Radiology: Artificial Intelligence

Deep Learning–based Automated Segmentation of the Left Ventricular Trabeculations and Myocardium on Cardiac MR Images:

A Feasibility Study

Axel Bartoli, MD

Joris Fournel

Zakarya Bentatou

Gilbert Habib, MD, PhD

Alain Lalande, PhD

Monique Bernard, PhD

Loïc Boussel, MD, PhD

François Pontana, MD, PhD

Jean-Nicolas Dacher, MD, PhD

Badih Ghattas, MCU

Alexis Jacquier, MD, PhD

From the Departments of Radiology (A.B., A.J.) and Cardiology (G.H.), Hôpital de la Timone Adultes, AP-HM, 264, rue Saint-Pierre 13385 Marseille Cedex 05, France; CRMBM-UMR CNRS 7339, Medical Faculty, Aix-Marseille University, Marseille, France (A.B., J.F., Z.B., M.B., A.J.); I2M-UMR CNRS 7373, Aix-Marseille University, CNRS, Centrale Marseille, Marseille, France (J.F., B.G.); ImVia laboratory and University Hospital of Dijon, Bourgogne-Franche Comté University, Dijon, France (A.L.); Department of Radiology, Hôpital de la Croix-Rousse, Hospices Civils de Lyon, Lyon, France (L.B.); Department of Cardiovascular Imaging, Lille University Hospital, Lille, France (F.P.); and Department of Diagnostic Imaging, Rouen University Hospital, Rouen, France (J.N.D.). Received XXX; revision requested XXX; revision received XXX; accepted XXX; final version accepted XXX. **Address correspondence to** A.B. (e-mail: *axel.bartoli01@gmail.com*).

https://doi.org/10.1148/radiol.2020200021

Purpose: To develop and evaluate a complete deep learning pipeline that allows fully automated end-diastolic left ventricle (LV) cardiac MRI segmentation including trabeculations and automatic quality control of the predicted segmentation.

Materials and Methods: This multicenter retrospective study includes training, validation, and testing datasets of 272, 27, and 150 cardiac MRI, respectively, collected between 2012 and 2018. The reference standard was the manual segmentation of four LV anatomic structures performed on end-diastolic short-axis cine cardiac MRI: LV trabeculations (LVT), LV myocardium (LVM), LV papillary muscles, and the LV blood cavity (LVC). The automatic pipeline was composed of five steps using a DenseNet architecture. Intraobserver agreement, interobserver agreement, and interaction time were recorded. The analysis includes the correlation between the manual and automated segmentation, a reproducibility comparison and Bland-Altmann plots.

Results: The automated method achieved Dice coefficients of 0.96 ± 0.01 for LVC, 0.89 ± 0.03 for LVM., and 0.62 ± 0.08 for LVT (mean absolute error: 3.63 g ± 3.4). Automatic quantification of LV end-diastolic volume, LVM mass, LVT, and trabeculation-mass-to-total-myocardial-mass ratio (T/TMM) showed a significant correlation with the manual measures (r = 0.99, 0.99, 0.90, and 0.83, respectively; all *P* < .01). On a subset of 48 patients, LVT Dice was 0.63 ± 0.10 and higher compared with the human interobserver (0.44 ± 0.09; *P* < .01) and intraobserver measures (0.58 ± 0.09; *P* < .01). Automatic quantification of the T/TMM had a higher correlation (0.92) compared with the intra-and interobserver measures (0.74, and 0.39, respectively; both *P* < .01).

Conclusion: Automated deep learning framework can achieve reproducible and quality-controlled segmentation of cardiac trabeculations outperforming inter-and intraobserver analyses.

A fully automated deep learning pipeline was developed to produce fast, reproducible, and automated quality-controlled left ventricle volume, mass, and trabeculation segmentation on short-axis cardiac MRI, to define the diagnostic criteria of excess of trabeculations.

Abbreviations

CNNs = convolutional neural networks, DCM = dilated cardiomyopathy, DFCNNs = dense fully convolutional neural networks, ET = excessive trabeculation, ETCM = excessive trabeculation cardiomyopathy, HCM = hypertrophic cardiomyopathy, LV = left ventricle, LVC = left ventricle blood cavity, LVEDV = left ventricle end-diastolic volume, LVM = left ventricle myocardium label, LVMM = left ventricle myocardial mass, LVPM = left ventricle papillary muscles, LVT = left ventricle trabeculation, MAE = mean absolute error, MVSF = mean volume similarity fraction, T = trabeculation mass, T/TMM = trabeculation-mass-to-total-myocardial-mass ratio

Key Points

A deep learning pipeline, based on a convolutional neural network, achieved segmentation Dice scores of 0.96 ± 0.01 for the left ventricular (LV) cavity, 0.89 ± 0.03 for the LV myocardium and 0.62 ± 0.08 the LV trabeculations. On a subset of 48 patients, the automated method achieved superior Dice scores for trabeculation segmentation (0.63 ± 0.10) compared with human intraobserver (0.58 ± 0.09, *P* < .01) and interobserver measures (0.44 ± 0.09; *P* < .01).

For the clinical parameters, automated computation of the ratio of trabeculation mass over the total myocardial mass showed higher correlation with the reference method compared with human intra-and interobserver measures (0.92 versus 0.74 [*P* < .01] and 0.39 [*P* < .01], respectively).

The pipeline includes automatic quality control and showed a high classification accuracy of 94.5% with a three-dimensional Dice coefficient mean absolute error of 0.05 ± 0.06 for the trabeculation segmentation.

Conflicts of interest are listed at the end of this article.

Trabeculations are a physiologic component of the human left ventricle (LV). Excessive LV trabeculations are observed in both pediatric and adult populations (1). However, controversies remain regarding the incidence, embryogenesis, physiopathology, classification, genetically associated disorders, diagnosis, and prognosis of excessive trabeculations (ET) (2). Several authors recommend the use of the term excessive trabeculation cardiomyopathy (ETCM) rather than LV noncompaction. Noncompaction implies a defect of the embryologic process during myocardial compaction, which remains uncertain (3). ETCM has been described in a wide variety of clinical and pathophysiological situations (4). Initially, it was mostly described in association with dilated cardiomyopathy (DCM) (5). However, many authors showed that an ET myocardial phenotype could be found in other clinical patterns such as hypertrophic cardiomyopathy (HCM), cardiac congenital diseases, as well as in health patients (4,6,7).

Cardiac MRI is the most reproducible diagnostic tool to explore ET phenotypes (8). Several parameters are described to reach a diagnosis, including the assessment of the noncompacted-to-compacted thickness ratio using end-diastolic cine cardiac MRI to define the ET phenotype (9), quantification of the whole LV trabecular mass (although that technique was too time consuming for broad use in clinical practice) (10), and more recently, new criteria have been proposed based on fractal dimensions (11). The definition of ETCM is still controversial due to the variability in clinical presentation and various diagnostic criteria. Moreover, trabeculations have a substantial impact on LV volume, mass, and ejection fraction measurements (12). Hence, there is a need for a fast, reproducible, and fully automated LV trabeculation segmentation method that can be used on a large cohort to facilitate the diagnosis of ETCM and to assess the impact of ET on patient prognosis.

Deep learning techniques, notably with the use of convolutional neural networks (CNNs), are very promising in the automation of fastidious measurements in medical imaging (13). These techniques, based on multilayer neural networks, are capable of learning and building hierarchical representation of data (14). They have demonstrated great capabilities, often outperforming humans on image segmentation tasks (15). Numerous publications have validated the interest and effectiveness of CNNs for the segmentation of the short-and long axis left and right ventricles for the myocardium, papillary muscles, or blood pool (16–18). The goal of the study was to develop and evaluate a complete deep learning pipeline that allows *(a)* fully automated end-diastolic LV segmentation including trabeculations, *(b)* automatic quality control of the predicted segmentation results, and *(c)* computation of LV clinical parameters.

Materials and Methods

Data Sources and Patient Descriptions

This multicenter retrospective study was approved by the local Institutional Review Board in accordance with the guidelines outlined in the Declaration of Helsinki. The study was approved by a French national ethical committee (N° IRB CRM-1907–02è).

Training dataset.—

We retrospectively selected a sample of 299 cardiac MRI examinations performed between October 2012 and November 2018 at three different French university hospitals (CHU Rouen, CHU Dijon, and CHU Marseille). The inclusion criterion was age between 18 and 85 years old. The exclusion criteria were congenital heart disease defined by the International Pediatric and Congenital Cardiac Code (19), ventricular postischemic remodeling, known amyloidosis, or iron overload. All the patients were classified depending on four cardiac phenotypes (healthy, DCM, HCM, or ETCM) by an experienced investigator (A.J. with 20 years of experience in cardiac imaging) to train the model in all situations where ET can be found (20). Healthy participants were included if the patients had no known risk factors or history of cardiac disease, no cardiac symptoms, and normal cardiovascular examination and parameters results. DCM, HCM and

ETCM used criteria are detailed in Appendix A3 (9,21,22). The dataset was randomly divided into a training set (*n* = 272) and a validation set (*n* = 27). A flow diagram of the procedure is shown in Figure 1.

External testing dataset.—

We retrospectively selected 150 consecutive cardiac MRI examinations, performed between November 2018 and April 2019, at three clinical hospitals to provide external validation of our model: 50 examinations each from the Cardiologic Hospital (Lille, France), Croix-Rousse Hospital (Lyon, France), and our institution (Marseille, France). The data were extracted from consecutive clinical care. The inclusion criteria, exclusion criteria, and cardiac phenotype classifications were similar to those in the training dataset with a predefined objective to include 90 healthy individuals and 20 patients each with DCM, HCM and ETCM. To assess the inter-and intraobserver reproducibility for the LV technical and clinical parameters, we established a dedicated subset of 48 examinations which represents approximately the third of the test sample (*n* = 150, 90 healthy; 60 cardiomyopathies). To preserve the original balance of the testing dataset between healthy and nonhealthy subjects, we randomly selected 30 healthy patients among the 90, and 18 patients with cardiomyopathies among the 60. We obtained 5 patients with ETCM, 5 with DCM and 8 with HCM.

Cardiac MRI Data

The cardiac MRI examinations were performed on two different types of scanners. One scanner was a 1.5T Ingenia scanner (Philips Health System, Best, the Netherlands) and a 1.5T Avanto scanner (Siemens Healthcare, Erlangen, Germany) using a multichannel body array coil combined with a spine array coil with the patients in a supine position. A stack of images using balanced steady-state-free precession sequence was acquired. Only the ED frame at each imaging level was retained (8). The patient records and information were deidentified prior to analysis.

Manual Segmentation and Reference Measures

Manual image segmentation was undertaken for the complete training and testing datasets by a fellowship-trained observer (Observer 1 was A.B., 5 years of experience in cardiac imaging). Manual segmentation and the derived clinical measures were considered the reference standard to train the model (training) and to evaluate its performance (testing). For each cardiac MRI, all the ED images of the short-axis stack were imported in Digital Imaging and Communications in Medicine format. Reference manual segmentation was performed on a postprocessing software developed by Bricq et al and previously validated (23–25). The LV structures were manually segmented to obtain four labels (23): blood cavity (LVC), myocardium (LVM), papillary muscles (LVPM), and trabeculations (LVT). The nonsegmented part of the image was considered a fifth label: background (Fig 2). The clinical parameters from manual segmentation were obtained as follows: LV end-diastolic volume (LVEDV) extracted from the LVC label, the LV myocardial mass (LVMM) extracted from the LVM label, the papillary muscles mass (PM) extracted from the LVPM label and the trabeculation mass (T) extracted from the LVT label. The total myocardial mass (TMM) was calculated as the sum of the papillary, myocardial muscle, and trabeculation masses. The trabeculation-mass-to-total-myocardial-mass ratios (T/TMMs) are expressed as a percentage. The user interaction time was recorded for manual segmentation.

