricular arrhythmias due to myocardial scar[2]; a risk that may be managed through device therapy (implantable cardiac defibrillator) or catheter ablation of scar-related re-entrant circuits. Patient selection and delivery of such therapies may be improved through indi-vidualized, noninvasive modeling of this arrhythmia sub-strate.[3] In particular, computational modeling of hearts with ischemic cardiomyopathy has emerged as a promising tool to guide patient-specific diagnosis and the treatment of associ-ated rhythm disorders.[4,5] To accurately represent patient-spe-cific structural remodeling, computational models must incorporate accurate reconstructions of left ventricular (LV) scar geometry.[6–8]

Two-dimensional late gadolinium enhancement magnetic resonance (2D LGE-MR) imaging has been established as

the reference standard imaging tool for the identification and quantification of myocardial scar in clinical practice.[9] Using 2D LGE-MR techniques, each image slice is acquired sequentially and the voxel size is anisotropic and large in the long axis direction. However, computational models of the heart need to incorporate high-resolution reconstructions of scar geometry to resolve the electrical activation front.[10] Due to recent advancements in imaging physics, 3D LGE-MRI has emerged as an alternative which provides isotropic and higher spatial resolution, enabling more accurate spatial reconstruction of scar geometry.[11,12] Moreover, free-breathing 3D whole-heart LGE imaging technique may reduce the image acquisition time in comparison with the breath holds 2D imaging.[13] However, segmentation of scar from 3D LGE-MR images is a challenging and tedious task, due to the large size of the image dataset. Therefore, automated image analysis techniques are critical to migrate this into clinical practice.

Myocardial scar may form in both left ventricles and atria of the heart. However as compared to segmentation of scar in LV, segmentations of scar in left atrium (LA) is a more challenging task due to multiple factors including the thinner walls of the LA wall, more complicated geometry, contrast variation due to inversion time, signal-to-noise ratio, motion blurring, and artefacts.[14] In this study, our focus is on determining the boundary of myocardial scar in LV from LGE-MRI. While alternative noncontrast techniques for the detection and spatial modeling of scar are being explored, such as the use of cine cardiac MR images to provide surrogates of scar distribution using feature learning,[15] these techniques for LGE-MRI require further development and validation. Therefore, 3D LGE-based segmentation remains a leading focus for the guidance of procedures where scar extent and its volumetric distribution are of primary interest.

Most of the previous studies developed for semiautomated segmentation of scar from presegmented myocardium in 2D or 3D LGE-MR images were based on image intensity thresholding-based methods, including full width at half maximum (FWHM), signal threshold to reference mean (STRM), and region growing (RG). The FWHM method considers myocardial scar as the reference tissue region, defining all tissue that exceeds half of its maximal signal intensity as scar.[16] In the STRM approach, the normal (i.e., remote) myocardium is considered the reference tissue with scar defined according to its mean value plus two (STRM2), three (STRM3), four (STRM4), five (STRM5), or six (STRM6) standard deviations.[17] The RG method is a greedy algorithm, where some initial seeds are chosen by the user and their neighboring pixels are examined for eligibility to be added to the region.[18] These types of scar segmentation methods, based solely on image intensity thresholding, are highly influenced by image noise.[19] Furthermore, these methods are subject to high interoperator variability due to the expert input required to choose the maximum intensity, remote myocardium and initial seed point for FWHM, STRM, and RG methods, respectively. There are also several energy minimization methods that have been developed for scar

segmentation. Ukwatta et al.[20] proposed a method for scar segmentation as a continuous min-cut optimization problem and solved it using the dual formulation of the problem, called the continuous max flow. Y. Lu et al. proposed a graph cuts-based method for myocardial infarct segmentation.[19] Both these methods use image intensity-based criteria as the data term in the segmentation. Rajchl et al.[21] described an interactive method based on hierarchical max flow (HMF) for segmenting scar in LV myocardium, which used intensity log-likelihood criteria as the data term. Although the method does not need manually presegmented myocardial borders, it required about 7 min of expert time to interactively segment LV myocardial scar,[22] which is subject to high-operator variability.

In this work, we propose a semiautomated method based on convolutional neural network (CNN) to segment LV myocardial scar from 3D LGE-MR images. Using our proposed CNN approach features are automatically identified that best contribute to the successful completion of a segmentation task using associations that may not be intuitive to humans. To our knowledge, this is the first application of CNN to the segmentation of scar in 3D LGE-MR images. Preliminary results of this work have been published in a previous conference paper.[23] Our approach is substantially different from this previously published work as it includes the use of a 3D CNN approach, a substantially larger dataset and provides a comparison to alternative methods. Previous studies by other investigators have shown that CNN-based methods outperform conventional image analysis techniques for segmentation problems across a variety of organs, such as brain tumours,[24,25] other brain structures,[26] liver lesions,[27] pulmonary nodules,[28] and multiorgan segmentation applied to the normal whole body (MRI).[29] We, therefore, anticipate expansion of this paradigm across a variety of cardiovascular applications.

2. MATERIALS AND METHODS

In this research work, we performed a binary classification with two possible labels including scar and healthy tissue. We used a 3D CNN-based approach to segment LV myocardial scar from 3D LGE-MR images. Figure 1 shows an overview of the developed approach. For each 3D LGE-MR image, the myocardial boundary of the LV and the scar were manually segmented separately by an expert using a 3D brush in a multiplanar view in ITK-SNAP Software.[30] The segmentation of myocardium requires about 54 min of expert time, while segmentation of scar needs 42 min.[21] The presegmented myocardium was considered as a region of interest and the myocardial scar was segmented from this region. The dataset was then divided into a training set (N = 10) and a testing set (N = 24). We have varied the number of training images and experimentally observed that the performance of the algorithm did not substantially change after 10 training datasets. This way we did not further increase the size of the training dataset, as we also aimed to maximize the number of testing images from the entire dataset to make our validation
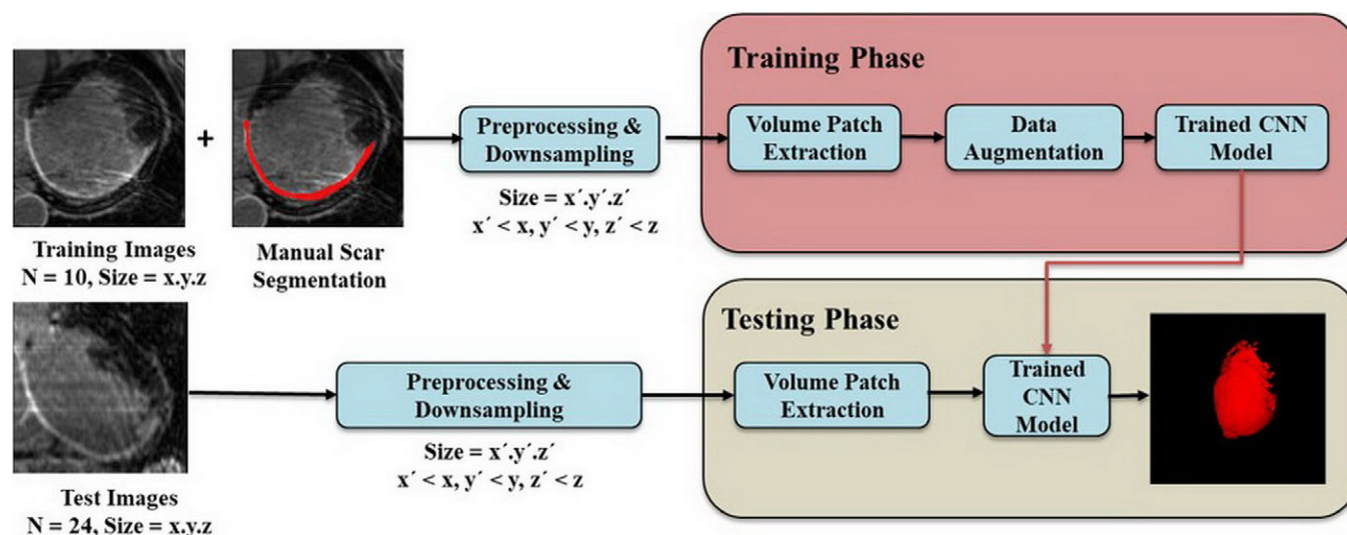
FIG. 1. Block diagram of the proposed method for left ventricular myocardial scar segmentation, where major steps of training and test phases are schemed. [Color figure can be viewed at wileyonlinelibrary.com]

