Medical image segmentation automatic quality control: A multi-dimensional approach

Joris Fournel, Axel Bartoli, David Bendahan, Maxime Guye, Monique Bernard, Elisa Rauseo, Mohammed Khanji, Steffen Petersen, Alexis Jacquier, Badih Ghattas

Medical Image Analysis (2021)

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Medical image segmentation automatic quality control: a multi-dimensional approach

Joris Fournel[a], Axel Bartoli[a], David Bendahan[b], Maxime Guye[b], Monique Bernard[b], Elisa Rauseo[d,e], Mohammed Y. Khanji[f,d,e], Steffen E. Petersen[d,e,g,h], Alexis Jacquier[a], Badih Ghattas[c]

[a] *Department of Radiology, Hôpital de la Timone Adultes, A.P.H.M. 264, rue Saint-Pierre 13385 Marseille Cedex 05, France*
[b] *C.N.R.S., C.R.M.B.M., Medical Faculty, Aix-Marseille University, 27, Boulevard Jean Moulin, 13385 Marseille Cedex 05, France*
[c] *Aix Marseille Univ, CNRS, I2M, Marseille, France*
[d] *William Harvey Research Institute, NIHR Barts Biomedical Research Centre, Queen Mary University London, Charterhouse Square, London, EC1M 6BQ, UK*
[e] *Barts Heart Centre, St Bartholomew's Hospital, Barts Health NHS Trust, West Smithfield, EC1A 7BE, London, UK*
[f] *Department of Cardiology, Newham University Hospital, Barts Health NHS Trust, Glen Road, London E13 8SL, UK*
[g] *Health Data Research UK, London, UK*
[h] *Alan Turing Institute, London, UK*

ARTICLE INFO

Article history:
Received ?
Received in final form ?
Accepted ?
Available online ?

Communicated by ?

2000 MSC: 41A05, 41A10, 65D05, 65D17

Keywords: Medical Image Segmentation Automatic Quality Control, Multi-dimensional quality control, CMR image segmentation, Deep Learning

ABSTRACT

In clinical applications, using erroneous segmentations of medical images can have dramatic consequences. Current approaches dedicated to medical image segmentation automatic quality control do not predict segmentation quality at slice-level (2D), resulting in sub-optimal evaluations. Our 2D-based deep learning method simultaneously performs quality control at 2D-level and 3D-level for cardiovascular MR image segmentations. We compared it with 3D approaches by training both on 36540 (2D) / 3842 (3D) samples to predict Dice Similarity Coefficients (DSC) for 4 different structures from the left ventricle, i.e., trabeculations (LVT), myocardium (LVM), papillary muscles (LVPM) and blood (LVC). addReviewerThe 2D-based method outperformed the 3D method. At the 2D-level, the mean absolute errors (MAEs) of the DSC predictions for 3823 samples, were 0.02, 0.02, 0.05 and 0.02 for LVM, LVC, LVT and LVPM, respectively. At the 3D-level, for 402 samples, the corresponding MAEs were 0.02, 0.01, 0.02 and 0.04. The method was validated in a clinical practice evaluation against semi-qualitative scores provided by expert cardiologists for 1016 subjects of the UK BioBank. Finally, we provided evidence that a multi-level QC could be used to enhance clinical measurements derived from image segmentations.

1. Introduction

Manual segmentation of medical images is time consuming and subject to inherent between- and within-operator variabilities. Artificial intelligence (AI)-based methods have been developed with the aim of reducing the amount of time dedicated to the segmentation task and speeding up the segmentation process without compromising the segmentation accuracy (Litjens et al., 2017).

In clinical practice, quantification derived from image segmentation is often used for diagnostic and prognostic purposes. In clinical research, image segmentation is a primary tool used to validate clinical hypotheses. The quality

e-mail: jorisfournell@gmail.com (Joris Fournel)

control (QC) of segmentations is thus critical in order to validate the robustness of the corresponding methods.

On that basis, Medical Image Segmentation Automatic Quality Control (MISAQC) methods have been growingly developed over the last few years. MISAQC methods intend to automatically evaluate the quality of a given segmentation and to track and potentially discard the segmentations which do not fulfill the quality criteria.

The quality of AI-based segmentation can be assessed using various metrics such as Dice Similarity Coefficient (DSC), Hausdorff distance, volume/mass error, etc ... This quality is considered to be good if these indices lie within a commonly acceptable range in the corresponding field. The computation of those metrics is based on a comparative analysis between the predicted segmentation and the ground truth manual segmentation. On the contrary, MISAQC methods aim at predicting a quality metric (e.g. DSC) for a given image and its predicted segmentation, and thus do not use the ground truth manual segmentation.

So far, very few approaches have been reported regarding MISAQC ((Albà et al., 2017; Audelan and Delingette, 2019; Kohlberger et al., 2012; Valindria et al., 2017)). A regression model was used to assess the quality for automated segmentation of lung and liver images (Kohlberger et al., 2012). The DSC was used as a QC index and variable correlations between the predicted and real DSCs for the lung (r=0.85) and the liver (r=0.54) have been reported. (Audelan and Delingette, 2019) used an unsupervised bayesian approach and reported good correlations between the real and predicted DSC scores for brain tumors (r=0.69) and the left ventricle myocardium (r=0.78).

Furthermore, (Robinson et al., 2019) showed that the reverse classification accuracy (RCA, (Valindria et al., 2017)) approach could provide good DSC predictions without the need of a large annotated dataset. However, the time required for the RCA analysis of a single segmentation (11 min) was considered prohibitive for real-time applications (Robinson et al., 2018). For the segmentation of cardiac magnetic resonance images, (Robinson et al., 2018) used a 3D convolutional neural network (CNN) in order to predict the DSC values of 3D segmentations. The corresponding results were very promising in terms of predictive accuracy, with mean absolute errors (MAEs) between the real and predicted DSCs of 0.03 ± 0.04 for the whole heart and a very short processing time (less than a second).

However, the segmentation quality assessment was based on the 3D DSC and this approach can actually be questioned. The 3D DSC is a volume-related information that is very useful to evaluate the global quality of a segmentation but it does not contain any specific information related to the individual segmented slices which are part of the volume of interest. In other words, errors cannot be accurately localized using a 3D DSC-based method. Significant segmentation errors can occur at the 2D level and be fully ignored at a 3D level thereby overestimating the segmentation quality. The localization of segmentation errors at a slice-level could offer the opportunity for corrective processes thereby providing more accurate volumetric measurements and potentially improved diagnosis. On that basis, we hypothesized that a MISAQC tool would largely benefit from both 3D and 2D evaluations.

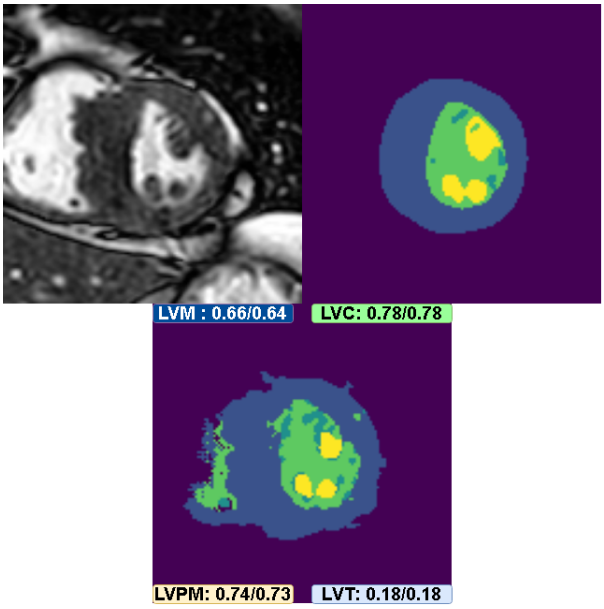


Fig. 1. First row: a CMR image from our testset and related manual segmentation; the four left ventricle structures are depicted illustrating the difference in size and shape for the left ventricle trabeculations (LVT) and left venctricle papillary muscles (LVPM) structures compared to left ventricle myocardium (LVM) and left venctricle cavity (LVC). Second row: a related automated segmentation to evaluate; the displayed scores represent for each class the real DSC on the left and the predicted DSC with our method on the right.

In the present paper, our goal was to present a new MISAQC tool having the following properties:

- Easily trainable using only 2D CNNs.
- Provides quality assessment at 2D-level, thus localizes slices where the segmentation is erroneous.
- Helps to automatically correct clinical measurements biases induced by erroneous segmentations.
- Provides quality assessment at 3D-level from a mathematical combination of 2D-level predictions.

2. Materials and Methods

2.1. Study design

This multicenter retrospective study was approved by the local institutional review board (N ° IRB CRM -1907-02è) in accordance with the guidelines outlined in the Declaration of Helsinki.

A CNN-based method was developed for the automated assessment of segmentation quality for the LV in the end-diastolic (ED) phase from CMR images (Bartoli et al., 2020).

The initial dataset included 4290 images from 449 different subjects. Patients were classified as healthy, dilated, hypertrophic, hypertrabeculation by the most experienced investigator (A.J. – 20 years of experience). Healthy refers to a normal cardiovascular examinations. Dilated indicated a dilated cardiomyopathy (DCM) as previoulsy described (Japp et al., 2016). Hypertrophic refers to hypertrophic cardiomyopathy (HCM) which was defined based on the myocardial thickness on the ED short-axis (Elliott et al., 2014). Excessive trabeculation cardiomyopathy (ETCM) was defined according to (Petersen et al., 2005) criteria i.e. a double-layer myocardium aspect and a non-compacted layer to compacted layer thickness > 2.3 on ED long-axis. In our dataset, 259 subjects were healthy subjects, 39 were DCM, 92 were HCM and 59 were ETCM. In Table 1 are displayed the patient characteristics for this initial dataset.

2.2. CMR Data

The CMR examinations were performed at 1.5T using 2 different scanners i.e. Ingenia ® (Philips Health System, Best, the Netherlands) and Avanto ® (Siemens Healthcare, Erlangen, Germany). Patients were positioned supine in the scanner with a multi-channel body array coil positioned on the top and a spine array coil positioned on the bottom. MR acquisition was gated to ECG and occurred during an inspiratory breath hold. A balanced turbo field-echo sequence and a balanced steady-state-free precession sequence was used with the Ingenia and Avanto scanners respectively. In both cases, images were acquired in the short-axis view in order to cover the LV from base to apex. The following parameters were used : in-plane resolution : 1.5 mm^2 (Ingenia) and 2.35 mm^2,(Avanto) slice thickness : 7-8 mm, gap between slices :10 mm, flip angle : 30. The short-axis image stack consisted in 8 to 17 slices depending on the scanner, patient height, cardiac anatomy and morphology. The ED frame at each imaging level was retained for further image analysis. Data of each patient was de-identified before the analysis.

