Computer Methods and Programs in Biomedicine

Automated Left and Right Ventricular Chamber Segmentation in Cardiac Magnetic Resonance Images Using Dense Fully Convolutional Neural Network

--Manuscript Draft--

Manuscript Number:	CMPB-D-20-00702R2
Article Type:	Full Length Article
Section/Category:	Biomedical signal and image processing methods
Keywords:	convolutional neural networks; cardiac segmentation; Cine Cardiac Magnetic Resonance; Dense skip connections
Corresponding Author:	Enrico Gianluca Caiani Politecnico di Milano Milan, MIlan ITALY
First Author:	Marco Penso
Order of Authors:	Marco Penso
	Sara Moccia
	Stefano Scafuri
	Giuseppe Muscogiuri
	Gianluca Pontone
	Mauro Pepi
	Enrico Caiani
Manuscript Region of Origin:	Europe
Abstract:	Background and Objective: Segmentation of the left ventricular (LV) myocardium (Myo) and RV endocardium on cine cardiac magnetic resonance (CMR) images represents an essential step for cardiac-function evaluation and diagnosis. In order to have a common reference for comparing segmentation algorithms, several CMR image datasets were made available, but in general they do not include the most apical and basal slices, and/or gold standard tracing is limited to only one of the two ventricles, thus not fully corresponding to real clinical practice. Our aim was to develop a deep learning (DL) approach for automated segmentation of both RV and LV chambers from short-axis (SAX) CMR images, reporting separately the performance for basal slices, together with the applied criterion of choice. Method: A retrospectively selected database (DB1) of 210 cine sequences (3 pathology groups) was considered: images (GE, 1.5 T) were acquired at Centro Cardiologico Monzino (Milan, Italy), and end-diastolic (ED) and end-systolic frames (ES) were manually segmented (gold standard, GS). Automatic ED and ES RV and LV segmentation were performed with a U-Net inspired architecture, where skip connections were redesigned introducing dense blocks to alleviate the semantic gap between the U-Net encoder and decoder. The proposed architecture was trained including: A) the basal slices where the Myo surrounded the LV for at least the 50% and all the other slice; B) all the slices where the Myo completely surrounded the LV. To evaluate the clinical relevance of the proposed architecture in a practical use case scenario, a graphical user interface was developed to allow clinicians to revise, and correct when needed, the automatic segmentation. Additionally, to assess generalizability, analysis of CMR images obtained in 12 healthy volunteers (DB2) with different equipment (Siemens, 3T) and settings was performed. Results: The proposed architecture outperformed the original U-Net. Comparing the performance on DB1 between the two criteria, no significant differences were measured when considering all slices together, but were present when only basal slices were examined. Automatic and manually-adjusted segmentation performed similarly compared to the GS (bias±95%LoA): LVEDV -1±12 ml, LVESV -1±14 ml, RVEDV 6±12 ml, RVESV 6±14 ml, ED LV mass 6±26 g) , ES LV mass 5±26 g. Also, generalizability showed very similar performance, with Dice scores of 0.944 (LV), 0.908 (RV) and 0.852 (Myo) on DB1, and 0.940 (LV), 0.880 (RV), and 0.856 (Myo) on DB2.

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

	Conclusions: Our results support the potential of DL methods for accurate LV and RV contours segmentation and the potential of dense skip connections to alleviate the semantic gap generated when high level features are concatenated with lower level feature. The evaluation on our dataset, considering separately the performance on basal and apical slices, reveals the potential of DL approaches for fast, accurate and reliable automated cardiac segmentation in a real clinical setting.
Suggested Reviewers:	
Opposed Reviewers:	
Response to Reviewers:	See attached file!

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

Dipartimento di Elettronica, Informazione e Bioingegneria

Politecnico di Milano

20133 Milano (Italia)
Piazza Leonardo da Vinci, 32

Prof. Filippo Molinari, PhD.
Editor *Computer Methods and Programs in Biomedicine*

Department of Electronics and Telecommunications, Politecnico di Torino,
Corso Duca degli Abruzzi 24, 10129 Torino, Italy.
E-mail: filippo.molinari@polito.it

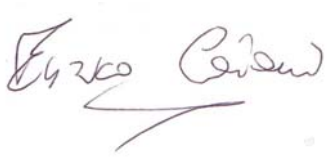
Milan, 28th December 2020

Dear Prof. Molinari,

We would like to thank you and the Reviewers for the additional helpful suggestions that we believed have contributed to improve the quality of our manuscript "*Automated Left and Right Ventricular Chamber Segmentation in Cardiac Magnetic Resonance Images Using Dense Fully Convolutional Neural Network*", of which now we are submitting the new revised version.

In our reply to the Reviewers' comments you can find our point-by-point responses, where both in the reply and in the revised manuscript, the text that was added was highlighted.
We hope that this revised version will be considered acceptable, and that our responses have adequately addressed the additional comments.