of the algorithm more robust. The images were then normalized and denoised to enhance the contrast of the image. During the supervised training phase, volume patches of the training set were up-sampled until the number of samples in both classes became equal. Then the training samples were introduced to the CNN to learn the image characteristics of scarred versus nonscarred myocardium. We reserved 20% of the training datasets (2 cases) for validation purposes. Extracted patches from each image were passed to the training model to be labeled as scar or healthy tissue. A segmentation map was then created from the predicted labels of each image and compared with the corresponding expert manual segmentation for network parameter adjustment. During the testing phase, the trained CNN model was applied to the unobserved volume patches extracted from the test set to be labeled as scar or healthy tissue. Thus, the input and output of the trained model was a volume patch extracted from a 3D LGE-MR images and a label, scar or healthy tissue, respectively. A segmentation map was then created from those labels and compared with the corresponding manual segmentation to evaluate the accuracy of the proposed method for myocardial scar segmentation.

2.A. Study subjects and image acquisition

Our study consisted of 34 subjects selected from patients with known chronic ischemic cardiomyopathy. By definition, these patients were a minimum of 3 months from any clinically recognized acute coronary syndrome. Therefore, all myocardial scar visualized by LGE imaging was reasonably assumed to be "mature" (i.e., healed). All subjects provided written informed consent and the study protocol was approved by the Research Ethics Board at Western University. All subjects underwent a standardized imaging protocol inclusive of cine imaging, contrast-enhanced coronary MRI and post-contrast 3D LGE at Robarts Research Institute (ON,

Canada). The latter two image datasets were acquired using whole-heart, respiratory navigated, 3-D inversion-recovery gradient echo pulse sequence performed during and 30 min following infusion of 0.2 mmol/kg Gadovist (Bayer, Toronto, ON, Canada), as previously described.[31] The voxel size of the acquired images was 0.625 mm × 0.625 mm × 1.3 mm. Table I describes imaging hardware and pulse sequence parameters.

Characteristics of the 34 patients, all completing the imaging protocol, are shown in Table II. The mean age of the population was 51.5 ± 12.6 years with the mean heart rate and left ventricle ejection fraction being 67.1 ± 11.2 beats per minute and $32.1\% \pm 12.7\%$, respectively. The mean total scar volume for the entire population, calculated from manual 3D segmentation, was 25.57 ± 16.45 ml (range 2.36–59.87 ml).

2.B. Volume patch extraction and augmentation

We first applied contrast stretching to the images by normalizing the range of voxel intensity values between 0 and 255. A median filter of size 3×3 was then used to remove noise in the image. Finally, the 3D images were downsampled in all three directions by a factor of two to reduce image size, because creating volume patch around each voxel in the original image required more than 5 h of computational time. To investigate the effects of the downsampling on the segmentation accuracy, the CNN was trained using a small subset of training dataset, where volume patches were extracted from the original image. We observed that the segmentation accuracy of the test set was approximately the same for both cases.

In this paper, segmentation is essentially a classification of each voxel into a binary assignment of scar or normal myocardium.[32] A volume patch was extracted around each voxel in a sliding window fashion and was then labeled according

TABLE I. Image acquisition parameters.

Manufacturer	Siemens medical
Model	MAGNETOM Trio
Field strength	3 Tesla
Echo time	1.3 ms
Flip angle	20˚
Pulse sequence	Inversion-recovery gradient echo
Voxel size	0.625 × 0.625 × 1.3 mm

TABLE II. Characteristics of the study subjects.

Characteristic	Mean ± SD
Age (years)	51.5 ± 12.6
Weight (Kg/m^2)	28.2 ± 4.7
Male (n)	29
Heart rate (BPM)	67.1 ± 11.2
LVEF (%)	32.1 ± 12.7
LV EDV (ml)	228.7 ± 69.3
LV ESV (ml)	160.5 ± 72.5
Scar volume (ml)	25.57 ± 16.45

LVEF, left ventricle ejection fraction, LV EDV, left ventricle end-diastolic volume, LV ESV, left ventricle end-systolic volume.
Plus–minus values are mean ± SD.

to its ratio of scar to healthy voxels in the corresponding volume patch extracted from manually segmented image. If the ratio of scar to healthy voxel was equal or greater than 0.5, the volume patch was labeled as scar. We extracted a volume patch around each voxel if the voxel was located within the myocardium. The volume patches were then used to train the CNN-based model in the training phase. In patch-based segmentation, choosing a proper size for patch and stride is very important. We chose the patch size based on the size, position, and orientation of the segmented target area. Since the shape of myocardial scar in the LV is highly heterogeneous choosing small size of volume patch helps to reconstruct the segmented image from the predicted classes more precisely, especially at the edges. However, there are factors, which limit the minimum possible size of these patches as each should be large enough to enclose sufficient information (i.e., patterns) to enable the classifier to distinguish between scar and healthy structures. In our experiments, we witnessed that extracting volume patches of size 4 × 4 × 4 yields the best results for LV scar segmentation from 3D LGE-MR images. We selected the stride of one voxel for patch extraction. Consequently, there was an overlap between adjacent volume patches in the training set.

Since scar comprises only a small fraction of the myocardium in a majority of the images, the number of volume patches labeled as scar was substantially less than the number of volume patches labeled as normal myocardium, leading to class imbalance in the training dataset. To have an approximately equal number of volume patches in both

classes the data of under-represented class (i.e., scar) in the training phase was oversampled by reproducing volume patches labeled as scar and augmented to the training dataset.[33,34] A total of 265,220 labeled patches extracted from ten 3D LGE-MR training images were used for supervised training of the CNN. During the testing phase, 450,454 volume patches were extracted from the total 3D LGE-MR test images and applied to the CNN to investigate its performance for LV myocardial scar segmentation. The average number of volume patches derived from each test image was 18769 ± 6661 depending on the test image size.

2.C. Architecture of CNN

For CNN design, we applied a growing approach, where a simple CNN including a convolutional layer along with a fully connected network with one hidden layer of 32 nodes was trained. To gradually improve the accuracy of scar segmentation, some layers were incrementally added and the parameters were tuned manually to achieve higher accuracy in myocardial scar segmentation. Figure 2 shows the architecture of the proposed CNN, which is made up of three convolutional layers with a different number of kernels in each layer along with two pooling layers followed by a fully connected (FC) layer. The corresponding size and number of kernels in convolutional layers along with the size of pooling layers and number of nodes in the FC layer are shown in Table III.

The first layer of our proposed CNN architecture was a convolutional layer with 64 kernels of size 2 × 2 × 2 voxels. The kernels were linearly convolved with the volume patches with the stride of size one to produce feature maps. The size of the feature map was a function of zero-padding, kernel size, and stride and the number of feature map was equal to the number of kernels. Two other convolutional layers with the same size of the kernel were used after the first layer, where the number of kernels was increased to 128 and 256, respectively. We performed zero-padding in each layer to retain the original size of volume patches and avoid losing information at the edges.

