Deep Learning Techniques for Automatic MRI Cardiac Multi-structures Segmentation and Diagnosis: Is the Problem Solved?

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohé, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jäger, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Chrisitan F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Išgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, and Pierre-Marc Jodoin

***Abstract*—Delineation of the left ventricular cavity, myocardium and right ventricle from cardiac magnetic resonance images (multi-slice 2D cine MRI) is a common clinical task to establish diagnosis. The automation of the corresponding tasks has thus been the subject of intense research over the past decades. In this paper, we introduce the "Automatic Cardiac Diagnosis Challenge" dataset (ACDC), the largest publicly-available and fully-annotated dataset for the purpose of Cardiac MRI (CMR) assessment. The dataset contains data from 150 multi-equipments CMRI recordings with reference measurements and classification from two medical experts. The overarching objective of this paper is to measure how far state-of-the-art deep learning methods can go at assessing CMRI, *i.e.* segmenting the myocardium and the two ventricles as well as classifying pathologies. In the wake of the 2017 MICCAI-ACDC challenge, we report results from deep learning methods provided by nine research groups for the segmentation task and four groups for the classification task. Results show that the best methods faithfully reproduce the expert analysis, leading to a mean value of 0.97 correlation score for the automatic extraction of clinical indices and an accuracy of 0.96 for automatic diagnosis. These results clearly open the door to highly-accurate and fully-automatic analysis of cardiac CMRI. We also identify scenarios for which deep learning methods are still failing. Both the dataset and detailed results are publicly available on-line, while the platform will remain open for new submissions.**

***Index Terms*—Cardiac segmentation and diagnosis, deep learning, MRI, left and right ventricles, myocardium.**

O. Bernard and F. Cervenansky are with the University of Lyon, CREATIS, CNRS UMR5220, Inserm U1044, INSA-Lyon, University of Lyon 1, Villeurbanne, France. E-mail: olivier.bernard@creatis.insa-lyon.fr.

C. Zotti and P.-M. Jodoin are with the Computer Science Department, University of Sherbrooke, Sherbrooke, Canada.

A. Lalande is with the Le2i laboratory, CNRS FRE 2005, University of Burgundy, Dijon, France and with the MRI department, University Hospital of Dijon, Dijon, France.

O. Humbert is with the TIRO-UMR E 4320 laboratory, University of Nice, Nice, France and with the department of Nuclear Medicine, Centre Antoine-Lacassagne, Nice, France.

X. Yang and P.A. Heng are with the department of computer science and engineering, the Chinese University of Hong Kong, Hong Kong, China.

I. Cetin, K. Lekadir, O. Camara, M. A. Gonzalez Ballester and G. Sanroma are with the Barcelona Centre for New Medical Technologies (BCN-MedTech), Universitat Pompeu Fabra, Barcelona, Spain.

M.A. Gonzalez Ballester is also with ICREA Barcelona, Spain.

S. Napel is with Stanford University School of Medicine, Department of Radiology, Stanford, CA, USA.

S. Petersen is with the Queen Mary University of London, William Harvey Research Institute, UK

G. Tziritas and E. Grinias are with the Department of Computer Science, University of Crete, Greece

M. Khened, V.A. Kollerathu and G. Krishnamurthi are with the Department of Engineering Design, IIT-Madras, Chennai-600036, India

M.M. Rohé,X. Pennec and M. Sermesant are with the Inria-Asclepios Project, BP 93 06902 Sophia Antipolis, France

F. Isensee, P. Jäger and K. H. Maier-Hein are with Division of Medical Image Computing German Cancer Research Center Heidelberg, Germany

I. Wolf and S. Engelhardt are with the Department of Computer Science, Mannheim University of Applied Sciences, Mannheim, Germany

P. M. Full is with the Department of Computer Science, Mannheim University of Applied Sciences, Mannheim, Germany and the Department of Cardiac Surgery, Heidelberg University Hospital, Heidelberg, Germany

C. F. Baumgartner is with the Computer Vision Laboratory, ETH Zürich, Switzerland

L. M. Koch is with the Computer Vision and Geometry Group, ETH Zürich, Switzerland

J.M. Wolterink and I. Išgum are with the Image Sciences Institute, University Medical Center Utrecht, Utrecht, The Netherlands

Y. Jang and Y. Hong are with Integrative Cardiovascular Imaging Research Center, Yonsei University College of Medicine, South Korea

J. Patravali and S. Jain are with qure.ai company, Mumbai, India

I. Introduction

Analysis of cardiac function plays an important role in clinical cardiology for patient management, disease diagnosis, risk evaluation, and therapy decision [1], [2], [3]. Thanks to digital imagery, the assessment of a set of complementary indices computed from different structures of the heart is a routine task for cardiac diagnostics. Because of its well-known capacity for discriminating different types of tissues, Cardiac MRI (CMR) (built from series of parallel short axis slices) is considered as the gold standard of cardiac function analysis through the assessment of the left and right ventricular ejection fractions (EF) and stroke volumes (SV), the left ventricle mass and the myocardium thickness. This requires accurate delineation of the left ventricular endocardium and epicardium, and of the right ventricular endocardium for both end diastolic (ED) and end systolic (ES) phase instances. In clinical practice, semi-automatic segmentation is still a daily practice because of the lack of accuracy of fully-automatic cardiac segmentation methods. This leads to time consuming tasks prone to intra- and inter-observer variability [4].

The difficulties of CMR segmentation have been clearly identified [5]: *i)* presence of poor contrast between myocardium and surrounding structures (conversely, there is a high contrast between blood and the myocardium); *ii)* brightness heterogeneities in the left ventricular/right ventricular

cavities due to blood flow; *iii)* presence of trabeculae and papillary muscles with intensities similar to the myocardium; *iv)* non-homogeneous partial volume effects due to the limited CMR resolution along the long-axis; *v)* inherent noise due to motion artifacts and heart dynamics; *vi)* shape and intensity variability of the heart structures across patients and pathologies; *vii)* presence of banding artifact.

In order to gauge performances of state-of-the-art CMR segmentation methods, four international challenges (all with a unique dataset) have been organized over the last decade [6], [7], [8], [9]. As mentioned in section II, three of those datasets focus on the left ventricle and one on the right ventricle. Since three of those challenges were organized before 2012, none of the participants implemented a deep learning approach. As for the fourth one, since the dataset only contains the ground-truth for the ED and ES ventricular volume (and not for the contour) it is difficult to ascertain which cardiac segmentation method was the most accurate and where it failed.

In this paper, we propose a new dataset called ACDC (Automatic Cardiac Diagnosis Challenge) which led to the organization of an international MICCAI challenge in 2017. The richness of the dataset as well as its tight bound to every-day clinical issues has the potential to enable machine learning methods to fully analyze cardiac MRI data. ACDC has a larger scope than previous cardiac datasets as it includes manual expert segmentation of the right ventricle (RV) and left ventricle (LV) cavities, and the myocardium (epicardial contour more specifically). ACDC also contains patients from five different medical groups namely : dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), myocardial infarction with altered left ventricular ejection fraction (MINF), abnormal right ventricle (ARV) and patients without cardiac disease (NOR).

The overarching objective of this paper is to provide answers to the following four questions :

1) How accurate recently proposed segmentation methods are at delineating the LV, RV and myocardium given clinical MR images?
2) How accurate recently proposed classification methods are at predicting the pathology of a patient given clinical MR images?
3) When methods fail, where do they fail?
4) How far are we from "solving" the problem of automatic CMRI analysis?

With those questions in mind, we first go through a detailed description of the previous MRI cardiac datasets as well as the CMRI segmentation methods in section II. We then describe our evaluation framework as well as the evaluated deep learning architectures in sections III and IV. We analyze the results obtained during the MICCAI-ACDC challenge in section V and finally draw conclusions in sections VI and VII.

II. Previous works

Previous MRI cardiac datasets

Four large datasets of clinical CMRI data have been broadly accepted by the community in the last decade. These datasets were released in conjunction with an international challenge allowing the organizers to benchmark state-of-the-art methods.

The Sunnybrook Cardiac MR Left Ventricle Segmentation challenge - MICCAI 2009[1] provides a database of 45 cardiac cine-MR images from four different pathological groups namely: heart failure with ischemia, heart failure without ischemia, hypertrophic cardiomyopathy, and normal subjects. The data is provided with two manually-drawn contours, one for the endocardium and one for the epicardium [6]. Although the database is still publicly available, neither collated results nor comparative study have been published thus reducing the impact of this event. However, recent papers [5], [10], [11] reported results from several automatic and semi-automatic segmentation methods published since the 2009 challenge. According to those results, the top performing methods (many of which being only focused on the endocardium segmentation) report Dice scores between 0.90 and 0.94 for the endocardium and/or the epicardium and an average perpendicular distance of less than 2.0 mm and an average 2D Hausdorff distance between 3.0 and 5.0 mm.