Model Architecture and Pipeline

The complete pipeline (Fig 3) was composed of five steps: basal to end-apical LV stack selection (step 1), LV detection and cropping (step 2), LV structure automatic segmentation (step 3), automatic quality control as in (27) (step 4) and computation of clinical parameters from the predicted segmentations (step 5).Dense fully CNNs (DFCNNs), similarly to those used by Khened et al were used (26). DFCNN has a DenseNet based architecture and contains way less parameters and tends to overfit less than U-Nets.

All the details regarding the dedicated dataset used and the network architecture of each step are presented in Appendices A and B, respectively. The details about image preprocessing, training parameters and images postprocessing are presented in Appendices C, D, and E respectively.

Performance Evaluation

The details regarding the evaluation of the preliminary steps 1 and 2 are presented in Appendix F.

Evaluation of the LV Segmentation and Obtained Clinical Parameters

To assess the segmentation accuracy, the automatic segmentation results were compared with the manual segmentations for the technical metrics and clinical parameters. We used Dice coefficients as a technical metric. The Dice coefficient measures the overlapping region between the predicted and manual contours (28). Dice coefficient and clinical parameters measures were presented in mean measures with standard deviation. The efficiency, defined as the user interaction time comparison was also evaluated. The reproducibility of the automatic method was compared with the inter-and intraobserver segmentation performances. The segmentation performed by Observer 1 on the 48 patients randomly extracted from the testing dataset was defined as manual 1a. Observer 1 performed a second analysis called manual 1b at least 1 month after the first segmentation, in randomized order to minimize the recall bias. A second independent observer (A.J.) performed segmentation on the same dataset of 48 patients, called manual 2. The observers were blinded to the patient characteristics (age, sex, and cardiac phenotype), the results from the other observers, and the reference segmentations. To assess the additional value of the 3D CNN in the model, the LVT Dice was measured separately for 3D DFCNNs, 2D DFCNNs and the combination of 2D and 3D DFCNNs on the 48 patients of the testset.

Evaluation of the Automatic Quality Control

The precision of the automatic quality control module was evaluated by predicting the Dice coefficient on a test set of 402 unseen 3D-segmented patients and comparing it with the reference Dice coefficient for all the classes. The main goal was to predict the quality of the LVT classification. The testing set was stratified using the LVT Dice coefficient with an equal number of cases in each of the following ranges of the LVT Dice score: 0–0.2, 0.3–0.4, 0.4–0.5, 0.5–0.6, 0.6–0.7 and 0.7–1.

Statistical Analysis

For the clinical parameters (LVEDV, LVMM, PM, T, and T/TTM), we reported the mean absolute errors (MAEs) and biases of the measures derived from automatic segmentation with the measures derived from manual. To establish significance of the biases, Wilcoxon signed-rank tests were used whenever the data within the subgroups were not distributed normally. Subgroups analyses were performed for each cardiac phenotype group also using Wilcoxon signed-rank tests, as each group had at least 20 patients. The manual and automatic clinical parameters were also compared by means of linear regression to determine the correlations. For efficiency, we used a two-factor repeated analysis of variance. The automatic segmentation performance was compared with the manual segmentation performances using paired *t* tests for the Dice coefficients. The following comparisons were performed: *(a)* manual 1a versus automatic, *(b)* manual 1a versus manual 1b, and *(c)* manual 1a versus manual 2. Bland-Altman plots were calculated to assess the bias and limits of agreement between the manual and automatic contours.

To evaluate the performance of the automatic quality control module, we reported the MAE between automated predicted values and the manually delineated reference values and the classification accuracy of binary Dice threshold prediction (the thresholds were 0.7, 0.7, 0.35 and 0.4 for the LVM, LVC, LVT, and LVPM respectively). A *P* value < 0.01 was considered statistically significant.

Code Availability Statement

The present code is protected (DSO2019018721) and could be shared upon the signature of a collaboration agreement.

Results

The patient characteristics and clinical parameters for the training, testing, and reproducibility datasets extracted from the manual reference-standard segmentation are shown in Table 1. The manual and automatic segmentation contours of the cardiac labels for the different cardiac phenotypes are shown in Figure 4. The results of LV stack selection and LV center detection and cropping are presented in Appendix E.

Automatic Quality Control Module Accuracy

The results are presented in Table 2. For the LVT, 3D Dice MAE was 0.05 ± 0.06, with a classification accuracy of 94.5%. For the LVM and LVC classes, the MAEs were 0.03 ± 0.03 and 0.02 ± 0.02, with binary threshold classification accuracies of 97.2% and 96.5%, respectively.

LV Segmentation Accuracy

The Dice coefficients and clinical parameters for the automatic and manual segmentation results evaluated on the testing dataset are shown in Table 3. The correlations between the automatic and manual measurements of the clinical parameters on the whole test set are presented on Figure 5. The Dice coefficient was 0.96 ± 0.01 for the LVC, 0.89 ± 0.03 for the LVM, 0.79 ± 0.11 for the LVPM, and 0.62 ± 0.08 for the LVT in the overall dataset. In the subgroups analysis for ETCM patients, the Dice coefficient for the LVT was 0.66 ± 0.08 between the manual and automatic segmentations. The LVC Dice was always greater than 0.96 in the overall and subgroups analyses.

Concerning the clinical parameters, the MAEs for the LVEDV values ranged from 4.90 ± 4.5 mL for the healthy subgroup to 9.45 ± 7.5 mL for the ETCM subgroup, with an MAE of 5.86 ± 5.8 mL for the overall test set. The biases for the LVEDV ranged from −0.2 ± 6.6 mL for the healthy subgroup to −2.7 ± 12.1 mL for the DCM subgroup and −0.4 ± 8.3 mL for the whole cohort; none of the biases were found to be significant. The manual and automated measurements correlated strongly with correlation coefficients between 0.99 for the healthy subgroup to 0.99 for the ETCM subgroup and 0.99 globally.

For the LVMM, the overall MAE was 9.46 ± 7.3 g with a bias of −6.3 ± 10.2 g (*P* < .01). In terms of the correlation, all measurements for LVMM were highly accurate with values greater than 0.97 for all subgroups and 0.99 overall.

For the T measure, the MAEs ranged from 2.60 ± 2.0 g for the healthy group to 6.79 ± 5.0 g for the ETCM group with an overall MAE of 3.63 ± 3.4 g. The biases ranged from −0.7 ± 3.2 g for the healthy group to −3.7 ± 7.7 g for the ETCM group and were not found to be significant for the ETCM, DCM and healthy groups. The overall correlation was estimated to be 0.90.

Concerning the T/TMM measure, the MAE was 1.95% ± 1.5% for the overall group, ranging between 1.32% ± 1.0% for the HCM group to 2.94% ± 1.7% for the ETCM group in the subgroup analysis. No biases were found to be significant, ranging from-0.3% ± 2.3% for the healthy group to −0.9% ± 2.4% for the DCM subgroup with a mean value of-0.5% ± 2.4% for the whole test set. The overall correlation between the automatic and manual measures was 0.83. Figure 5 shows a fitted linear regression line (slope: 0.908; intercept: 1.321; R^2:0.687) for this measure.

In terms of the segmentation efficiency, the mean interaction time was 15 ± 3.6 minutes per patient for complete LV manual segmentation. The mean interaction time was 5 seconds to automatically assess the segmentation for 50 patients.

LV Segmentation Reproducibility

Table 4 reports the comparison of the model performances in terms of the Dice coefficient and clinical parameters (MAE and bias) with the intra-and interobserver agreement measures. For the LVT group, the Dice coefficient for the automatic compared with manual 1a was 0.63 ± 0.10 and was higher compared with intra-and interobserver agreement (0.58 ± 0.09; *P* < .01 and 0.44 ± 0.09; *P* < .01, respectively). Furthermore the LVT Dice with the presented combined 2D and 3D DFCNNs (0.63 ± 0.10) was higher than the LVT Dice measure using 2D DFCNNs (0.62 ± 0.10; *P* = .08) or 3D DFCNNs (0.62 ± 0.10; *P* < .01).

For T, the MAE was 1.70 g ± 1.30 (bias: 0.5 g ± 2.1) for automatic versus manual 1a, 3.54 g ± 2.4 (bias:-1.7 g ± 3.9) for the manual intraobserver agreement and 5.84 g ± 4.5 (bias: 1.0 g ± 7.3) for the interobserver agreement. The T measure correlation scores (0.95) were superior to both the intraobserver (0.82; *P* < .01) and interobserver (0.38; *P* < .01) agreement scores (*P)*. The T/TMM MAE was 1.10% ± 0.90 (bias: 0.6% ± 1.3) for automatic versus manual 1a analysis, 2.07% ± 1.4 (bias:-1.0% ± 2.3) for the manual intraobserver agreement and 3.36% ± 3.2 (bias: 1.0% ± 4.6) for the manual interobserver agreement. The T/TMM measure correlation scores (0.92) were superior to both the intraobserver (0.74; *P* < .01) and interobserver (0.39; *P* < .01) agreement. For the LVMM measure, automatic method was not superior compared with human intra-and interobserver variabilities.

Figure 6 shows Bland-Altman plots for the main clinical parameters between automatic versus manual 1a, manual 1a versus manual 1b, and manual 1a versus manual 2. Bland-Altman plots show better agreement for the automatic versus manual performance than for the inter-and intraobserver performance for the T and T/TMM measures.

Discussion

The main findings of the study were that our model pipeline *(a)* provided accurate and reproducible results, matching the human performance for LV structures and trabeculations segmentations; *(b)* was a fast (5 seconds) and fully automatic segmentation tool; *(c)* had a similar accuracy for all cardiac phenotypes included; and *(d)* provided an automatic quality assessment of the predicted segmentation results.

Data from multiple centers and imaging parameters were used in the training and testing datasets in the present study, as recommended by Bluemke et al (29). The training dataset showed an overrepresentation of the ETCM population (12.3%) that does not reflect the overall population for which the ETCM population has an estimated prevalence of 0.26% (30). This phenomenon could be explained because only tertiary reference centers were included in the study. This study examined most of the clinical situations where excessive trabeculations can be found. Regarding the clinical characteristics of healthy patients in the training dataset, the T/BSA was 4.52 g/m^2 ± 2.69 was in line with that from Bentatou et al (4.9 g/m^2 ± 2.4) (24). The overall measure was superior (5.39 g/m^2 ± 2.69) due to the inclusion of DCM, HCM and mostly ETCM patients.