We used an activation layer after each convolutional layer to increase the nonlinearity properties of the decision function. This layer alters the values of the feature maps, but the size of the feature map stays constant. Available activation functions in our selected open source neural network library, Keras, were softmax, exponential linear unit, scaled exponential linear unit, rectified linear unit (ReLU), softplus, softsign, hyperbolic tangent (tanh), sigmoid, hard-sigmoid, and linear. Here, ReLU

$$f(x) = \max(0, x) \qquad (1)$$

was used as an activation function after convolutional layer, where x is the input to a neuron. The ReLU activation function is zero when x < 0 and then linear with slope of 1 when x > 0. We chose ReLU function, as it leads to faster and more straightforward in implementation than training the equivalent model with tanh function.[35]

After the second and third convolutional and activation layers, pooling layer of size 2 × 2 ×2 voxels was added to
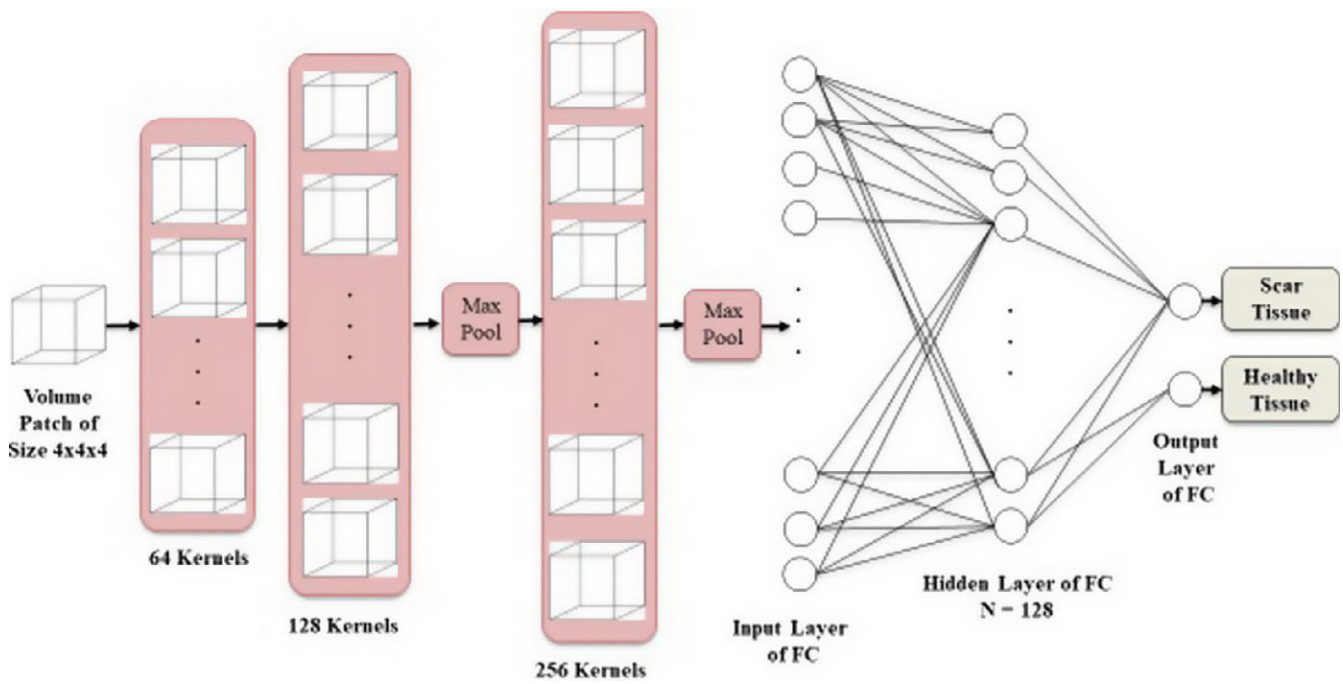
FIG. 2. Architecture of the proposed convolutional neural network, which is made up of a combination of convolutional, pooling, and FC layers. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE III. Configuration of the proposed convolutional neural network.

Layer	Type	Size	No. of Kernels	FC units
1	Conv.	$2 \times 2 \times 2$	64	–
2	Conv.	$2 \times 2 \times 2$	128	–
3	Max. Pooling	$2 \times 2 \times 2$	–	–
4	Conv.	$2 \times 2 \times 2$	256	–
5	Max. Pooling	$2 \times 2 \times 2$	–	–
6	Dropout	–	–	–
7	FC	–	–	128
8	Dropout	–	–	–
9	FC	–	–	2
10	Softmax	–	–	–

compute the maximum value of eight output nodes of the convolutional layer. The advantage of this layer is that it makes the model less sensitive to the local translation. Additionally, it yields a simpler model through downsampling, which reduces the computational complexity of the next set of layers.[36]

In spite of the convolutional layer, where neurons are connected to a local region in the input, in FC layer neurons are fully connected to all activations in the previous layer. In fact, this layer is identical to a regular multilayer perceptron, which typically consists of an input layer, one or more hidden layers and an output layer. Our CNN was configured using a two-layer perceptron including one hidden and one output layer with 128 and 2 nodes, respectively.

In this research work, softmax was used as an activation function after the FC layer. The softmax function is

a generalization form of logistic regression and is defined as[37]:

$$softmax(z)_i = \frac{\exp(z_i)}{\sum_j \exp(z_i)} \qquad \text{for j = 1 or 2} \qquad (2)$$

where z and j represent input vector and the number of possible classes respectively. The output of this function is considered as a categorical distribution over different possible classes. The softmax function highlights the largest values while it suppresses values that are significantly smaller than the maximum value.

We optimized the weights of the CNN using gradient descent method, which is one of the most widely used algorithms for optimization, particularly in neural networks. This method uses backpropagation for calculating gradient of the loss function, which defines the cost incurred for an error in prediction. When the class label of a sample needs to be determined by the CNN, the difference between the predicted class label by the CNN and its expected class label is computed in the backpropagation step.[38] To perform optimization, an adaptive learning rate method called AdaDelta was used, which does not require manual tuning of a learning rate and is robust to noisy gradient information, differences in CNN architecture, various data modalities, and selection of hyper-parameters.[39] We used categorical cross-entropy as a loss function to train the CNN because it is a more granular way for error computation. In spite of classification error or mean squared error, in cross-entropy the closeness of the computed output to its expected output takes into account.

To maintain a low generalization error in our CNN model, we used techniques, such as dataset augmentation as explained

in the previous section, early stopping and Dropout as a regularization strategy.[37] We terminated the training procedure early (50 epochs) to avoid overfitting and keep the train and test errors close to each other. The key idea in Dropout technique is to randomly drop some units along with their connections from neural network in training phase.[40] In this way, we form many numbers of thinned networks to be trained. During the testing phase, we obtained the average of all thinned networks' predictions using a single unthinned network with smaller weights. Moreover, by nature, an extensive parameter sharing occurs in CNN, which means that in the CNN, each kernel in the convolutional layer is used at every position of the input. Thus, rather than learning a separate set of parameters for every location, we learn only one set. Parameter sharing not only regularizes the parameters but also leads to a significant reduction in the required memory by storing only a subset of the parameters.[37]

To understand the relative influence of each layer for our CNN model on its accuracy of segmentation, we conducted layer-by-layer ablation, in which some layers were removed, and performance of the network revaluated. Table IV shows the obtained DSCs on the test dataset for each observation, where the removed layer is indicated with a star symbol. It is apparent that removing a single or a combination of layers leads to a drop in accuracy of the network for myocardial scar segmentation.