Thank you for your consideration.
Sincerely,



Prof. Enrico G Caiani
on behalf of all co-authors.

We would like to thank the Reviewers for their additional helpful suggestions. In the following document you can find our point-by-point responses to the general and specific comments

Reviewer #1: The authors' motivation for this work is to address the potential inaccuracy of mislabelling the basal slices on cardiac MRI. They use CMRs with full coverage of the ventricles and single-observer ground truth labels. They use a conventional network architecture with modified skip connections. They found no benefit in terms of volumetric analysis but did identify a difference in basal slice accuracy.
Adding external validation has strengthened the study - but the performance should be quoted in the abstract in preference to internal validation.
Major concerns:
(1). My main concern is that they are trying to address a problem which no longer truly exists in a clinical setting. Modern proprietary software (using later versions of CMR42 for instance) use ML for segmentation and automated valve tracking to address the basal slice issue. Any adjustments are also simple to perform. It is incorrect to say this process is "tedious, time-consuming, and error prone" in current practice.
*We agree with the Reviewer that in the field of commercial software for ventricular segmentation over the last few years there have been significant improvements, and the CMR42 software (as others) is able to handle the automated segmentation process quite well. Although the time and effort required to the clinician to complete the analysis of the acquired images have been reduced, we have observed in our reference center (Centro Cardiologico Monzino IRCCS, Milan, Italy) that ventricular segmentation still requires about 20% of the total time. In fact, manual correction, not only at the basal or apical level, but also in the medial section, is still required (e.g., for the adjustment of myocardial mass). Valve tracking (on the long-axis view) certainly improves but does not solve the basal slice issue, and this is evident from the high inter-observer variability that still exists [*Bernard *et al*. *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514-2525. (2018); Bai W et al, *J Cardiovasc Magn Reson* **20,** 65 (2018)*]. Taking into account the Reviewer comment, the relevant sentence has been modified as follows:*
"Although this operation can be performed using software for automated segmentation, visual inspection and manual correction are still required, in particular for the right ventricular endocardium and for the LV basal and apical slices [3].

(2). Expert-level performance using related FCN architectures have been previously reported (Bai et al). The authors refer to discordant identification of the basal slices between observers in that study - but the current paper uses only a single expert for annotations which is far from a "gold standard" (https://doi.org/10.1186/s12968-017-0327-9 - used 8 observers with strict QC procedures).
Following a previous observation in the first review, we were focusing on this aspect of the study of Bai et al to highlight that, when compared with expert annotations, segmentations of FCN architectures work well but with differences in performance due to the different slice level, and in particular with the worst performance for basal slices. In that study, as reported in Table 4 of Bai et al, two experienced image analysts (with over ten and four years experiences in cardiovascular image analysis, respectively) visually assessed the correspondence between the manual and automated contours in a subsample of 250 subjects. Specifically, as we reported in our manuscript, the correspondence was around 85%and 92%% in the mid-ventricular slices, while decreased down to

40% and 33% for basal slices, and down to 44% and 81% for apical slices, respectively for the first and second observer. These results, in our opinion, underline the still existing need to further investigate FCN performance for basal (and apical) slices, and to report FCN performance separately for different levels of short-axis slices.

Conversely, the work of Petersen et al 2017 now referred by the Reviewer, is focusing on the assessment of normality ranges by manual tracing of LV and RV contours, evaluating inter-observer variability among several annotators in different centers, with no purpose of validating automated detection methods.

As regards the Reviewer's observation that we used a single expert only, we would like to remark that the traced contours were part of the original analyses performed for the clinical evaluation of the patient, and not the result of a dedicated analysis aiming at multiple manual tracings comparisons, that was beyond the scope of our study. As a result, in case of doubts in the interpretation and tracings, in the clinical routine the expert cardiologist could count on the second opinion of additional experts in the group. We believe this way constitutes the realistic gold standard in clinical practice, as no one provides independent multiple tracings for the same patient, unless for a specific research study. However, in relation to the Reviewer's comment, we further specified this point in the Experimental protocol as follows:

"In case of doubts in the interpretation and tracings, the expert cardiologist could count on the second opinion of additional experts in the group"

In addition, we added into the Limitation section the fact that our gold standard was not based on consensus after multiple experts independent annotation:

"One possible limitation of our study is that the gold standard was not based on independent multiple tracings but on a single expert annotation, and only in cases with uncertainty in the image interpretation, the contours were reviewed through group consensus."

(3). So my fundamental concern is that the authors are not making a fair comparison between their network and current clinical practice. It would be reasonable to compare CMR42 v5.12 (or equivalent) with a prototype algorithm trained on expert consensus ground truth labels for instance. But I don't feel the current study is justified in saying it "represents a concrete step toward a more realistic clinical application of DL algorithm."

As explained as a reply to comment 2), we think our comparison between the proposed network and the ground truth utilized, based on current clinical practice and not on a specific research protocol, was fair.