The LV Segmentation Dataset and Challenge, MICCAI-STACOM 2011[2] focuses on the comparison of LV segmentation methods [12]. The database is made of CMR acquisitions from 200 patients with coronary artery disease and prior myocardial infarction (100 for training and 100 for testing). In this study, the authors introduced the concept of objective ground truth based on the evidence from the contribution of several raters. In particular, ground truths computed for the 100 patients of the testing set were generated from an Expectation-Maximization framework (the STAPLE algorithm) [13] using the results of two fully-automated methods (automated raters) and three semi-automated approaches with manual input (manual raters). No 100% manually annotated ground truth were involved in this study. From the derived ground truths, the best results in terms of segmentation accuracy were obtained by a guide-point modeling technique (manual rater) which obtained an average Jaccard score of 0.84 [14].

The Right Ventricle Segmentation Dataset - MICCAI 2012[3] aims at comparing RV segmentation methods based on a set of 48 cardiac cine-MR data with contours drawn by one cardiac radiologist (16 for training, 32 for testing) [8]. Three fully-automatic and four semi-automatic methods were evaluated through this challenge. Back in 2012, the outcome of the challenge revealed that the best scores were obtained by semi-automatic methods like the graph-cut method by Grosgeorge *et al.* [15] which reached an average Dice score of 0.78 and an average 2D Hausdorff distance of 8.62. In a recent publication, Phi Vu Tran [16] showed how a fine-tuned fully-convolutional neural network [17] can out-perform every semi-automatic method with an average Dice score of 0.85.

The 2015 Kaggle Second Annual Data Science Bowl[4] is a challenge for which more than 190 teams competed to win the $200,000 grand price. The goal of this event was to automatically measure ED and ES volumes from CMR.

[1] http://smial.sri.utoronto.ca/LV_Challenge/Home.html
[2] www.cardiacatlas.org/challenges/lv-segmentation-challenge/
[3] www.litislab.fr/?projet=1rvsc
[4] www.kaggle.com/c/second-annual-data-science-bowl

TABLE I
SUMMARY OF THE FULL SET OF EXISTING CARDIAC MRI DATASETS WHICH ARE PUBLICLY AVAILABLE FOR COMPARISON PURPOSES.

CMRI datasets								
Name	**Year**	**Nb Subjects**		**Ground truth**				**Active website**
		train	test	LV	RV	Myo	Pathology	
Sunnybrook	2009	45	—	✔	✘	✔	✔	✘
STACOM	2011	100	100	✔	✘	✔	✘	✘
MICCAI RV	2012	16	32	✘	✔	✘	✘	✘
Kaggle	2015	500	200	✘	✘	✘	✘	✘
ACDC	2017	100	50	✔	✔	✔	✔	✔

Challengers were given a database composed of 500 patients for training and 200 patients for testing. The training images came only with the ED and ES reference volumes and not a manually segmented ground truth as for the other three datasets. The outcome of the challenge revealed that the top-performing methods relied on deep learning technologies, in particular fully convolutional networks (fCNN) [17] and U-Net [18]. Unfortunately, no summary paper was provided in the wake of this challenge.

Table I summaries the MRI cardiac datasets mentioned above. Let us also mention that other fully-annotated cardiac datasets have been released such as HVSMR 2016[5] and the Multi-Modality Whole Heart Segmentation dataset[6]. Although interesting, these datasets contain images that a clinically atypical, a topic that goes beyond the focus of this study. Furthermore, without being bound to a challenge, the UK Biobank [19] corresponds to the largest existing CMR database which could be used to train and test deep learning methods whenever the manual annotations of these images will be rendered public. However, one limit of this database is that it is not free, which inevitably limits its access by research teams, and thus does not correspond to open science initiatives such as challenges.

Non-deep learning methods

In parallel to those challenges, Petitjean *et al.* proposed in 2011 a complete review of segmentation methods for delineating the LV and/or the RV in short axis cardiac MR images [20]. In this study, the authors listed the results published in more than 70 peer-reviewed publications. As for the four challenges cited above, the reported methods can be divided in two main categories: weak prior and strong prior methods. The first group involves weak assumptions such as spatial, intensity or anatomical information. It includes image-based techniques (threshold, dynamic programming) [21], pixel classification methods (clustering, Gaussian mixture model fitting) [22], deformable models (active contour, level-set) [23] and graph-based approaches (graph-cut) [24]. The second group uses methods with strong prior including shape prior based deformable models [5], active shape and appearance models [25] and atlas based methods [26], all requiring a training dataset with manual annotations. Although this huge work provides a complete picture of the performance of the state-of-the-art methods in LV/RV segmentation, it does benchmark these techniques with a unique dataset. Such comparison thus remains a glaring issue in our community.

Deep learning methods

To our knowledge, before 2013 no deep learning techniques was used to analyze CMRI. However, a drastic change occurred in 2015 during the Kaggle Second Annual Data Science Bowl during which the undeniable power of deep learning methods was revealed to the community. Since then, a dozen deep learning papers have been published on the topic of CMRI analysis. Most papers used 2D convolutional neural networks (CNNs) and analyzed the MRI data slice by slice.

Three papers used deep learning framework to extract relevant features for segmentation. Emad *et al.* [27] used a patch-wise CNN to localize the LV in CMRI slices. Kong *et al.* [28] developed a temporal regression framework to identify end-diastolic and end-systolic instances from the cardiac cycle by integrating a 2D CNN with a recurrent neural network (RNN). The CNN was used to encode the spatial information while the RNN was used to decode the temporal information. Finally Zhang *et al.* [29] used a simple CNN to automatically detect missing slices (apical and basal) in cardiac exams to assess the quality of MRI acquisitions.

Four papers used deep learning methods combined with classical cardiac segmentation tools. Rupprecht *et al.* [30] integrated a patch-based CNN into a semi-automatic active contour (a snake) to segment cardiac structures. Ngo *et al.* [31] used a deep belief network (DBN) to accurately initialize and guide a level-set model to segment the left ventricle. Yang *et al.* [32] developed a combined approach between CNN and multi-atlas to perform LV segmentation. In particular, a deep architecture was trained to learn deep features achieving optimal performance for the label fusion operation classically involved in multi-atlas segmentation. Alternatively, Avendi *et al.* [10] proposed a combined deep-learning and deformable-model approach to automatically segment the left ventricle. The method works as follows: *i)* a simple CNN locates and crops the LV; *ii)* a stack of autoencoders pre-segment the LV shape; *iii)* the pre-segmented shape is refined with a deformable model. Although the authors report almost perfect results on Sunnybrook 2009, it is not clear how their method generalizes to more than one cardiac region.

Finally, three papers used standalone deep learning techniques to segment cardiac structures from CMR data. Poudel *et al.* [33] proposed a recurrent fully-convolutional network (RFCN) that learns image representations from the full stack

[5]http://segchd.csail.mit.edu/
[6]http://stacom2017.cardiacatlas.org/

of 2D slices. The derived architecture allows leveraging inter-slice spatial dependences through internal memory units. Tran *et al.* [16] developed a deep fully convolutional neural network architecture to segment both LV and RV structures. Finally, Oktay *et al.* [34] proposed an image super-resolution approach based on a residual convolutional neural network model. Their key idea is to reconstruct high resolution 3D volumes from 2D image stacks for more accurate image analysis.

For more details on deep learning methods applied to medical image analysis (including cardiac MRI segmentation) please refer to Litjens *et al.* [35] and Havaei *et al.* [36].

III. Evaluation Framework

A. CMR data

1) Patient selection: The ACDC dataset was created from real clinical exams acquired at the University Hospital of Dijon (France). Our dataset covers several well-defined pathologies with enough cases to properly train machine learning methods and clearly assess the variability of the main physiological parameters obtained from cine-MRI (in particular diastolic volume and ejection fraction). The targeted population is composed of 150 patients evenly divided into 5 classes with well-defined characteristics according to physiological parameters. These examinations were initially classified according to medical reports. Patients with ambiguous clinical indices were excluded from this study. The different subgroups are given hereunder:

- NOR: Examination with normal cardiac anatomy and function. The ejection fraction is greater than 50%, the wall thickness in diastole is lower than 12 mm, the LV diastolic volume is below 90 mL/m^2 for men and 80 mL/m^2 for women [37]. The RV is normal for each patient (RV volume less than 100 mL/m^2 and RV ejection fraction above 40%). The visual analysis of the segmental LV and RV myocardial contraction is normal.
- MINF: Patients with a systolic heart failure with infarction. Subjects have an ejection fraction below 40% and abnormal myocardial contractions. Some subjects have a high diastolic LV volume due to a remodeling of the LV to compensate for the myocardial infarction.
- DCM: Patients with dilated cardiomyopathy have an ejection fraction below 40%, a LV volume greater than 100 mL/m^2 and a wall thickness in diastole smaller than 12 mm. As a consequence of dilated LV, some patients of this category have a dilated RV and/or a high LV mass.
- HCM: Patients with hypertrophic cardiomyopathy, *i.e.* a normal cardiac function (ejection fraction greater than 55%) but with myocardial segments thicker than 15 mm in diastole. In this category, patients can present abnormal cardiac mass indices with values above 110 g/m^2.
- ARV: Patients with abnormal right ventricle have a RV volume greater than 110 mL/m^2 for men, and greater than 100 mL/m^2 for women [38], or/and a RV ejection fraction below 40%. Almost every subject in this subgroup has a normal LV.

