Automated Detection of Left Ventricle in Arterial Input Function (AIF) Image Series for Cardiac MR Perfusion Imaging: A Large Study on 13K Patients

Hui Xue
hxue@stanford.edu

Abstract

Accurate detection of left ventricle (LV) blood pool from arterial input function (AIF) image series is a pre-requisite for fully automated quantification of cardiac perfusion imaging. Due to the invasive nature of stress cardiac exam, the AIF LV detection must be very accurate to enable clinical deployment. This study developed a U-net based neural net (NN) with ResNet modules as building block for highly accurate detection of AIF blood pool. A large data cohort of 25,027 perfusion scans from 12,984 patients was collected from three hospitals for training. A bootstrap strategy was implemented to speedup data labelling. Resulting NN achieved high detection rate of 99.56%, 99.856% and 99.77% for Tra, Dev and test sets. This solution was integrated to clinical MR scanner and running in collaborative hospitals.

Introduction

Cardiac vascular diseases (CAD), especially coronary diseases, is a leading cause of death in the United States and worldwide. Early diagnosis of potential CAD is of great importance to improve public health and reduce healthcare cost. Among different techniques to image the heart, cardiac magnetic resonance (CMR) imaging is a leading modality, for its non-invasive nature, flexible imaging planning ability and rich contrast mechanism [1]. Specific for CAD, cardiac perfusion imaging is the top tool which today's clinicians rely on for ischemic and non-ischemic heart disease diagnosis [2]. Recent progress of CMR perfusion imaging allows free-breathing imaging with robust single-shot acquisition. This opens the avenue to fully quantification, meaning that if the input function to the myocardium during contrast uptake can be measured from the arterial input function (AIF) image series, the absolution myocardial blood flow (MBF, in the unit of ml/min/g) can be computed (Fig. 1) [3].

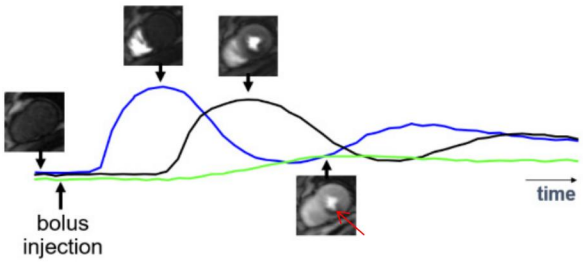

Figure 1. A illustration of CMR perfusion imaging. The contrast bolus is injected and pass through the heart. For myocardium where blood supply is limited, the image shows reduced intensity. This is a sensitive way to diagnose CAD.

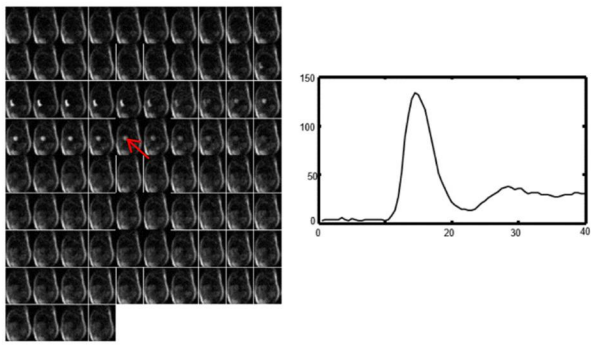

Figure 2. Left panel shows a perfusion AIF image series. While the contrast bolus passes through, the LV is enhanced. If the LV can be segmented, the signal intensity curve (right panel) can be extracted. This is a pre-requisite to quantify myocardial blood flow in myocardium.

One barrier of computing MBF is to extract AIF signal from images. Start-of-the-art approach relies on imaging fast and acquiring low resolution perfusion image series. While the contrast agent flows in, the left ventricular (LV) will be enhanced (Fig. 2) [3].

While researchers are manually drawing AIF LV for input function or detect it with heuristic methods, I propose to improve the workflow by developing a neural net based solution for automated AIF LV detection with high accuracy. The input is 2D+T AIF image series and output is a mask for LV.

The goal of this study is to achieve very high success rate of LV detection (e.g. 1 failure in 500 scans), to enable the clinical deployment of NN based perfusion flow mapping. The reason for not tolerating failure is the high cost and patient discomforts of perfusion stress imaging. To conduct a cardiac perfusion stress imaging, patients are administrated with adenosine which dilates the coronary to maximal level and causes elevated blood pressure. This leads to noticeable patient discomforts. Under this condition, the contrast agent (Gd, to shorten the T1 relaxation time)

is injected as a bolus and imaging is conducted by running MR imaging sequence for about 1min. A failed LV detection will fail perfusion mapping, which in turn voids the stress study. This often means a call-back of patient and interrupted clinical workflow.