Also, we would like to stress that our aim was not to validate the performance of our proposed FCN with that of commercial automated contour detection software, as this was used only to obtain the manual tracings.

As regards the last sentence in the abstract where the sentence "represents a concrete step toward a more realistic clinical application of DL algorithm." recalled by the Reviewer was present, according to the Reviewer's comment this was modified as follows:

"The evaluation on our dataset, considering separately the performance on basal and apical slices, reveals the potential of DL approaches for fast, accurate and reliable automated cardiac segmentation in a real clinical setting."

Minor concerns:

(1). Abstract - Reporting the correlation between two methods for analysing the same data isn't useful. Please report the Bland-Altman mean difference and limits of agreement. Use LVEDV etc.

We thank the Reviewer for this comment and, as suggested, we modified the abstract as follows, also inserting some information about the generalizability of our results:

Automatic and manually-adjusted segmentation performed similarly compared to the GS (bias±95%LoA): LVEDV -1±12 ml, LVESV -1±14 ml, RVEDV 6±12 ml, RVESV 6±14 ml, ED LV mass 6±26 g) , ES LV mass 5±26 g. Also, generalizability showed very similar performance, with Dice scores of 0.944 (LV), 0.908 (RV) and 0.852 (Myo) on DB1, and 0.940 (LV), 0.880 (RV), and 0.856 (Myo) on DB2.

(2). A number of items have been missed from the checklist including the population demographics (please provide as a table).

Thanks for this observation. We updated the checklist table. Some items were omitted as our manuscript does not generally apply to them.

As suggested, we added demographics and clinical characteristics of the two dataset as Table 1 in the section 2.1 Patient Population:

The population demographic, together with the GS measurements for DB1 (considering both all 210 patients altogether and separately in the three HCM, VA and DCM groups) and DB2 are presented in Table 1.

Characteristics	DB 1				DB 2
	All (n = 210)	VA (n = 70)	DCM (n = 70)	HCM (n = 70)	All (n = 12)
Age, years	56 (42-65)	43 (31-59)	60 (50-69)	57 (44-64)	35 (28-43)
Female, n (%)	77 (37%)	33 (47%)	25 (36%)	19 (27%)	6 (50%)
Body surface area, m^2	1.9 (1.7-2.0)	1.8 (1.7-2.0)	1.9 (1.7-2.0)	1.9 (1.8-2.0)	1.9 (1.9-2.0)
LVEDV index (mL/m^2)	89 (77-113)	90 (74-100)	108 (83-133)	85 (76-92)	92 (81-101)
LVESV index (mL/m^2)	42 (33-56)	40 (32-49)	58 (42-89)	36 (29-41)	32 (28-35)
LVSV (mL/m^2)	47 (40-52)	48 (44-55)	45 (37-52)	48 (43-52)	60(53-67)
LVEF (%)	53.0 (44.0-60.0)	55.0 (50.5-60.0)	41.0 (32.8-53.0)	58.0 (52.0-63.0)	65.1 (64.7-66.0)
RVEDV index (mL/m^2)	81 (69-94)	86 (73-92)	80 (67-101)	76 (69-91)	104 (88-120)
RVESV index (mL/m^2)	36 (28-45)	36 (30-42)	39 (28-51)	32 (27-40)	46 (33-53)
RVSV (mL/m^2)	46 (40-51)	48 (41-53)	44 (36-50)	46 (41-50)	59 (53-62)
RVEF (%)	56.0 (51.0-61.0)	56.0 (51.5-62.0)	54.0 (49.8-59.0)	58.0 (53.3-62.0)	58.1 (52.4-61.5)

Table 1. Values are reported as median (25th-75th percentile), or n (%). LV: left ventricular; RV: right ventricular; EDV: end diastolic volume; ESV: end systolic volume; SV: stroke volume; EF: ejection fraction; HCM: hypertrophic cardiomyopathy; DCM: dilated cardiomyopathy; VA: ventricular arrhythmias.

(3). A link to the code has been provided but there is no documentation. Please provide a basic README showing how to use the code. Please also create a DOI.

As requested, we inserted a README, as well as the LICENSE and a DOI.

(https://github.com/Monzino/Ventricle_segmentation)

Reviewer #2: This revision has better readability than the original version. Some of my questions were answered carefully. The following are some of my minor comments:

(1). In the Abstract, "12 healthy volunteers" is mentioned in assessing generalizability, why not include some patients?

We thank the Reviewer for this comment. As previously explained, we tried to get access to patients' data from shared databases providing both MRi images (with basal slices) and ground truth, in order to provide external testing for our method, but without success. As a consequence, our choice was to test our method on a population of 12 normal volunteers for which we had both the ownership of the original MRI data (acquired with a 3T system at DLR, Koln, Germany) and the availability of the traced contours by an expert observer. This is why the inclusion of patients acquired externally was not possible. In addition, as the FCN was trained on a patient population, assessing its good performance on normal subjects seemed a good strategy to further assess external validity.