2) Acquisition protocol: Acquisitions were obtained over a 6 year period with two MRI scanners of different magnetic strengths (1.5 T - Siemens Area, Siemens Medical Solutions, Germany and 3.0 T - Siemens Trio Tim, Siemens Medical Solutions, Germany). Cine MR images were acquired with a conventional SSFP sequence in breath hold with a retrospective or prospective gating [39]. After the acquisitions of long axis slices, a series of short-axis slices covering the LV from the base to the apex was acquired, with a slice thickness from 5 mm to 10 mm (in general 5 mm) and sometimes an inter-slice gap of 5 mm. The spatial resolution varies from 1.34 to 1.68 $mm^2/pixel$. Depending on the patient, 28 to 40 volumes were acquired to cover completely (retrospective gating) or partially (prospective gating) one cardiac cycle. In the latter case, only 5 to 10% of the end of the cardiac cycle was omitted. The full dataset was acquired in clinical routine, leading to natural variability in the image quality (intrinsic noise, patient movement, banding artifacts, MRI low-frequency intensity fluctuation, etc.), variable field-of-view and integral or almost integral covering of the LV. Finally, to be in compliance with previous cardiac MRI segmentation challenges, the long axis slices were not provided. Even though the use of long axis slices could provide extra information about the base, the apex and the longitudinal motion of the ventricles, the analysis of short and long-axis slices are generally independent and outside the scope of this project.

3) Training and testing dataset: The data for each subject was converted to a general 4D image representation format (nifti) without loss of resolution. ED and ES frames were identified based on the motion of the mitral valve from the long axis orientation by a single expert. Both training and testing data contain whole short-axis slices. The identification of the most basal and apical slices is also not provided, while the diastolic and systolic phases are indicated. In order for challengers to normalize the physiological parameters (mainly the LV and RV volumes and the MYO mass) with the body surface area (BSA), the weight and height of each patient are included in the dataset. For instance, the BSA can be calculated from the formula of Dubois and Dubois [40], *i.e.* $BSA = 0.007184 \cdot (weight^{0.425} \cdot height^{0.725})$ and normalized parameters can be computed by simply dividing their values with the corresponding BSA. The training database is composed of 100 patients, *i.e.* 20 patients for each group. For all these data, the corresponding manual references as well as the patient group are provided. The testing dataset is composed of 50 patients, *i.e.* 10 patients per group. The manual references and group labels of the testing data are kept private.

B. Reference segmentation and contouring protocol

The expert references are manually-drawn 3D volumes of the LV and RV cavities as well as the myocardium, both at the ED and ES gates. The epicardial border of the RV was not considered because its accurate position next to the septum is difficult to establish, and the myocardial thickness of the RV is of the same order of magnitude than the spatial resolution. The contours were drawn and double-checked by two independent

experts (10 and 20 years of experience) who had to reach consensus in case of discordance.

The following annotation rules were retained: the LV and RV must be completely covered, the papillary muscle are included into the cavity and there is no interpolation of the muscle at the base of the LV (the contours follow the limit defined by the aortic valve). The main difficulty when annotating RV corresponds to correctly localize the pulmonary infundibulum area. This area must not be included into the RV annotation and a clear separation must be seen between the RV cavity and the root of the pulmonary artery. Due to the systolic shortening of the RV, the first basal slice is not mandatory being the same in diastole and systole. Another difficulty is to accurately separate the RV from the right atrium on the systolic image. As such, we defined the RV as the region on the right of heart with a significant contraction between ventricular diastole and systole, *i.e.* the surface area of the RV must be higher in ventricular diastole than in ventricular systole. For an easier understanding, illustrations of the annotation rules are provided in the supplementary materials (available in the supplementary files /multimedia tab).

The ground truth label images were stored in nifti format. The label values vary from 0 to 3 and represent voxels belonging to the background (0), the RV cavity (1), the myocardium (2) and the LV cavity (3).

C. Evaluation metrics

In order to evaluate the tested methods in a fair and reproducible manner, we customized a dedicated Girder[7]online platform[8]. This platform is now available and will be maintained and kept open as long as the data remains relevant for clinical research. Based to this platform, the performance of state-of-the-art methods are compared both from a geometrical and a clinical standpoint. This implies the use of a complementary set of metrics as described hereunder [41].

1) Geometrical metrics: In order to measure the accuracy of the segmentation output (LV endocardium, myocardium or RV endocardium) provided by a given method, the Dice metric and the 3D Hausdorff distance were used.

Dice similarity index: The Dice similarity index is defined as $D = 2\left(|V_{user} \cap V_{ref}|\right) / \left(|V_{user}| + |V_{ref}|\right)$ and is a measure of overlap between the segmented volume V_{user} extracted from a method and the corresponding reference volume V_{ref}. The Dice index gives a measurement value between 0 (no overlap) and 1 (full overlap).

Hausdorff surface distance (d_H): The Hausdorff distance d_H, measures the local maximum distance between the two surfaces S_{user} and S_{ref}. This is carried out efficiently using the Proximity Query Package (PQP) [42] which we slightly modified to compute point-to-triangle distances. Moreover, in order to minimize the difference between sampling densities of S_{user} and S_{ref}, we apply a linear subdivision operator to the surface containing the lowest number of vertices. As opposed to several MRI cardiac segmentation papers which report 2D Hausdorff distances [5], [10], [11], we report the 3D d_H, which allows an intrinsic management of the missing segmentation problem on the end slices.

2) Clinical performance: We also implemented three indices for the clinical parameters, namely the correlation ($corr$), the bias and the standard deviation (std) values. These three metrics are computed from the measurements of: *i)* the ED volumes (LV_{EDV} and RV_{EDV} expressed in mL/m^2 for the LV and RV, respectively); *ii)* the ejection fractions (LV_{EF} and RV_{EF} expressed in percent for the LV and RV, respectively); *iii)* the myocardium mass (MY_{Mass} expressed in g/m^2 and calculated in diastole). The combination of the bias and standard deviation also provides useful information on the corresponding limit of agreement values.

Let us mention that these geometrical and clinical metrics are complementary in the sense that a good score on one metric does not inevitably imply a good score on other metrics. This property is fundamentally important to prevent our system from unexpectedly favoring some methods over others. For instance, a low EF error does not always mean a good delineation of the ED and ES ventricle since EF relies on the difference between the ED and ES volumes. As such, a method that would systematically over- or under-estimate the size of a ventricle in the same order at both ED and ES would potentially have a low EF bias, a low mean average error and a high EF correlation, but at the same time a low Dice score and a large Hausdorff distance.

3) Classification performance: For the classification context, a prediction accuracy measure was provided. This accuracy was calculated for the whole examinations of the testing database, and also per disease. Confusion matrix was created in order to highlight the results.

D. MICCAI 2017 framework

The evaluation framework was launched during the "Automatic Cardiac Diagnosis Challenge (ACDC)" workshop held in conjunction with the 20th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), on September 10th, 2017 in Quebec City, Canada. After having publicly invited people to participate to this challenge, 106 accounts were created on the challenge website. Ten teams uploaded meaningful results within the allotted time for the segmentation contest, while 4 teams participated in the diagnosis contest.

IV. Evaluated architectures

In this section, we describe the different architectures involved in the segmentation contest as well as the methods proposed for the classification contest.

A. Architectures for cardiac multi-structure segmentation

A summary of the ten architectures involved in this study is provided in Table II. Nine methods implemented a deep convolutional architecture, most of which a U-Net like networks [18] analyzing the 3D data slice by slice. The only exception is the method by Tziritas and Grinias [49] which implemented a

[7]https://girder.readthedocs.io/en/latest/
[8]http://acdc.creatis.insa-lyon.fr/

TABLE II
OVERVIEW OF METHODS EVALUATED DURING THE ACDC CHALLENGE.