To achieve required high accuracy, a large data cohort was collected from three collaborative hospitals. This dataset consists of ~13K patients with ~25K perfusion scans. **To my best knowledge, this is the largest cardiac perfusion MR data cohort ever assembled for AI study.**

Related work

Although there is no prior study to apply NN for perfusion AIF LV detection, the problem of imaging segmentation using NN is well studied. The first well-know algorithm is the fully convolutional network (FCN) [4], which modified output layer and loss function upon classification networks for pixel-wise segmentation. This network was extended by adding encoding-decoding structures and expanding the filter depth while reducing the spatial resolution [5,6]. More importantly, these NNs used skip connections between corresponding down/upsampling layers, to allow detailed pixel-level segmentation. This led to major improvement in performance, as demonstrated by U-net [5] or Seg-net [6]. The U-net architecture was further extended by using recurrent [7] and residual modules [8]. More recent variation included V-net [8] and TernausNet [9], which kept the skip connections, but had different number of convolution filters and kernel size in each down/upsampling layers. These segmentation NNs had been used for cardiac MR image segmentation. The first well-known work utilized a modified FCN network to segment myocardium from cine imaging [10]. Recent published work included automated analysis of T1 mapping image [11] and cardiac flow images [12].

Dataset

Imaging and data collection Patients underwent perfusion imaging in three hospitals (As requested by collaborators, their names are masked out). Scans were performed in accordance with local protocol and patients were asked to refrain from caffeine for at least 12 hours before the scan. Consecutive enrollment was used for all centers. This is, all patients who underwent perfusion scan and gave consent were included in either training or test datasets, without exclusion. This is to ensure the same distribution of training and test sets and capture the real-world data distribution.

Same imaging sequence were used in all hospitals on all patients. The study was performed using the eight scanners. Five scanners were at Center A, including three 1.5T MAGNETOM Area and two MAGNETOM Prisma, Siemens AG Healthcare, Erlangen, Germany. Two 1.5T MAGNETOM Area was used in Center B. Center C has a MAGNETOM Prisma.

Data collection A total of N=**25,027** stress and rest scans were collected for training. This dataset was split into trai ning set (Tra) (N=22,941) and development set (Dev, N=2,086). These data were acquired from N=12,984 patients.

Following table gives data distribution from all centers:

Centers	#patients	#scans stress	#scans rest	#scanners	Duration
A	10,128	8,988	11,358	5	20160429 - 20190215
B	2,273	765	2,502	2	20160609 - 20190214
C	583	596	818	1	20160524 - 20190214
Total	**12,984**	10,349	14,678	8	-

As an example, a total of 2,273 patients were scanned at Center C, contributing 3,267 scans with 765 stress and 2,502 rest perfusion scans. The first scan was performed on 20160609 and the last were performed on 20190214.

To test the neural net, a separate test set was assembled from 20190216 to 20190314. This cohort consists of N=429 patients (A, 377 patients; B, 25; C, 27) with a total of 880 perfusion scans (430 stress and 450 rest).

Methods

A modified U-net architecture was developed and implemented for the detection of LV from AIF time series. Input data was formatted to be a 2D+T array of 64x48x64. Output of NN can be either pixel-wise probability maps of LV (binary segmentation, background and LV) or both LV and RV (multi-class segmentation, background, LV and RV). While drafting this report, a large training dataset has been collected from three hospitals with IRBs' approval for retrospective anonymous technical development studies. Anonymized data were analyzed at NIH with approval by the NIH Office of Human Subjects Research OHSR (Exemption #13156). This study is described in detail as following:

Bootstrap data labelling Given the large datasets, a bootstrap strategy was implemented to speed up data labelling. First, the input AIF image series were resampled to 64x48 matrix. Along the temporal direction, one image was sampled per heart beat. To compensate the influences of varied heart rate, the temporal series was resampled to 0.5s grid. This allows a maximal heart rate of 120 beat per minute. A heuristic method was developed to help data labelling. First, the respiratory motion was corrected from AIF image series using a motion correction (MOCO) algorithm [13]. The corrected images series were inputted for LV and RV labelling. The noise background was removed using a simple thresholding with signal-to-noise ratio being <3. The temporal signal curves were analyzed to automatically detect the first significant peak (first-pass of contrast bolus). The signal slope and area-under-curve (AUC) were computed, together with foot and peak time. A simple threshold was applied to pick 10% pixels with high upslope and AUC. These pixels were analyzed using the temporal k-means, first to two classes for LV and RV signals. Since myocardium and other soft tissue can contaminate the k-means signal. A second round of clustering was performed to separate myocardium from blood. The final LV and RV centroid was picked to give tentative data labelling. This heuristic algorithm was applied to all training samples. All results were manually checked to either accept or correct tentative labels. Approx. 200 hours were spent on labelling entire data cohort. This process is illustrated in Figure 3.

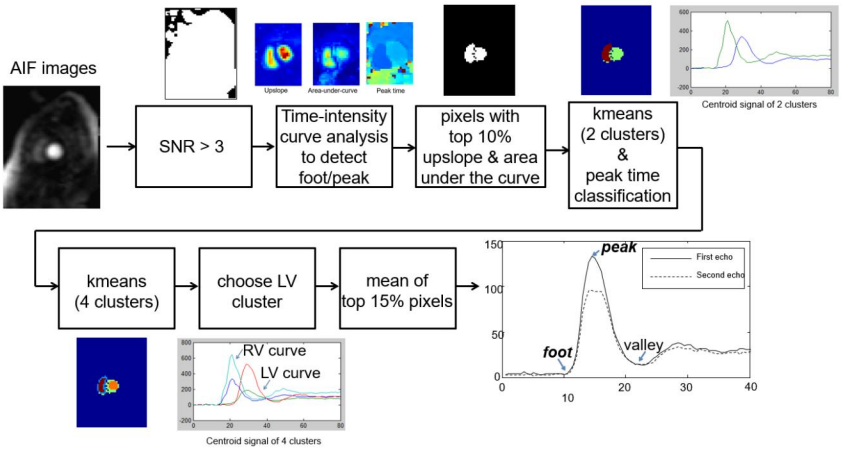

Figure 3. The heuristic AIF LV labelling scheme. This method utilizes simple thresholding and kmeans clustering to find centroids corresponding to LV and RV. The LV signal intensity curve is further analyzed for foot, peak and valley feature points.

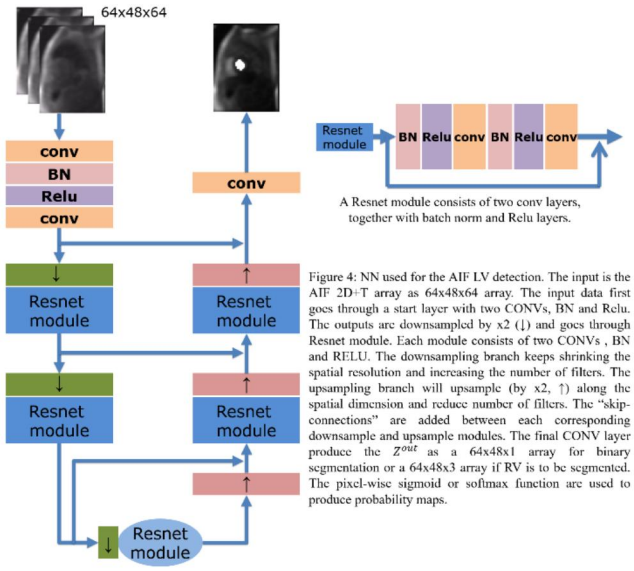

Figure 4: NN used for the AIF LV detection. The input is the AIF 2D+T array as 64x48x64 array. The input data first goes through a start layer with two CONVs, BN and Relu. The outputs are downsampled by x2 (↓) and goes through Resnet module. Each module consists of two CONVs , BN and RELU. The downsampling branch keeps shrinking the spatial resolution and increasing the number of filters. The upsampling branch will upsample (by x2, ↑) along the spatial dimension and reduce number of filters. The "skip-connections" are added between each corresponding downsample and upsample modules. The final CONV layer produce the Z^{out} as a 64x48x1 array for binary segmentation or a 64x48x3 array if RV is to be segmented. The pixel-wise sigmoid or softmax function are used to produce probability maps.