We implemented the CNN in Python using Keras library, which is built on top of Theano. The network was trained for 50 epochs on the shared hierarchical academic research computing network (SHARCNET) of Canada, which is a high-performance computing cluster for research. Training the CNN required approximately 124 min. Using our developed CNN model, the average time required for segmenting scar tissue from a 3D LGE-MR image is 79.29 ± 19.64 s. This will be in addition to the approximately 54 min of expert user time required for manual segmentation of the LV myocardium.

2.D. Testing phase

Testing phase contained three main steps: (a) image preprocessing and volume patch extraction, (b) patch labeling,

and (c) 3D binary image reconstruction using assigned labels. The image was normalized to the range of voxel intensity values between 0 and 255 and then a median filter of size 3×3 was applied to remove the image noise. The image was then down-sampled by the factor of two to reduce the image size. A volume patch of size $4 \times 4 \times 4$ was extracted around each voxel in a sliding window fashion with the stride of one and passed to the trained CNN to be labeled as a myocardial scar or healthy tissue. Since we used the softmax function in the proposed CNN, the output label of CNN for each patch was one or zero for myocardial scar or healthy tissue. Each pixel in the test image was replaced by its corresponding label and the binary 3D image was then compared with the manual ground truth for evaluation purpose.

2.E. Comparison with previously reported techniques

In order to provide a comprehensive evaluation of the proposed method, we compared the results of several alternative methods to manual segmentations. In particular, we applied K-nearest neighbor (KNN, where K = 7) as another machine learning technique, HMF as an energy minimization method and several intensity thresholding-based methods, including FWHM and STRM to the same dataset. We chose FWHM and STRM as those methods have been widely used in clinical studies.[16,17]

2.F. Evaluation metrics for segmentation

We used region-based and volume-based metrics to evaluate of the accuracy of algorithm-generated segmentations of the scar geometry. For region-based methods, we used the Dice similarity coefficient (DSC) and the Jaccard index (JI). The DSC measures the spatial overlap of the algorithm-generated boundary and manual segmentation by the expert and is defined as follow:

$$\text{DSC} = \frac{2(A \cap M)}{A + M} \tag{3}$$

where A and M denote algorithm and manually segmented regions. Similarly, JI,

TABLE IV. Ablation experiment results.

	Conv64	Conv128	Max. Pooling	Conv256	Max. pooling	Dropout (0.5)	FC (128)	Dropout (0.5)	FC(2)	Mean of DSC
Net 1		*	*	*						83%
Net 2		*								88%
Net 3				*	*					89%
Net 4	*									88%
Net 5			*							89%
Net 6					*					90%
Net 7						*		*		90%
Proposed										93%

$$JI = \frac{A \cap M}{A \cup M} \qquad (4)$$

was reported as a measure of similarity between algorithm and manual segmentation using equation 4. As volume-based metrics, we used absolute volume difference (AVD) and normalized absolute volume difference (NAVD).

$$AVD = |V_A - V_M| \qquad (5)$$

$$NAVD = \frac{|V_A - V_M|}{V_M} \times 100\% \qquad (6)$$

V_A is the volume of the scar determined by the algorithm and V_M is the volume of the manually segmented scar. We also used Bland–Altman plots to analyze the agreement between the volumes of scar determined by several methods for scar segmentation and manually segmentation.

To investigate the statistical significant differences between the algorithm-generated and manually generated measurements, the Wilcoxon signed rank test was used with a 5% level of significance. This statistical test was chosen because a normal distribution for the sample population cannot be assumed.

3. RESULTS

Example results of CNN-based scar segmentation for 3D LGE-MR images from three patients are shown in Fig. 3. Volume rendered images are presented in an anterior–posterior projection to demonstrate a large MI of the anterior wall and apical segments, consistent with the vascular territory of the left anterior descending artery. Visually, the algorithm-generated boundaries closely match the boundaries of manual expert delineations. In order to provide a quantitative evaluation of the proposed method, the accuracy of segmentation using the trained model was evaluated on a test dataset of 24 3D LGE-MR images. Our developed method yielded an average DSC and JI of 93.63% ± 2.61% and 88.13% ± 4.70% as the accuracy of segmentation versus manually expert segmented images. The CNN-based technique reported a mean AVD and NAVD of 2.08 ± 1.81 ml and 16.71% ± 14.31%.

We also compared the results of our method to that of several previous methods, including K-nearest neighbors
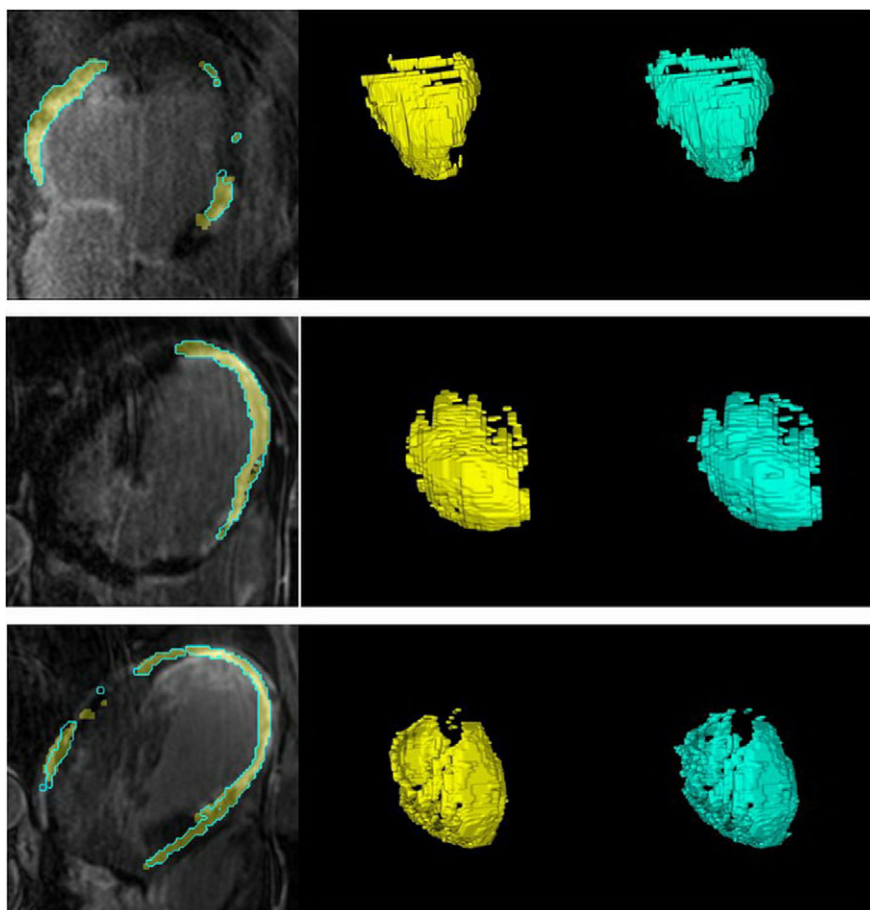

FIG. 3. Exemplary results of segmented scar tissue from 3D late gadolinium enhancement magnetic resonance (LGE-MR) image of three patients. In the first column, the expert manual segmentation of the scar for one slice extracted from 3D LGE-MRI is shown in yellow, and the contours computed by the convolutional neural network (CNN)-based method are displayed in cyan. The second and third columns depict expert manual segmentation of myocardial scar and the CNN generated scar region, respectively. [Color figure can be viewed at wileyonlinelibrary.com]

(KNN), HMF, FWHM, and STRM. Example segmentation results for all methods are shown in Fig. 4 for a single 3D LGE-MR image. Three orthogonal slice views of the original image along with the manual segmentation of scar are illustrated on the first row and segmented images obtained using CNN, KNN, HMF, FWHM, STRM + 2SD, STRM + 3SD, STRM + 4SD, STRM + 5SD, and STRM + 6SD are shown in the corresponding row. It can be seen that FWHM, STRM + 5SD, and STRM + 6SD methods underestimated scar volume. In contrast, HMF, KNN, STRM + 2SD, STRM + 3SD, and STRM + 4SD methods overestimated scar volume. These results showed that our method yielded higher accuracy for segmentation.