Reference *	Contest	Method	Remarks
Baumgartner *et al.* [43]	S	2D U-Net	Tested several architectures, the best one being a 2D U-Net with a cross-entropy loss
Isensee *et al.* [44]	S	2D+3D U-Net	Ensemble of 2D and 3D U-Net architectures with a Dice loss
Jang *et al.* [45]	S	2D M-Net	Use of a weighted cross-entropy loss function
Khened *et al.* [46]	S	Dense U-Net	2D U-Net with dense blocks and an inception first layer
Patravali *et al.* [47]	S	2D U-Net	Tested several architectures, the best one being a 2D U-Net with a Dice loss
Rohé *et al.* [48]	S	SVF-Net	Multi-atlas strategy where the registration module is realized using an encoder-decoder network
Tziritas and Grinias [49]	S	Levelset+MRF	Chan-Vese levelset followed by graph cut and a B-Spline fitting to smooth out results
Wolterink *et al.* [50]	S	Dilated CNN	Feed-forward CNN but with dilated convolution operations
Yang *et al.* [51]	S	3D U-Net	Use of 3D U-Net but with residual connections instead of the usual concatenation operator
Zotti *et al.* [52]	S	2D Grid-Net	Use of a Grid-Net architecture with an automatically-registered shape prior
Cetin *et al.* [53]	C	SVM	Use of physiological and radiomic (shape, intensity and texture) features
Isensee *et al.* [44]	C	RF	Extract a series of instant and dynamic features; use an ensemble of 50 multilayer perceptrons
Khened *et al.* [46]	C	RF	Extract 11 features from seg. results + patient height/weight; trained a 100-trees RF classifier
Wolterink *et al.* [50]	C	RF	Extract 14 features from seg. results + patient height/weight; trained a 1000-trees RF classifier

** S: Segmentation contest; C: Classification contest; SVM: Support Vector Machine; RF: Random Forest; MRF: Markov Random Field.*

Chan-Vese level-set method followed by a MRF graph cut segmentation method and spline fitting to smooth out the resulting boundaries.

Four papers re-used the U-Net architecture. Baumgartner *et al.* [43] tested the U-Net and the FCN architectures with various hyper parameters. They also tested the impact of using 2D and 3D convolution layers as well as a training Dice loss versus a cross-entropy loss. Their best architecture ended up being a U-Net with 2D convolution layers trained with a cross-entropy loss. Isensee *et al.* [44] implemented an ensemble of 2D and 3D U-Net architectures (with residual connections along the upsampling layers). Concerning the 3D network, due to large inter slice gap on the input images, pooling and upscaling operations are carried out only in the short axis plane. Moreover, due to memory requirements, the 3D network involves a smaller number of feature maps. Both networks were trained with a Dice loss. Similar to Baumgartner' study, Patravali *et al.* [47] tested a 2D and 3D U-Net trained with different Dice and cross entropy losses. From their experiments, the best performing architecture was a 2D U-Net with a Dice loss. Finally, Yang *et al.* [51] implemented a 3D U-Net but with residual connections instead of the usual concatenation operator. They also used pre-trained weights for the downsampling path using the C3D network known to work well on video classification tasks [54]. Their network was trained with a multi-class Dice loss.

Four papers used a modified version of the U-Net. Jang *et al.* [45] implemented a "M-Net" [55] architecture whose main difference with U-Net resides in the feature maps of the decoding layers which are concatenated with those of the previous layer. The corresponding network was trained with a weighted cross-entropy loss. Khened *et al.* [46] implemented a dense U-Net. Their method starts by finding the region of interest with a Fourier transform followed by a Canny edge detector on the first harmonic image and compute an approximate radius and center of the LV with a circular Hough transform on the edge map previously generated. They then use a U-Net with dense blocks instead of basic convolution block to make the system lighter. The first layer of this network also corresponds to an inception layer. The system was trained with a sum of Dice and cross-entropy losses. Rohé *et al.* [48] developed a multi-atlas algorithm that first registers a target image with all images in the training dataset. The registered label fields are then merged with a soft fusion method using pixel-wise confidence measures. The registration module implements an encoder-decoder network called SVF-Net [56]. Finally, Zotti *et al.* [52] implemented a "Grid Net" architecture which corresponds to a U-Net with convolutional layers along the skip connections. The architecture also registers a shape prior which is used as additional features map before performing the final decision. The model was trained with a four term loss function.

Wolterink *et al.* [50] is the only team that implemented a CNN without an encoder-decoder architecture. Instead, they used a sequence of convolutional layers with increasing levels of kernel dilation to ensure that sufficient image context was used for each pixel's label prediction. This CNN was fed simultaneously with spatially corresponding ED and ES 2D slices while the output of the network was split in two, one softmax for ED and one for ES.

B. Solutions for automatic cardiac diagnosis

Three participants of the segmentation challenge used their segmentation result to extract features for cardiac diagnosis. Isensee *et al.* [44] extracted a series of instants and dynamic features from the segmentation maps and used an ensemble of 50 multilayer perceptrons (MLP) and a random forest to perform classification. Khened *et al.* [46] used 11 features, 9 derived from their segmentation map in addition to the patient weight and height. From those features, they trained a 100-trees random forest classifier. Wolterink *et al.* [50] extracted 14 features (12 from the segmentation maps + patient weight and height) and used a five-class random forest classifier with 1,000 decision trees.

Cetin *et al.* [53] were the only one to involved a semi-automatic segmentation method to manually extract the contours of the cardiac structures. Based on those contours, they computed 567 features including physiological features

TABLE III
SEGMENTATION ACCURACY OF THE 10 EVALUATED METHODS ON THE TESTING DATASET. **RED** IS THE BEST METHOD, AND BLUE ARE THE METHODS WITHIN THE RANGE OF AGREEMENT (DICE INDEX OF 0.02 AND HAUSDORFF DISTANCE OF 2.26 MM FROM THE BEST).

Methods *	ED						ES					
	LV		*RV*		*Myo*		*LV*		*RV*		*Myo*	
	D	d_H	D	d_H	D	d_H	D	d_H	D	d_H	D	d_H
	val.	mm	val.	mm	val.	mm	val.	mm	val.	mm	val.	mm
Isensee *et al.* [44]	**0.968**	7.4	**0.946**	**10.1**	**0.902**	8.7	**0.931**	**6.9**	**0.899**	12.2	**0.919**	**8.7**
Baumgartner *et al.* [43]	0.963	**6.5**	0.932	12.7	0.892	**8.7**	0.911	9.2	0.883	14.7	0.901	10.6
Jang *et al.* [45]	0.959	7.7	0.929	12.9	0.875	9.9	0.921	7.1	0.885	**11.8**	0.895	8.9
Zotti *et al.* [52]	0.957	6.6	0.941	10.3	0.884	8.7	0.905	8.7	0.882	14.1	0.896	9.3
Khened *et al.* [46]	0.964	8.1	0.935	14.0	0.889	9.8	0.917	9.0	0.879	13.9	0.898	12.6
Wolterink *et al.* [50]	0.961	7.5	0.928	11.9	0.875	11.1	0.918	9.6	0.872	13.4	0.894	10.7
Jain *et al.* [47]	0.955	8.2	0.911	13.5	0.882	9.8	0.885	10.9	0.819	18.7	0.897	11.3
Rohé *et al.* [48]	0.957	7.5	0.916	14.1	0.867	11.5	0.900	10.8	0.845	15.9	0.869	13.0
Tziritas-Grinias [49]	0.948	8.9	0.863	21.0	0.794	12.6	0.865	11.6	0.743	25.7	0.801	14.8
Yang *et al.* [51]	0.864	47.9	0.789	30.3	N/A	N/A	0.775	53.1	0.770	31.1	N/A	N/A

* *ED:End diastole; ES: End systole; LV: Endocardial contour of the left ventricle; RV: Endocardial contour of the right ventricle; Myo: Epicardial contour of the left ventricle (myocardium); D: Dice Index;* d_H*: Hausdorff distance.*

TABLE IV
CLINICAL METRICS FOR THE 10 EVALUATED METHODS ON THE TESTING DATASET. **RED** IS THE BEST METHOD, AND BLUE ARE THE METHODS WITHIN A P-VALUE LARGER THAN 0.05 ACCORDING TO BIAS AND STD MEASUREMENTS.