Trabeculations have not been clearly defined in medical imaging and trabeculation segmentation is associated with poor interobserver reproducibility. Our method produced similar results to those of Bai et al concerning the LV cavity (0.96 ± 0.01 versus 0.94 ± 0.04) and LV myocardium segmentation (0.89 ± 0.03 versus 0.88 ± 0.03) (31). Bernard et al evaluated the performance of average deep learning based cardiac segmentation methods versus the Automated Cardiac Diagnosis Challenge ground-truth dataset (13). For the myocardium mass, they achieved an MAE of 10.4 g, in line with the presented results. The largest MAE for the LVMM was found in the HCM subgroup. This finding is expected, considering that an elevated total myocardial mass is the definitive criteria of this subgroup (Table 1). However, the automated measurements for the LVMM tend to be slightly underestimated compared with the manual measurements as the biases, which were all significant, were all negative for the overall dataset and all the subgroups. For the LVEDV measurement, we found an inferior MAE compared with Bernard et al (7.1 mL) for the average deep learning methods. Chaung et al evaluated the correlation between cardiac trabeculations and muscles and their impact on the LV anatomy and function but did not separate the papillary muscles from trabeculations in their measures (2). Lu et al achieved a global correlation coefficient between the measures derived from automated and manual segmentation of greater than 0.89 for LVEDV, and we found a correlation of 0.99 for the same measure (33). In the Bland-Altman analysis, there was good agreement between the measures of the clinical parameters with negligible bias. Lu et al achieved limits of agreement (± 1.96 SD) of ± 25 mL and 28.8 g for the LVEDV and LVMM, respectively, while we achieved limits of agreement of ± 5.8 mL and 7.3 g, respectively.

The results concerning automatic quality control are comparable to those obtained by Robinson et al for the LVM and LVC with an MAE of 0.03 and a high binary threshold classification accuracy (98.3% for the LVM and 97.5% for the LVC) (27). Regarding the LVT and LVPM segmentation labels, the literature offered no previous example of an automatic quality control method, and hence, we are unsure whether the methods that were successful for a larger class could also be successful for smaller and noisy classes.

Trabeculations are normal components of the LV and discrimination between normal and excessive trabeculations remains challenging and controversial (34). Cardiac MRI has been indicated to be the best diagnostic tool to evaluate the distribution and measures of trabeculations (8). Kholi et al showed that segmental measurements of the trabeculation thickness in transthoracic US resulted in an overdiagnosis of ETCM (35). In a cardiac MRI study, more than 40% of healthy individuals presented a trabeculated-over-compact-myocardium thickness ratio greater than 2.3 (36). However, Bentatou et al showed that trabeculations were not homogeneously distributed over the LV cavity and that quantification of the total amount of LV trabeculations was preferable to the thickness criteria (24). Lu et al published a fully automated LV segmentation and measurement method including the papillary muscles and trabeculations but without separating them (33). Hence, to quantify the total number of LV trabeculations by calculating the trabeculation mass, there was a need to measure these two distinct cardiac structures automatically on cardiac MRI.

One of the major strengths of the presented pipeline is the minimal human interaction needed. The only human interaction is to load the short-axis MRI sequence into the pipeline interface, then the system automatically complete the evaluation of the first frame of each short-axis cine cardiac MRI. In case of manual (visual) or automatic rejection of the segmentation results, physicians can directly perform manual segmentation on the cardiac MRI images with reference manual segmentation software (25).

Our study had limitations. This study relied on retrospective data to train, validate, and test the deep learning algorithm. Only two cardiac MRI systems at 1.5 T from two different manufacturers were used in the present study. ED frame selection for segmentation may have varied between patients. Technical parameters, such as the Hausdorff distance, can also be used to strengthen the obtained results. However, because of anatomic particularities of trabeculations that have been described above, it did not seem relevant to use the Hausdorff distance in this case, and the Dice coefficient was the only metric used. LV trabeculation is a small structure with some amount of partial volume effect on cardiac MRI giving an inherent interobserver and intraobserver variability. We intend to teach the algorithm with the segmentation of only one expert reader, to avoid interobserver bias. The drawback of that choice is that we privilege the segmentation of the manual 1a observer.

A fully automated deep learning based framework was proposed for segmentation of LV trabeculations and cardiac structures for short-axis ED cardiac MRI analysis including automatic quality control of the obtained segmentation results. The proposed automatic cardiac MRI analysis of trabeculations can be used in further research to help improve our understanding of excessive trabeculation cardiomyopathy.

Disclosures of Conflicts of Interest: A.B. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed no relevant relationships. Other relationships: institution has patents: SATT-SUD-EST, date of deposit 12/13/2019, demand number: DSO000042312, record number: DSO2019018721, depositors: AMU/CNRS/AP-HM. **J.F.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed no relevant relationships. Other relationships: institution has patents: SATT-SUD-EST, patent deposit number: DSO2019018721. **Z.B.** disclosed no relevant relationships. **G.H.** disclosed no relevant relationships. **A.L.** disclosed no relevant relationships. **M.B.**

disclosed no relevant relationships. **L.B.** disclosed no relevant relationships. **F.P.** disclosed no relevant relationships. **J.N.D.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: author is consultant for Microport; author paid for lectures by Bayer Healthcare, Siemens, Shire Takeda, and MSD; author received travel accommodations from GE Healthcare. Other relationships: disclosed no relevant relationships. **B.G.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed no relevant relationships. Other relationships: institution (Aix Marseille University) has issued patents: A software is actually protected by the Satt Sud Est, deposit number DSO2019018721. **A.J.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed no relevant relationships. Other relationships: institution has issued patents: CNRS/AMU/APHM, date of deposit 12/13/2019, demand number: DSO000042312, record number: DSO2019018721, depositors: AMU/CNRS/AP-HM.

References

1. Jenni R, Oechslin E, Schneider J, Attenhofer Jost C, Kaufmann PA. Echocardiographic and pathoanatomical characteristics of isolated left ventricular non-compaction: a step towards classification as a distinct cardiomyopathy. Heart 2001;86(6):666–671.

2. Jenni R, Oechslin EN, van der Loo B. Isolated ventricular non-compaction of the myocardium in adults. Heart 2007;93(1):11–15.

3. Jensen B, Agger P, de Boer BA, et al. The hypertrabeculated (noncompacted) left ventricle is different from the ventricle of embryos and ectothermic vertebrates. Biochim Biophys Acta 2016;1863(7 Pt B):1696–1706.

4. Habib G, Charron P, Eicher JC, et al. Isolated left ventricular non-compaction in adults: clinical and echocardiographic features in 105 patients. Results from a French registry. Eur J Heart Fail 2011;13(2):177–185.

5. Elliott P, Andersson B, Arbustini E, et al. Classification of the cardiomyopathies: a position statement from the European Society Of Cardiology Working Group on Myocardial and Pericardial Diseases. Eur Heart J 2008;29(2):270–276.

6. Gati S, Chandra N, Bennett RL, et al. Increased left ventricular trabeculation in highly trained athletes: do we need more stringent criteria for the diagnosis of left ventricular non-compaction in athletes? Heart 2013;99(6):401–408.

7. Gati S, Papadakis M, Papamichael ND, et al. Reversible de novo left ventricular trabeculations in pregnant women: implications for the diagnosis of left ventricular noncompaction in low-risk populations. Circulation 2014;130(6):475–483.

8. Thuny F, Jacquier A, Jop B, et al. Assessment of left ventricular non-compaction in adults: side-by-side comparison of cardiac magnetic resonance imaging with echocardiography. Arch Cardiovasc Dis 2010;103(3):150–159.

9. Petersen SE, Selvanayagam JB, Wiesmann F, et al. Left ventricular non-compaction: insights from cardiovascular magnetic resonance imaging. J Am Coll Cardiol 2005;46(1):101–105.

10. Jacquier A, Thuny F, Jop B, et al. Measurement of trabeculated left ventricular mass using cardiac magnetic resonance imaging in the diagnosis of left ventricular non-compaction. Eur Heart J 2010;31(9):1098–1104.

11. Captur G, Muthurangu V, Cook C, et al. Quantification of left ventricular trabeculae using fractal analysis. J Cardiovasc Magn Reson 2013;15(1):36.

12. Weinsaft JW, Cham MD, Janik M, et al. Left ventricular papillary muscles and trabeculae are significant determinants of cardiac MRI volumetric measurements: effects on clinical standards in patients with advanced systolic dysfunction. Int J Cardiol 2008;126(3):359–365.

13. Bernard O, Lalande A, Zotti C, et al. Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? IEEE Trans Med Imaging 2018;37(11):2514–2525.

14. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.

15. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June 27–30, 2016. Piscataway, NJ: IEEE, 2016; 770–778.

16. Tao Q, Yan W, Wang Y, et al. Deep Learning-based Method for Fully Automatic Quantification of Left Ventricle Function from Cine MR Images: A Multivendor, Multicenter Study. Radiology 2019;290(1):81–88.

17. Ngo TA, Carneiro G. Left ventricle segmentation from cardiac MRI combining level set methods with deep belief networks. In: 2013 IEEE International Conference on Image Processing, Melbourne, Australia, September 15–18, 2013. Piscataway, NJ: IEEE, 2013; 695–699.

18. Tran PV. A Fully Convolutional Neural Network for Cardiac Segmentation in Short-Axis MRI. arXiv:1604.00494 [cs]. http://arxiv.org/abs/1604.00494. Published April 2, 2016. Accessed April 29, 2018.

19. Franklin RCG, Béland MJ, Colan SD, et al. Nomenclature for congenital and paediatric cardiac disease: the International Paediatric and Congenital Cardiac Code (IPCCC) and the Eleventh Iteration of the International Classification of Diseases (ICD-11). Cardiol Young 2017;27(10):1872–1938.

20. Stöllberger C, Wegner C, Finsterer J. Left ventricular hypertrabeculation/noncompaction, cardiac phenotype, and neuromuscular disorders. Herz 2019;44(7):659–665.

21. Japp AG, Gulati A, Cook SA, Cowie MR, Prasad SK. The Diagnosis and Evaluation of Dilated Cardiomyopathy. J Am Coll Cardiol 2016;67(25):2996–3010.

22. Authors/Task Force members; Elliott PM, Anastasakis A, et al. 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 2014;35(39):2733–2779.

23. Bricq S, Frandon J, Bernard M, et al. Semiautomatic detection of myocardial contours in order to investigate normal values of the left ventricular trabeculated mass using MRI. J Magn Reson Imaging 2016;43(6):1398–1406.

24. Bentatou Z, Finas M, Habert P, et al. Distribution of left ventricular trabeculation across age and gender in 140 healthy Caucasian subjects on MR imaging. Diagn Interv Imaging 2018;99(11):689–698.

25. Frandon J, Bricq S, Bentatou Z, et al. Semi-automatic detection of myocardial trabeculation using cardiovascular magnetic resonance: correlation with histology and reproducibility in a mouse model of non-compaction. J Cardiovasc Magn Reson 2018;20(1):70.

26. Khened M, Kollerathu VA, Krishnamurthi G. Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. Med Image Anal 2019;51:21–45.

27. Robinson R, Oktay O, Bai W, et al. Real-time Prediction of Segmentation Quality. arXiv:1806.06244 [cs]. http://arxiv.org/abs/1806.06244. Published June 16, 2018. Accessed November 22, 2019.

28. Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. Radiology 2018;286(3):800–809.

29. Bluemke DA, Moy L, Bredella MA, et al. Assessing Radiology Research on Artificial Intelligence: A Brief Guide for Authors, Reviewers, and Readers-From the *Radiology* Editorial Board. Radiology 2020;294(3):487–489.

30. Sandhu R, Finkelhor RS, Gunawardena DR, Bahler RC. Prevalence and characteristics of left ventricular noncompaction in a community hospital cohort of patients with systolic dysfunction. Echocardiography 2008;25(1):8–12.

31. Bai W, Sinclair M, Tarroni G, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. J Cardiovasc Magn Reson 2018;20(1):65.