(2). In the introduction, it is suggested to describe some drawbacks of the existing segmentation methods.

We appreciate your constructive comment and, as suggested, we included in the introduction some drawbacks of the existing deep-learning segmentation methods, as follows:

"similarly, Savioli et al. [16] adopted a Gated Recurrent Unit to exploit information of consecutive CMR cine frames, that might affect the backpropagation through time, limiting the number of consecutive frames that can be analyzed. Differently, Zhen et al. [17] developed a multi-scale CNN followed by random forests for a direct estimation of cardiac biventricular volumes, while Yang et al. [18] implemented a 3D U-Net with residual connections for ventricular volumetric segmentation in cine images. However, using a 3D network, the amount of training images is drastically reduced and a consistent downsampling of the data is required due to memory constraints, leading to a potential loss of information [19].

(3). At page 18, in a pre-processing step, how to crop a 2D region of interest (ROI) centered with respect to the heart? Manually or automatically?

Thank you for posing this question. Image cropping was fully automatic, and was performed using the circular Hough Transform. To clarify better this part of our pre-processing step, we modified the paragraph in Sec. 2.2 as follows:

"Image cropping to obtain a squared 2D ROI was fully automatic, and was performed on a mid-ventricular slice for each patient, using the circular Hough Transform for the identification of the LV cavity centroid, and then transposed to all other slices by centering the crop area with the detected LV center.."

(4). At page 19, in the sentence of "After each pooling, the number of feature maps is doubled", can the pooling double the number of feature maps?

The question gives us the opportunity to clarify this point better. The number of feature maps is doubled because there is a convolutional layer, with doubled number of filters, right after the max pooling.

To better specify this point, the sentence has been modified as follows:

"At each downsampling step, the number of feature maps is doubled, and this process is repeated four times."

HIGHLIGHTS

- **Modified U-Net with dense skip connections improves left and right ventricular segmentation in cine MRI.**
- **Basal slices, according to the guidelines defined by the Society for Cardiovascular Magnetic Resonance, and different pathological conditions were included in the analysis.**
- **Criteria of definition of basal slice, as well as specific algorithm performance, are reported: this choice should be adopted to allow better comparison of segmentation algorithms performance in clinical scenarios.**
- **Comparison with gold standard, and with manual correction of the automatically obtained contours, was performed to understand the impact on accuracy of computed volumetric clinical values.**

CONFLICT OF INTEREST

There are no conflicts of interest that could have inappropriately influenced this research work.

Automated Left and Right Ventricular Chamber Segmentation in Cardiac Magnetic Resonance Images Using Dense Fully Convolutional Neural Network

Marco Penso[a], Sara Moccia[b,c], Stefano Scafuri[a], Giuseppe Muscogiuri[a], Gianluca Pontone[a], Mauro Pepi[a], Enrico Gianluca Caiani[d,e]

[a] Department of Cardiovascular Imaging, Centro Cardiologico Monzino IRCCS, Milan, Italy.

(Marco.Penso@cardiologicomonzino.it, scafuristefano90@gmail.com, Giuseppe.Muscogiuri@cardiologicomonzino.it, gianluca.pontone@cardiologicomonzino.it, Mauro.Pepi@cardiologicomonzino.it)

[b] Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy.

(s.moccia@staff.univpm.it)

[c] Department of Advanced Robotics, Istituto Italiano di Tecnologia, Genoa, Italy.

[d] Department of Electronics, Information and Biomedical engineering, Politecnico di Milano, Milan, Italy.

(enrico.caiani@polimi.it)

[e] Consiglio Nazionale delle Ricerche, Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni, Milan, Italy.

Corresponding author: Prof. Enrico Gianluca Caiani

Politecnico di Milano

Department of Electronics, Information and Biomedical engineering

P.zza L. da Vinci 32

20133 Milan, Italy

Office: +39.02.2399.3390

Abstract

Background and Objective: Segmentation of the left ventricular (LV) myocardium (Myo) and RV endocardium on cine cardiac magnetic resonance (CMR) images represents an essential step for cardiac-function evaluation and diagnosis. In order to have a common reference for comparing segmentation algorithms, several CMR image datasets were made available, but in general they do not include the most apical and basal slices, and/or gold standard tracing is limited to only one of the two ventricles, thus not fully corresponding to real clinical practice. Our aim was to develop a deep learning (DL) approach for automated segmentation of both RV and LV chambers from short-axis (SAX) CMR images, reporting separately the performance for basal slices, together with the applied criterion of choice.