Methods *	LV_{EDV}			LV_{EF}			RV_{EDV}			RV_{EF}			MY_{Mass}		
	corr	**bias**$\pm\sigma$	**mae**	**corr**	**bias**$\pm\sigma$	**mae**	**corr**	**bias**$\pm\sigma$	**mae**	**corr**	**bias**$\pm\sigma$	**mae**	**corr**	**bias**$\pm\sigma$	**mae**
	val.	ml.	ml.	val.	%	%	val.	ml.	ml.	val.	%	%	val.	g.	g.
Khened *et al.* [46]	0.997	0.6 ± 5.5	4.2	0.989	-0.5 ± 3.4	2.5	0.982	-2.9 ± 12.6	8.4	0.858	-2.2 ± 6.9	5.3	0.990	-2.9 ± 7.5	6.3
Isensee *et al.* [44]	0.997	2.7±5.7	5.1	0.991	0.2 ± 3.1	2.1	0.988	4.4±10.8	7.9	0.901	-2.7 ± 6.2	4.7	0.989	-4.8 ± 7.6	7.3
Zotti *et al.* [52]	0.997	9.6±6.4	10.3	0.987	-1.2 ± 3.6	2.7	0.991	-3.7 ± 9.2	7.4	0.872	-2.2 ± 6.8	5.4	0.984	-12.4±9.0	13.1
Jain *et al.* [47]	0.997	9.9±6.7	10.8	0.971	1.7±5.5	4.1	0.945	5.6±22.2	15.0	0.791	6.8±8.1	8.3	0.989	11.6±8.1	11.9
Wolterink *et al.* [50]	0.993	3.0±8.7	6.8	0.988	-0.5 ± 3.4	2.5	0.980	3.6 ± 15.2	10.9	0.852	-4.6±6.9	6.6	0.963	-1.0 ± 14.6	10.0
Jang *et al.* [45]	0.993	-0.4 ± 8.7	6.0	0.989	-0.3 ± 3.3	2.3	0.986	-10.8±11.6	12.1	0.793	-3.2 ± 8.3	6.3	0.968	11.5±12.9	14.1
Baumgartner *et al.* [43]	0.995	1.4 ± 7.6	6.1	0.988	0.6 ± 3.4	2.6	0.977	-2.3 ± 15.1	11.1	0.851	1.2 ± 7.3	5.7	0.982	-6.9±9.8	9.8
Rohé *et al.* [48]	0.993	4.2±8.6	7.5	0.989	-0.1 ± 3.2	2.6	0.983	7.3±13.4	11.7	0.781	-0.7 ± 9.9	7.8	0.967	-3.4 ± 13.3	10.3
Tziritas-Grinias [49]	0.992	2.0 ± 11.7	8.5	0.975	-1.6 ± 5.0	4.3	0.930	18.6±25.4	24.8	0.758	-0.5 ± 9.1	7.1	0.942	-28.9±28.0	30.3
Yang *et al.* [51]	0.894	12.2±32.0	27.5	0.926	1.5 ± 8.7	6.1	0.789	47.3±41.9	48.7	0.576	8.8±23.2	15.7	N/A	N/A	N/A

* LV_{EDV}*: End diastolic left ventricular volume;* LV_{EF}*: Left ventricular ejection fraction;* RV_{EDV}*: End diastolic right ventricular volume;* RV_{EF}*: Right ventricular ejection fraction;* MY_{Mass}*: Myocardial mass in diastole; mae: mean absolute error*

(*e.g.* height and weight) and radiomic features such as shape-based features, intensity statistics, and various texture features. To prevent their method from overfitting, they selected the most discriminative features and used SVM for classification.

V. RESULTS

A. Segmentation Challenge

For a detailed analysis of the results, a set of segmentation outputs are provided in the supplementary materials (available in the supplementary files /multimedia tab). This should help better assess the quality of the best approaches. Table III shows the segmentation testing accuracy (50 patients) for all 10 algorithms. The red values correspond to the best scores for each metric while the blue values correspond to the methods that are one pixel away from the top method. We use this color code to underline the closeness between the involved methods. This one-pixel criterion is a range of agreement of 2.3 mm for the Hausdorff distance (the maximum in-plane diagonal distance between two pixels: $\sqrt{(1.66^2) * 2}$) and 0.02 for the Dice metric (the average Dice score between the segmentation map of a method and the same segmentation map dilated or eroded by 1 pixel). This one pixel criterion comes from the fact that the two experts gave themselves a one pixel error margin such that two annotations were considered identical when their 2D Hausdorff distance was smaller or equal than one pixel.

From these results, one can see that the 2D-3D U-Net ensemble model proposed by Isensee *et al.* [44] is overall the top performing method (the corresponding code is publicly available through the following link[9]). This approach is closely followed by other methods which are less than one pixel away from it, especially for the LV and RV at ED. For instance, Baumgartner *et al.*, Jang *et al.*, Zotti *et al.*, and Khened *et al.* are within the range of agreement of the top performing method for 9 of the 12 metrics. As for the none deep-learning method by Tziritas and Grinias, it is relatively far away from the top, especially for the RV and the MYO.

Table IV contains the clinical metrics for all 10 methods. As for the segmentation part, red values correspond to the best

[9]https://github.com/MIC-DKFZ/ACDC2017

TABLE V
PERCENTAGE OF PATIENTS WITH AN EF ERROR LOWER THAN 5%.

Methods	*LV*	*RV*
Isensee *et al.* [44]	92 %	68%
Jang *et al.* [45]	88 %	60%
Rohe *et al.* [48]	88 %	34%
Zotti *et al.* [52]	84 %	60%
Khened *et al.* [46]	84 %	56%
Baumgartner *et al.* [43]	84 %	54%
Wolterink *et al.* [50]	80 %	38%
Jain *et al.* [47]	68 %	54%
Tziritas-Grinias [49]	66 %	38%
Yang *et al.* [51]	58 %	32%

TABLE VI
RESULTS ON THE CLASSIFICATION CHALLENGE.

Methods		*Accuracy*
Authors	Architectures	
Khened *et al.* [46]	Random Forest	**0.96**
Cetin *et al.* [53]	SVM	0.92
Isensee *et al.* [44]	Random Forest	0.92
Wolterink *et al.* [50]	Random Forest	0.86

scores for each metric. Blue values correspond to the methods with a p-value larger than 5% compared to the best method (we used an unequal variances two-sample t-test).

For the clinical indices, Khened *et al.* [46] globally outperforms the other approaches with 14 metrics out of 20 close to the top performing method (*i.e.* red and blue metrics). In terms of correlation metrics, most of the methods obtained highly accurate results with values above 0.96 for the volumes. Methods also get good LV_{EF} results with high correlation scores, a bias close to zero (0.8% on average), a small mean absolute error (3.2% on average) and small standard deviations (4.3%). The most difficult clinical metric to estimate is the EF of the RV with a correlation score of 0.9 for the best method.

A joint analysis of Table III and Table IV reveals that results on the myocardium (especially at ES) are those that vary the most. This may be partially explained by the fact that an accurate myocardium segmentation implies the precise delineation of two walls instead of one for the LV and RV. Methods also struggle with the RV. The RV often has the highest Hausdorff distances, the lowest Dice scores, the lowest correlation values, and the largest biases. To further underline this observation, we recorded in Table V the percentage of patients for which the predicted EF is less than 5% away from the ground-truth (5% is often considered as an acceptable error margin [57]). While the top six methods accurately predict the LV ejection fraction for $\approx 87\%$ of the patients, that number drastically goes down to $\approx 59\%$ for the RV.

B. Classification Challenge

Table VI presents an overview of the classification performance of the 4 evaluated methods. Due to the small number of samples (50 patients), the scores have to be considered with care since a miss-classification causes an accuracy drop of 2%. From this table, one can see that Khened *et al.* [46] obtained nearly perfect results with 48 patients correctly classified. The confusion matrix of this approach is shown in Fig. 1. Please note that the best approach is closely followed by the next two methods which obtained an accuracy of 92%.

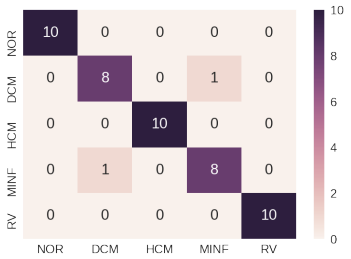

Fig. 1. Confusion matrix of the winner of the classification challenge [46].

Let us mention that although MINF and DCM are visually similar, MINF implies a local lack of myocardial contraction as opposed to DCM. Moreover, for DCM, the LV must exceeds 100 mL/m2. This is why machine learning methods have been able to successfully differentiate these pathologies.

C. Discussion

1) How far are we from solving the CMRI analysis problem? Automatic classification results (healthy subjects and patients with 4 different pathologies), showed that the best methods are very close to each other with an accuracy above 92%. Although these observations have to be validated on more patients, it appears from this study that well designed machine learning techniques can reach near perfect classification scores.

However, conclusions are not so straightforward for the segmentation task. While results obtained on the LV are competitive, it appears that the same level of accuracy is still difficult to obtain for the RV and the MYO. It is thus important to assess the performance of the top methods relatively to the experts variability. Unfortunately, the actual version of the ACDC dataset comes with one expert annotation per subject and does not provide any inter- or intra-observer error margin.

In order to evaluate the inter- and intra-observer variabilities, we asked the two experts O_1 and O_2 that jointly annotated the ACDC ground-truths to independently relabel the images of the 50 test subjects. O_1 annotated twice the images (we call those annotations O_{1a} and O_{1b}) one month apart while O_2 annotated the images once. The average geometric distance between O_{1a}, O_{1b} and O_2 are given in the first three lines of Table VII. As one can see, the Dice scores oscillates between 0.86 and 0.96 and the HD between $4\ mm$ and $14.1\ mm$. Without much surprise, the RV at ES is the most difficult region to annotate, even for experimented observers. It is also interesting to note that the Dice variations (especially for the inter-observer) are very close to that reported in a recent publication by Wenjia *et al.* [58]. As for the d_H values, the ones reported in Table VII are larger than those in Wenjia *et al.*'s paper due to the fact that our implementation of d_H accounts for the 3D structures of the heart. With an inter-slice thickness of $10\ mm$ (in average), any slight lateral shift between two annotations greatly increases the d_H score (please refer to the supplementary material available in the supplementary files /multimedia tab for more details).