32. Chuang ML, Gona P, Hautvast GL, et al. Correlation of trabeculae and papillary muscles with clinical and cardiac characteristics and impact on CMR measures of LV anatomy and function. JACC Cardiovasc Imaging 2012;5(11):1115–1123.

33. Lu YL, Connelly KA, Dick AJ, Wright GA, Radau PE. Automatic functional analysis of left ventricle in cardiac cine MRI. Quant Imaging Med Surg 2013;3(4):200–209.

34. Arbustini E, Weidemann F, Hall JL. Left ventricular noncompaction: a distinct cardiomyopathy or a trait shared by different cardiac diseases? J Am Coll Cardiol 2014;64(17):1840–1850.

35. Kohli SK, Pantazis AA, Shah JS, et al. Diagnosis of left-ventricular non-compaction in patients with left-ventricular systolic dysfunction: time for a reappraisal of diagnostic criteria? Eur Heart J 2008;29(1):89–95.

36. Kawel N, Nacif M, Arai AE, et al. Trabeculated (noncompacted) and compact myocardium in adults: the multi-ethnic study of atherosclerosis. Circ Cardiovasc Imaging 2012;5(3):357–366.

37. Gulati A, Jabbour A, Ismail TF, et al. Association of fibrosis with mortality and sudden cardiac death in patients with nonischemic dilated cardiomyopathy. JAMA 2013;309(9):896–908 [Published correction appears in JAMA 2013;310(1):99.].

38. Milletari F, Navab N, Ahmadi SA. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. arXiv:1606.04797 [cs]. http://arxiv.org/abs/1606.04797. Published June 15, 2016. Accessed January 10, 2020.

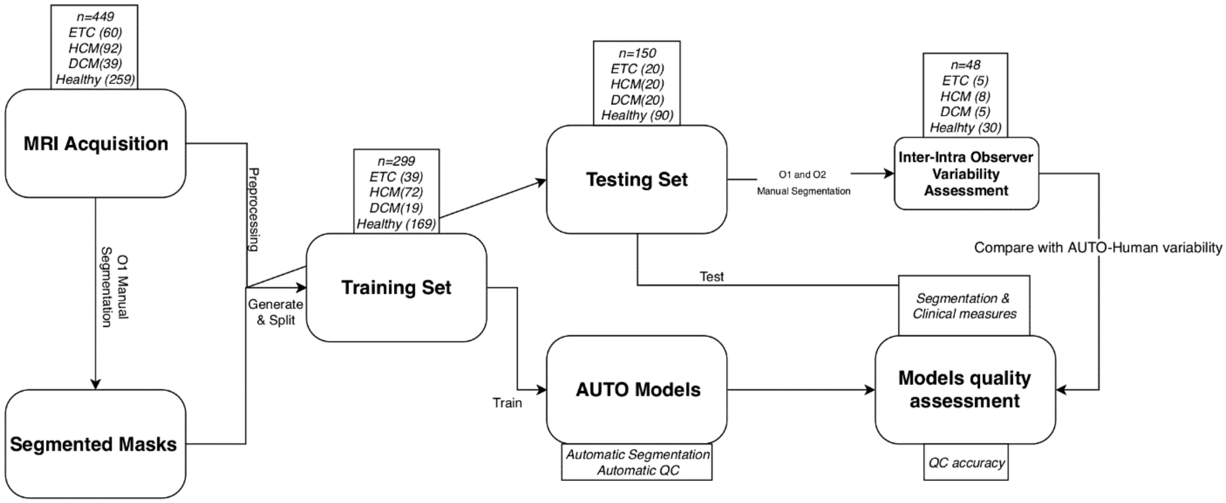


Figure 1: Overview of the study design and data flow. DCM = dilated cardiomyopathy, ETC = excessive trabeculation cardiomyopathy, HCM = hypertrophic cardiomyopathy, QC = quality control.

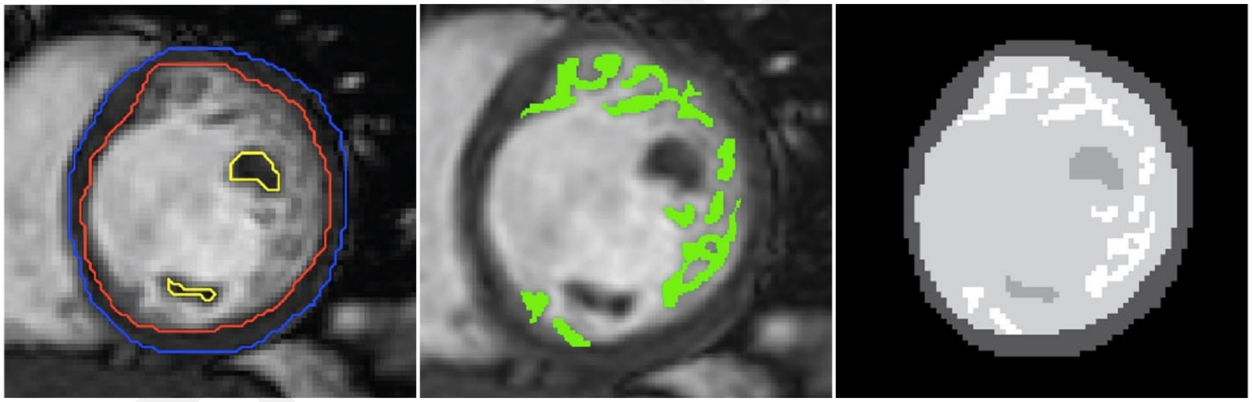
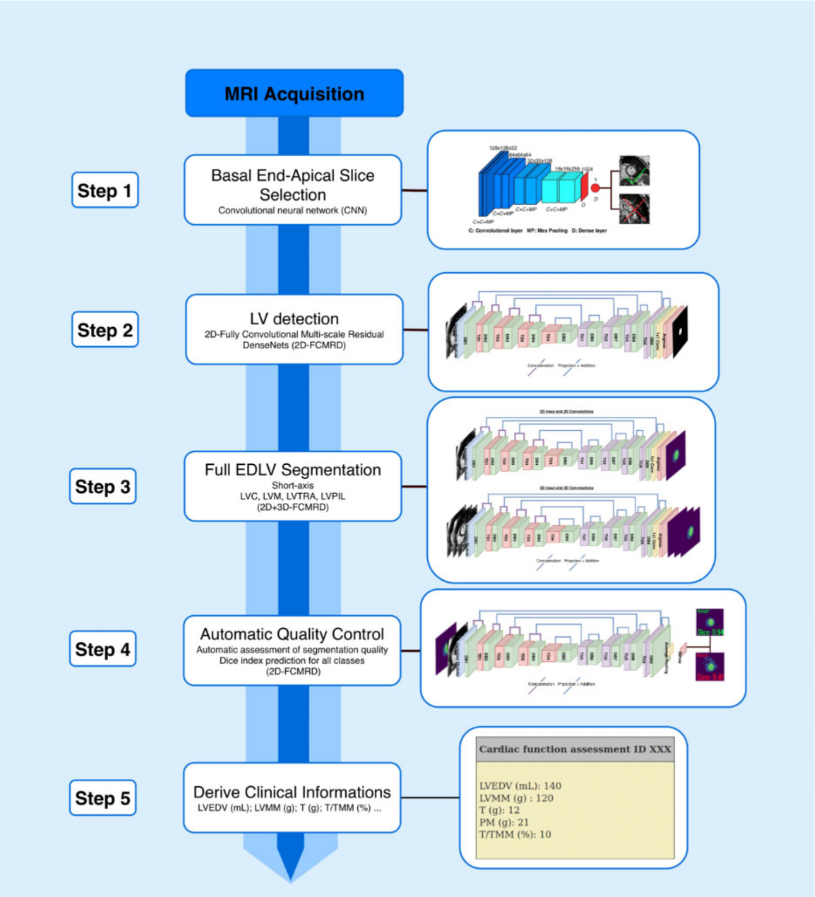


Figure 2: Left ventricle (LV) structures after manual reference segmentation. One mid ventricular end-diastolic short-axis cine cardiac MRI, using NCP processing software (23). **(a)** Epicardial border (blue layer), endocardial border (red layer), and the borders of the papillary muscles (yellow layer) is shown. **(b)** The trabeculation segmentation (light green) is shown. **(c)** LV labels for the same cardiac MRI: background (black); LV myocardium (dark gray); LV cavity (gray); LV trabeculations (light gray); LV papillary muscles (white).


MRI Acquisition
Step 1
Basal End-Apical Slice Selection
Convolutional neural network (CNN)
128x128x32
64x64x64
32x32x128
16x16x256
1024
1
C+C+MP
C+C+MP
C+C+MP
C+C+MP
D
D
C: Convolutional layer MP: Max Pooling D: Dense layer
Step 2
LV detection
2D-Fully Convolutional Multi-scale Residual DenseNets (2D-FCMRD)
Step 3
Full EDLV Segmentation
Short-axis
LVC, LVM, LVTRA, LVPIL
(2D+3D-FCMRD)
Step 4
Automatic Quality Control
Automatic assessment of segmentation quality
Dice index prediction for all classes
(2D-FCMRD)
Step 5
Derive Clinical Informations
LVEDV (mL); LVMM (g); T (g); T/TMM (%) ...

Cardiac function assessment ID XXX

LVEDV (mL): 140
LVMM (g) : 120
T (g): 12
PM (g): 21
T/TMM (%): 10

Figure 3: Complete pipeline and model description in five steps. LVEDV = left ventricle end-diastolic volume, LVMM = left ventricle myocardial mass; T = trabeculation mass, PM = papillary muscles, T/TMM = trabeculation-mass-to-total-myocardial-mass ratio.

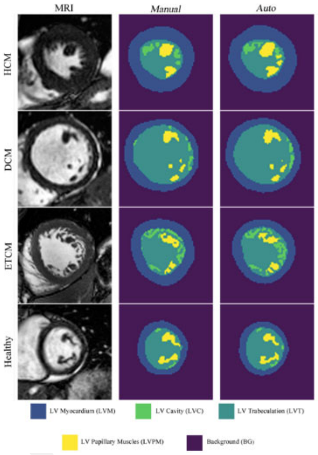


Figure 4: Examples of the obtained automatic left ventricular segmentations compared with the cardiac MRI and manual reference segmentation for different cardiac phenotypes. Short-axis cardiac MRI (left), manual reference segmentation (middle), and

obtained automated segmentations predicted by the model (right) for the four main cardiac phenotypes. DCM = dilated cardiomyopathy, ETCM = excessive trabeculations cardiomyopathy, HCM = hypertrophic cardiomyopathy.