2.3. Manual segmentation and reference measures

Manual image segmentation was undertaken by a trained observer (A.B., who has 5 years of experience) using a post-processing software previously validated by Bricq et al. (Bentatou et al., 2018; Bricq et al., 2015; Frandon et al., 2018). The LV structures were manually segmented so as to obtain four labels i.e: blood cavity (LVC), myocardium (LVM), papillary muscles (LVPM) and trabeculations (LVT). The non segmented part of each image was assigned to a fifth label: background (BG). The corresponding results (Manual) were used as the reference standard.

3. Methods

The aim of a MISAQC approach is to learn to predict the quality of a segmentation. In order to train a supervised model for that purpose, we first need to build a dataset containing couples of images/segmentations with varying qualities for the segmentations as shown in figure 2. Once this dataset is built, different CNN models are trained and compared for the 2D and 3D DSC prediction. Our approach (illustrated in figure 3) consists of two steps. First, we predict 2D indices of quality for each segmented slice. Second, we use the 3D DSC formula to estimate its components using the 2D indices predicted in the first step. We compare our approach to a direct 3D approach.

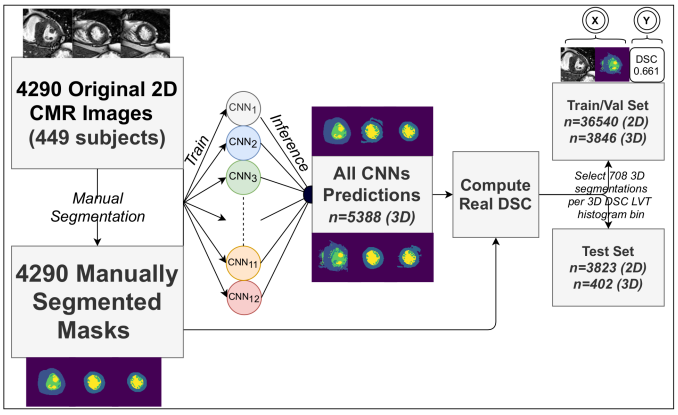


Fig. 2. Illustration of MISAQC dataset construction. First, 4290 cardiac MR images were segmented by an expert and used to train segmentation models in order to produce segmentations of varying quality of the 4290 MR images. Then, the DSC values were computed for those segmentations based on the ground truth manual segmentations. Finally, LVT DSC-based subsampling was applied before splitting the data into training/validation and test sets.

3.1. MISAQC dataset construction

To generate the MISAQC dataset, we used an approach similar to the one proposed by Robinson et al. (Robinson et al., 2018) illustrated in figure 2. Twelve versions of a same CNN model (see (Bartoli et al., 2020) for details about the architecture) differing by the number of feature maps per convolutional layer, the proportion of the original dataset used for training, and the number of training epochs are implemented. The three hyperparameters of the CNNs were selected randomly, resulting in models producing segmentations of variable qualities. To train the CNNs we used the 4290 2D CMR images for which manually segmented masks were available. Each of these models was then used in inference to generate segmentations for the whole original dataset. The corresponding 2D and 3D DSC scores were computed for all the segmentations. Thus for each original 2D image, twelve segmentations are available together with their corresponding DSC. This resulted in a total of 5388 3D segmentations, corresponding to 51480 2D segmentations.

The next step guarantees that the obtained dataset is balanced with respect to the quality of the segmentations. Therefore, we sample from this artificial dataset triplets (image, mask, DSC) stratifying over LVT 3D DSC distribution, using the following bins: [0, 0.2], [0.2, 0.3], [0.3,

Table 1. Baseline Patient Characteristics for the initial dataset.

Characteristics					
Parameter	HCM	DCM	ETCM	Healthy	Overall
Total no.	92	39	59	259	449
Age (y)	53 ± 15	54 ± 12	44 ± 16	40 ± 17	45 ± 17
No. men	59	21	37	147	264
LVEDV (mL)	151.5 ± 38.1	226.5 ± 86.7	189.0 ± 90.3	146.4 ± 42.3	160.0 ± 60.3
LVEDV-to-BSA (mL/m²)	80.0 ± 17.4	118.8 ± 41.0	101.4 ± 39.0	80.5 ± 19.7	86.5 ± 27.8
LVMM (g)	169.8 ± 53.8	163.9 ± 48.1	124.3 ± 66.8	113.9 ± 34.3	131.0 ± 51.2
LVMM-to-BSA (g/m²)	89.6 ± 26.2	86.2 ± 22.4	66.6 ± 29.2	62.3 ± 14.7	70.5 ± 23.5
Trabeculation mass (g)	11.3 ± 5.8	17.0 ± 10.0	18.9 ± 13.3	9.5 ± 4.7	11.7 ± 7.9
Trabeculation mass-to-BSA (g/m²)	6.0 ± 3.0	8.8 ± 4.7	10.1 ± 6.3	5.2 ± 2.4	6.3 ± 3.9
Trabeculation mass-to-TMM (%)	6.0 ± 2.4	8.6 ± 3.8	12.7 ± 5.2	7.1 ± 2.6	7.6 ± 3.7

0.4], [0.4, 0.5], [0.5, 0.6], and [0.7, 1]. The number of samples taken from each interval is the minimum number observed within these intervals. Hence, 708 3D segmentations were randomly selected in each bin, resulting in a total of 4248 3D segmentations, corresponding to 40 363 2D segmentations. The final dataset was randomly split into training/validation (90%) and testing (10%) samples, stratifying over the LVT 3D DSC for both.

For each segmentation, 5 one-hot-encoded masks were generated, one for each class (BG, LVM, LVC, LVT and LVMP), and concatenated with the CMR input. Segmentations and CMR inputs were previously cropped around the LV (see (Bartoli et al., 2020) for more details) and the CMR inputs were pre-processed with Contrast Limited Adaptive Histogram Equalization (CLAHE). The 2D inputs size was 128x128x6 while the 3D inputs were resized to a 128x128x8x6 using nearest neighbour interpolation in order to account for the variable number of slices per subject. The 2D inputs are used for our method and the 3D inputs for the other approaches for comparative purposes. This process resulted in a total of 36540 2D and 3840 3D samples for training and validation. The testing phase was performed on 3823 2D and 402 3D samples.

3.2. 3D DSC as a function of 2D indices

We will show how the 3D DSC may be recovered from the 2D DSC and the mean volume similarity fraction (MVSF). Let Y and Z be two 3D regions (Y the ground-truth 3D segmentation and Z the corresponding automated segmentation) composed of n 2D stacks, or equivalently:

$$Y = (y_i)_{i\in[\![1,n]\!]}$$

and:

$$Z = (z_i)_{i\in[\![1,n]\!]}$$

where y_i and z_i are 2D regions. The 3D DSC and MVSF between Y and Z are defined as:

$$DSC(Y,Z) = \frac{2|Y \cap Z|}{|Y| + |Z|} \quad (1)$$

$$MVSF(Y,Z) = \frac{2(|Y| - |Z|)}{|Y| + |Z|} \quad (2)$$

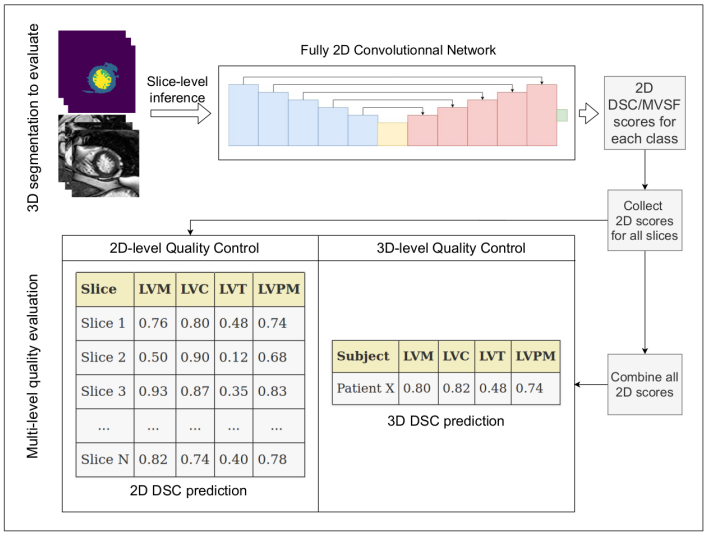


Fig. 3. Illustration of the methodology at inference time. For a given segmented MRI to evaluate, the inference is done slice-wise by sequentially providing the segmentation and related image as the input of the two models. The predicted 2D DSCs are stored and provide a quality evaluation for all slices and classes. When slice-wise inference is over, the stored predictions are combined to compute the 3D DSC prediction for all classes. Errors can be detected at every scale.

Similarly, let DSC(y_i,z_i) and MVSF(y_i,z_i) be the DSC and MVSF between the 2D slices y_i and z_i. DSC(Y,Z) is used as the quality index of the subject-level automated segmentation Z, while DSC(y_i,z_i) serves the same purpose for the slice-level segmentation z_i. The goal is to predict these metrics in the absence of the ground truth, Y.

Each example from the learning sample was composed of an MR image (x_i), the corresponding automated (z_i) and ground-truth segmentation (y_i). In order to evaluate the quality of Z, we wish to predict both scores (DSC and MSVF) relative to each pair $(y_i, z_i)_{i\in[\![1,n]\!]}$.

A model, denoted by M, is trained such that:

$$M(x_i, z_i) = [\widehat{DSC}(y_i, z_i), \widehat{MVSF}(y_i, z_i)]$$

This model can be used to predict, for each new pair consisting of an MR image and its automated segmentation,

a value for both scores without requiring the ground-truth segmentation. We will now see how the recovery of the 3D DSC can be achieved. Considering the expression for the 2D DSC:

$$DSC(y_i, z_i) = \frac{2|y_i \cap z_i|}{|y_i| + |z_i|} \tag{3}$$

we obtain the following estimation:

$$\widehat{|y_i \cap z_i|} = \frac{1}{2}\widehat{DSC(y_i, z_i)}(\widehat{(|y_i|} + |z_i|) \tag{4}$$

And considering the 2D MVSF expression:

$$MVSF(y_i, z_i) = \frac{2(|y_i| - |z_i|)}{|y_i| + |z_i|} \tag{5}$$

we can estimate $|y_i|$:

$$\widehat{|y_i|} = |z_i|\frac{2 + \widehat{MVSF(y_i, z_i)}}{2 - \widehat{MVSF(y_i, z_i)}} \tag{6}$$

The 3D DSC predictions can be recovered as follows:

$$\widehat{DSC(Y, Z)} = \frac{2\widehat{|Y \cap Z|}}{\widehat{|Y|} + |Z|} \tag{7}$$

Where $\widehat{|Y|} = \sum_{i \in [\![1,n]\!]} \widehat{|y_i|}$, $|Z| = \sum_{i \in [\![1,n]\!]} |z_i|$ and $\widehat{|Y \cap Z|} = \sum_{i \in [\![1,n]\!]} \widehat{|y_i \cap z_i|}$.