Method: A retrospectively selected database (DB1) of 210 cine sequences (3 pathology groups) was considered: images (GE, 1.5 T) were acquired at Centro Cardiologico Monzino (Milan, Italy), and end-diastolic (ED) and end-systolic frames (ES) were manually segmented (gold standard, GS). Automatic ED and ES RV and LV segmentation were performed with a U-Net inspired architecture, where skip connections were redesigned introducing dense blocks to alleviate the semantic gap between the U-Net encoder and decoder. The proposed architecture was trained including: A) the basal slices where the Myo surrounded the LV for at least the 50% and all the other slice; B) all the slices where the Myo completely surrounded the LV. To evaluate the clinical relevance of the proposed architecture in a practical use case scenario, a graphical user interface was developed to allow clinicians to revise, and correct when needed, the automatic segmentation. Additionally, to assess generalizability, analysis of CMR images obtained in 12 healthy volunteers (DB2) with different equipment (Siemens, 3T) and settings was performed.

Results: The proposed architecture outperformed the original U-Net. Comparing the performance on DB1 between the two criteria, no significant differences were measured when considering all slices together, but were present when only basal slices were examined. Automatic and manually-adjusted segmentation performed similarly compared to the GS (bias±95%LoA): LVEDV -1±12 ml, LVESV -1±14 ml, RVEDV 6±12 ml, RVESV 6±14 ml, ED LV mass 6±26 g) , ES LV mass 5±26 g. Also, generalizability showed very similar

performance, with Dice scores of 0.944 (LV), 0.908 (RV) and 0.852 (Myo) on DB1, and 0.940 (LV), 0.880 (RV), and 0.856 (Myo) on DB2.

Conclusions: Our results support the potential of DL methods for accurate LV and RV contours segmentation and the potential of dense skip connections to alleviate the semantic gap generated when high level features are concatenated with lower level feature. The evaluation on our dataset, considering separately the performance on basal and apical slices, reveals the potential of DL approaches for fast, accurate and reliable automated cardiac segmentation in a real clinical setting.

Keywords: Convolutional Neural Networks, Cardiac Segmentation, Cine Cardiac Magnetic Resonance, Dense skip connections

1. Introduction:

Cardiovascular disease represents one of the main causes of death in Europe [1]. Cardiac imaging is used to obtain structural and functional cardiac parameters for both diagnosis and disease evaluation [2]. In particular, Cardiac Magnetic Resonance (CMR) is the reference technique for estimating parameters such as left (LV) and right (RV) ventricular end-diastolic (EDV) and end-systolic (ESV) volumes, stroke volume (SV), ejection fraction (EF), and myocardial mass, as a result of an endocardial and epicardial contour segmentation. Although this operation can be performed using software for automated segmentation, visual inspection and manual correction are still requested, in particular for the right ventricular endocardium and for the LV basal and apical slices [3]. In the last decade, several model-based approaches to automatize the segmentation process were developed and tested, such as deformable models, cardiac atlas-based registration, level sets, statistical shape and active appearance models [4-7]. More recently, machine learning (ML) algorithms have been introduced to deal with the huge variability of clinical data. Specifically, Convolutional Neural Networks (CNNs) are emerging as the new state of the art for medical-image segmentation [8-12]. One of the advantages of the CNNs is the ability to learn more robust features through a hierarchical structure composed of learnable filters, without relying on hand-crafted features.

CNNs in different configurations have been applied to CMR images for segmenting both myocardial fibrosis and cardiac chambers. For example, Moccia et al. [13] obtained promising performance in LV scar tissue segmentation on LGE images using an encoder-decoder fully convolutional network (FCN). On cine images, Isensee et al. [14] implemented an ensemble of 2D and 3D U-Net architectures for LV and RV segmentation, while Poudel et al. [15] proposed a recurrent neural network for LV segmentation in multi-slice CMR cine images to capture global anatomical structure of the heart; similarly, Savioli et al. [16] adopted a Gated Recurrent Unit to exploit information of consecutive CMR cine frames, that might affect the backpropagation through time, limiting the number of consecutive frames that can be analyzed. Differently, Zhen et al. [17] developed a multi-scale CNN followed by random forests for a direct estimation of cardiac biventricular volumes, while Yang et al. [18] implemented a 3D U-Net with residual connections for ventricular volumetric segmentation in cine images. However, using a 3D network, the amount of

training images is drastically reduced and a consistent downsampling of the data is required due to memory constraints, leading to a potential loss of information [19]. The majority of this approaches are typically based on U-Net [20], an architecture made of a down-sampling (encoder) and up-sampling (decoder) path, and its variations. The success of U-Net may be attributed to the introduction of the skip connection for combining coarse-grained with fine-grained feature maps, recovering lost details of the target structures to be segmented [21]. A possible limitation of the original skip connections presented in [20] can be seen in the direct fusion of feature maps that are semantically different. Currently, this feature fusion is not supported by a solid theory which guarantees it to be the best solution [22]. Hence, feature maps of the up-sampling path, which are computed from the deepest layers of the network, are supposed to encode high-level information, while the maps of the down-sampling path refer to more general, low-level features. This generates a semantic gap, which could affect the segmentation accuracy [23].