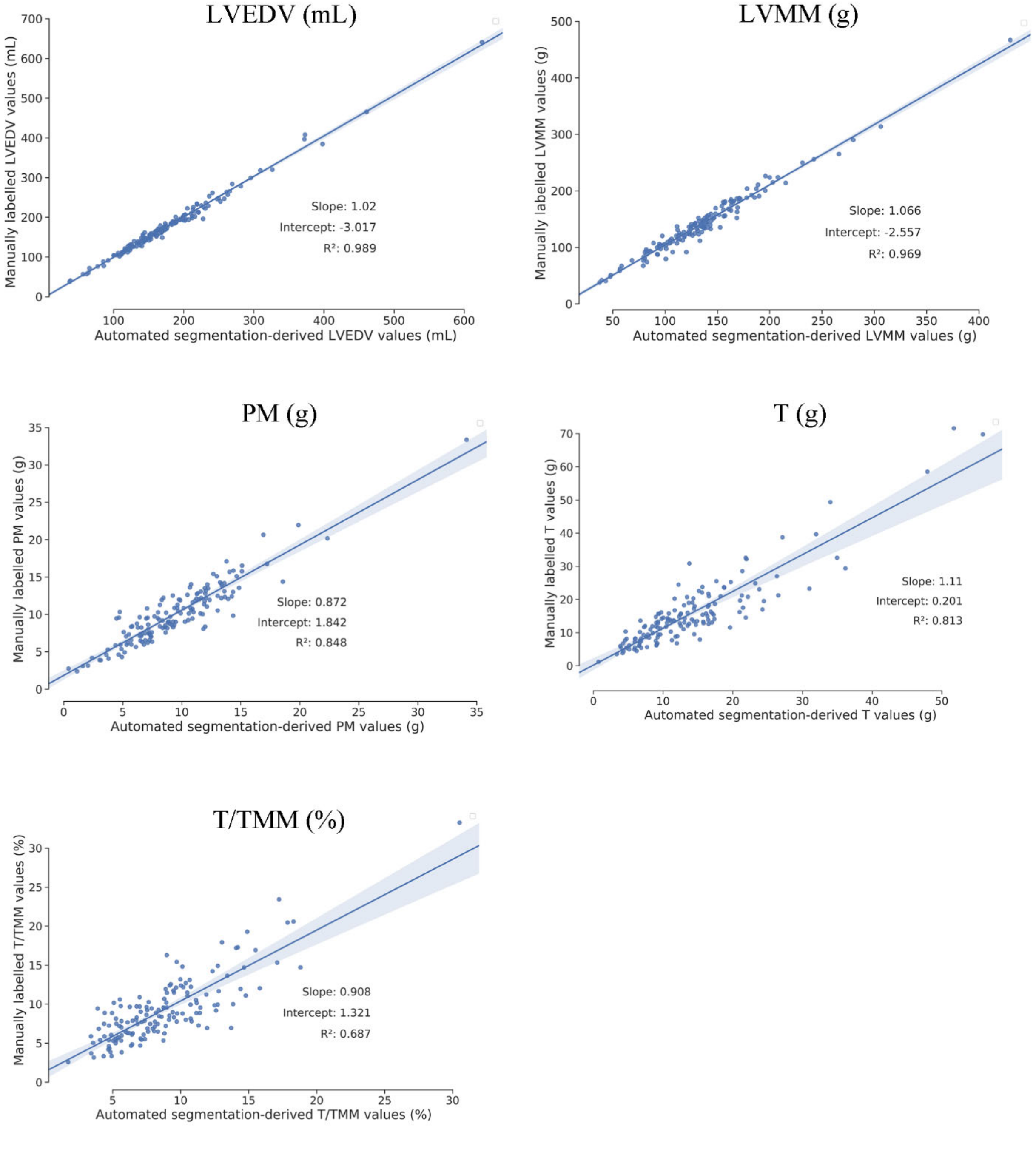


Figure 5: Correlation of clinical parameters measures between manual and automatic on the testing dataset (*n* = 150). Blue line is the fitted regression line. Blue surface represent confidence bands around regression line. Blue cross represent paired *t* test for each value. LVEDV = left ventricle end-diastolic volume, LVMM = left ventricle

myocardial mass; T = trabeculation mass; T/TMM = trabeculation-mass-to-total-myocardial-mass ratio.

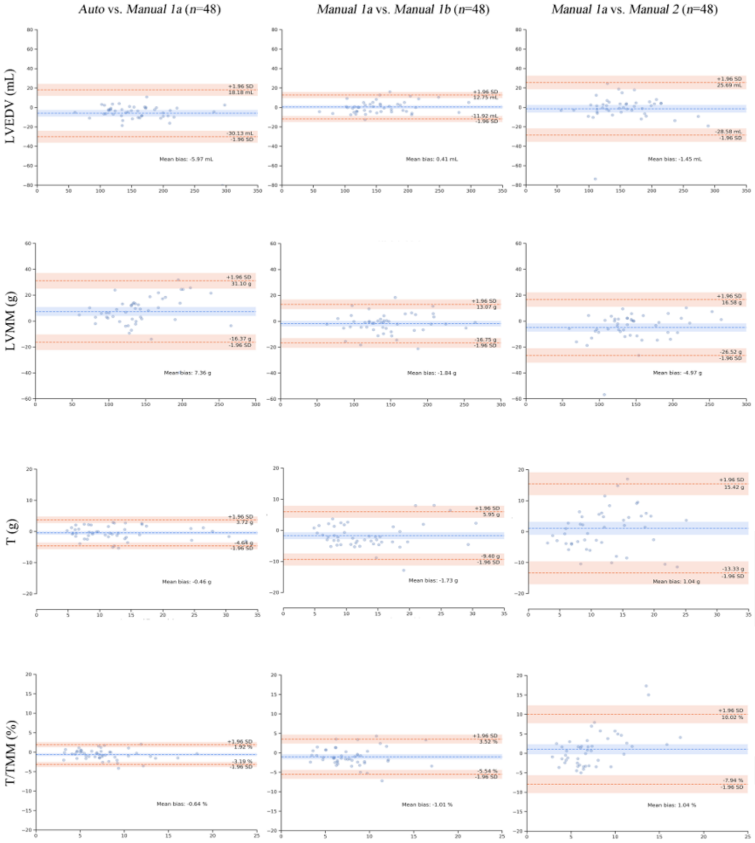


Figure 6: Bland-Altman analysis for the clinical parameters between the automatic and manual measurements. Evaluation between automatic versus manual 1a (left), manual 1a versus manual 1b (middle), and manual 1a versus manual 2 (right) on the 48 patients randomly selected from the testing dataset. In each Bland-Altman plot, the x-axis denotes the average of two measurements, and the y-axis is the difference between them. The blue dashed line denotes the mean difference (bias), and the two orange dashed lines denote ± 1.96 standard deviations from the mean. LVEDV = left ventricle end-diastolic volume, LVMM = left ventricle myocardial mass, T = trabeculation mass, T/TMM = trabeculation-mass-to-total-myocardial-mass ratio.

Table 1

Baseline Patient Characteristics for the Three Datasets

Parameter	HCM	DCM	ETCM	Healthy	Overall
A. Training Dataset (*n* = 299)					
Total No.	72	19	39	169	229
Age, yrs	53 ± 13	55 ± 13	44 ± 13	50 ± 13	50 ± 13
No. Men	43	11	26	94	174
LVEDV (mL)	150.5 ± 38.2	225.3 ± 113.4	175.3 ± 58.2	139.5 ± 36.4	151.6 ± 48.4
LVEDV/BSA (mL/m^2)	79.9 ± 17.4	115.0 ± 43.3	96.5 ± 27.0	77.9 ± 16.0	83.2 ± 23.0

LVMM (g)	168.1 ± 56.4	162.7 ± 45.0	112.9 ± 43.4	107.4 ± 31.1	126.3 ± 49.0
LVMM/BSA (g/m^2)	89.2 ± 28.0	87.2 ± 23.6	62.3 ± 21.1	59.6 ± 12.6	68.8 ± 23.2
T (g)	10.5 ± 5.0	13.2 ± 7.6	14.4 ± 6.6	8.1 ± 3.6	9.8 ± 5.2
T/BSA (g/m^2)	5.6 ± 2.7	6.9 ± 3.3	8.0 ± 3.4	4.5 ± 1.8	5.4 ± 2.7
T/TMM (%)	5.7 ± 2.3	6.9 ± 3.3	11.3 ± 4.5	6.6 ± 2.3	7.0 ± 3.2
B. Testing Dataset (*n* = 150)					
Total No.	20	20	20	90	150
Age, yrs	56 ± 15	54 ± 12	43 ± 15	45 ± 19	47 ± 17
No. Men	16	10	11	53	90
LVEDV (mL)	154.9 ± 38.7	237.7 ± 91.6	215.8 ± 130.4	159.4 ± 49.3	176.7 ± 76.4
LVEDV/BSA (mL/m^2)	80.3 ± 18.0	122.5 ± 39.5	111.0 ± 55.1	85.5 ± 24.6	93.1 ± 34.71
LVMM (g)	176.1 ± 43.8	165.0 ± 52.1	146.6 ± 95.2	126.0 ± 37.0	140.6 ± 54.3
LVMM/BSA (g/m^2)	91.1 ± 19.6	85.3 ± 21.9	75.2 ± 40.0	67.3 ± 17.1	73.9 ± 23.9
T (g)	14.4 ± 7.3	20.6 ± 10.6	27.7 ± 18.2	12.1 ± 5.5	15.6 ± 10.6
T/BSA (g/m^2)	7.4 ± 3.6	10.6 ± 5.2	14.3 ± 8.4	6.4 ± 2.8	8.2 ± 5.1
T/TMM (%)	7.0 ± 2.6	10.3 ± 3.5	15.4 ± 5.4	8.1 ± 2.9	9.2 ± 4.2
C. Reproducibility Dataset (*n* = 48)					
Total No.	8	5	5	30	48
Age, yrs	57 ± 18	51 ± 17	32 ± 17	47 ± 19	44 ± 19
No. Men	6	3	4	18	31
LVEDV (mL)	150.39 ± 22.6	254.9 ± 91.0	158.5 ± 70.3	146.8 ± 41.5	159.8 ± 57.9
LVEDV/BSA (mL/m^2)	78.1 ± 11.1	126.5 ± 33.5	83.2 ± 30.3	79.2 ± 23.1	84.4 ± 27.1
LVMM (g)	174.7 ± 45.6	180.9 ± 63.2	115.0 ± 41.0	137.4 ± 35.1	145.81 ± 44.4
LVMM/BSA (g/m^2)	90.4 ± 21.3	90.0 ± 24.2	61.2 ± 18.6	73.4 ± 16.1	76.7 ± 19.8
T (g)	10.63 ± 4.2	15.84 ± 5.0	16.7 ± 6.7	10.2 ± 4.9	11.5 ± 5.4
T/BSA (g/m^2)	5.5 ± 2.2	8.7 ± 3.1	9.0 ± 3.5	5.5 ± 2.7	6.2 ± 3.0
T/TMM (%)	5.4 ± 1.9	7.6 ± 0.8	11.9 ± 3.0	6.4 ± 2.2	6.9 ± 2.8

Note.—The means ± SDs are reported. BSA = body surface areas, HCM = hypertrophic cardiomyopathy, DCM = dilated cardiomyopathy, ETCM = excessive trabeculations cardiomyopathy, LVEDV = left ventricular end-diastolic volume, LVMM = left ventricular myocardial mass, PM = papillary muscles mass, T = trabeculation mass, T/TMM = trabeculation-mass-to-total-myocardial-mass ratio.

Table 2

3D Dice MAE for the Automatic Quality Control Module

Labels	**MAE**	**Binary** Accuracy
LVM	0.03 ± 0.03	97.2%
LVC	0.02 ± 0.02	96.5%
LVT	0.05 ± 0.04	94.5%
LVPM	0.05 ± 0.06	97%

Note.—The means ± SDs of the metrics are reported. MAE = mean absolute error, LVC = left ventricular cavity, LVM = left ventricular myocardium; LVPM = left ventricular papillary muscles, LVT = left ventricular trabeculation.

Table 3

Model segmentation performance for the technical and clinical parameters.