Below the inter- and intra-observer results given in table VII, we provide *i)* the average geometrical metrics obtained

TABLE VII

DICE AND HAUSDORFF DISTANCES FOR *i)* INTER- AND INTRA- OBSERVERS *ii)* THE AVERAGE OF EVERY SUBMITTED DEEP LEARNING (DL) METHODS AND *iii)* THE WINNER OF THE SEGMENTATION CHALLENGE. **RED** CORRESPONDS TO RESULTS WITHIN OR ABOVE THE INTER-OBSERVER VARIATION. THE LAST 5 LINES CORRESPOND TO METRICS COMPUTED WITHOUT THE APICAL AND THE BASAL SLICES.

Methods *	**ED**						**ES**					
	LV		***RV***		***MYO***		***LV***		***RV***		***MYO***	
	D	d_H	***D***	d_H	***D***	d_H	***D***	d_H	***D***	d_H	***D***	d_H
	val.	mm	val.	mm	val.	mm	val.	mm	val.	mm	val.	mm
O_{1a} vs O_2 (inter-obs)	0.956	5.6	0.930	12.6	0.870	6.7	0.898	8.1	0.866	14.0	0.891	7.6
O_2 vs O_{1b} (inter-obs)	0.950	6.2	0.931	12.1	0.868	7.2	0.895	8.5	0.861	14.1	0.886	8.0
O_{1a} vs O_{1b} (intra-obs)	0.967	4.0	0.957	7.6	0.900	5.1	0.941	5.4	0.930	9.1	0.917	6.0
Average DL methods vs GT	**0.965**	7.6	**0.947**	13.2	**0.906**	10.1	**0.927**	9.2	**0.886**	15.2	**0.898**	10.9
Isensee *et al.* [44] vs GT	**0.968**	7.4	**0.946**	**10.1**	**0.902**	8.7	**0.931**	**6.9**	**0.906**	**12.1**	**0.919**	8.7
O_{1a} vs O_2 (inter-obs)	0.956	4.4	0.938	7.7	0.867	5.0	0.913	5.5	0.890	8.7	0.894	5.5
O_2 vs O_{1b} (inter-obs)	0.953	4.9	0.937	8.6	0.864	5.5	0.905	5.8	0.898	9.4	0.886	6.1
O_{1a} vs O_{1b} (intra-obs)	0.971	3.1	0.960	5.8	0.905	3.6	0.950	3.9	0.940	6.9	0.923	4.4
Average DL methods vs GT	**0.972**	**3.7**	**0.951**	**8.1**	**0.896**	**5.2**	**0.929**	**4.2**	**0.899**	9.9	**0.915**	**6.1**
Isensee *et al.* [44] vs GT	**0.972**	**3.7**	**0.969**	**6.4**	**0.910**	**4.6**	**0.945**	**4.2**	**0.912**	**8.6**	**0.930**	**5.1**

** ED: End diastole; ES: End systole; LV: Endocardial contour of the left ventricle; RV: Endocardial contour of*
** the right Ventricle; Myo: Myocardium contours; D: Dice Index; d_H: Hausdorff distance; GT: Ground-truth.*

TABLE VIII

INTER- AND INTRA-OBSERVER VARIATION OF THE MEAN ABSOLUTE ERROR OF THE LV_{EDV}, RV_{EDV} AND MY_{Mass}. BELOW, THE WINNER OF THE CHALLENGE AND THE AVERAGE DEEP LEARNING METHODS COMPARED WITH THE ACDC GROUND-TRUTH. **RED** ARE RESULT BETWEEN THE INTER- AND INTRA-OBSERVER VARIANCE.

	LV_{EDV}	RV_{EDV}	MY_{Mass}
	ml.	ml.	g.
O_{1a} vs O_2 (inter-obs)	10.4	9.2	12.6
O_2 vs O_{1b} (inter-obs)	10.8	9.5	11.5
O_{1a} vs O_{1b}(intra-obs)	4.6	5.7	6.2
Average methods vs GT	**7.1**	10.6	**10.4**
Isensee et al. vs GT	**5.1**	**7.9**	**7.3**

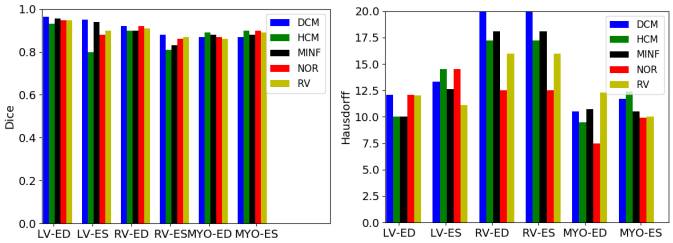

Fig. 2. Average Dice index and Hausdorff distances for every method reported in Table III broken down for every pathology.

by the deep learning methods involved in the challenge and *ii)* the scores obtained by Isensee *et al.*, the winner of the segmentation challenge. Interestingly, their Dice scores are all between the inter-observer and intra-observer scores. This suggests that state-of-the-art deep learning techniques have reached a plateau in the light of this metric. Although further investigations shall be made to validate this assertion (especially for images acquired from a set of more heterogeneous settings), the obtained results tend to show that, when properly trained, deep learning techniques are able to improve the Dice scores all the way to those of an expert. As for the d_H scores, methods are slightly above the inter-observer scores, but by only 2 to 3 mm.

In table VIII, we put the inter- and intra-observer mean absolute errors computed from the LV_{EDV}, RV_{EDV} and MY_{Mass} metrics. From the given numbers, one can see that the inter- and intra-observer scores are very close to that reported by Wenjia *et al.* [58]. Moreover, the results obtained by Isensee *et al.* and the average deep learning methods are between the inter- and intra-observer scores.

2) Where do methods fail?

In the light of the results reported so far, it appears that top deep learning segmentation methods are in the range of human expects according to the Dice scores and the clinical metrics but still 2 to 3 mm away from experts in regards of the 3D Hausdorff distance. One may thus wonder where do methods fail? One hypothesis can be that hearts suffering from a pathology may be more difficult to segment. To verify this assumption, we broke down in Fig. 2 the average Dice and Hausdorff metrics for each pathology obtained by the challengers on the test set (we remind that each pathological case corresponds the same amount of patients, both for the training and the testing phases). As one can see, there is no pathology for which methods systematically fail. For instance, while the HCM Dice score is somewhat low for the LV-ES (certainly due to the difficulty to see the cardiac cavity), it is larger than the other pathologies for MYO-ES and MYO-ED. Also, contrary to what one might think, images from healthy subjects (NOR) are not easier to segment than those from pathological cases as the scores relative to this group get the largest Hausdorff distances for the LV-ED and LV-ES.

Another hypothesis would be that 1.5T images are more difficult to segment than 3T CMR images due to an intrinsic lower SNR. However, after careful analysis of segmentation results, we found no particular differences between 1.5T and 3T results, as illustrated in table IX. One reason for this could be explained by the fact that both 1.5T and 3T images were included in the training set thus allowing neural networks to learn a representation specific to both magnetic fields. In

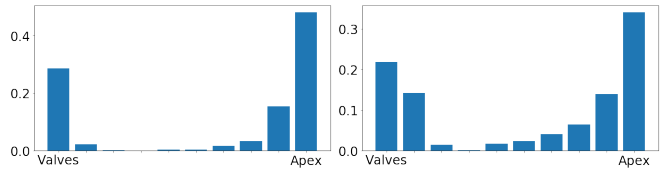

Fig. 3. Histogram of degenerated slices ED (left), and ES (right).

TABLE IX
DICE SCORES OF THE WINNER OF THE SEGMENTATION CHALLENGE [44] ON THE 1.5T AND 3T CMR IMAGES TAKEN FROM THE TESTSET.

	ED			ES		
	LV	**RV**	**MYO**	**LV**	**RV**	**MYO**
1.5T	0.97	0.95	0.90	0.93	0.90	0.92
3T	0.97	0.94	0.91	0.94	0.88	0.92

order to allow visual inspection of the difference between 1.5T and 3T CMR images, we putted in the supplementary materials (available in the supplementary files /multimedia tab) an example of such images as well as their corresponding MRI histograms.

Another hypothesis commonly accepted in the community is that slices next to the valves and/or the apex of the ventricle are more difficult to segment due to partial volume effect with surrounding structures. To investigate this assumption, we computed the total number of 2D segmentation results produced by each method for which the LV, MYO or RV had a Dice score below 0.70. The corresponding results are summarized through the histogram in Fig. 3, where the x-axis stands for the slice position (from the valves on the left to the apex on the right). Please note that since the number of slices varies from one patient to another, we stacked the 2D segmentation result of each method and made a 3D volume. Each volume was then resized to 10 slices with a nearest neighborhood interpolation method. From this figure, one can see that segmentation results obtained next to the valves and the apex are far more error prone. In particular, we notice almost 50% of results with very low Dice score at the apex (often because LV/MYO/RV are very small at that position). As for the base, we observe that methods often struggle to differentiate between the RV, the LV, the atria and the surrounding structures (c.f. Fig. 4). We also put in Table VII the Dice and Hausdorff metrics computed without the apical and basal slices. While the Dice scores are almost identical with and without the end slices, the Hausdorff distance decreases significantly, sometimes by a factor of two for the learning methods. Interestingly, the learning methods fall within the inter- and intra-observer variabilities (apart for Hausdorff metric for the RV at ES) which shows that segmenting apical and basal slices is far more difficult, even for experts.