CNNs achieved remarkable success as described in the literature, but an accurate segmentation of the myocardium in cine CMR images remains a challenging task in the present clinical routine. This is due to the presence of papillary muscles and trabeculations, and to the shape variability of ventricles across slices, cardiac phases, and pathologies [24]. This is particularly true at the ventricular basal and apical level: the former, with the presence of different anatomical structures such as the aortic and tricuspid valves and atrial chambers, and the latter due to the reduced size of the ventricles. Even if commercial software for automated ventricular boundary detection and computation of clinical parameters are available [25], the accuracy of the detected contours still needs to be improved to reach that of manual segmentation in all slices, and in particular for the basal and apical ones, usually more prone to discrepancies compared to manual tracings. As an example, in [26] the correspondence between manual and automated contours evaluated by two different observers in a subsample of 250 patients resulted as 85% and 92% in the mid-ventricular slices, while decreased down to 40% and 33% for the basal slices, and down to 44% and 81% for apical slices, respectively for the first and second observer.

To possibly develop and compare the performance of deep learning (DL) algorithms, over the past years different CMR image datasets were made available through several challenges [27]. However, they

generally did not include the most apical and basal slices, reference contours were limited to only one of the two ventricles, or the patient population was relevant to a specific clinical conditions, thus resulting in a non-optimal reference for comparison when referred to real clinical practice [15,28-31]. At the same time, the variability encoded in such datasets is limited when compared to the real clinical scenario, posing issues to the development of DL algorithms.

In this work, we hypothesized that CNNs could achieve segmentation accuracy comparable to that of expert physicians in the tracing of LV and RV endocardium and epicardium even when including the more basal and apical slices. Accordingly, our contributions can be summarized as follows:

1) Developing an FCN architecture based on U-Net [20], trained end-to-end for semantic segmentation of LV, RV and LV myocardium (Myo) on short-axis (SAX) cine CMR images. To attenuate the semantic gap that occurs in the original implementation of skip connections and improve segmentation accuracy, a novel skip connections with dense block is proposed.

2) Performing an experimental analysis on a large clinical dataset including both LV and RV chambers of patients with different pathologies and evaluating how the performance of the proposed FCN changes according to the definition criterion of the basal slices (i.e., following current recommendations of the Society for Cardiovascular Magnetic Resonance (SCMR) [32], or adopting a simplification criteria as in publicly-available datasets).

3) Evaluating the extent of manual correction needed after automated segmentation for a real use case scenario, and the relevant clinical impact on the obtained measurements of LV and RV volumes and LV mass.

4) Assessing the generalizability of the proposed approach by applying it to a different CMR image database obtained using different settings and equipment.

2. Materials and Methods:

2.1 Patient population

A retrospectively selected database (DB1) consisting of SAX cine (steady-state free-precession) CMR images (size 512x512 pixels, in-plane pixel resolution from 1.2 to 2.0 mm, acquired (GE, 1.5 T) in 210 patients referred to CMR evaluation at IRCCS Centro Cardiologico Monzino (Milan, Italy) along a 6 months period (September 2019-February 2020) was analyzed. The dataset included patients with the following diagnoses: 70 patients with hypertrophic cardiomyopathy (HCM), 70 with dilated cardiomyopathy (DCM) and 70 with ventricular arrhythmias (VA). Each stack of SAX images was acquired during breath hold, including both the LV and RV from base to apex, with a slice thickness of 8 mm, no inter-slice gap, and a field of view ranging from 300 to 360 mm^2. For each patient, images were acquired with 30 phases/cardiac cycle. The manual tracings of the LV endocardium and epicardium and RV endocardium (cvi software, ver. 5.1.1, Circle Cardiovascular Imaging Inc., Calgary, Canada), performed as part of the clinical evaluation at the time of the exam by one of the experienced cardiologists (10 years of experience) in the team at IRCCS Centro Cardiologico Monzino (Milan, Italy), were used as gold standard (GS). To determine where to trace the contours, the expert cardiologist usually examines all the acquired datasets, including other views (apical 2-, 3-, 4-chambers), to appreciate the dynamic motion and contiguity in space of structures visible in adjacent slices. In case of doubts in the interpretation and tracings, the expert cardiologist could count on the second opinion of additional experts in the group. For each patient, the manual tracing was provided only at the end-diastolic (ED) and at the end-systolic (ES) frames, which were chosen as the image with the largest and smallest LV blood volume, respectively, by examining the full image stack. The LV papillary muscles and trabeculae were included as part of LV and RV blood cavity, in agreement with what indicated in the guidelines of the SCMR [32].

Additionally, to investigate the generalization power of the proposed network, a second database (DB2) in our availability, acquired with different settings and equipment, was used as additional testing platform. DB2 consisted of SAX cine CMR images (size from 208x162 to 208x210 pixels, in-plane pixel resolution 1.63 mm, slice thickness 7 mm, 25 frames/cardiac cycle) acquired during breath hold, including both the LV and

RV from base to apex, using a Biograph mMR 3-Tesla scanner (Siemens, Erlangen, Germany) at the :envihab facility of the Institute of Aerospace Medicine at the German Aerospace Center (DLR, Koln, Germany) in 12 healthy volunteers in the context of a head-down bed rest study organized by the European Space Agency [33]. Also for this database, GS manual contours of the LV endocardium and epicardium and RV endocardium traced by an experienced cardiologist for the ED and ES phases were available.