3.3. CNN architectures for DSC prediction

To implement our approach, we suggest a CNN architecture for the prediction of both 2D DSC and MVSF. We use then the formulas from the previous section to obtain the 3D DSC values. In order to show that the performance of our approach is not dependant on a specific architecture, two types of backbone architectures were used i.e QC ResNet and QC U-Net (figure 4). The rationale behind the QC U-Net backbone type was to build an architecture that would implicitly reproduce the steps performed by a human expert to produce a DSC for a given segmentation and related medical image. The expert would (1) segment the image manually and then (2) compute the DSC between his segmentation (considered as ground truth) and the segmentation to evaluate. Hence, we considered the image segmentation task as a necessary step in the process of a DSC computation task for a human expert. Having this in mind, the QC U-Net backbone architecture was designed as a two-stages process, i.e., a U-Net encoder-decoder block (originally designed for image segmentation) augmented with two layers aiming at directly predicting the DSC values. Based on empirical preliminary measurements, the QC ResNet version was designed with a fewer number of features maps than in the original ResNet-50 (He et al., 2016). Finally, to compare our approach to methods directly predicting the 3D DSC (Robinson et al., 2018), we used similar networks adapted for 3D data structure.

The 2D-based method was identified as $2D_R$, where R stands for reconstruction given that the 3D DSCs were reconstructed from 2D indices. For each class, the $2D_R$ method performed a multi-level quality control: it predicted a 2D DSC for each segmented slice together with a reconstructed 3D DSC for the whole volume.

For each implementation of our $2D_R$ method, 2D images and the corresponding segmentations were taken as inputs while a 2D CNN was trained to predict both the 2D DSC and the 2D MVSF scores for each class (BG, LVM, LVC, LVT and LVPM). On that basis, the last layer was composed of ten units: five for the 2D DSC scores and five for the 2D MVSF scores. For the 3D direct approach, taking 3D images and their segmentations as inputs, networks were trained to predict the five 3D DSC resulting in five units for the last layer. As indicated in figure 4, the remaining differences between the 2D and 3D networks for a given backbone architecture type were the kernel size of the convolutional layers, the pooling size in pooling layers and the input shape.

3.4. Implementation details

Each 2D network was trained for 80 epochs (until no significant gain was observed on the validation set) using Adam optimizer at a learning rate of 0.001. Similarly, each 3D network was trained for 50 epochs using Adam optimizer at a learning rate of 0.001. For each network the loss function was the mean squared error (MSE). No data augmentation was used.

3.5. Evaluation

3.5.1. Slice-level evaluation

The performances of the $2D_R$ method implemented with each backbone (QC ResNet and QC U-Net) were initially assessed at the slice-level using 3823 segmented 2D images of the test dataset. For each class, the evaluation was carried out on the basis of the mean absolute error (MAE) which was calculated between the predicted 2D DSCs and the ground-truth values.

For each class, the accuracy indicated the discriminative capacity of a given method to distinguish "bad quality" from "good quality" 2D segmentations based on the predicted 2D DSC values. To compute the accuracy, predicted and observed DSC values were encoded as "bad quality" or "good quality" according to the following thresholds: 0.7 for LVM and LVC according to (Robinson et al., 2018). Considering that LVPM and LVT (as illustrated in figure 5) are prone to lower DSC values (Bartoli et al., 2020) (as illustrated in figure 5), a lower threshold was used, i.e., 0.45 for LVPM and 0.35 for LVT.

Pearson correlation coefficients were also computed in order to further evaluate the performances at the slice-level.

3.5.2. Subject-level evaluation

The performance of the $2D_R$ method was similarly evaluated for each backbone at a 3D-level using the 402 segmented 3D images of the testing dataset. The correspond-

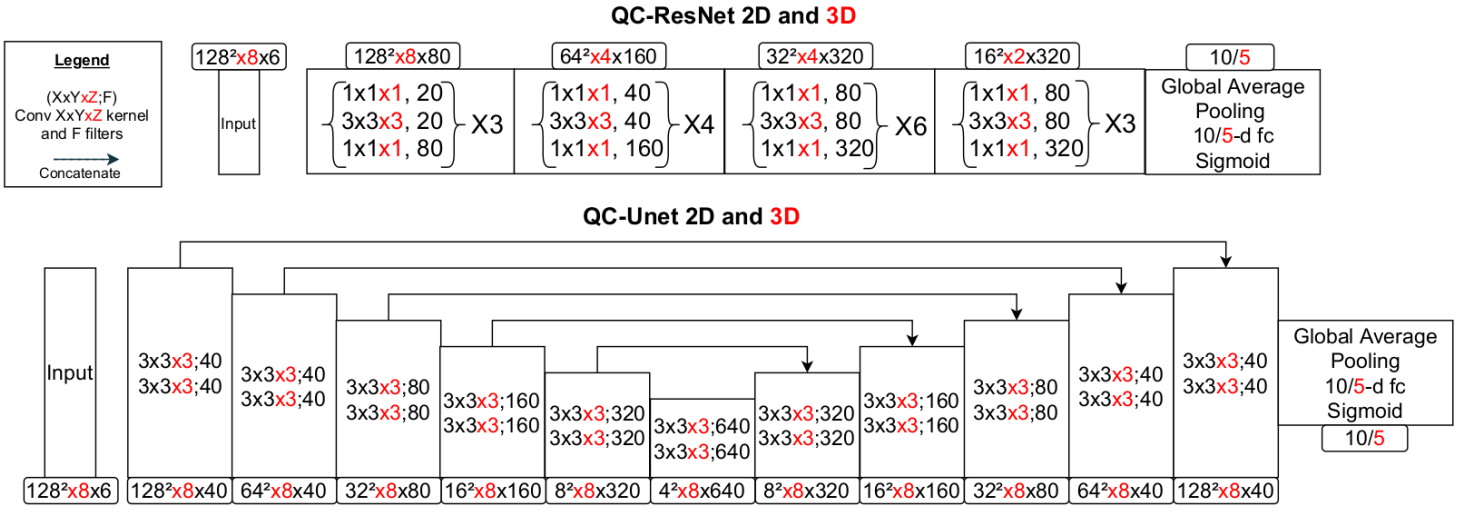


Fig. 4. Display of the 4 used architectures, for each backbone type the numbers in red are only present for the 3D version. Downsampling was always performed with a max pooling layer while the upsampling (only in QC U-Net) was performed with transposed convolutions. The pooling size can be deduced from the difference in dimensions between consecutive layers. Transposed convolution kernels were always 2x2 (2x2x1 in 3D version). QC ResNet 2D and 3D had 2 and 3 million parameters respectively, while QC Unet 2D and 3D had 19 and 21 million parameters respectively.

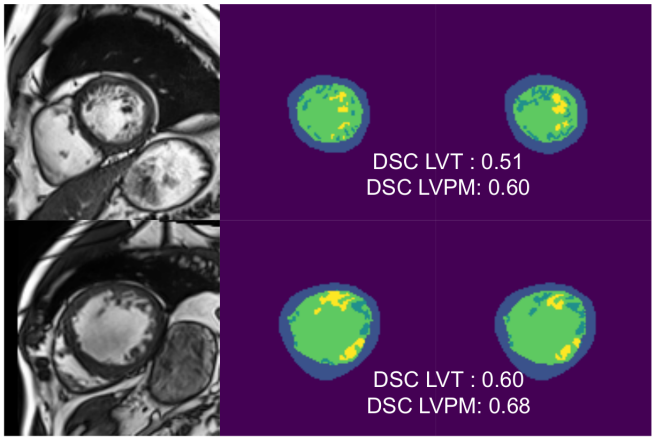


Fig. 5. Each row displays an CMR image and a pair of related segmentations produced (with a delay of one month between each segmentation) by the same human operator (Bartoli et al., 2020), the 2D DSC between the two segmentations is displayed for left ventricle trabeculations (LVT) and left ventricle papillary muscles (LVPM). The nature of those anatomical regions and appearance on a CMR image justify a lower quality DSC threshold than for LVM and LVC.

ing results were compared to those obtained with a 3D direct approach.

3.5.3. Clinical practice evaluation

Our trained $2D_R$ QC U-Net model was communicated to the research team of Steffen Petersen that used it to control the quality of 1016 segmented subjects from the UK BioBank. For this dataset, trained cardiologists also assigned a semi-qualitative quality control score (QC 1=Good, QC 2=Sub-optimal but still usable, QC 3=Poor) for each class and slice. At the 3D level, QC 2 and QC 3 are merged and patient is classified as QC 2/3 if at least one slice belongs to either QC 2 or QC 3. The agreement between these human scores and our model predictions was assessed.

3.6. Statistical tests

Statistical comparison were performed (for mean differences) using Wilcoxon signed-rank tests for paired samples, unpaired t-test for unpaired samples. Two-sample Kolmogorov-Smirnov test was used for testing whether two independent samples are drawn from the same continuous distribution. Differences were considered as significant for p-values lower than 0.05.

4. Results

4.1. Slice-level results

The comparative analysis between the different backbones is summarized in Table 2. As illustrated in Table 2, regardless of the backbone and for the whole set of classes but LVT, the MAEs computed at the slice-level i.e. 2D were systematically lower than 0.023. The MAE related to the LVT class was slightly larger i.e. 0.047.

For the QC U-Net backbone, the MAE associated with the $2D_R$ method was 0.005 for the BG 2D DSC prediction, 0.023 for LVM, 0.019 for LVC, 0.047 for LVT and 0.020 for LVPM. Of note the largest error was found for a small and scattered class i.e. LVT.

When comparing backbones, MAEs values were found significantly smaller for the QC U-Net $2D_R$ for all the classes but LVT.

Accuracies in distinguishing bad from good quality slice-level segmentations were larger than 96.1% for 6 out of the 8 tested classes and larger than 92.7% for the LVT class.

Table 2. DSC MAEs, accuracies and correlation coefficients at slice-level for each $2D_R$ version.

MAE				
Method	LVM	LVC	LVT	LVPM
QC ResNet $2D_R$	0.025 ± 0.027	0.021 ± 0.025	0.047 ± 0.048	0.023 ± 0.044
QC U-Net $2D_R$	**0.023 ± 0.024**	**0.019 ± 0.022**	0.047 ± 0.051	**0.020 ± 0.040**
p-values	3.62e-05	2.18e-07	0.47	6.01e-30

Accuracy and correlation								
	LVM		LVC		LVT		LVPM	
Method	Acc.	Corr.	Acc.	Corr.	Acc.	Corr.	Acc.	Corr.
QC ResNet $2D_R$	96.1 %	0.991	97.3%	0.992	**92.8%**	**0.960**	98.5%	0.991
QC U-Net $2D_R$	**96.6 %**	**0.993**	**97.4 %**	**0.994**	92.7 %	0.957	**98.8 %**	**0.993**

Table 3. DSC MAEs at subject-level for each backbone and method. Regardless of the backbone the $2D_R$ method consistently outperforms the equivalent 3D direct approach.