Summary of the results of quantitative evaluation for the developed algorithm is shown in Table V. The highest DSC and lowest AVD achieved were obtained using the CNN technique (93.63% ± 2.6% and 2.08 ± 1.81 ml). The next best DSC was achieved using the KNN technique (77.85% ± 5.01%), whereas the HMF technique reported a lower error with regard to AVD (3.45 ± 2.85 ml).

To visualize the distance-based accuracy of employed algorithms for myocardial scar segmentation, the output of the segmentation results of CNN, HMF, and STRM + 4SD, which reported the highest DSC for each group of methods were selected. Euclidean distance between each vertex of algorithm-generated scar tissue and closest vertices of manual scar tissue was computed. The color-coded map was then generated from the mean distance error between the two surfaces. Figure 5 shows the segmentation error between surfaces generated from label maps of CNN, HMF, and STRM + 4SD and the expert manual delineations for an example patient image. The color-coded map indicates that as compared to color-code maps for other methods, the CNN method has smaller distance errors.

The Wilcoxon signed rank test was conducted to assess the statistical significant difference between the algorithm-generated and manually segmented scar volumes. As shown in Table VI, the *P*-value is larger than 0.05 for the CNN, KNN, FWHM, and STRM + 4SD, which indicates weak evidence against the null hypothesis. Therefore, we fail to reject the null hypothesis at the 0.05% significance level and one can say that the mean volume of segmented scar obtained using CNN, KNN, FWHM, and STRM + 4SD methods is not significantly different from manually segmented scar volume.

We performed Wilcoxon signed rank test on DSC obtained from different segmentation methods using the testing dataset of 24 3D LGE-MR images. The median value of DSC produced by CNN was compared pair-wise with that generated by KNN, HMF, FWHM, STRM + 2SD, STRM + 3SD,
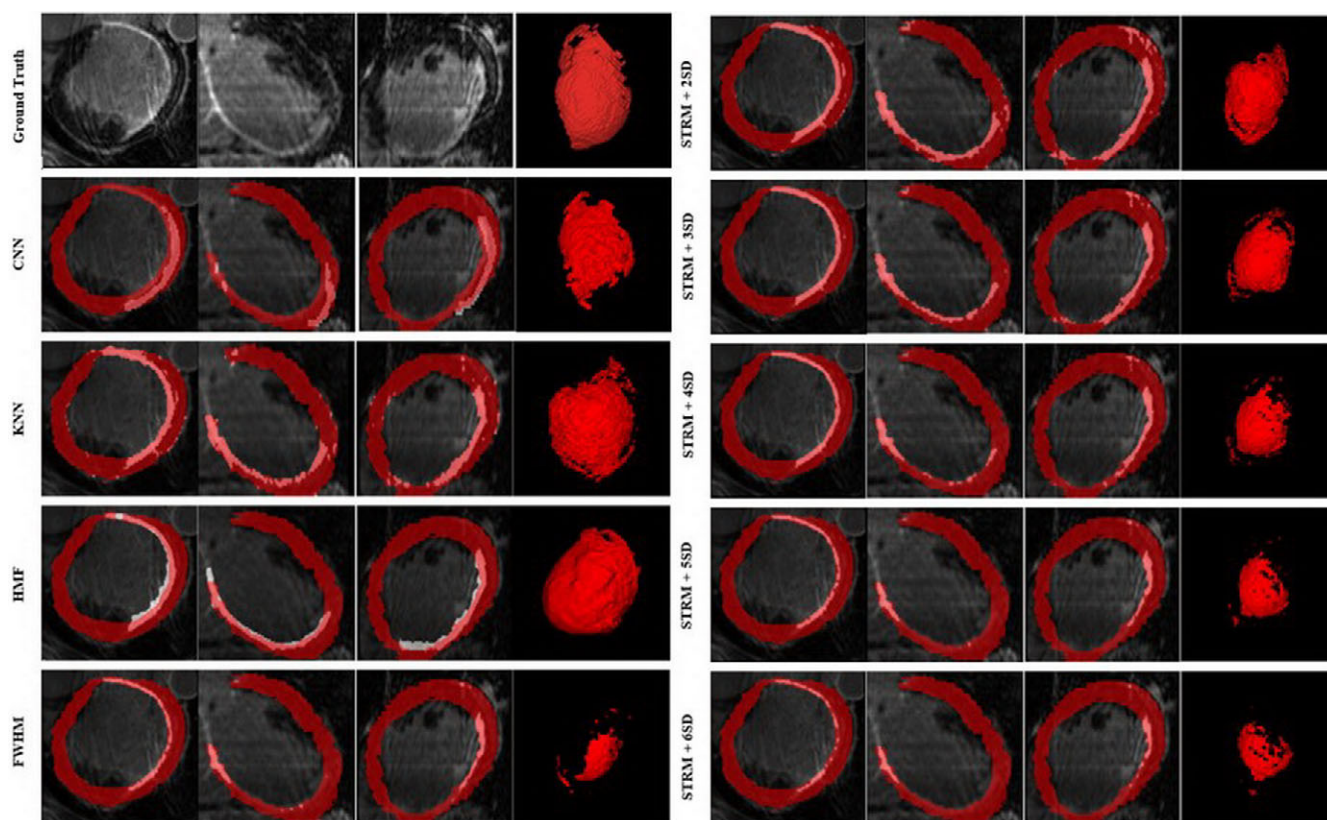

FIG. 4. Exemplary results of 3D scar segmentation methods applied for an example 3D late gadolinium enhancement magnetic resonance image, presented in orthogonal views (Column 1–3) and using a surface rendering. In these images, expert manual delineation of the myocardium is shown in red, the segmented scar computed by respective methods is displayed in white. Results of each method are shown in the corresponding row. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE V. Performance of all 3D scar segmentation methods against expert manual segmentation.