NN architecture The U-net semantic segmentation architecture [5,8] was modified for the AIF LV detection. The main change was to use ResNet module as a building block. One module consists of two CONV layers with same number of output filters. As shown in Figure 4, the downsampling and upsampling branches were connected with "Skip-connections". The input AIF array went through the input layer and got downsampled a few times. While the spatial resolution was reduced, the number of filters increased. The upsampling branch increased the spatial resolution and reduced the number of filters. Final CONV layer outputted Z^{out} as either 64x48x1 or 64x48x3 array for binary segmentation (2CS) or multi-class segmentation (3CS, background, LV and RV). Z^{out} was inputted into pixel-wise

sigmoid or softmax to get the probability maps. All CONV layers used 3x3 with stride 1 and padding 1. To increase the number of CONV layers, more Resnet module can be inserted in each downsample/upsample level.

Cost function and data augmentation Inspired by the MICCAI 2017 Robotic Instrument Segmentation [9], the cost function was a weighted sum of cross-entropy loss and the Intersection Over Union (IoU or Jaccard index). For the binary mask y^i and predicted probability $\hat{y}^i$, the IoU was computed as :

$$J_{IoU} = \frac{1}{N}\sum_{i=0}^{N-1}\frac{y^i\hat{y}^i}{y^i + \hat{y}^i - y^i\hat{y}^i}$$

N is the number of pixels. The total cost was the weighted sum:

$$J = J_{cross-entropy} - \lambda \cdot \log(J_{IoU})$$

Each AIF data was corrected for MR receiver inhomogeneity using the PD images and scaled to have maximal value to be 1 before inputting for training or testing:

$$\boldsymbol{AIF = AIF/np.max(AIF)}$$

Since MR images can have different orientation due to imaging planning, random flip along the first and second dimensions were applied as data augmentation steps during the training. The training code is available at *https://github.com/xueh2/cmr_perf*.

Training and hyperparameter search All Tra and Dev datasets was processed with heuristic method and manually checked for success. For all failed cases, manual contouring was performed to label LV and RV. Figure 5 illustrates AIF image variation caused by anatomy and imaging. The NN model was implemented using PyTorch and training was performed on an Ubuntu 18.04 PC with one NVIDIA GTX 1080Ti GPU card with 11GB RAM. With initial experiments, it was determined that number of ResNet modules and number of CONV filters in the NN had the most significant impact on the detection accuracy. Therefore, a hyperparameter search was conducted to test different combinations (5 to 9 ResNet modules, number of CONV filters from 64 to 256). For all tests, the ADAM optimization was used with initial learning rate being 0.001. The betas are 0.9 and 0.999. Epsilon was 1e-8. Learning rate was reduced by x2 for every 10 epochs. Training took 40 epochs and best model was selected as the one giving best performance on Dev set.

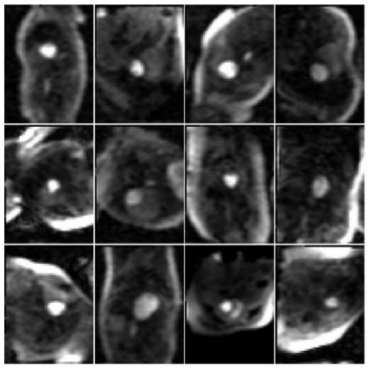

Figure 5. Examples of AIF image series from different patients, which illustrates the variation among the data cohort for heart location, orientation, shape and intensity etc.

Evaluation metric Dice ratio was selected as the metric for success of LV detection. If the Dice ratio between ground-truth mask and NN result was over 0.5, the detection was defined as success. The detection rate was computed and reported for Tra, Dev and Test sets.

MR scanner integration The best model found during training was saved for deployment. An open-source framework (Gadgetron [14]) was used to load and apply the Pytorch model in MR scanners. The MR scanner imaging workflow was modified to first read in saved NN model and later apply to AIF image series. The detected mask was used to extract LV intensity curves. Both mask location and AIF curves were sent back to scanner dicom database. Figure 6 gives a screen snap of MR scanner with detected AIF mask and intensity curves.

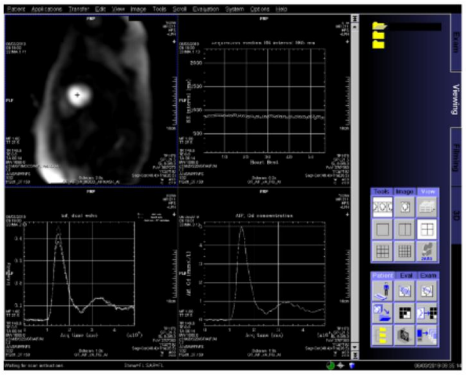

Figure 6. AIF LV mask detection on the MR scanner. This is a screen snap of clinical cardiac perfusion scan, using the trained NN model in this study. The upper left panel showed a AIF image with mask center marked. The lower left and right panels showed AIF intensity curves and converted contrast concentration curve.