Auto versus Manual (*n* = 150)					
	HCM (*n* = 20)	DCM (*n* = 20)	ETCM (*n* = 20)	Healthy (*n* = 90)	Overall (*n* = 150)

A. Technical Metric: Dice Coefficient					
LVC	0.96 ± 0.01	0.97 ± 0.01	0.96 ± 0.01	0.96 ± 0.02	0.96 ± 0.01
LVM	0.91 ± 0.02	0.86 ± 0.02	0.88 ± 0.03	0.89 ± 0.03	0.89 ± 0.03
LVPM	0.82 ± 0.05	0.79 ± 0.07	0.72 ± 0.19	0.79 ± 0.09	0.79 ± 0.11
LVT	0.61 ± 0.07	0.62 ± 0.09	0.66 ± 0.08	0.62 ± 0.09	0.62 ± 0.08
B. Clinical Parameters					
Left ventricular end-diastolic volume					
MAE (mL)	3.60 ± 3.3	8.90 ± 8.5	9.45 ± 7.5	4.90 ± 4.5	5.86 ± 5.8
Bias (mL)	1.0 ± 4.8	–2.7 ± 12.1	–0.8 ± 12.2	–0.2 ± 6.6	–0.4 ± 8.3
P value	0.37	0.58	0.6	0.99	0.84
Correlation	0.99	0.99	0.99	0.99	0.99
Left ventricular myocardial mass					
MAE (g)	13.15 ± 8.4	9.75 ± 6.7	10.82 ± 10.2	8.27 ± 6.1	9.46 ± 7.3
Bias (g)	–12.5 ± 9.4	–8.6 ± 8.2	–10.8 ± 10.2	–3.5 ± 9.7	–6.3 ± 10.2
P value	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Correlation	0.98	0.99	0.99	0.97	0.99
Papillary muscle mass					
MAE (g)	1.52 ± 1.1	1.45 ± 1.1	0.97 ± 0.7	1.39 ± 1.1	1.36 ± 1.1
Bias (g)	–0.9 ± 1.6	–0.2 ± 1.9	–0.1 ± 1.2	–0.8 ± 1.7	–0.6 ± 1.7
P value	0.04	0.41	0.79	< 0.01	< 0.01
Correlation	0.90	0.87	0.99	0.85	0.92
Trabeculation mass					
MAE (g)	3.52 ± 3.1	5.19 ± 4.5	6.79 ± 5.0	2.60 ± 2.0	3.63 ± 3.4
Bias (g)	–2.8 ± 3.8	–3.1 ± 6.2	–3.7 ± 7.7	–0.7 ± 3.2	–1.7 ± 4.7
P value	< 0.01	0.04	0.06	0.02	< 0.01
Correlation	0.90	0.82	0.92	0.83	0.90
Trabeculation-mass-to-total-myocardial-mass ratio					
MAE (g)	1.32 ± 1.0	2.16 ± 1.3	2.94 ± 1.7	1.82 ± 1.5	1.95 ± 1.5
Bias (g)	–0.8 ± 1.5	–0.9 ± 2.4	–0.7 ± 3.4	–0.3 ± 2.3	–0.5 ± 2.4
P value	0.05	0.14	0.37	0.18	0.01
Correlation	0.83	0.74	0.80	0.66	0.83

Note.—The means ± SDs of the metrics are reported. DCM = dilated cardiomyopathy, ETCM = excessive trabeculations cardiomyopathy, LVC = left ventricular cavity, LVM = left ventricular myocardium; LVPM = left ventricular papillary muscles, LVT = left ventricular trabeculation, MAE: mean absolute error.

Table 4

Automatic Segmentation Performance Compared with Human Reproducibility

	Manual 1a versus Auto (*n* = 48)					Manual 1a versus Manual 1b (*n* = 48)					Manual 1a versus Manual 2 (*n* = 48)				
	HCM (*n* = 8)	DCM (*n* = 5)	ETCM (*n* = 5)	Healthy (*n* = 30)	Overall (*n* = 48)	HCM (*n* = 8)	DCM (*n* = 5)	ETCM (*n* = 5)	Healthy (*n* = 30)	Overall (*n* = 48)	HCM (*n* = 8)	DCM (*n* = 5)	ETCM (*n* = 5)	Healthy (*n* = 30)	Overall (*n* = 48)
A. Technical Metric: Dice Coefficient															
LVC	0.96 ± 0.01	0.94 ± 0.06	0.96 ± 0.01	0.96 ± 0.02	0.96 ± 0.02	0.96 ± 0.02	0.97 ± 0.01	0.97 ± 0.02	0.96 ± 0.01	0.96 ± 0.01	0.96 ± 0.02	0.96 ± 0.02	0.97 ± 0.03	0.95 ± 0.05	0.96 ± 0.04
LVM	0.92 ± 0.01	0.87 ± 0.04	0.90 ± 0.02	0.90 ± 0.02	0.90 ± 0.03	0.94 ± 0.04	0.90 ± 0.05	0.91 ± 0.05	0.90 ± 0.03	0.91 ± 0.04	0.94 ± 0.03	0.90 ± 0.06	0.93 ± 0.06	0.88 ± 0.06	0.89 ± 0.06
LVPM	0.83 ± 0.03	0.79 ± 0.06	0.76 ± 0.11	0.81 ± 0.06	0.81 ± 0.07	0.83 ± 0.04	0.75 ± 0.05	0.78 ± 0.11	0.76 ± 0.13	0.77 ± 0.11	0.75 ± 0.07	0.72 ± 0.10	0.79 ± 0.11	0.72 ± 0.08	0.74 ± 0.04
LVT	0.66 ± 0.05	0.61 ± 0.09	0.72 ± 0.08	0.61 ± 0.11	0.63 ± 0.10	0.60 ± 0.07	0.58 ± 0.05	0.65 ± 0.09	0.56 ± 0.09	0.58 ± 0.09	0.46 ± 0.12	0.50 ± 0.04	0.56 ± 0.09	0.42 ± 0.13	0.44 ± 0.09
B. Clinical Parameters															
Left ventricular end-diastolic volume															
MAE (mL)	4.4 ± 4.5	19.6 ± 34.1	4.3 ± 3.5	6.3 ± 4.8	7.1 ± 11.7	5.2 ± 4.9	7.8 ± 3.0	5.2 ± 5.0	4.5 ± 3.3	5 ± 3.8	7.2 ± 6.2	11.9 ± 6.1	3 ± 1.2	8.4 ± 13.7	8.04 ± 11.3
Bias (mL)	0.5 ± 6.5	18.6 ± 34.8	4.3 ± 3.5	5.6 ± 5.5	6 ± 12.3	4.2 ± 5.9	3 ± 8.5	3 ± 6.8	–1.5 ± 5.4	0.4 ± 6.3	1.5 ± 9.8	–7.2 ± 12.2	1.7 ± 3.0	–1.8 ± 16.1	–1.4 ± 13.8
Left ventricular myocardial mass															
MAE (g)	13.8 ± 8.8	17.4 ± 14.9	6.4 ± 7.0	9.9 ± 7.7	10.9 ± 8.9	5.4 ± 4.6	8.3 ± 8.5	7.7 ± 7.1	5 ± 4.7	5.67 ± 5.3	5.3 ± 4.6	5.9 ± 3.9	4.7 ± 4.3	9.2 ± 11.1	7.87 ± 9.2
Bias (g)	–7.1 ± 1 5.5	0.0 ± 24.5	–4.9 ± 8.3	–9.1 ± 8.6	–7.4 ± 12.1	–5.4 ± 4.6	–0.5 ± 12.6	–0.4 ± 11.2	–1.4 ± 6.7	–1.8 ± 7.6	–2 ± 7.0	–4.6 ± 5.7	–3 ± 6.0	–6.2 ± 13.1	–5 ± 11.0
Papillary muscle mass															
MAE (g)	1.4 ± 1.3	2.5 ± 3.1	1.2 ± 1.3	1.3 ± 1.3	1.4 ± 1.6	1.5 ± 1.2	3.2 ± 3.4	1.5 ± 1.7	1.4 ± 1.2	1.6 ± 1.6	2.6 ± 1.9	2.4 ± 2.2	2.2 ± 2.2	2.8 ± 1.5	2.65 ± 1.7
Bias (g)	–1.0 ± 1.6	–2.3 ± 3.3	–0.9 ± 1.6	–0.3 ± 1.8	–0.7 ± 2.0	1.5 ± 1.2	2.5 ± 4.1	1 ± 2.1	1 ± 1.6	1.2 ± 1.9	–2.1 ± 2.5	–2.4 ± 2.2	–2.2 ± 2.2	–2.8 ± 1.5	–2.5 ± 1.8
Trabeculation mass															
MAE (g)	1.9 ± 1.5	1.9 ± 1.6	1.4 ± 0.9	1.7 ± 1.3	1.7 ± 1.3	3.3 ± 2.5	5.9 ± 4.7	4.8 ± 2.6	3 ± 1.4	3.54 ± 2.4	5.1 ± 5.0	8.7 ± 2.0	4 ± 1.4	0.9 ± 4.9	5.84 ± 4.5
Bias (g)	–0.2 ± 2.5	0.9 ± 2.5	1.0 ± 1.4	0.5 ± 2.1	0.5 ± 2.1	–2.6 ± 3.3	–1.4 ± 8.0	1.7 ± 5.6	–2.1 ± 2.6	–1.7 ± 3.9	3.2 ± 6.6	–2.5 ± 9.5	1.8 ± 4.2	0.9 ± 7.6	1 ± 7.3
Trabeculation-mass-to-total-myocardial-mass ratio															
MAE (%)	0.8 ± 0.8	1.4 ± 1.1	1.3 ± 0.5	1.1 ± 1.0	1.1 ± 0.9	1.6 ± 1.4	3.5 ± 3.1	2.6 ± 0.7	1.9 ± 0.9	2.07 ± 1.4	2.4 ± 2.3	4.1 ± 1.2	2.5 ± 1.4	3.7 ± 3.8	3.36 ± 3.2
Bias (%)	0.3 ± 1.1	0.5 ± 1.8	1.2 ± 0.5	0.7 ± 1.4	0.6 ± 1.3	–1.1 ± 1.9	–1.5 ± 4.7	0.3 ± 3.0	–1.1 ± 1.8	–1 ± 2.3	1.4 ± 3.1	0.2 ± 0.47	1.1 ± 2.9	1.1 ± 5.2	1 ± 4.6

Note.—The means ± SDs of the metrics are reported. DCM = dilated cardiomyopathy, ETCM = excessive trabeculations cardiomyopathy, HCM = hypertrophic cardiomyopathy, LVC = left ventricular cavity, LVM = left ventricular myocardium, LVPM = left ventricular papillary muscles, LVT = left ventricular trabeculation, MAE = mean absolute error.

Appendix

This appendix explains in detail the 5 steps of the proposed algorithm.

Step 1: Basal to End-Apical LV Stack Selection.—A dedicated CNNs with a dedicated dataset was trained to classify whether each image as belonging to the LV basal to end-apical range in short-axis orientation or not.

Step 2: LV Detection and Cropping.—To facilitate the visualization of the LV before segmentation, we trained a distinct DFCNNs to segment the LV as a whole in each image and then to extract a 128 × 128 pixel region around the computed center of the LV mass.

Step 3: LV Structure Segmentation.—Automatic image segmentation was performed on cropped images to segment the LVM, LVC, LVPM, LVT, and background. The automated segmentations obtained were named *Auto*.