QC ResNet backbone					
Method	BG	LVM	LVC	LVT	LVPM
3D	0.011 ± 0.022	0.037 ± 0.035	0.029 ± 0.039	0.044 ± 0.038	0.055 ± 0.049
$2D_R$	**0.004 ± 0.005**	**0.016 ± 0.028**	**0.011 ± 0.016**	**0.022 ± 0.022**	**0.042 ± 0.053**
p-values	9.96e-30	2.47e-34	3.02e-31	5.55e-25	5.12e-07

QC U-Net backbone					
Method	BG	LVM	LVC	LVT	LVPM
3D	0.010 ± 0.018	0.036 ± 0.035	0.026 ± 0.031	0.045 ± 0.041	0.042 ± 0.039
$2D_R$	**0.004 ± 0.006**	**0.016 ± 0.028**	**0.012 ± 0.017**	**0.023 ± 0.023**	**0.040 ± 0.054**
p-values	2.42e-24	6.67e-34	7.57e-26	2.08e-21	0.02

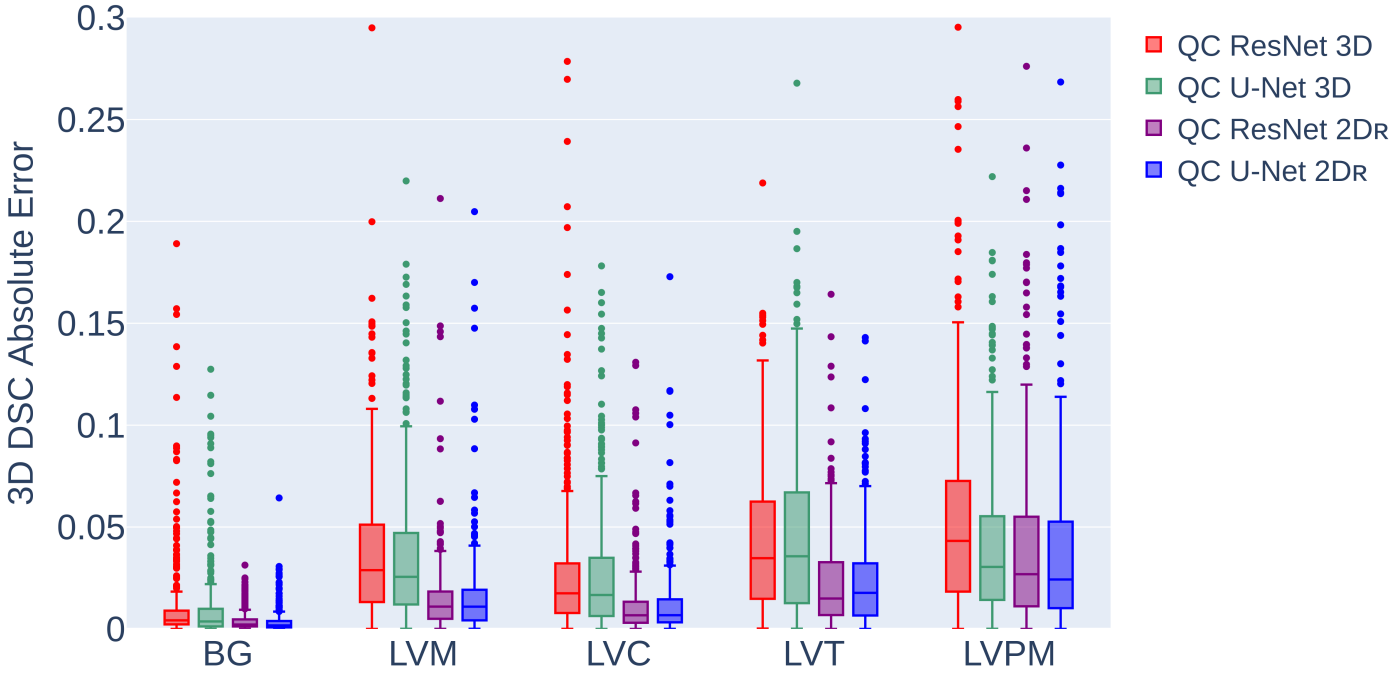


Fig. 6. Boxplots of the 3D DSC absolute errors (AE) for each class and method.

The correlation coefficients between the ground-truth and the predicted 2D DSC values were larger than 0.957, with the lowest value for QC U-Net LVT (0.957) and the largest for QC U-Net LVC (0.994).

Table 4. Accuracies and correlation coefficients at subject-level for backbone and method.

QC ResNet backbone								
	LVM		LVC		LVT		LVPM	
Method	Acc.	Corr.	Acc.	Corr.	Acc.	Corr.	Acc.	Corr.
3D	92.7 %	0.983	94.7 %	0.977	90.7 %	0.955	96.5 %	0.961
$2D_R$	**97.7%**	**0.992**	**97.0%**	**0.996**	**97.2%**	**0.987**	**97.2%**	**0.965**
QC U-Net backbone								
	LVM		LVC		LVT		LVPM	
Method	Acc.	Corr.	Acc.	Corr.	Acc.	Corr.	Acc.	Corr.
3D	94.4 %	0.982	96.0 %	0.986	89.7 %	0.951	97.7 %	**0.971**
$2D_R$	**97.5 %**	**0.992**	**97.0 %**	**0.996**	**96.5 %**	**0.987**	97.7 %	0.965

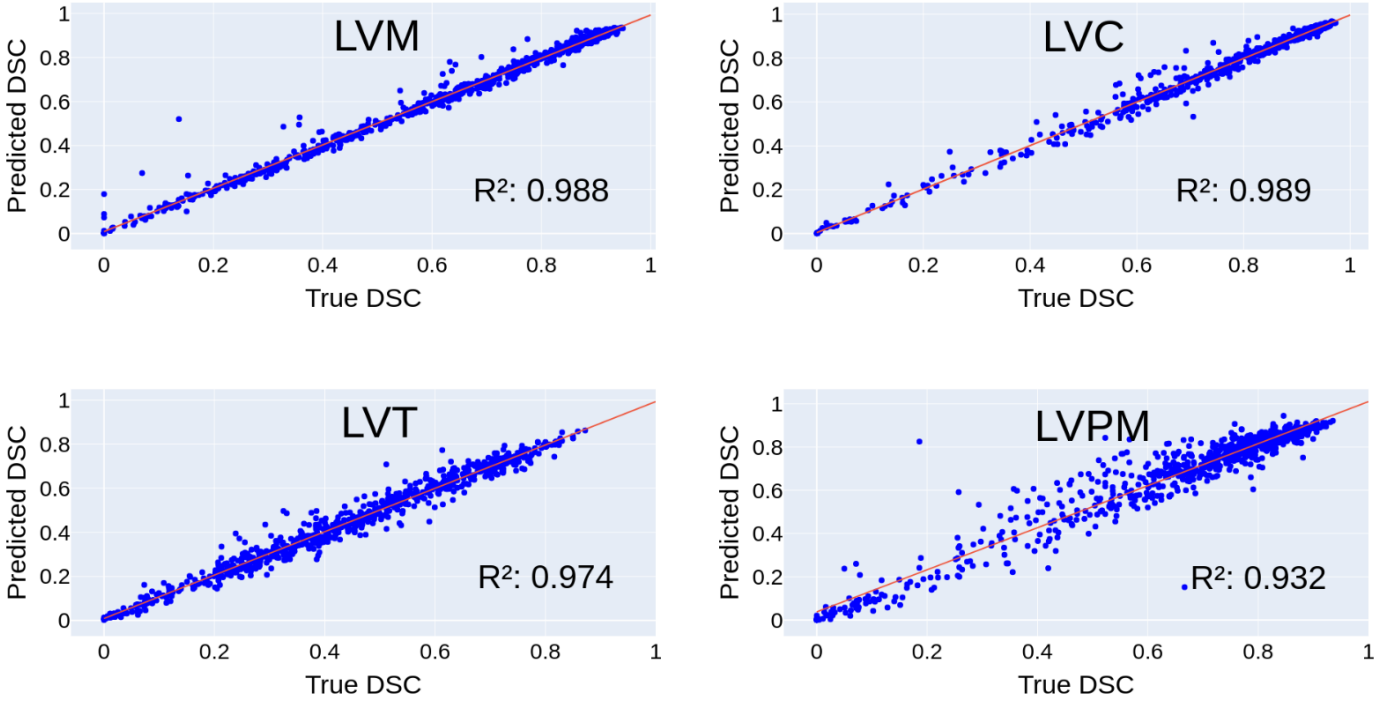


Fig. 7. Scatter plots of the true 3D DSC values against the predicted values for all classes except BG. The R^2 scores and red lines are derived from a fitted regression model. The model used here is QC U-Net $2D_R$.

Table 5. MAE of the predicted 3D DSC with the QC U-Net $2D_R$ method with regard to the patient classification.

MAE				
Group	LVM	LVC	LVT	LVPM
DCM (n=30)	0.014 ± 0.016	0.010 ± 0.008	0.032 ± 0.032	0.056 ± 0.064
HCM (n=68)	0.019 ± 0.025	0.019 ± 0.025	0.022 ± 0.023	0.046 ± 0.060
ETCM (n=53)	0.024 ± 0.036	0.016 ± 0.029	0.026 ± 0.025	0.050 ± 0.090
Healthy (n=250)	0.015 ± 0.028	0.010 ± 0.010	0.022 ± 0.020	0.034 ± 0.036
Non-Healthy (n=152)	0.020 ± 0.029	0.016 ± 0.025	0.026 ± 0.026	0.049 ± 0.072
p-values Healthy/Non-Healthy	0.08	7.51e-4	0.16	8.64e-3

Overall, the segmentation quality predictions at the slice-level were characterized by low MAEs values regardless of the backbone and the class.

4.2. Subject-level results

As illustrated in Table 3 and figure 6, the MAEs computed from the ground-truth and the predicted DSC values were systematically and significantly lower for the $2D_R$ variant and so regardless of the backbone used. A high level of predictive accuracy was consistently achieved for the whole set of classes. Regarding the QC ResNet backbone, the $2D_R$ (3D) method provided a MAE of 0.004 (0.011) for the BG 3D DSC prediction, 0.016 (0.028) for LVM, 0.011 (0.016) for LVC, 0.022 (0.022) for LVT and

Table 6. Mean processing time for each network at subject-level and at slice-level additionally for $2D_R$ versions.

Inference time		
Method	Subject-wise	Slice-wise
QC ResNet 3D	100 ms	∅
QC U-Net 3D	120 ms	∅
QC ResNet $2D_R$	314 ms	32 ms
QC U-Net $2D_R$	312 ms	31 ms

0.042 (0.053) for LVPM. Regardless of the backbone, errors of the $2D_R$ method were at least twice smaller than those from the 3D direct approach and so for the whole set of classes but LVPM. The results of the corresponding statistical analysis are summarized in (Table 3). For the QC ResNet backbone, all the MAEs values were significantly smaller for the $2D_R$ version. For the QC U-Net backbone it was the case for the whole set of class but LVPM.