Results

Figure 7a gives the cost vs. the number of iterations for a typical training session (3 class segmentation, 3CS). The ADAM optimizer was very efficient to drive down the total cost and increase the detection accuracy. After

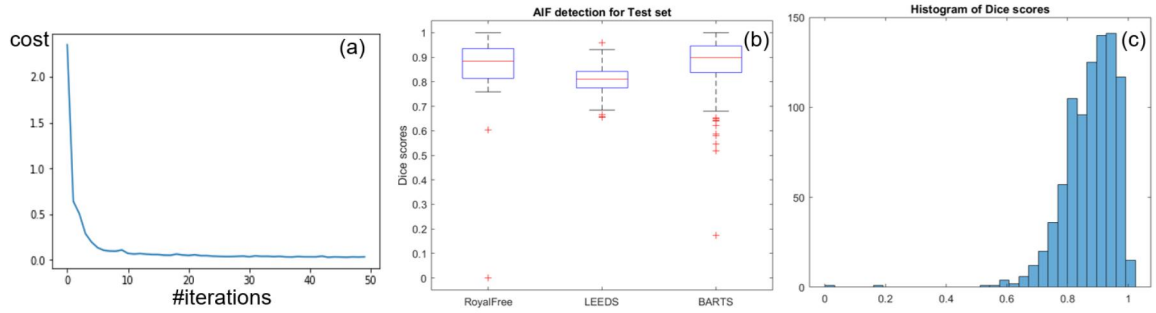

Figure 7. AIF LV detection results with 3CS. (a) Cost vs. number of iterations; (b) Box plot of dice scores of Test set. (c) Histogram of dice scores. All test cases but two were detected with dice being higher than 0.5 and 93.1% cases have dice higher than 0.75.

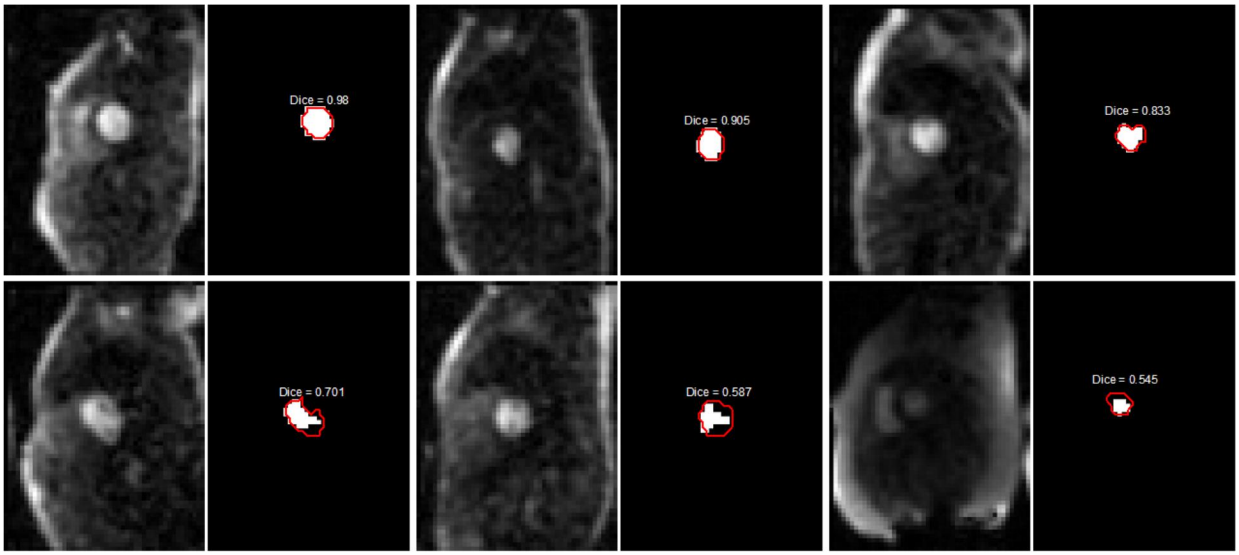

Figure 8. Examples of AIF LV detection using NN. The binary mask is the ground-truth established with boot-strapping labelling strategy. The red contour overlay is the AI detection results.