Step 4: Automatic Quality Control.—An automatic quality control module was built to predict the automatic segmentation quality.

Step 5: Clinical Parameters.—The last step of the pipeline is the automated computation of the clinical parameters from the predicted segmentation: LVEDV, LVMM, PM, T, and T/TMM.
Appendix A

Dataset

A.1. Dataset for basal end-apical stack selection (step 1)

To automatically detect the stacks of interest inside all the acquired stacks, we retrospectively selected 224 consecutive cardiac MRI examinations with their ED short-axis cine cardiac MRI images separated into those between the most basal slice and most apical slice and those outside of the left ventricle. A mostly basal slice was a slice that showed the LV myocardium extending over at least 50% of the myocardial circumference. A mostly apical slice was the last slice where the LV myocardium was seen.

The data acquisition was performed between November 2018 and April 2019 from our institution. Inclusion criteria and exclusion criteria were similar to the training test. There was no cardiac phenotype classification for this dataset. Cardiac MRI examinations were performed on 1.5 T Avanto scanner (Siemens Healthcare, Erlangen Germany) with the same technical parameters as described in the manuscript. The dataset was separated into 184 patients for training and validation and 40 for testing the model.

A.2. Dataset for LV region of interest detection and cropping (step 2)

The same dataset as in step 3 was used, except that each pixel in the dataset was classified as belonging to the LV or not, hence resulting in a pixel-wise binary classification problem.

A.3. Dataset for LV structures segmentation (step 3)

This dataset main characteristics are described in manuscript.

Concerning patients characteristics, DCM was defined as dilation of the LV with cardiac dysfunction as described by Gulati et al (21). HCM was defined based on myocardial thickness on end-diastolic short-axis cardiac MRI of greater than 15 mm following the European Society of Cardiology recommendations (22). ET was defined according to the criteria outline by Petersen et al: double-layer myocardial aspect and a noncompacted layer over a compacted layer thickness of greater than 2.3 on the end-diastolic long axis (9).

Concerning the Cardiac MRI characteristics. For the 1.5T Ingenia scanner (Philips Health System, Best, the Netherlands): in-plane spatial resolution was set between 2 mm^2 and 2.35 mm^2, the slice thickness was set to 8 mm, and the gap between slices was set to 0 mm. For the 1.5T Avanto scanner (Siemens Healthcare, Erlangen, Germany): in-plane spatial resolution was set between 1.37 mm^2 and 1.48 mm^2, the slice thickness was set to 6 mm, and the gap between slices was 0 mm. The flip angle and resolution were adapted to the morphologies of the patients. The acquisitions were electrocardiogram synchronized and were performed under breath-hold. The short-axis image stack consisted of 8 to 17 slices depending on the scanner, patient height, cardiac anatomy and morphology. Each cardiac slice consists of 30 time frames.

A.4. Dataset for segmentation automatic quality control (step 4)

The dataset was built by training CNN (to see CNN used for step 2, refer to §B.2.) models with varying number of feature maps per convolutional layer, number of training epochs (1, 2, 10, 100) and randomly choosing subsets from the dataset used in step 3 as training data. The resulting trained models were used to segment the 449 cardiac MRI sequences of the dataset used in step 3. To have a score-balanced dataset we computed the histogram of the Dice of all the predictions obtained from the different models and selected an equal number of predictions in each bin (corresponding to the minimum counts-per-bin) of the Dice value distribution for the LVT class for the following bins: [0, 0.2], [0.2, 0.3], [0.3, 0.4], [0.4, 0.5], [0.5, 0.6], [0.7, 1].

The inputs are constructed as in Robinson et al where for each segmentation 5 one-hot encoded masks are generated for each class: background, LVM, LVC, LVT and LVMP and concatenated with the cardiac MRI input (27).
Appendix B

Networks Architecture

B.1. Network architecture for basal end-apical stack selection (step 1)

The 14-layers CNN used for basal end-apical stack selection is illustrated in Figure E1.

It is composed of 4 convolution-convolution-max pooling blocks followed by one Dense layer with 1024 units, all with ReLU activations and with batch normalization used before every convolutional layer. The last layer is a Dense layer with one unit and sigmoid activation. It takes as input 128 × 128 images and returns a value between 0 and 1 that corresponds to the prediction of the probability for the input of belonging to the basal to end-apical range or not.

B.2. Network architecture for LV region of interest detection and cropping (step 2)

For LV detection, we used a Deep Fully Convolutional Neural Network (DFCNN), introduced by Khened et al (26). We refer to this architecture as $DFCNN_{2DSeg}$*. This network is almost identical to the network used by Khened et al (26) on the ACDC dataset, we only added a pooling step and used k = 12 feature maps per convolutional layer. The network takes as input 256 × 256 images and returns 256 × 256 segmented maps where each pixel is predicted as belonging to the LV or not, as illustrated in Figure E2.*

A dense block (DB) is composed of several dense block layers, a composition of batch normalization (BN), exponential linear unit (ELU), 3 × 3 convolution and a dropout layer with a drop-out rate of $P = .2$. Our network has 4 dense blocks in the downsampling path with, in the ascending order, two, three, four and five dense block layers; bottleneck dense block has 6 DB layers; the dense blocks properties in the up-sampling path are mirrored from the down-sampling path. Transition down (TD) block reduces the spatial resolution of the feature maps as the depth increases and is composed of BN, ELU, 1 × 1 convolution, dropout ($P = .2$) and 2 × 2 max-pooling layers. A transition up (TU) increases the resolution of the feature maps and is composed of a 3 × 3 transposed convolution with a stride of 2. Each convolutional layer in the network has k = 12 feature maps for step 2.

B.3. Networks architecture for LV structures segmentation (step 3)

For the segmentation of the LV structures, we combined a 2D and a 3D network, both illustrated in Figure E3:

One $DFCNN_{2DSeg}$ as in step 2 except it has k = 36 feature maps per convolutional layer.

It also takes as input 128 × 128 images cropped around the LV and returns 128 × 128 segmented maps where each pixel is predicted as belonging to background, LVM, LVC, LVT or LVPM.

One network referred to as $DFCNN_{3DSeg}$. It differs from $DFCNN_{2DSeg}$ only because the convolutional layers have 3 × 3 × 3 filters and the inputs are also in 3D. The inputs of $DFCNN_{3DSeg}$ correspond to 3 consecutive (in a patient, slice-wise) 128 × 128 images concatenated together, it then returns the corresponding segmented map, were each voxel is predicted as belonging to background, LVM, LVC, LVT or LVPM.

At test time, for each slice-level image, the corresponding output of the softmax layers of the two networks are added together and the argmax is taken on the last axis to obtain 128 × 128 segmented maps.

B.4. Networks architecture for segmentation quality control (step 4)

Two networks with the same architecture (referred to as $DFCNN_{QC}$*) were used for segmentation quality prediction.*

Both networks take as input 128 × 128 × 6 tensors, corresponding to the concatenation of a 128 × 128 cropped around the LV image and 5 one-hot encoded 128 × 128 masks which are generated, given a segmentation, for each class: background, LVM, LVC, LVT and LVMP. Both networks return a vector of length 5.

For a given input tensor, one network predicts the Dice value for each class while the other predicts the MVSF value for each class.

$DFCNN_{QC}$ differs from $DFCNN_{2DSeg}$ in the following way: the last 1 × 1 convolutional layer is replaced by an average pooling layer with a pooling size of 128 × 128 followed by a dense layer with 5 units and a sigmoid activation to predict quality values between 0 and 1. An illustration is given in Figure E4.

At test time, the Dice value predictions for each class at patient level (3D) are reconstructed from the predicted Dices and MVSF values for each 2D image and corresponding segmentation in the sequence of a patient, this mathematical reconstruction is detailed in Appendix E.4.
Appendix C

Preprocessings

C.1. Image preprocessing algorithms

Three different algorithms were used for this step, illustrated for a cardiac MRI input on Figure E5.

Algorithm 1. Min-max normalization:

$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$, where X is the original voxel intensity in an image, the minimum and maximum values are taken with regard to all voxel intensities in the image.

Algorithm 2. Contrast Limited Adaptive Histogram Equalization (CLAHE).

Local contrast enhancement, that uses histograms computed over different tile regions of the image. It was implemented as in the skimage Python library with default parameters. It was always applied on the output of the min max normalization.

Algorithm 3. Global histogram equalization with 256 bins.

As implemented in the skimage Python library. Always applied on the output of the min max normalization.

C.2. Preprocessing for basal end-apical stack selection (step 1)

The images were first resized to a 128 × 128 resolution and Algorithm 2 was applied.

C.3. Preprocessing for region of interest detection (step 2)

Algorithms 1, 2 and 3 were applied on original images separately and their outputs concatenated together, resulting in 3-channels 2D images. The resulting images were resized to a 256 × 256 × 3 resolution, while manually annotated maps were resized to a 256 × 256 resolution. If an image had a horizontal or a vertical resolution smaller than 256 zero padding was used. If an image had a horizontal or a vertical resolution larger than 256 then center cropping was used. In each manual segmentation, all the pixels belonging to the LV region were converted to have the same value.

C.4. Preprocessing for LV structures segmentation (step 3)

Algorithms 1, 2 and 3 were applied on cropped around the LV 128 × 128 images separately and their outputs concatenated together, resulting in 3-channels 2D images. The resulting images had a 128 × 128 × 3 resolution, while manually annotated maps had a 128 × 128 resolution.

C.4. Preprocessing for segmentation quality control (step 4)

As said in §B.4. the networks in this step take as input 128 × 128 × 6 tensors, corresponding to the concatenation of a 128 × 128 cropped around the LV image, on which only Algorithm 2 was applied, and 5 one-hot encoded 128 × 128 masks.

Appendix D

Training Parameters

D.1. Training parameters for basal end-apical stack selection (step 1)

Binary cross entropy loss function was used. Training was performed for 300 epochs with the Adam optimizer and a learning rate of 0,001. Random flips, croppings, Gaussian blurs and affine transformations where applied during training. The best model was selected based on the loss on the validation set.

D.2. Training parameters for region of interest detection (step 2)

The Dice loss, as defined by Milletari et al was used (38). The network was trained for 150 epochs with the Adam optimizer with a learning rate of 0,001. Random rotations between-1 and 1 radius were applied during training. The network which gave the smallest loss on the validation set was kept.

D.3. Training parameters for LV structures segmentation (step 3)

Weighted Dice-cross entropy loss, as defined by Khened et al was used for both networks used in this step (26). The 2D network and 3D network were trained for 400 and 200 epochs respectively both with the Adam optimizer with a learning rate of 0,001. Random rotations between-1 and 1 radius and random croppings were applied during training. The networks which gave the smallest loss on the validation set were kept.

D.4. Training parameters for segmentation quality control (step 4)

Mean squared error loss was used for both networks used in this step. Both networks were trained for 160 epochs with the Adam optimizer and a learning rate of 0,001. The networks which gave the smallest loss on the validation set were kept.

Appendix E

Postprocessings

E.1. Post processing for basal end-apical stack selection (step 1)

At test time, all the images in the sequence of a patient are classified as belonging to the basal to end-apical range or not. If a given image was classified as not belonging to the range of interest while the previous image and the next image in the sequence were classified as belonging to it, this given image will be classified as belonging to the basal to end-apical range.

E.2. Post processing for region of interest detection (step 2)

No post processing is applied on the predicted segmentation maps.