Finally, it is worth pointing that the use of a larger database than the one involved in this project might help in resolving the listed remaining issues. For instance, the UK Biobank [19] may be a serious candidate for this purpose. We thus see the UK Biobank and our database as complementary with the strong potential to offer materials for upcoming research studies.

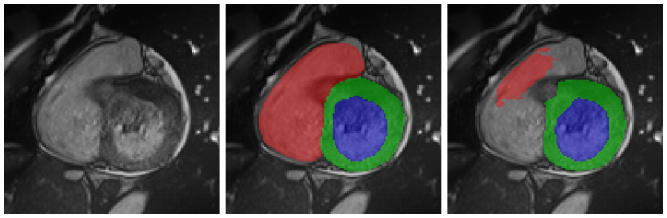

Fig. 4. Typical degenerated result at the base of the heart. [Left] input image; [Middle] ground truth; [Right] prediction.

3) For the need of a new metric

Results reported so far suggest that top deep learning methods are very close to the inter-observer variability. However, the visual inspection of their segmentation results reveal that unlike experts, deep learning methods sometimes generate anatomically impossible results as shown in Fig. 4. Interestingly, the metrics used to gauge performances seem resilient to such abnormalities. In order to measure the number of anatomically impossible results, one of our expert visually screened the test results by Isensee *et al.* [44]. This revealed that results for 41 patients out of 50 had at least one slice with an anatomically impossible segmentation such as the RV disconnected from the MYO or the LV cavity in contact with the background (several detailed examples are given in the supplementary materials available in the supplementary files /multimedia tab). Those 41 patients had problematic results for 1.6 slices on average, most of them located next to the valves or the apex. This clearly underlines the fact that clinical and geometrical metrics used to assess results have important limits and that methods within the inter-observer variability may still be error-prone. This suggests the need for new evaluation metrics before one may claim that methods have reached the accuracy of an expert.

VI. CLINICAL IMPLICATIONS

Results presented thus far suggest that we are at the eve of cracking the nut of fully automatic CMRI analysis. This would allow to reduce the time spent on analyzing raw data so conclusions of the examination could be provided to the patient before leaving the radiology department. In todays clinical practices, the latest systems provide pre-filled radiologic reports with an integrated automatic speech recognition technology so doctors can dictate the various physiological and technical parameters. An automatic CMRI analysis software could thus easily be integrated within this framework. That being said, further investigations are still required before such software gets approved by accreditation agencies (CE mark, FDA, ISO, etc.) and get integrated in MRI consoles. Also, although classification software get near-perfect results, the use of a ”diagnostic black box” could not be integrated as-is in a clinical practice. Along with the pathology prediction, a medical report must always contain the physiological reasons for which the patient was diagnosed in a certain way. This calls for cardiac parameters such as EF, volumes, and mass

estimated by a segmentation method which, in the context of deep learning approaches, may sometimes fail at the apex and the base and even produce anatomically impossible results. One shall also perform further analysis on images acquired by a wider variety of MRI scanners with different acquisition protocols to better assess the true generalization accuracy of machine learning algorithms.

Further research is also required on patient data suffering from other pathologies. Although we believe that some other pathologies such as inflammatory cardiomyopathy could be successfully diagnosed with the proposed machine learning methods, other (yet more complex) diseases such as congenital heart diseases or heart defect, would need dedicated studies.

VII. Conclusions

ECG-gated sequences such as Cine-MRI allow for accurate analysis of left and right ventricular functions. The delineation of ventricular endocardium and epicardium allows the calculation of different parameters, such as LV_{EF}, RV_{EF}, myocardial mass, myocardial thickness, tele-systolic and tele-diastolic ventricular volumes. These measurements are an integral part of the exam interpretation by the radiologist and are necessary for the diagnosis of many cardiomyopathies. In this paper, we have shown that state-of-the-art machine learning methods can successfully classify patient data and get highly accurate segmentation results. Results also reveal that the best convolutional neural networks get accurate correlation scores on clinical metrics and low bias and standard deviation on the LV_{EDV} and LV_{EF}, two of the most commonly-used physiological measures. However, methods are still failing at the base and the apex, especially when considering the Hausdorff distance.

References

[1] H. D. White, R. M. Norris, M. A. Brown, P. W. Brandt, R. M. Whitlock, and C. J. Wild, "Left ventricular end-systolic volume as the major determinant of survival after recovery from myocardial infarction," *Circulation.*, vol. 76, pp. 44–51, 1987.

[2] R. Norris, H. White, D. Cross, C. Wild, and R. Whitlock, "Prognosis after recovery from myocardial infarction: the relative importance of cardiac dilatation and coronary stenoses," *Eur. Heart J.*, vol. 13, pp. 1611–1618, 1992.

[3] P. M. Elliott, A. Anastasakis, M. A. Borger, M. Borggrefe, F. Cecchi, P. Charron, A. A. Hagege, A. Lafont, G. Limongelli *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.*, vol. 35, no. 39, pp. 2733–2779, 2014.

[4] C. A. Miller, P. Jordan, A. Borg, R. Argyle, D. Clark, K. Pearce, and M. Schmitt, "Quantification of left ventricular indices from SSFP cine imaging: Impact of real-world variability in analysis methodology and utility of geometric modeling," *J. Magn. Reson. Imaging*, vol. 37, no. 5, pp. 1213–1222, 2013.

[5] S. Queirós, D. Barbosa, B. Heyde, P. Morais, J. L. Vilaça, D. Friboulet, O. Bernard, and J. Dhooge, "Fast automatic myocardial segmentation in 4D cine CMR datasets," *Med. Image Anal.*, vol. 18, no. 7, pp. 1115 – 1131, 2014.

[6] P. Radau, Y. Lu, K. Connelly, G. Paul, A. Dick, and G. Wright, "Evaluation framework for algorithms segmenting short axis cardiac MRI," in *The MIDAS Journal - Cardiac MR Left Ventricle Segmentation Challenge*, 2009.

[7] A. Suinesiaputra, B. R. Cowan, J. P. Finn, C. G. Fonseca, A. H. Kadish, D. C. Lee, P. Medrano-Gracia, S. K. Warfield, W. Tao, and A. A. Young, "Left ventricular segmentation challenge from cardiac MRI: A collation study," in *Proc. STACOM*, 2011, pp. 88–97.

[8] C. Petitjean, M. A. Zuluaga, W. Bai, J.-N. Dacher, D. Grosgeorge, J. Caudron, S. Ruan, I. B. Ayed, M. J. Cardoso *et al.*, "Right ventricle segmentation from cardiac MRI: A collation study," *Med. Image Anal.*, vol. 19, no. 1, pp. 187–202, 2015.

[9] "The 2015 kaggle second annual data science bowl," www.kaggle.com/c/second-annual-data-science-bowl.

[10] M. Avendi, A. Kheradvar, and H. Jafarkhani, "A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI," *Med. Image Anal.*, vol. 30, pp. 108–119, 2016.

[11] L. K. Tan, Y. M. Liew, E. Lim, and R. A. McLaughlin, "Cardiac left ventricle segmentation using convolutional neural network regression," in *Proc. IECBES*, 2016, pp. 490–93.

[12] A. Suinesiaputra, B. R. Cowan, A. O. Al-Agamy, M. A. E. e, N. Ayache, A. S. Fahmy, A. M. Khalifa, P. Medrano-Gracia, M.-P. Jolly, A. H. Kadish, D. C. Lee, J. Margeta, S. K. Warfield, and A. A. Young, "A collaborative resource to build consensus for automated left ventricular segmentation of cardiac mr images," *Med. Image Anal.*, vol. 18, no. 1, pp. 50–62, 2014.

[13] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation," *IEEE Trans. Med. Imaging*, vol. 23, no. 7, pp. 903–921, July 2004.

[14] B. Li, Y. Liu, C. J. Occleshaw, B. R. Cowan, and A. A. Young, "In-line automated tracking for ventricular function with magnetic resonance imaging." *JACC. Cardiovasc. Imaging*, vol. 3 8, pp. 860–6, 2010.

[15] D. Grosgeorge, C. Petitjean, J.-N. Dacher, and S. Ruan, "Graph cut segmentation with a statistical shape model in cardiac MRI," *Comput. Vis. Image Underst.*, vol. 117, no. 9, pp. 1027 – 1035, 2013.

[16] P. V. Tran, "A fully convolutional neural network for cardiac segmentation in short-axis MRI," *arXiv:1604.00494*, 2017.

[17] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, 2014, pp. 3431–3440.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2016, pp. 234–241.