Method	DSC (%)	JI (%)	AVD (ml)	NAVD (%)
CNN	93.63 ± 2.61	88.13 ± 4.70	2.08 ± 1.81	16.71 ± 14.31
KNN	77.85 ± 5.01	64.00 ± 6.59	10.66 ± 10.24	32.52 ± 31.26
HMF	75.96 ± 3.61	61.38 ± 4.69	3.45 ± 2.85	15.80 ± 14.54
FWHM	61.77 ± 9.81	45.38 ± 9.79	7.70 ± 7.43	33.22 ± 28.16
STRM ± 2SD	48.33 ± 17.68	33.62 ± 15.14	31.66 ± 20.77	31.66 ± 20.77
STRM ± 3SD	57.71 ± 14.42	41.94 ± 13.74	16.08 ± 12.67	91.54 ± 81.55
STRM ± 4SD	62.44 ± 10.67	46.22 ± 10.75	7.29 ± 7.35	36.96 ± 47.57
STRM ± 5SD	61.16 ± 9.95	44.75 ± 9.85	8.03 ± 7.89	34.01 ± 26.31
STRM ± 6SD	54.08 ± 11.94	37.95 ± 10.91	11.78 ± 10.84	46.30 ± 21.25

AVD, absolute volume difference; CNN, convolutional neural network; DSC, dice similarity coefficient; JI, Jaccard Index; KNN, K-nearest neighbor; HMF, hierarchical max flow; FWHM, full width at half maximum; NAVD, normalized absolute volume difference; STRM, signal threshold to reference mean.
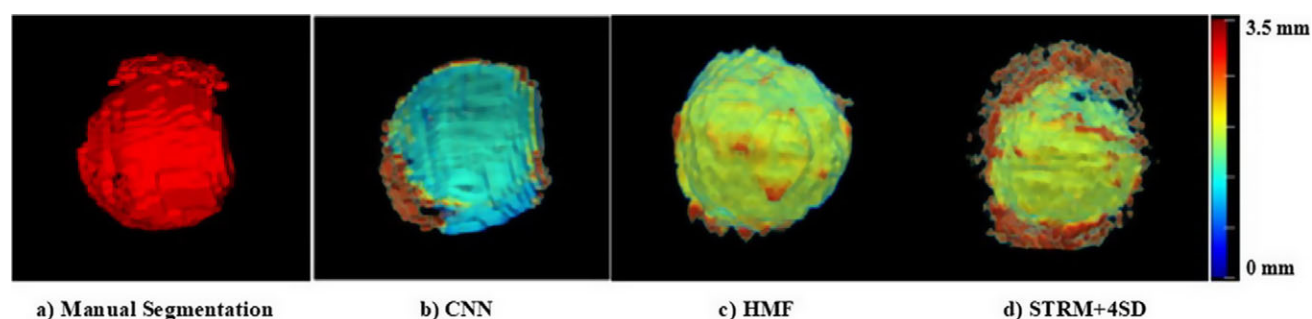

FIG. 5. First image from left is (a) manual segmentation of myocardial scar of one patient accompanied with mean Euclidean distance color-coded maps computed from the label map generated using (b) convolutional neural network, (c) hierarchical max flow, and (d) signal threshold to reference mean + 4SD from manual segmentation. [Color figure can be viewed at wileyonlinelibrary.com]

STRM + 4SD, STRM + 5SD, and STRM + 6SD and the p-value was computed in each case. In this test, the alternative hypothesis was that the performance of CNN is higher than that of either KNN, HMF, FWHM, or STRM for segmenting myocardium scar from 3D LGE-MR images. The computed p-value in all cases were <0.0001. Therefore, the null hypothesis was rejected, which indicated that the developed CNN-based method significantly outperformed other previous techniques. The results of statistical tests further demonstrated that, as compared to other methods, CNN is more accurate and precise in segmenting the LV myocardial scar.

To investigate agreement between algorithm-generated and manually determined scar volumes, we used a Bland–Altman plot, which graphs the mean difference of measured scar volume versus manual scar volume and constructs limits of agreement.[41] From the nine methods, we chose the best methods in each category based on the order of ranking in accuracy of segmentation in terms of AVD. Figure 6 shows the Bland–Altman plots for CNN, HMF, STRM + 4SD versus manual scar volume, respectively. The solid dark blue line, upper and lower dashed lines represent the mean value of the difference, the upper and lower limits of agreement, respectively. The smallest range for the limits of agreement was reported for the CNN method.

TABLE VI. *P*-value of Wilcoxon rank test performed to compare the median value of total scar volume generated from different algorithms and expert manually segmentations.

Method	*P*-value
CNN	0.1451
KNN	0.1161
HMF	0.0129
FWHM	0.0556
STRM ± 2SD	1.8215e-05
STRM ± 3SD	0.0012
STRM ± 4SD	0.1702
STRM ± 5SD	0.0152
STRM ± 6SD	3.8803e-05

CNN, convolutional neural network; KNN, K-nearest neighbor; HMF, hierarchical max flow; FWHM, full width at half maximum; STRM, signal threshold to reference mean.

The mean of scar volume difference $(V_A - V_M)$ between CNN, HMF, and STRM + 4SD techniques versus expert manually segmentation were 0.76, 2.45, and 2.55 ml, respectively. As bias values were all positive, each method was felt to slightly overestimate total scar volume. This overestimation
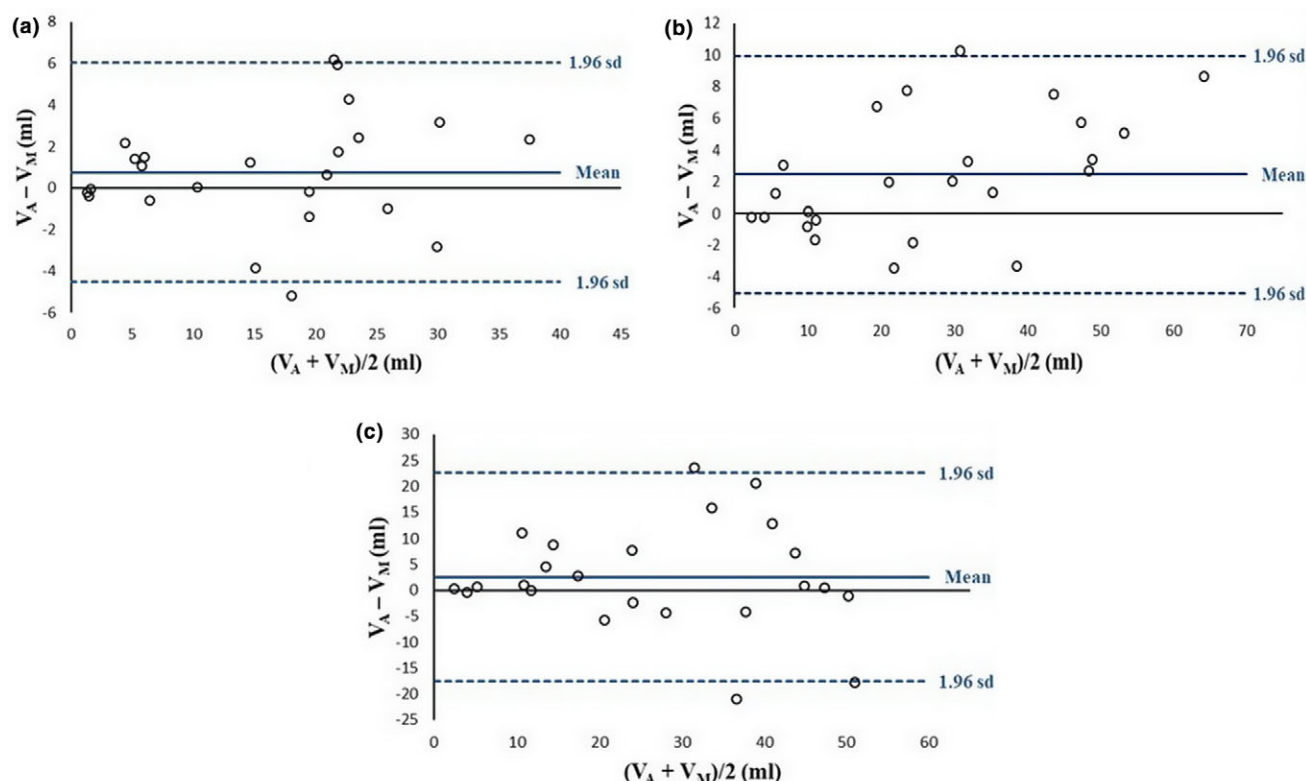
FIG. 6. Bland–Altman plot of scar volume obtained from four techniques, which reported highest dice similarity coefficient for each category (a) convolutional neural network, (b) hierarchical max flow, (c) signal threshold to reference mean + 4SD. The plot graphs the difference between the algorithm-generated and manual scar volume as a function of their mean. [Color figure can be viewed at wileyonlinelibrary.com]

of volume may reflect incorrect labeling of tissue regions with high intensity as scar, which are not visually labeled as scar by the expert.