The center of mass of the pixels predicted as belonging to the LV is computed.

The distance between all predicted centers in the slices of a patient are computed to detect the 3 centers that are the closest to each other. The mean of those 3 centers is computed and taken as a reference center, also the mean distance of those 3 points to the reference center is taken as the reference distance

Then a post processing is applied is the following way: for each predicted center in a patient if the distance of that center to the reference center is superior to 23 times the reference distance, then this center is rejected and the reference center is taken instead for that slice.

Those centers are then used to obtain centered around the LV images.

E.3. Post processing for LV structures segmentation (step 3)

Slice-wise 2D connected component analysis was made and only one (the largest) connected component of the nonbackground predicted pixels is kept as nonbackground.

E.4. Post processing for segmentation quality control (step 4)

No post processing is done on the predictions for this step.

Appendix F

Results

F.1. Results for basal end-apical stack selection (step 1)

For each patient we computed the patient-level accuracy by dividing the number of correctly predicted slices by the total number of slices for that patient.

We report a mean patient accuracy of 95.4% for the 40 patients in the test set for step 1.

Additionally the average number of slices that were classified differently by the human operator and the CNN on our test set was 0.675, for an average number of slices per patient of 14.3.

The majority (55%) of the patients in the test set had no difference with the human operator.

F.1. Results for region of interest detection (step 2)

We report a mean Euclidean distance between the predicted centers and labeled centers of 1.25 pixels (1.41 without postprocessing), standard deviation of 1.87 (3.45 without postprocessing) for the 1578 2D images composing the 150 patients in the testing for step 2.

For all the 2D images, 100% of the LV were completely inside the 128 × 128 square centered on the predicted center of mass.

Appendix G

Additional Information

G.1. How the cropped around the LV images are obtained (step 2)

To obtain cropped around the LV 128 × 128 images from the 256 × 256 segmentation maps, the center of mass of the pixels classified as belonging to the LV is computed and a 128 × 128 patch centered on it is extracted.

G.2. How the patient level (3D) Dice values are obtained (step 4)

Let suppose X *and* Y *are two 3D regions with* n *stacks, or equivalently* $X=(x_i)_{i\in 1,n}$ *and* $Y=(y_i)_{i\in 1,n}$ *where* x_i *and* y_i *are 2D regions.*

Therefore we can express the Dice between X and Y as:

$$Dice(X,Y)=\frac{2|X\cap Y|}{|X|+|Y|}.$$

When we only have the segmentation of a given cardiac MRI sequence and not the manual segmentations considered as ground truth, it is similar to get $X=(x_i)_{i\in 1,n}$ without having $Y=(y_i)_{i\in 1,n}$ in this context.

One model denoted as M_1 is used to estimate $Dice(x_i,y_i)$ from x_i and related cardiac MRI mri_{x_i}:

$$M_1(x_i,mri_{x_i})=\widehat{Dice(x_i,y_i)}.$$

A second model denoted as M_2 is used to estimate $MVSF(x_i,y_i)=\frac{2(|x_i|-|y_i|)}{|x_i|+|y_i|}$:

$$M_2\left(x_i, mri_{x_i}\right)=\widehat{MVSF\left(x_i,y_i\right)}.$$

Finally predictions of the 3D Dice are recovered:

$$\widehat{Dice(X,Y)}=\frac{2\widehat{|X\cap Y|}}{|X|+\widehat{|Y|}},$$

with $|X|=\sum_{i\in1,n}|x_i|$, $\widehat{|Y|}=\sum_{i\in1,n}\widehat{|y_i|}$, $\widehat{|X\cap Y|}=\frac{1}{2}\sum_{i\in1,n}\widehat{Dice(x_i,y_i)}\left(|x_i|+\widehat{|y_i|}\right)$, and

$$\widehat{|y_i|}=|x_i|\frac{2+\widehat{MVSF(x_i,y_i)}}{2-\widehat{MVSF(x_i,y_i)}}.$$

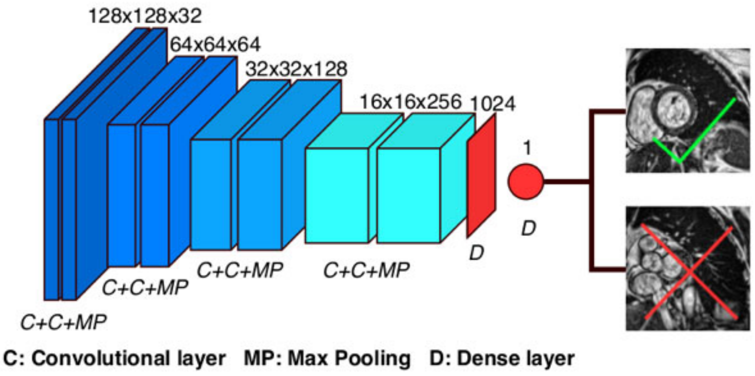


Figure E1: Illustration of 14-layer CNN architecture developed for step 1. C: convolutional layer; MP: max pooling; D: dense layer.

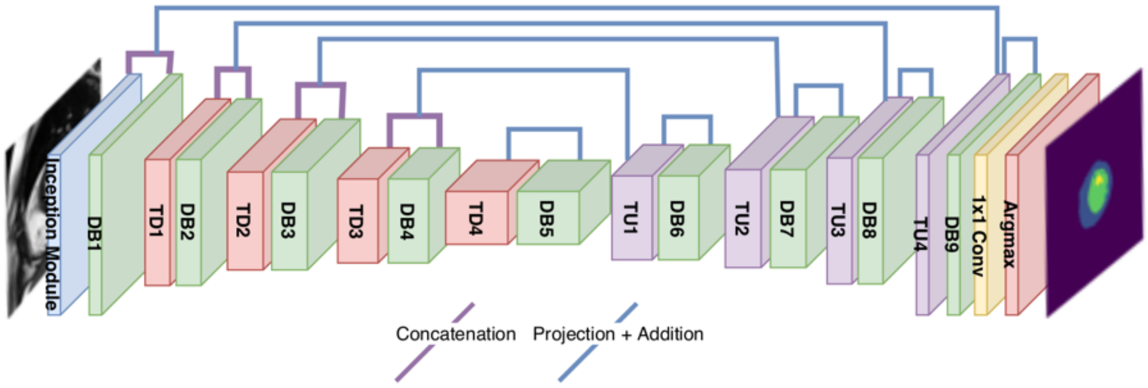


Figure E2: Illustration of CNN architecture developed for step 2. Purple line represents a concatenation. Blue line represents a projection and addition.

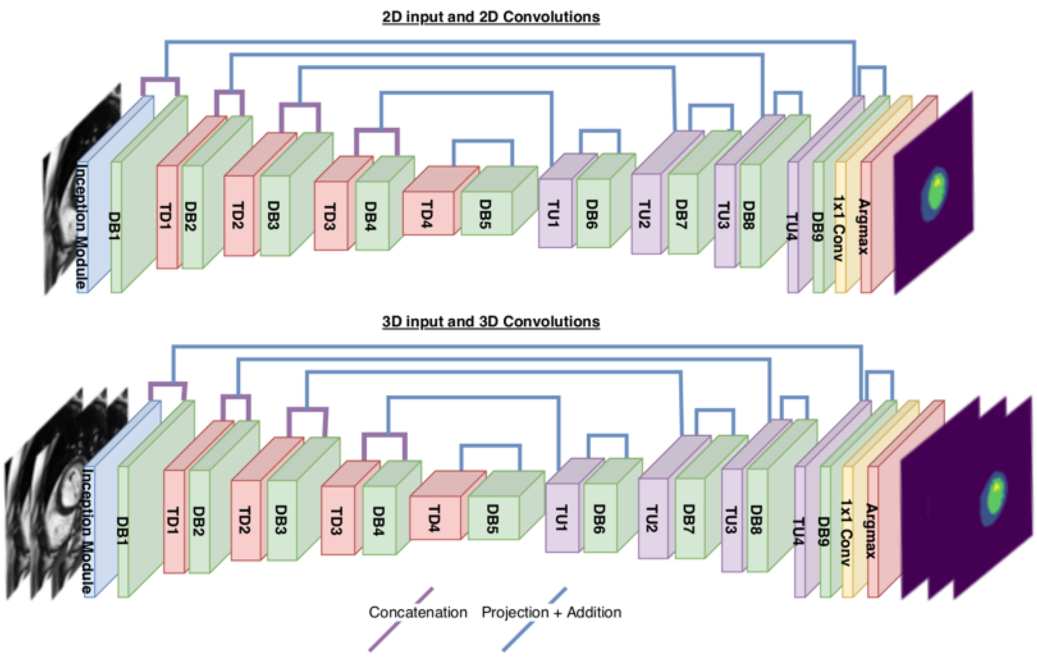


Figure E3: Illustration of both 2D and 3D CNN architecture developed for step 3. Purple line represents a concatenation. Blue line represents a projection and addition.

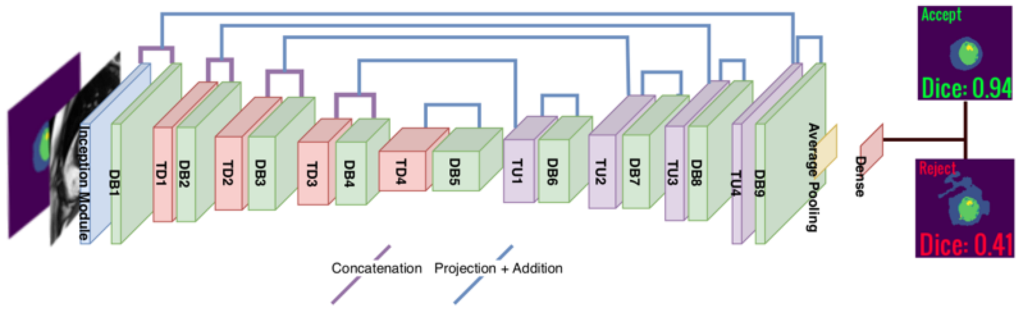


Figure E4: Illustration of CNN architecture developed for step 4. Purple line represents a concatenation. Blue line represents a projection and addition.

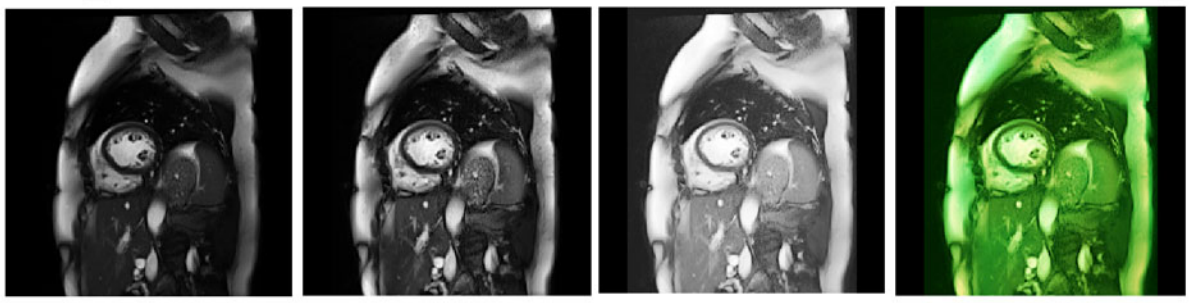


Figure E5: The three algorithms in ascending order and their concatenation together on a single end-diastolic short-axis cardiac MRI (from left to right).