Accuracy values and pearson correlation coefficients for each backbone and method are summarized in Table 4. Considering the QC ResNet backbone, the accuracy values ranged from 90.7% (LVT) to 96.5% (LVPM) for the 3D method, while they ranged from 97.0% (LVC) to 97.7% (LVM) for the $2D_R$ method. The correlation coefficients ranged from 0.955 (LVT) to 0.983 (LVM) for the 3D method while they ranged from 0.965 (LVPM) to 0.996 (LVC) for the $2D_R$ method. These values were systematically superior for the $2D_R$ version as compared to the 3D (Table 4). As a matter of example, for the LVM class the accuracy related to the $2D_R$ method was 97.7% i.e. 5.0% larger than the corresponding 3D method's value (92.7%).

Similar results were found for the QC U-Net backbone and for the whole set of structures but LVPM (Table 4). The accuracy values from the two implementations of the $2D_R$ method were larger than 96.5% regardless of the evaluated class while 6 out of 8 values were lower than 96.5% for the two implementations of the 3D direct approach.

MAEs at the 3D-level were generally inferior to the corresponding values at the slice-level for the $2D_R$ method. As an example, the MAE value for the LVC class was 0.012 at the 3D-level for QC U-Net $2D_R$ and 0.019 at the 2D-level. Of note, this was not the case for the 3D direct approach for which MAEs at the 3D-level were generally inferior than those at the 2D-level. As an illustration, the MAE for the LVM (LVC) class was 0.037 (0.029) at the 3D-level for QC U-Net 3D whereas it was 0.023 (0.024) at the 2D-level for QC U-Net $2D_R$.

Scatter plots, fitted regression line and R^2 scores of the true 3D DSC against the predicted values from the QC U-Net $2D_R$ are illustrated in Figure 7. The scatter plots illustrate the capacity of the method to predict the 3D DSC with similar accuracy in the low, middle and high range DSC values. R^2 scores are significantly high for all classes (from 0.932 for LVPM to 0.989 for LVC).

Table 5 displays the 3D DSC MAE on the test dataset for the QC U-Net $2D_R$, for each patient class (DCM, HCM, ETCM, Healthy, Non-Healthy). The Non-Healthy group was composed of the patients that were not in the Healthy group. The predictions for the Healthy group generally presented smaller MAE values, but the differences were only significant for the LVC (0.010 for Healthy subjects against 0.016 for Non-Healthy, p=7.51e-4) and LVPM (0.034 against 0.049, p=8.64e-3). Overall, the method presented low MAE values for all groups and classes.

Table 6 gathers the mean processing time of each method at both 3D and 2D-levels for the $2D_R$ version. The $2D_R$ method had a 300 ms inference time while the 3D direct approach took about 100ms per inference.

Finally, in order to assess the impact of the training sample size, our $2D_R$ QC-UNet approach was trained with 10%, 33%, 66% and 100% of the total training dataset. The corresponding 3D DSC MAE for LVM, LVC, LVT and LVPM are illustrated in figure 8. One can observe a substantial 3D DSC MAE reduction when the learning dataset size increases.

4.3. Clinical practice results

Figure 9 overlays the density plots of the predicted 2D DSCs for the three semi-qualitative groups (QC 1, QC 2 and QC 3) provided by the expert for each class.

First, the distribution of the predicted 2D DSCs for QC 1 always significantly (according to Kolmogorov-Smirnov test, p-values < 0.05) differed from those of the lower quality groups (QC 2 and QC 3) and so for all classes. This was especially the case for LVM and LVC which are classes that are generally present in all slices contrarily to LVT and LVPM. When a class is absent in a segmented slice, the predicted 2D DSC is zero for that slice: this explains the small mode in zero appearing for QC 1, much more present for LVT and LVPM. Secondly, the shift between the distributions of the different groups was in the same order as the scores given by the human expert (QC 3 at the left, QC 2 in the middle and QC 1 at the right in the DSC range of values) except for the LVT class for which the QC 2 and QC 3 groups had a comparable distribution shape. Moreover, all the distributions significantly differed (QC 1 vs QC 2, QC 1 vs QC 3 and QC 2 vs QC 3) except for LVT and LVPM (for QC 2 vs QC 3) where p-values were of 0.99 and 0.64 respectively.

Figure 10 shows the corresponding density plots of the predicted 3D DSC for each class. As expected, the differences between the distributions were less important than those of 2D DSCs. However, regardless of the class, the 3D DSC distribution of the QC 2/3 group was always found to be significantly different, more shifted to the left and with

Fig. 8. 3D DSC MAE (averaged over all LV classes) as a function of the percentage of total training data used during training. The model used here is QC U-Net $2D_R$.

larger standard deviation compared to the distribution of the QC 1 group.

5. Automatic correction of segmentation model failure

We conducted an additional analysis in order to understand how erroneous segmentations in the mid-slice section could affect the 3D DSC values and the derived volumetric measurements. A simple approach is then suggested to correct the measurements in this case.

We consider the scenario of a single short-axis slice segmentation with a very poor quality in the mid-slice section. In this particular scenario, the 3D DSC value might be acceptable whereas the 3D volumetric measurement would not. In fact, measurements derived from mid-slices, where the largest section of the LV is commonly located, are likely to contribute the most to the overall volumetric quantification.
In an attempt to address this issue, we used of the CNN-predicted segmentations for the test dataset reported in (Bartoli et al., 2020), which was composed of 150 subjects. These segmentations were of high quality and the corresponding clinical quality metrics were accessible in (Bartoli et al., 2020). For each subject, the original 3D predicted segmentation was corrupted based on a random replacement of the mid-slice segmentation by another segmentation from our MISAQC dataset. As a result, we put together two sets of 150 3D segmentations : the original dataset without deterioration (w/o D), and the dataset with deteriorations (D), for which the mid-slices were systematically erroneous. The comparative analysis between the two datasets in terms of 3D DSC values and relative volume errors (RVE) allowed to quantify the effect of mid-slice segmentation errors on those two metrics and to check if the metrics were divergently affected.

Furthermore, this study provided an opportunity to investigate whether our 2D-based QC model could be used to correct the overall volume measurement. In order to do so, we used our $2D_R$ QC U-Net model to predict the 2D DSC and 3D DSC for every segmented slice and subject, respectively. Each segmented slice was classified as "good" or "wrong" based on the predicted 2D DSC: segmented slices with a predicted 2D DSC lower than 0.3 for the BG or LVM or LVC class were classified as "wrong". We chose these classes as they were most likely present in all slices as opposed to the LVT and LVPM classes. The most basal and apical slices were also excluded. We then replaced the associated volume measurements in the segmented slices identified as "wrong" by the mean of measurements in the two adjacent slices identified as "good" and so for the whole set of classes. The corresponding measurements were referred to as deteriorated and corrected (D + C). Hence, by comparing the RVE values computed with and without correction, we could verify if this approach could rectify the overall volume measurements.

Averaged real 3D DSC values and mean relative volume errors (RVE) are reported in Table 7 for each dataset and each class. The density plots of the predicted 2D DSC for the mid-slice from the original and corrupted datasets are displayed in figure 11. The density plots of the 150

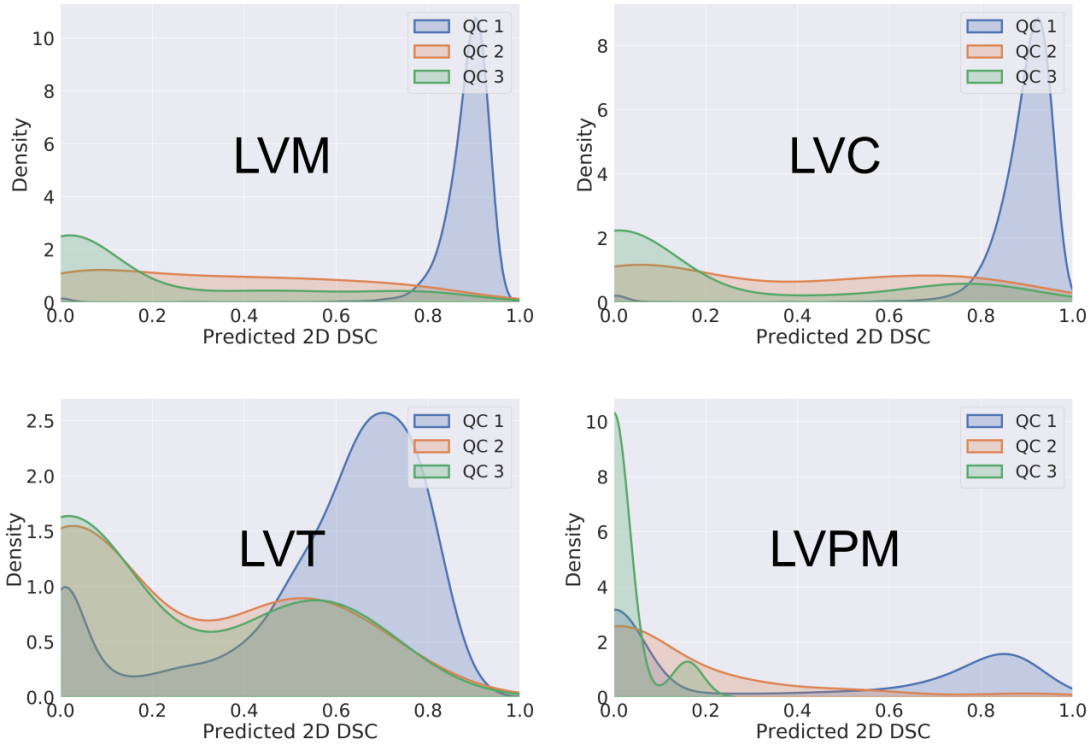


Fig. 9. Density plots of the predicted 2D DSC for the QC 1 and QC 2 and QC 3 groups in clinical evaluation. Within each class, all distributions were significantly different except QC 2 vs QC 3 for LVT and LVPM.