The population demographic, together with the GS measurements for DB1 (considering both all 210 patients altogether and separately in the three HCM, VA and DCM groups) and DB2 are presented in Table 1.

Characteristics	DB 1				DB 2
	All (n = 210)	VA (n = 70)	DCM (n = 70)	HCM (n = 70)	All (n = 12)
Age, years	56 (42-65)	43 (31-59)	60 (50-69)	57 (44-64)	35 (28-43)
Female, n (%)	77 (37%)	33 (47%)	25 (36%)	19 (27%)	6 (50%)
Body surface area, m^2	1.9 (1.7-2.0)	1.8 (1.7-2.0)	1.9 (1.7-2.0)	1.9 (1.8-2.0)	1.9 (1.9-2.0)
LVEDV index (mL/m^2)	89 (77-113)	90 (74-100)	108 (83-133)	85 (76-92)	92 (81-101)
LVESV index (mL/m^2)	42 (33-56)	40 (32-49)	58 (42-89)	36 (29-41)	32 (28-35)
LVSV (mL/m^2)	47 (40-52)	48 (44-55)	45 (37-52)	48 (43-52)	60(53-67)
LVEF (%)	53.0 (44.0-60.0)	55.0 (50.5-60.0)	41.0 (32.8-53.0)	58.0 (52.0-63.0)	65.1 (64.7-66.0)
RVEDV index (mL/m^2)	81 (69-94)	86 (73-92)	80 (67-101)	76 (69-91)	104 (88-120)
RVESV index (mL/m^2)	36 (28-45)	36 (30-42)	39 (28-51)	32 (27-40)	46 (33-53)
RVSV (mL/m^2)	46 (40-51)	48 (41-53)	44 (36-50)	46 (41-50)	59 (53-62)
RVEF (%)	56.0 (51.0-61.0)	56.0 (51.5-62.0)	54.0 (49.8-59.0)	58.0 (53.3-62.0)	58.1 (52.4-61.5)

Table 1. Values are reported as median (25th-75th percentile), or n (%). LV: left ventricular; RV: right ventricular; EDV: end diastolic volume; ESV: end systolic volume; SV: stroke volume; EF: ejection fraction; HCM: hypertrophic cardiomyopathy; DCM: dilated cardiomyopathy; VA: ventricular arrhythmias.

2.2 Pre-processing

Data imported as DICOM images underwent a pre-processing step, consisting in cropping a 2D region of interest (ROI) centered with respect to the heart, normalizing voxel intensities to [0, 1] and a down-sampling operation for DT1 or an up-sampling operation for DT2 to obtain a ROI of 176x176 pixels size, to comply with memory constraints.