hyperparameter search, best performance was achieved for 8 ResNet modules and 128 CONV filters for 3CS and 2CS. These parameters were selected for all following performance measurement. The multi-class segmentation outperformed binary segmentation. For the 2CS, success rate for Tra and Dev sets were 99.865% and 99.808% (4 failed detection). For the 3CS, the success rate on Tra set was higher to be 99.956%. Only 10 scans out of 22,941 cases gave dice ratio <0.5. The Dev set gave success rate of 99.856% (3 unsuccessful detection out of 2,086 scans). 3CS model successfully detect LV for 99.77% of all test set cases (878 out of 880 cases). Figure 7b gives the box plot for the Test set of three hospitals. Majority of test cases have high dice scores (93.1% cases with dice higher than 0.75). Figure 8 gives examples of AIF detection with different dice scores from Test set. While majority has very good detection, those low dice cases still locate the LV blood pool successfully, with some papillary muscle included in the mask.

Conclusion and Future Work

A neural net model was developed for the automated detection of LV blood pool in the arterial input function series of cardiac MR perfusion imaging. This model was based on the U-net architecture and utilized ResNet module as building block. A combined cross-entropy and log IoU was minimized as loss function. To achieve very high accuracy required by stress perfusion study, a large data cohort of 25,027 perfusion scans were collected for training and development, together with an independent test set. A boot-strap strategy was developed to speedup data labeling on such large cohort. Final detection rate after hyperparameter tuning was 99.956%, 99.856% and 99.77% for Tra, Dev and test sets. Next steps will continuously collect more testing datasets and monitor inference performance of trained models. Failed cases will be added to future training set and repeat offline and inline tests on bigger test set.

Reference

1. Pennell DJ. Cardiovascular Magnetic Resonance. Circulation. 2010;121:692–705.

2. Coelho-Filho OR, Rickers C, Kwong RY, Jerosch-Herold M. MR myocardial perfusion imaging. Radiology. 2013;266:701–15.

3. Kellman P, Hansen MS, Nielles-Vallespin S, Nickander J, Themudo R, Ugander M, et al. Myocardial perfusion cardiovascular magnetic resonance: optimized dual sequence and reconstruction for quantification. Journal of Cardiovascular Magnetic Resonance; 2017;19:43.

4. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2015;3431–40.

5. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Lect. Notes Comput. Sci. 2015;9351:234–41.

6. Badrinarayanan V, Kendall A, Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 2017;39:2481–95.

7. Alom MZ, Hasan M, Yakopcic C, Taha TM, Asari VK. Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation. https://arxiv.org/abs/1802.06955.

8. Zhang Z, Liu Q, Wang Y. Road Extraction by Deep Residual U-Net. IEEE Geosci. Remote Sens. Lett. 2018;1–5.

9. Shvets A, Rakhlin A, Kalinin AA, Iglovikov V. Automatic Instrument Segmentation in Robot-Assisted Surgery Using Deep Learning. 2018 17th IEEE Int. Conf. Mach. Learn.

10. Bai W, Sinclair M, Tarroni G, Oktay O, Rajchl M, Vaillant G, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. Journal of Cardiovascular Magnetic Resonance; 2017;1–12.

11. Fahmy AS, El-Rewaidy H, Nezafat M, Nakamori S, Nezafat R. Automated analysis of cardiovascular magnetic resonance myocardial native T1 mapping images using fully convolutional neural networks. Journal of Cardiovascular Magnetic Resonance; 2019;21:1–12.

12. Bratt A, Kim J, Pollie M, Beecy AN, Tehrani NH, Codella N, et al. Machine learning derived segmentation of phase velocity encoded cardiovascular magnetic resonance for fully automated aortic flow quantification. Journal of Cardiovascular Magnetic Resonance; 2019;21:1–11.

13. Xue H, Zuehlsdorff S, Kellman P, Arai A, Nielles-Vallespin S, Chefdhotel C, et al. Unsupervised inline analysis of cardiac perfusion MRI. Lect. Notes Comput. Sci. 2009;5762:741–9.

14. Hansen MS, Sørensen TS. Gadgetron: An open source framework for medical image reconstruction. Magn. Reson. Med. 2013;69:1768–76.