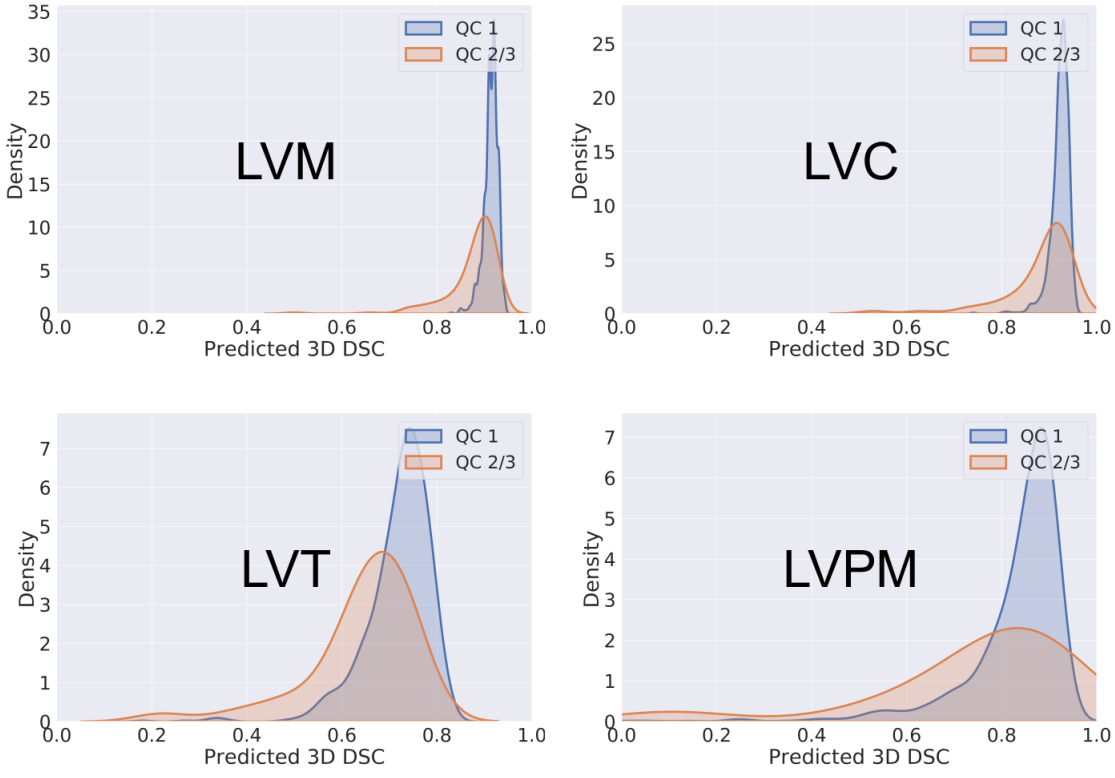


Fig. 10. Density plots of the predicted 3D DSC for the QC 1 and QC 2/3 groups in clinical evaluation. Within each class, all distributions were significantly different.

predicted 3D DSC, for each dataset are displayed in figure 12 while the density plot of the RVEs for our three set-

Table 7. Additional analyses: cases with and without deteriorated mid-slice segmentation.

3D DSC and RVE					
Class	3D DSC w/o D	3D DSC D	RVE w/o D	RVE D	RVE D + C
LVM	0.89 ± 0.26	0.73 ± 0.07	0.07 ± 0.05	0.19 ± 0.14	0.07 ± 0.05
LVC	0.92 ± 0.02	0.81 ± 0.09	0.05 ± 0.05	0.32 ± 0.29	0.05 ± 0.05
LVT	0.62 ± 0.08	0.45 ± 0.11	0.23 ± 0.16	0.65 ± 1.06	0.23 ± 0.15
LVPM	0.79 ± 0.11	0.51 ± 0.12	0.15 ± 0.13	0.51 ± 0.40	0.16 ± 0.14

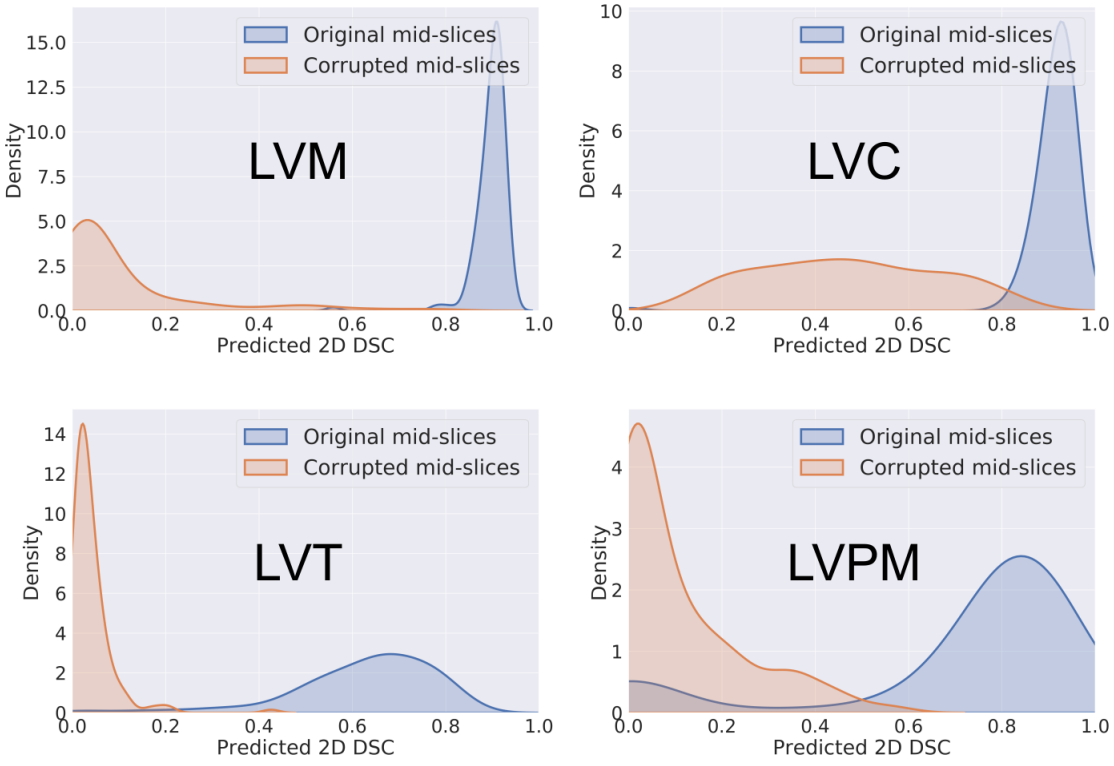


Fig. 11. Density plots of the predicted 2D DSC for the mid-slice segmentation, before and after deterioration, on 150 subjects in synthetic data. Within each class, all distributions were significantly different.

tings are displayed in figure 13. Regarding the impact of a poor quality segmentation in mid-slices on the 3D DSC values, we observed (Table 7) a substantial and significant 3D DSC reduction i.e. -18% for LVM (from 0.89 to 0.73), -12% for LVC, -27% for LVT and -35% for LVPM. Changes regarding the relative volume errors were much larger (and all significant) i.e. +171% for LVM (from 0.07 to 0.19), +540% for LVC, +182% for LVT and +240% for LVPM. These results strongly supported that a relatively small 3D DSC change e.g. from 0.92 to 0.81 (0.7 has been proposed in the litterature as a good quality threshold for LVC) might correspond to a very large change in relative volume errors e.g. from 0.05 to 0.32 for LVC. The capacity of our 2D-based QC model to correctly detect the mid-slice errors was supported by the significant (according to Kolmogorov-Smirnov test) distributions shifts of both 2D DSC for each class (figure 11) and 3D DSC (figure 12). The effect was more pronounced for 2D DSC than for 3D DSC distributions. Table 7 and figure 13 indicated that the shift in measurement errors distributions induced by the mid-slice errors could be almost totally suppressed using our proposed correction. This was further confirmed by Kolmogorov-Smirnov test as the RVE w/o D and RVE D distributions were always found significantly different while the RVE w/o D and RVE D + C distributions were not (p-values: 0.99 for LVM, 0.95 for LVC, 0.89 for LVT and 0.72 for LVPM).

6. Discussion

In the present study, we reported a MISAQC tool designed to jointly predict 2D and 3D DSC values for the segmented structures in cardiac MR images. The tool performance was compared with a 3D-direct approach. Regardless of the segmented structure and the chosen backbone architecture and based on the accuracy and errors metrics, the computation of 3D DSC values based on 2D indices systematically outperformed the 3D-direct approach.
The main contributions of our work are the following:

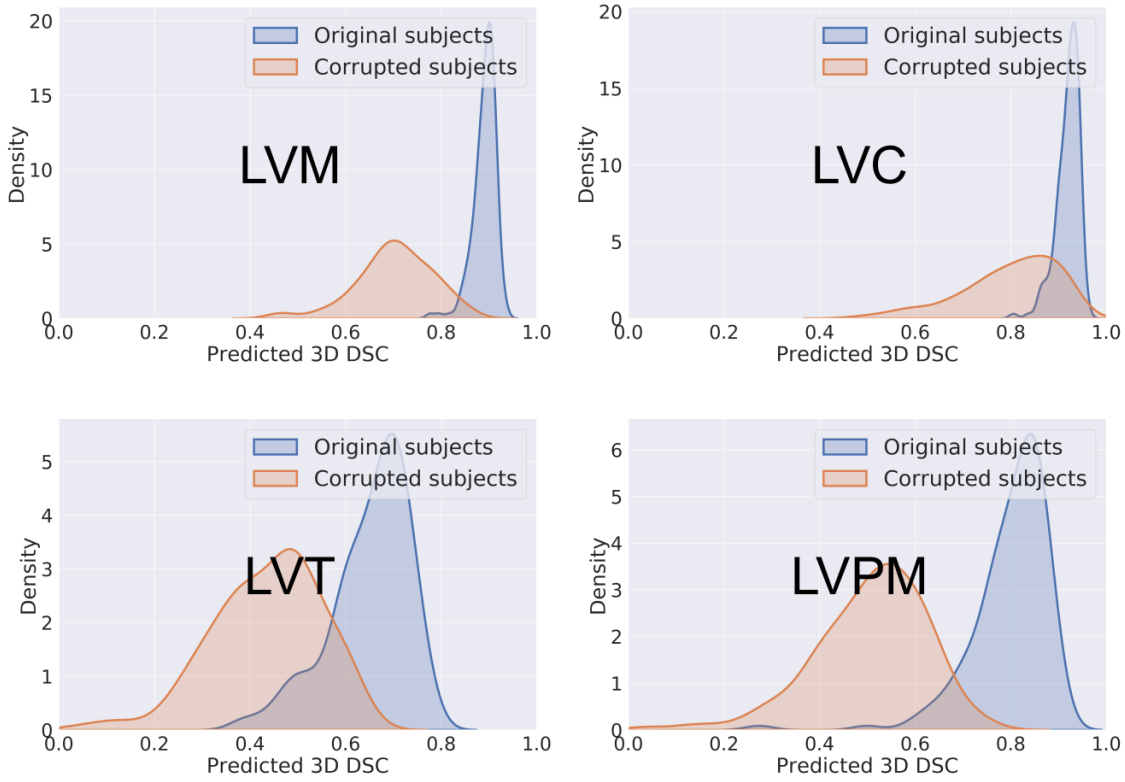


Fig. 12. Density plots of the predicted 3D DSC for 150 3D segmentations, before and after deterioration in synthetic data. Within each class, all distributions were significantly different.