4. DISCUSSION

In this paper, we described a novel, semiautomated methodology for 3D segmentation of LV myocardial scar from 3D LGE-MR images using a deep neural network based on 3D CNN, a model that yields hierarchies of features through training voxels-to-voxels. Our experimental results demonstrated that our method, as compared to other alternative techniques, yielded the highest accuracy for the tested evaluation metrics including DSC, JI, and AVD. To our knowledge, this is the first description of LV myocardial scar tissue segmentation from 3D LGE-MR images using a CNN-based method.

The potential for machine learning-based segmentation of 3D LGE-MR to impact patient care is rapidly expanding. However, this task is quite challenging due to a large variability of cardiac structures across patients, complexity in segmenting the apical and basal slice images that have similar intensity profile with that of surrounding structures, image noise and artifacts, and poor image contrast at the boundary layers.[42,43] Numerous studies have validated the potential of 3D LGE data to provide meaningful insights into scar architecture for the purpose of modeling ventricular arrhythmia,[6,44]

and in guiding invasive cardiac procedures aimed at targeting[45–47] or avoiding[48] myocardial scar to improve patient outcomes. Given immediate capacity for segmented datasets to be fused to live fluoroscopic images,[49,50] a paradigm of image-guided personalized therapeutics is emerging. Further, the work described in this sentinel study provides new opportunities to pursue rapid, standardized scar quantification for risk stratification in patients with ischemic cardiomyopathy, a field strongly justified by expanding publications from 2D LGE data.[1–3]

In designing our method, we needed to address several relevant challenges. As the scar tissue was a relatively small region of the global myocardium, the number of extracted patches with the label of scar was less than patches with a healthy label, causing class imbalance in the training dataset. To improve the performance of the CNN-based network the under-represented data were augmented artificially.[34] Although we could have randomly discarded some volume patches in the healthy tissue class, it was preferred to artificially increase the number of training samples due to the limited number of images in the training set. The CNN was built over convolutional and pooling layers with small sizes to allow for deeper architectures while reducing computational complexity. Our deep network was prone to overfitting because of having millions of trainable parameters. Accordingly, parameters were adjusted experimentally, and some regularization strategies were employed to avoid overfitting.

We compared the performance of the proposed algorithm with KNN, HMF, and several signal intensity thresholding-based methods including the FWHM and STRM approaches, which are widely employed in the literature due to their efficiency and ease of use. Compared to these methods, the CNN-based method does not rely on expert input for initialization, the primary source of variability for such techniques. We evaluated the mean values of DSC and AVD for all methods to estimate the accuracy of segmentation versus expert delineation and identified that CNN not only reported the highest DSC and the lowest AVD on average but also the lowest standard deviation. Accordingly, it was demonstrated that a CNN-based algorithm improves both accuracy and reproducibility in segmenting scar. Furthermore, the required time for delineating scar from presegmented myocardium does not exceed 2 min on a regular PC with an Intel i7 processor.

One of the main limitations of our proposed method is the requirement for manual delineation of the myocardial boundaries of the LV. Some emerging applications in cardiology, such as computational modeling of cardiac electrophysiology require fully segmented LV myocardium along with scar geometry. In such an application, when LV myocardium is already segmented our proposed method is useful in obtaining an accurate segmentation of the scar geometry. As future work, we aim to fully automate the scar segmentation by combining atlas-based LV segmentation methods with deep learning-based techniques.

Based on the currently restricted availability of 3D LGE-MR imaging in patients with ICM, we were restricted to a relatively small dataset upon which to train and validate our model. Additionally, while characterizing scar heterogeneity is of interest for ventricular tachycardia risk assessment in post-MI patients, there is no consensus on how to define border zone and infarct core on these uniquely high-resolution datasets. Accordingly, further work is required to develop and validate the role of multi-threshold scar segmentation using these images.

5. CONCLUSION

In conclusion, we described a novel deep learning-based method for the segmentation of LV myocardial scar from 3D LGE-MR images. Our results confirm that CNN-based techniques can perform accurate and automated segmentation of myocardial scar from these datasets relative to expert manual segmentation. To our knowledge, deep learning-based methods for myocardial scar segmentation from 3D LGE-MR images have not been previously described.

ACKNOWLEDGMENT

CONFLICT OF INTEREST

The authors of this paper have no conflict of interest to disclose.

[a)]Author to whom correspondence should be addressed. Electronic mail: fatemehzabihollahy@cmail.carleton.ca.

REFERENCES

1. Wu KC. Sudden Cardiac Death Substrate Imaged by Magnetic Resonance Imaging: from Investigational Tool to Clinical Applications. *Circ Cardiovasc Imaging*. 2017;10:7.
2. Schmidt A, Azevedo CF, Cheng A, et al. Infarct tissue heterogeneity by magnetic resonance imaging identifies enhanced cardiac arrhythmia susceptibility in patients with left ventricular dysfunction. *Circulation*. 2007;116:2006–2014.
3. Bello D, Fieno DS, Kim RJ, et al. Infarct morphology identifies patients with substrate for sustained ventricular tachycardia. *J Am Coll Cardiol*. 2005;45:1104–1108.
4. Arevalo HJ, Vadakkumpadan F, Guallar E, et al. Arrhythmia risk stratification of patients after myocardial infarction using personalized heart models. *Nat Commun*. 2016;7:11437.
5. Trayanova NA, Pashakhanloo F, Wu KC, Halperin HR. Imaging-Based Simulations for Predicting Sudden Death and Guiding Ventricular Tachycardia Ablation. *Circ Arrhythmia Electrophysiol*. 2017;10: e004743.
6. Ukwatta E, Arevalo H, Rajchl M, et al. Image-based reconstruction of three-dimensional myocardial infarct geometry for patient-specific modeling of cardiac electrophysiology. *Med Phys*. 2015;42: 4579–4590.
7. Trayanova NA. Computational Cardiology: the Heart of the Matter. *ISRN Cardiol*. 2012;20:1–15.
8. Vadakkumpadan F, Gurev V, Constantino J, Arevalo H, Trayanova N. Modeling of whole-heart electrophysiology and mechanics: Toward patient-specific simulations. In: *Patient-Specific Modeling of the Cardiovascular System: Technology-Driven Personalized Medicine*. Berlin, Germany: Springer Science & Business Media; 2010:145–165.
9. Stirrat J, White JA. The prognostic role of late gadolinium enhancement magnetic resonance imaging in patients with cardiomyopathy. *Can J Cardiol*. 2013;29:329–336.
10. Plank G, Zhou L, Greenstein JL, et al. From mitochondrial ion channels to arrhythmias in the heart: computational techniques to bridge the spatio-temporal scales. *Philos Trans R Soc A Math Phys Eng Sci*. 2008;366:3381–3409.
11. Kawaji K, Tanaka A, Patel MB, et al. 3D Late gadolinium enhanced cardiovascular MR with CENTRA-PLUS profile/view ordering: feasibility of right ventricular myocardial damage assessment using a swine animal model. *MRI*. 2017;39:7–14.
12. White JA, Fine N, Gula LJ, et al. Fused whole-heart coronary and myocardial scar imaging using 3-T CMR. Implications for planning of cardiac resynchronization therapy and coronary revascularization. *JACC Cardiovasc Imaging*. 2010;3:921–930.
13. Kecskemeti S, Johnson K, François CJ, Schiebler ML, Unal O. Volumetric late gadolinium-enhanced myocardial imaging with retrospective inversion time selection. *J Magn Reson Imaging*. 2013;38:1276–1282.
14. Peters DC, Wylie JV, Hauser TH, et al. Detection of pulmonary vein and left atrial scar after catheter ablation with three-dimensional navigator-gated delayed enhancement mr imaging: initial experience[1]. *Radiology*. 2007;243:690–695.
15. Xu C, Xu L, Gao Z, et al. Direct delineation of myocardial infarction without contrast agents using a joint motion feature learning architecture. *Med Image Anal*. 2018;50:82–94.