[19] S. Petersen and P. M. Matthews and J. M. Francis and M. D. Robson and F. Zemrak and R. Boubertakh and A. A. Young and others, "UK Biobank's cardiovascular magnetic resonance protocol," *Journal of Cardiovascular Magnetic Resonance*, vol. 18, no. 8, 2016.

[20] C. Petitjean and J.-N. Dacher, "A review of segmentation methods in short axis cardiac MR images," *Med. Image Anal.*, vol. 15, no. 2, pp. 169–184, 2011.

[21] H. Liu, H. Hu, X. Xu, and E. Song, "Automatic left ventricle segmentation in cardiac MRI using topological stable-state thresholding and region restricted dynamic programming," *Acad. Radiol.*, vol. 19, no. 6, pp. 723–731, 2012.

[22] J. Ulen, P. Strandmark, and F. Kahl, "An efficient optimization framework for multi-region segmentation based on lagrangian duality," *IEEE Trans. Med. Imaging*, vol. 32, no. 2, pp. 178–188, 2013.

[23] T. Chen, J. Babb, P. Kellman, L. Axel, and D. Kim, "Semiautomated segmentation of myocardial contours for fast strain analysis in cine displacement-encoded MRI," *IEEE Trans. Med. Imaging*, vol. 27, no. 8, pp. 1084–1094, Aug 2008.

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

[25] S. C. Mitchell, J. G. Bosch, B. P. F. Lelieveldt, R. J. van der Geest, J. H. C. Reiber, and M. Sonka, "3-d active appearance models: segmentation of cardiac mr and ultrasound images," *IEEE Trans. Med. Imaging*, vol. 21, no. 9, pp. 1167–1178, 2002.

[26] W. Bai, W. Shi, C. Ledig, and D. Rueckert, "Multi-atlas segmentation with augmented features for cardiac MR images," *Med. Image Anal.*, vol. 19, no. 1, pp. 98–109, 2015.

[27] O. Emad, I. A. Yassine, and A. S. Fahmy, "Automatic localization of the left ventricle in cardiac mri images using deep learning," in *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug 2015, pp. 683–686.

[28] B. Kong, Y. Zhan, M. Shin, T. Denny, and S. Zhang, "Recognizing end-diastole and end-systole frames via deep temporal regression network," in *Proc. MICCAI*, 2016, pp. 264–272.

[29] L. Zhang, A. Gooya, B. Dong, R. Hua, S. E. Petersen, P. Medrano-Gracia, and A. F. Frangi, "Automated quality assessment of cardiac mr images using convolutional neural networks," in *Proc. SASHIMI-MICCAI*, 2016, pp. 138–145.

[30] C. Rupprecht, E. Huaroc, M. Baust, and N. Navab, "Deep active contours," *arXiv:1607.05074*, 2016.

[31] T. A. Ngo, Z. Lu, and G. Carneiro, "Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance," *Med. Image Anal.*, vol. 35, pp. 159–171, 2017.

[32] H. Yang, J. Sun, H. Li, L. Wang, and Z. Xu, "Deep fusion net for multi-atlas segmentation: Application to cardiac MR images," in *Proc. MICCAI*, 2016, pp. 521–528.

[33] R. P. Poudel, P. Lamata, and G. Montana, "Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation," *arXiv:1608.03974*, 2016.

[34] O. Oktay, W. Bai, M. Lee, R. Guerrero, K. Kamnitsas, J. Caballero, A. de Marvao, S. Cook, D. O'Regan, and D. Rueckert, "Multi-input cardiac image super-resolution using convolutional neural networks," in *Proc. MICCAI*, 2016, pp. 246–254.

[35] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Snchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, no. Supplement C, pp. 60 – 88, 2017.

[36] M. Havaei, N. Guizard, H. Larochelle, and P.-M. Jodoin, "Deep learning trends for focal brain pathology segmentation in mri," in *book Machine Learning for Health Informatics, State-of-the-art and future challenges*, 2015, pp. 125–148, LNAI 9605, Springer.

[37] C. H. Lorenz, E. S. Walker, V. L. Morgan, S. S. Klein, and T. P. Graham, "Normal human right and left ventricular mass, systolic function, and gender differences by cine magnetic resonance imaging," *J Cardiovasc Magn Reson*, vol. 1, no. 1, pp. 1097–6647, 1999.

[38] F. I. Marcus, W. J. McKenna, D. Sherrill, C. Basso, B. Bauce, D. A. Bluemke, H. Calkins, D. Corrado, M. G. Cox *et al.*, "Diagnosis of arrhythmogenic right ventricular cardiomyopathy / dysplasia," *Circulation*, vol. 121, no. 13, pp. 1533–1541, 2010.

[39] K. Scheffler and S. Lehnhardt, "Principles and applications of balanced ssfp techniques," *European radiology*, vol. 13, no. 11, pp. 2409–2418, 2003.

[40] D. Dubois and E. F. Dubois, "A formula to estimate the approximate surface area if height and weight be known. archives of internal medicine,," *Archives of Internal Medicine*, vol. 17, pp. 863–871, 1916.

[41] A. Lalande, M. Garreau, and F. Frouin, "Evaluation of cardiac structure segmentation in cine magnetic resonance imaging," in *Multi-modality Cardiac Imaging: Processing and Analysis*. Iste, 2015, pp. 171–215.

[42] S. Gottschalk, M. C. Lin, and D. Manocha, "Obbtree: A hierarchical structure for rapid interference detection," in *Proc. SIGGRRAPH*, 1996, pp. 171–180.

[43] C. Baumgartner, L. M. Koch, M. Pollefeys, and E. Konukoglu, "An exploration of 2d and 3d deep learning techniques for cardiac mr image segmentation," in *Proc. STACOM-MICCAI, LNCS, volume 10663*, 2017, pp. 111–119.

[44] F. Isensee, P. Jaeger, P. Full, I. Wolf, S. Engelhardt, and K. H. Maier-Hein, "Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features," in *Proc. STACOM-MICCAI, LNCS, volume 10663*, 2017, pp. 120–129.

[45] Y. Jang, S. Ha, S. Kim, Y. Hong, and H.-J. Chang, "Automatic segmentation of lv and rv in cardiac mri," in *Proc. STACOM-MICCAI, LNCS, volume 10663*, 2017, pp. 161–169.

[46] M. Khened, V. Alex, and G. Krishnamurthi, "Densely connected fully convolutional network for short-axis cardiac cine mr image segmentation and heart diagnosis using random forest," in *Proc. STACOM-MICCAI, LNCS, volume 10663*, 2017, pp. 140–151.

[47] J. Patravali, S. Jain, and S. Chilamkurthy, "2d-3d fully convolutional neural networks for cardiac mr segmentation," in *Proc. STACOM-MICCAI, LNCS, volume 10663*, 2017, pp. 130–139.

[48] M.-M. Rohe, M. Sermesant, and X. Pennec, "Automatic multi-atlas segmentation of myocardium with svf-net," in *Proc. STACOM-MICCAI, LNCS, volume 10663*, 2017, pp. 170–177.

[49] G. Tziritas and E. Grinias, "Fast fully-automatic localization of left ventricle and myocardium in mri using mrf model optimization, substructures tracking and b-spline smoothing," in *Proc. STACOM-MICCAI, LNCS, volume 10663*, 2017, pp. 91–100.

[50] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Isgum, "Automatic segmentation and disease classification using cardiac cine mr images," in *Proc. STACOM-MICCAI, LNCS, volume 10663*, 2017, pp. 101–110.

[51] X. Yang, C. Bian, L. Yu, D. Ni, and P.-A. Heng, "Class-balanced deep neural network for automatic ventricular structure segmentation," in *Proc. STACOM-MICCAI, LNCS, volume 10663*, 2017, pp. 152–160.

[52] C. Zotti, Z. Luo, O. Humbert, A. Lalande, and P.-M. Jodoin, "Gridnet with automatic shape prior registration for automatic mri cardiac segmentation," in *Proc. STACOM-MICCAI, LNCS, volume 10663*, 2017, pp. 73–81.

[53] I. Cetin, G. Sanroma, S. E. Petersen, S. Napel, O. Camara, M. ngel Gonzlez Ballester, and K. Lekadir, "A radiomics approach to computer-aided diagnosis in cardiac cine-mri," in *Proc. STACOM-MICCAI, LNCS, volume 10663*, 2017, pp. 82–90.

[54] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. ICCV*, 2015.

[55] R. Mehta and J. Sivaswamy, "M-net: A convolutional neural network for deep brain structure segmentation," in *Proc. ISBI*, 2017, pp. 437–440.

[56] M. Rohe, M. Sermesant, and X. Pennec, "Svf-net: Learning deformable image registration using shape matching," in *Proc. MICCAI*, 2017.

[57] J. Bogaert, S. Dymarkowski, A. Taylor, and V. Muthurangu, "Cardiac Function," in *Clinical Cardiac MRI*. Springer, 2012, pp. 109–168.

[58] B. W. et al., "Human-level cmr image analysis with deep fully convolutional networks," *arXiv preprint:1710.09289*, 2017.