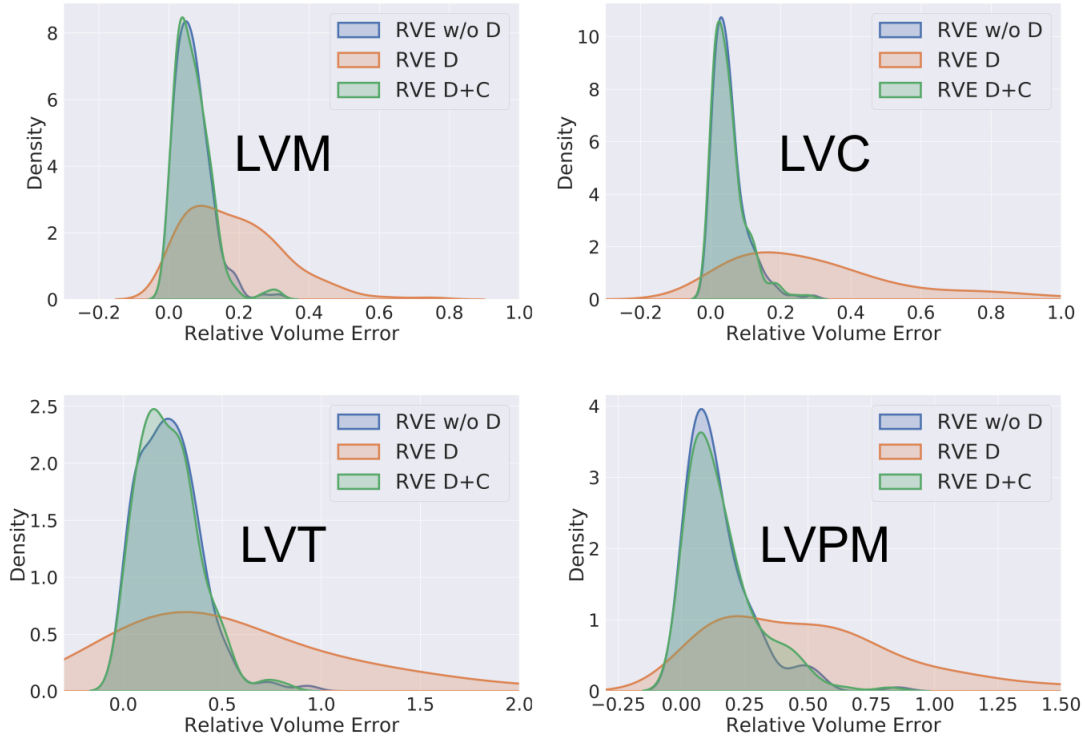


Fig. 13. Density plots of the relative volume errors (RVE) for 150 3D segmentations, before (w/o D), after (D) deterioration and after deterioration and correction using the $2D_R$ QC U-Net predictions (D + C) in synthetic data. For each class, RVE w/o D and RVE D distributions were always significantly different, while RVE w/o D and RVE D+C were not.

- A mathematical solution allowing to get 3D DSC predictions from 2D DSC and mean volume similarity fraction (MVSF) predictions.
- A localization of segmentation errors at a 2D/slice level with a possibility to correct the effect of segmentation errors on clinical measurements.
- A significant improvement of results obtained from state of the art approaches (for 3D DSC predictions).
- A general approach, implementable for any MRI or computed tomography (CT) segmentation task, even if classes are unbalanced.

6.1. Comparison with existing works

In most of the previous studies which have addressed the issue of MISAQC with machine-learning-based methods, evaluations were carried on at the 3D level whereas the 2D-level was ignored (Albà et al., 2017; Audelan and Delingette, 2019; Kohlberger et al., 2012; Robinson et al., 2019, 2018; Valindria et al., 2017). Based on DSC computation at the 2D level, the present method yielded correlation factors higher than 0.965 regardless of the segmented structures. In a previous study conducted in lung images, the correlation between the predicted and the real DSC values was lower (r = 0.85) (Kohlberger et al., 2012), while values for liver and brain tumors were much lower (r = 0.54 and r = 0.69, respectively) (Audelan and Delingette, 2019).

In the present study, the segmentation quality assessment was based on the DSC values, DSC being a continuous variable between 0 and 1. The corresponding evaluation allowed to score the segmentations quality on a continuous scale. Using 3D DSC quality thresholds, the reported accuracies were systematically superior to 96,5% for both the well represented (LVM and LVC) and the less represented classes (LVT and LVPM). As a matter of comparison, using a random forest classifier method for cardiac images, (Albà et al., 2017) used a 4 mm error threshold and reported a 96% accuracy for LVM and LVC, two classes that are very well represented in cardiac images. The utilization of such a threshold for classes such as LVT, is highly questionable given their small and scattered structure. In addition, the 4 mm threshold enabled the distinction of only two groups of errors (lower and higher than 4 mm) which is a critical limitation for accuracy computations.

The MAE values we reported in the present study are similar to those reported by (Robinson et al., 2018) who used a 3D-direct approach and a much larger dataset. More particularly, the MAE reported by these authors for LVM was 0.055 ± 0.064 as compared to 0.037 ± 0.035 in the present study using the QC ResNet 3D. Similarly, a value of 0.029 ± 0.039 was found in the present study for the LVC class using the QC ResNet 3D while Robinson *et al* reported a slightly larger value, i.e., 0.038 ± 0.040. On that basis, these results strongly support the similarity between the two 3D-direct approaches. In addition, our $2D_R$ approach clearly outperformed the 3D direct approach thereby indicating that a better performance can be achieved when MISAQC is initially performed at the 2D level. This new method was also very efficient in terms of computing time. The mean processing time for a subject was around 300 ms which permits real-time applications, on the contrary to other methods such as reverse classification accuracy where the corresponding reported time was 11 min (6000-fold larger) (Robinson et al., 2019).
We also addressed a limitation of previous works who used artificially degraded segmentations for learning and evaluation of their QC tools, and never confronted the automatic evaluation to human evaluation. For that we considered a clinical practice application using a real dataset having 1016 segmented subjects from the UK BioBank whose segmentation quality was assessed by a medical expert. Quality scores from our model were in very good agreement with those given by human operator, further supporting not only the robustness of our method but also it's ability to generalize to unseen data.

6.2. Advantages of a 2D-based multi-level MISAQC method

In the present method, training and inference were 2D-based whereas the quality assessment was performed at both 2D and 3D levels. In that respect, this MISAQC approach differs from those in which training, inference and quality evaluation have been performed at a 3D-level and provides multiple advantages which are listed below.

6.2.1. Possibilities opened by a multi-level QC

As indicated in section 5, the MISAQC performed at a slice-level allowed the detection of erroneous segmented slices (Table 2, figure 9 and 11). In addition, once detected, the corresponding erroneous measurement could be replaced by measurements from the neighbouring slices. In that respect volume errors could be drastically reduced (Table 7, figure 13). As a matter of example, the LVM mean RVE was 32% before the correction and was reduced by a factor five (5%) after the correction. This corrective process would not have been possible using a 3D-only quality evaluation. Being able to correct in real-time automatic clinical measurements can have an obvious diagnostic impact. Of interest, the impact of the mid-slice errors was moderate on the 3D DSC values and much larger on the volumetric errors (Table 7). This is a very important result indicating that the 3D DSC metric might not be sufficient in evaluating the quality of medical image segmentations and that morphological and/or functional metrics should be also considered. The additional 2D DSC evaluation at slice-level represents a clear advantage in that matter. Another application of the multi-level quality control would be to allow the clinicians to be able, in their diagnosis, to give more credit to the segmentation-derived measurements by being capable of directly visualizing the location

of the segmentation errors and eventually correct them immediately. Using the feedback brought by our models, the clinician would also identify more easily and reliably the profile of data for which the model's predictions should be carefully checked.

6.2.2. Enhanced predictive accuracy

We showed that by reconstructing the 3D DSC from 2D indices one could obtain 3D-level and 2D-level quality evaluation of segmentations with a single method. This would not be possible in the case of a 3D-direct approach. This methodology also encourages the robustness of the 3D-level evaluation: since (as illustrated in Table 2) the $2D_R$ versions can predict 2D indices with a very high predictive accuracy, the likelihood of large 3D DSC errors is substantially lower than what could be achieved using a 3D-direct approach.

6.2.3. Training and data efficiency

Annotation cost and computational cost are two recognized issues in the field of CNN-based segmentation. In that respect, using a 2D approach for a given dataset, the number of training examples is substantially increased (Baumgartner C.F., 2018), without any additional annotation cost. Indeed, when working in a 3D framework, the patients segmentations' quality is expressed by a unique score for each class, whereas this information is enriched in the 2D context and multiplied by the number of slices. Having a higher variability in the training sample is highly beneficial for the convergence of the stochastic gradient algorithm. As a matter of example, in the promising study of Robinson et al. (Robinson et al., 2018) the training dataset was the UK Biobank database with thousand of available segmented 3D volumes. Such a gigantic database can be considered as exceptional and not commonly available. On that basis, a 2D dataset of an hundred segmented MRIs could be comparable to the UK BioBank database from the 3D point of view in terms of number of training examples.

Obviously, computational costs are substantially reduced as well with our approach. This could be of particular relevance for computed tomography (CT) datasets for which the number of slices is of several hundreds: 3D approaches might face computational limits, whereas our approach will be easily implemented.

6.3. Quality control for small anatomical regions

Of particular interest, in addition to providing a quality control at both 2D and 3D levels, our method was insensitive to DSC values and size of the segmented structures. In other words, we were able to predict with the same accuracy both high and low DSC values. The issue of size structures was also addressed very efficiently. An interesting and challenging issue in the field of medical images segmentation is related to MISAQC for classes highly and poorly represented. Previous studies have mainly reported methods dedicated to the MISAQC of segmentations for relatively large and connected classes, such as the myocardium and blood cavity (Audelan and Delingette, 2019; Robinson et al., 2019, 2018). The suitability and performance of these methods for small and scattered anatomical structures such as trabeculations or papillary muscles (illustrated in figure 1), have not been addressed. The present results clearly demonstrate that our $2D_R$ method is suitable for both large and small classes. The corresponding MAEs were similar (QC ResNet $2D_R$: 0.016 (LVM), 0.011 (LVC), 0.022 (LVT) and 0.042 (LVPM)), illustrating that the CNNs were able to learn and extract relevant features for both types of classes. Our method was less efficient for the LVPM class and so most likely because papillary muscles do not appear in all slices.

6.4. Limitations

A few limitations should be acknowledged in the present work. One could wonder about the realistic features of the synthetic database we generated. The methodology we used to generate this database was comparable to what has been previously reported (Robinson et al., 2018). Actually, we replaced the random forests approach by CNNs and this slight change is not expected to introduce major changes. More importantly, the external validation of our model on the UK BioBank dataset supports the realistic nature of this synthetic dataset as the model was able to generalize on a new data source. We used a single type of network to build the MISAQC dataset and one could wonder whether other networks could have produced other "types" of segmentation errors. We think that the key issue for this dataset construction is to generate a sufficient number of segmentations fairly distributed over the entire range of possible DSCs (from 0 to 1).