16. Neizel M, Katoh M, Schade E, et al. Rapid and accurate determination of relative infarct size in humans using contrast-enhanced magnetic resonance imaging. *Clin Res Cardiol*. 2009;98:319–324.

17. Kolipaka A, Chatzimavroudis GP, White RD, O'Donnell TP, Setser RM. Segmentation of non-viable myocardium in delayed enhancement magnetic resonance images. *Int J Cardiovasc Imaging*. 2005;21:303–311.

18. Albà X, Figueras IVR, Lekadir K, Frangi AF. Healthy and scar myocardial tissue classification in DE-MRI. In: *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Heidelberg, Berlin: Springer-Verlag; 2013.

19. Lu Y, Yang Y, Connelly KKA, Wright GA, Radau PE. Automated quantification of myocardial infarction using graph cuts on contrast delayed enhanced magnetic resonance images. *Quant Imaging Med Surg*. 2012;2:81.

20. Ukwatta E, Arevalo H, Li K, et al. Myocardial infarct segmentation from magnetic resonance images for personalized modeling of cardiac electrophysiology. *IEEE Trans Med Imaging*. 2016;35:1408–1419.

21. Rajchl M, Yuan J, White JA, et al. Interactive hierarchical-flow segmentation of scar tissue from late-enhancement cardiac mr images. *IEEE Trans Med Imaging*. 2014;33:159–172.

22. Rajchl M, Stirrat J, Goubran M, et al. Comparison of semi-automated scar quantification techniques using high-resolution, 3-dimensional late-gadolinium-enhancement magnetic resonance imaging. *Int J Cardiovasc Imaging*. 2015;31:349–357.

23. Zabihollahy F, White JA, Ukwatta E. Myocardial scar segmentation from magnetic resonance images using convolutional neural network. SPIE. 2018.

24. Pereira S, Pinto A, Alves V, Silva CA. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans Med Imaging*. 2016;35:1240–1251.

25. Zhao L, Jia K. Multiscale CNNs for brain tumor segmentation and diagnosis. *Comput Math Methods Med*. 2016;2016:1–7.

26. Choi H, Jin KH. Fast and robust segmentation of the striatum using deep convolutional neural networks. *J Neurosci Methods*. 2016;274:146–153.

27. Christ PF, Elshaer MEA, Ettlinger F, Tatavarty S, Bickel M, Bilic P, et al. Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields. In: *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) MICCAI. Cham, Switzerland: Springer; 2016.

28. Teramoto A, Fujita H, Yamamuro O, Tamaki T. Automated detection of pulmonary nodules in PET/CT images: ensemble false-positive reduction using a convolutional neural network technique. *Med Phys*. 2016;43:2821–2827.

29. Lavdas I, Glocker B, Kamnitsas K, et al. Fully automatic, multiorgan segmentation in normal whole body magnetic resonance imaging (MRI), using classification forests (CFS), convolutional neural networks (CNNs), and a multi-atlas (MA) approach. *Med Phys*. 2017;44:5210–5220.

30. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *NeuroImage*. 2006;31:1116–1128.

31. Manlucu J, Gula LJ, Al-Admawi M, et al. Scar morphology in patients presenting with ventricular arrhythmias: implications for risk stratification in patients with chronic ischemic cardiomyopathy. *JACC*. 2010;55:E783.

32. Kayalibay B, Jensen G, van der Smagt P. CNN-based segmentation of medical imaging data. 2017.

33. García V, Sánchez JS, Mollineda RA, Alejo R, Sotoca JM. The class imbalance problem in pattern classification and learning. In: Congreso Espanol de Informatica. 2007.

34. Wahab N, Khan A, Lee YS. Two-phase deep convolutional neural network for reducing class skewness in histopathological images based breast cancer detection. *Comput Biol Med*. 2017;85:86–97.

35. Krizhevsky A, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Neural Inf Process Syst. 2012.

36. Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436–444.

37. Goodfellow I, Bengio Y, Courville A. "Deep Feedforward Networks", "Regularization for Deep Learning", and "Convolutional Networks" in Deep Learning, MIT Press, 2016, ch. 6, 7, and 9, sec. 6.2.2.3, 7.8, and 9.2 pp.180-181, 241-249, and 330-335, respectively.

38. Rosasco L, De Vito E, Caponnetto A, Piana M, Verri A. *Neural computation*. Vol. 16. London, UK: MIT Press; 2004:1063–1076.

39. Zeiler MD. ADADELTA: an adaptive learning rate method. arXiv:1212.5701v1, 2012;

40. Sutskever I, Hinton G, Krizhevsky A, Salakhutdinov RR. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15:1929–1958.

41. Giavarina D. Understanding bland altman analysis. *Biochem Medica*. 2015;25:141–151.

42. Zhuang X. Challenges and methodologies of fully automatic whole heart segmentation: a Review. *J Healthc Eng*. 2013;4:371–408.

43. Petitjean C, Dacher JN. A review of segmentation methods in short axis cardiac MR images. *Med Image Anal*. 2011;15:169–184.

44. Deng D, Arevalo H, Pashakhanloo F, et al. Accuracy of prediction of infarct-related arrhythmic circuits from image-based models reconstructed from low and high resolution MRI. *Front Physiol*. 2015;6:282.

45. Acosta J, Fernández-Armenta J, Penela D, et al. Infarct transmurality as a criterion for first-line endo-epicardial substrate-guided ventricular tachycardia ablation in ischemic cardiomyopathy. *Hear Rhythm*. 2016;13:85–95.

46. Dickfeld T, Lei P, Dilsizian V, et al. Integration of three-dimensional scar maps for ventricular tachycardia ablation with positron emission tomography-computed tomography. *JACC Cardiovasc Imaging*. 2008;1:73–82.

47. Cochet H, Komatsu Y, Sacher F, et al. Integration of merged delayed-enhanced magnetic resonance imaging and multidetector computed tomography for the guidance of ventricular tachycardia ablation: a pilot study. *J Cardiovasc Electrophysiol*. 2013;24:419–426.

48. Laksman Z, Yee R, Stirrat J, et al. Model-based navigation of left and right ventricular leads to optimal targets for cardiac resynchronization therapy: a single-center feasibility study. *Circ Arrhythmia Electrophysiol*. 2014;7:1040–1047.

49. Nguyên UC, Mafi-Rad M, Aben JP, et al. A novel approach for left ventricular lead placement in cardiac resynchronization therapy: intraprocedural integration of coronary venous electroanatomic mapping with delayed enhancement cardiac magnetic resonance imaging. *Hear Rhythm*. 2017;14:110–119.

50. Behar JM, Mountney P, Toth D, et al. Real-Time X-MRI-Guided Left Ventricular Lead Implantation for Targeted Delivery of Cardiac Resynchronization Therapy. *JACC Clin Electrophysiol*. 2017;3:803–814.