The performance of our approach was assessed using a rather small dataset. However, as it relies on a 2D analysis, it should be less sensitive to the dataset size than a 3D-based deep learning approach. This expectation is supported by the promising comparison between our results and those from (Robinson et al., 2018) with a smaller dataset, we reported lower MAE values.

6.5. Conclusion

In the present study, a multi-level (2D and 3D) deep-learning-based real-time automated quality control method for cardiovascular MR image segmentations was designed and the corresponding performance was assessed. Based on the MAE values and classification accuracy, it was clearly demonstrated that the proposed method was equally efficient for large and small cardiac anatomical structures. The QC provided by our method was shown to be consistent with scores generated by trained cardiologists. The 2D-based structure can be trained with a modestly sized dataset while enabling a very accurate real-time automatic quality prediction at both 2D and 3D levels. We

highlighted its' possible applications such as successfully rectifying erroneous clinical measurements derived from medical image segmentations.

7. Acknowledgments

Steffen E. Petersen (SEP) acknowledges support from the National Institute for Health Research (NIHR) Biomedical Research Centre at Barts. SEP acknowledge the British Heart Foundation for funding the manual analysis to create a cardiovascular magnetic resonance imaging reference standard for the UK Biobank imaging resource in 5000 CMR scans (www.bhf.org.uk; PG/14/89/31194). SEP acknowledges support from the "SmartHeart" EPSRC programme grant (www.nihr.ac.uk; EP/P001009/1). SEP and ER also acknowledge support by the London Medical Imaging and Artificial Intelligence Centre for Value Based Healthcare (AI4VBH), which is funded from the Data to Early Diagnosis and Precision Medicine strand of the government's Industrial Strategy Challenge Fund, managed and delivered by Innovate UK on behalf of UK Research and Innovation (UKRI). Views expressed are those of the authors and not necessarily those of the AI4VBH Consortium members, the NHS, Innovate UK, or UKRI. SEP provides consultancy to and owns stock of Cardiovascular Imaging Inc, Calgary, Alberta, Canada. Other authors did not receive any financial or material support from any industrial company.

References

Albà, X., Lekadir, K., Pereanez, M., Medrano-Gracia, P., Young, A., Frangi, A., 2017. Automatic initialization and quality control of large-scale cardiac mri segmentations. Medical Image Analysis 43.

Audelan, B., Delingette, H., 2019. Unsupervised quality control of image segmentation based on bayesian learning. MICCAI 2019, Lect. Notes in Comp. Sc. 11765.

Bartoli, A., Fournel, J., Bentatou, Z., Habib, G., Lalande, A., Bernard, M., Boussel, L., Pontana, F., Dacher, J., Ghattas, B., Jacquier, A., 2020. Deep learning–based automated segmentation of the left ventricular trabeculations and myocardium on cardiac mr images: A feasibility study. Radiology: Artificial Intelligence .

Baumgartner C.F., Koch L.M., P.M.K.E., 2018. An exploration of 2d and 3d deep learning techniques for cardiac mr image segmentation. Lecture Notes in Computer Science, Springer 10663.

Bentatou, Z., Finas, M., Habert, P., Kober, F., Guye, M., Bricq, S., Lalande, A., Frandon, J., Dacher, J., Dubourg, B., Habib, G., Caudron, J., Normant, S., Rapacchi, S., Bernard, M., Jacquier, A., 2018. Distribution of left ventricular trabeculation across age and gender in 140 healthy caucasian subjects on mr imaging. Diagnostic and Interventional Imaging 99, 689 – 698.

Bricq, S., Frandon, J., Bernard, M., Guye, M., Finas, M., Marcadet, L., Miquerol, L., Kober, F., Habib, G., Fagret, D., Jacquier, A., Lalande, A., 2015. Semiautomatic detection of myocardial contours in order to investigate normal values of the left ventricular trabeculated mass using mri. Journal of magnetic resonance imaging : JMRI 43.

Elliott, P., Anastasakis, A., Borger, M., Borggrefe, M., Cecchi, F., Charron, P., Hagège, A., Lafont, A., Limongelli, G., Mahrholdt, H., McKenna, W., Mogensen, J., Nihoyannopoulos, P., Nistri, S., Pieper, P., Pieske, B., Rapezzi, C., Rutten, F., Tillmanns, C., Wolpert, C., 2014. 2014 esc guidelines on diagnosis and management of hypertrophic cardiomyopathy: The task force for the diagnosis and management of hypertrophic cardiomyopathy of the european society of cardiology (esc). Eur Heart J. , 1–55.

Frandon, J., Bricq, S., Bentatou, Z., Marcadet, L., Barral, P., Finas, M., Fagret, D., Kober, F., Habib, G., Bernard, M., Lalande, A., Miquerol, L., Jacquier, A., 2018. Semi-automatic detection of myocardial trabeculation using cardiovascular magnetic resonance: Correlation with histology and reproducibility in a mouse model of non-compaction. Journal of Cardiovascular Magnetic Resonance 20.

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

Japp, A., Gulati, A., Cook, S., Cowie, M., Prasad, S., 2016. The diagnosis and evaluation of dilated cardiomyopathy. Journal of the American College of Cardiology 67, 2996–3010.

Kohlberger, T., Singh, V., Alvino, C., Bahlmann, C., Grady, L., 2012. Evaluating segmentation error with-out ground truth. MICCAI 2012, Lect. Notes in Comp. Sc. 7510, 528–536.

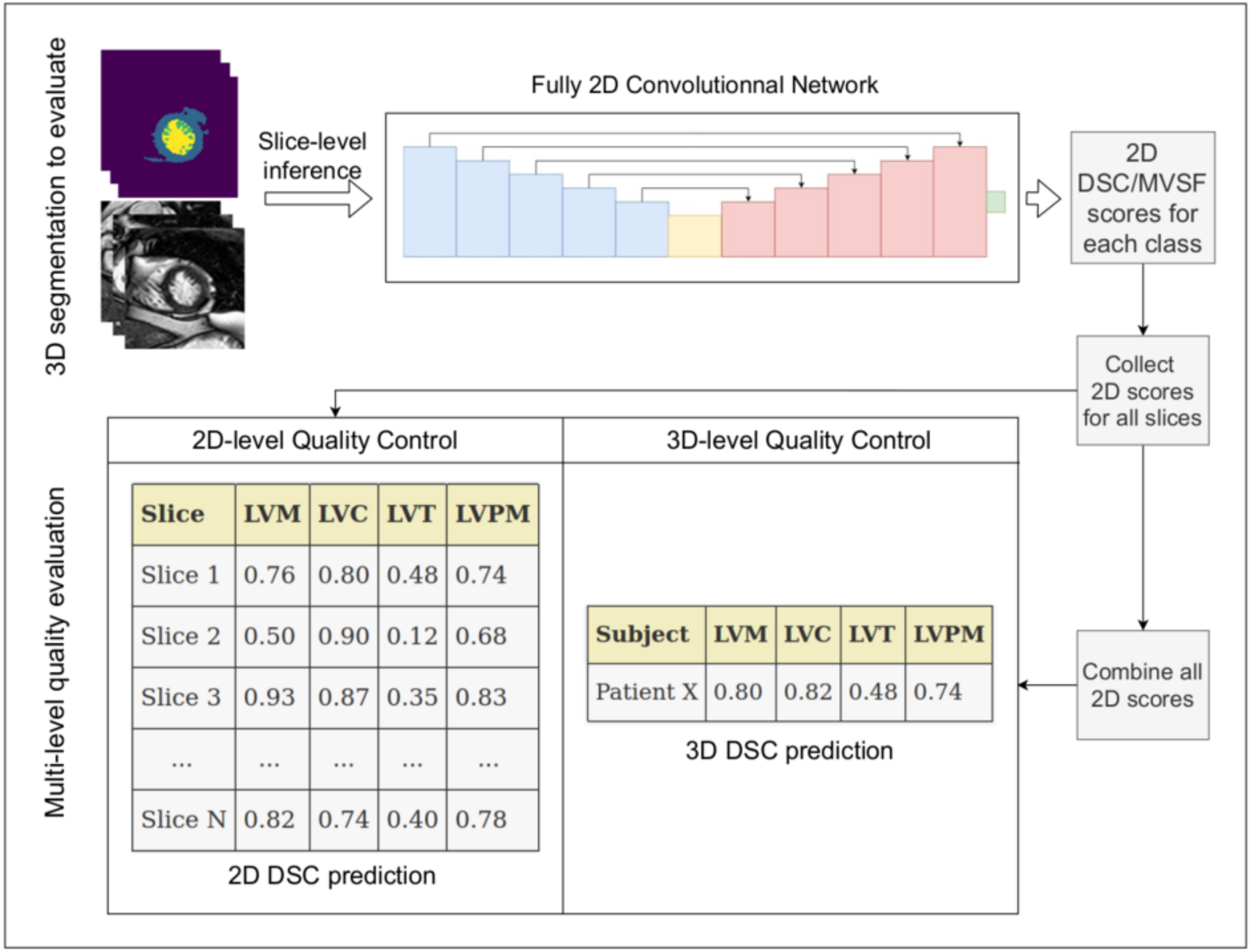
Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical Image Analysis 42, 60 – 88.

Petersen, S.E., Selvanayagam, J.B., Wiesmann, F., Robson, M.D., Francis, J.M., Anderson, R.H., Watkins, H., Neubauer, S., 2005. Left ventricular non-compaction: insights from cardiovascular magnetic resonance imaging. Journal of the American College of Cardiology 46, 101–105.

Robinson, R., Oktay, O., Bai, W., Valindria, V., Sanghvi, M., Aung, N., Paiva, J., Zemrak, F., Fung, K., Lukaschuk, E., Lee, A., Carapella, V., Kim, Y., Kainz, B., Piechnik, S., Neubauer, S., Petersen, S., Page, C., Rueckert, D., Glocker, B., 2018. Real-time prediction of segmentation quality .

Robinson, R., Valindria, V., Bai, W., Oktay, O., Kainz, B., Suzuki, H., Sanghvi, M., Aung, N., Paiva, J., Zemrak, F., Fung, K., Lukaschuk, E., Lee, A., Carapella, V., Kim, Y., Piechnik, S., Neubauer, S., Petersen, S., Page, C., Glocker, B., 2019. Automated quality control in image segmentation: Application to the uk biobank cardiovascular magnetic resonance imaging study. Journal of Cardiovascular Magnetic Resonance 21, 1597–1606.

Valindria, V.V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B., 2017. Reverse classification accuracy: Predicting segmentation performance in the absence of ground truth. IEEE Transactions on Medical Imaging 36, 1597–1606.



Slice	LVM	LVC	LVT	LVPM
Slice 1	0.76	0.80	0.48	0.74
Slice 2	0.50	0.90	0.12	0.68
Slice 3	0.93	0.87	0.35	0.83
...	...	...	...	...
Slice N	0.82	0.74	0.40	0.78

Subject	LVM	LVC	LVT	LVPM
Patient X	0.80	0.82	0.48	0.74