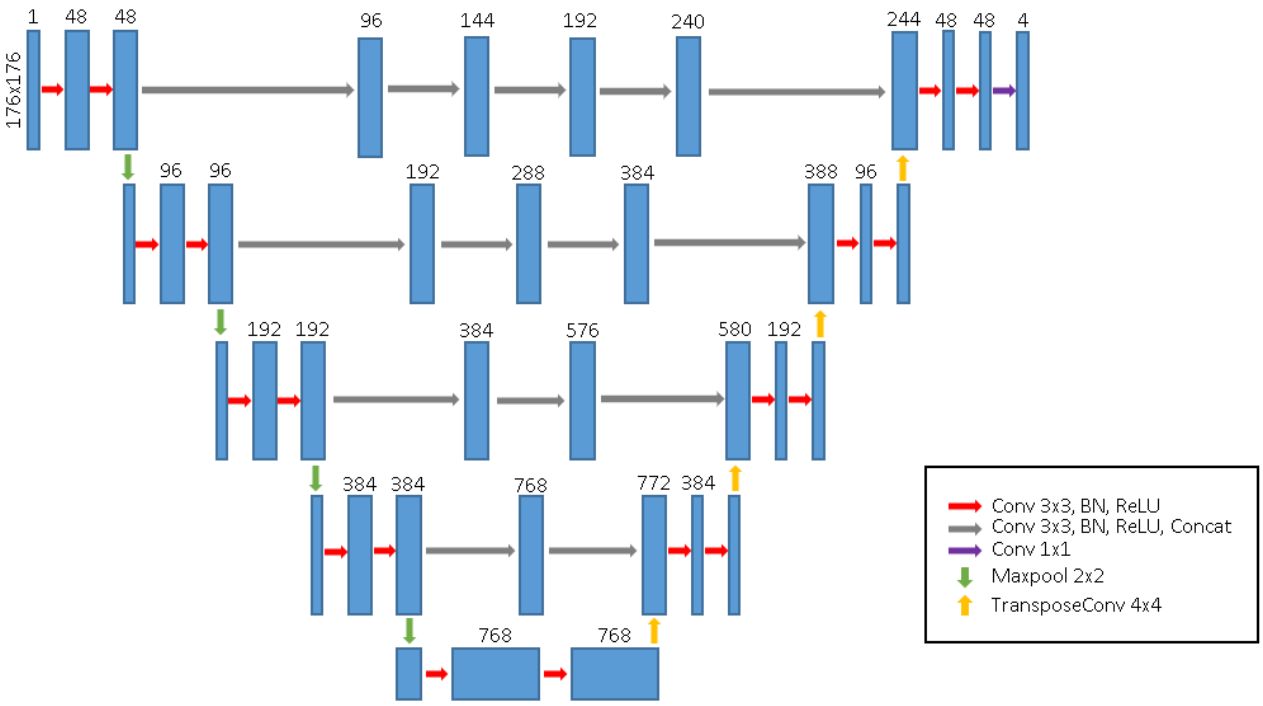


Fig. 1. Architecture of the proposed segmentation network for a 176x176 pixels input image. Each blue box corresponds to a multi-channel feature map. The number of channels is reported on top of each box.

Image cropping to obtain a squared 2D ROI was fully automatic, and was performed on a mid-ventricular slice for each patient, using the circular Hough Transform for the identification of the LV cavity centroid, and then transposed to all other slices by centering the crop area with the detected LV center. Selecting a ROI was useful in alleviating the class-imbalance problem between the analyzed structures (i.e. LV, RV and Myo) and background pixels [34]. The cine CMR slices were rotated and converted to the Right-Anterior-Head reference frame.

2.3 Network Architecture

The proposed FCN architecture takes inspiration from U-Net [20], which has symmetric encoder (down-sampling) and (up-sampling) decoder paths, with skip connections between them. With the down-sampling path, U-Net encodes context information that are transferred with skip connections to the up-sampling path, which is used to restore feature-map resolution and to produce the segmentation maps. Each step of the down-sampling path consists of two 3x3 convolutions with batch normalization [35] and activations with the rectified linear unit (ReLU) [36]. The sequence of convolution, batch normalization and ReLU is

called convolutional block (conv block), where each block is followed by a max-pooling operation with size 2x2 and stride of 2. At each downsampling step, the number of feature maps is doubled, and this process is repeated four times.

The bottom layer consists of a conv block linking the down-sampling to the up-sampling path. In this work, transposed convolution was implemented in the up-sampling path instead of up-convolution (used in the original U-Net implementation), as suggested in the literature [37]. A series of 4x4 transposed convolutions with stride 2 and conv block are repeated four times, similar to the down-sampling path. A final 1x1 convolution layer with softmax activation function, that maps the features maps to 4 classes, is used to generate the final segmentation maps (one for each target structure). The resulting network has a different number of filters in each conv block, from 48 (first layer) to 768 (bottom layer).

In this work, the original skip connections of U-Net were modified incorporating dense blocks [38] to reduce the semantic gap between the up- and down-sampling path [22,23]. Specifically, each layer of the dense block consisted of 3x3 convolutions with ReLU and batch normalization. Each layer concatenated its own feature maps with the feature maps of all the preceding layers, and transferred its own feature maps to all the following ones, thus adding and preserving information. The number of filters in each convolutional layer in a dense block was the same as in the corresponding convolutional layer in the down-sampling path. The number of layers used in each dense block were 4, 3, 2, and 1, respectively, along the four skip connections.

A further modification compared to the original U-Net was the reduction of the number of filters in the transposed convolutional layer, with a number of feature maps equal to the number of classes. As reported in [19], the full complexity of the up-sampling path is not necessary, with benefits in terms of reduction of learnable parameters. A schematic illustration of the proposed architecture is given in Figure 1.

To evaluate the accuracy and computations trade-off of the proposed FCN, an ablation study with two experiments was carried out: in the former, the deeper layer of the network was removed, while in the

latter the number of filters in each conv block was changed, from 32 in the first layer to 512 in the bottom layer.

2.4 Parameter setting and data augmentation

The weights of each convolution were initialized following the standard *He normal* initialization [39]. To deal with the class imbalance between the background and the foreground pixels, weighted cross-entropy (WCE) was adopted as loss function [20]: WCE measures the pixel-wise error probability between the predicted output and target class and accumulates summing the errors across all the pixels. The WCE is defined as:

$$WCE = -\sum_{x} w_{t(x)} \log(p(x, t(x))) \quad (1)$$

where w is the weight map to balance the class frequencies, and $p(x, t(x))$ is the probability prediction for each class at each pixel x after the softmax function. The targets class label at location x is denoted by $t(x)$. Class weights were set to be inversely proportional to their frequencies in the training data and normalized, so that for the background class a weight of 0.076 was used, whereas for all the classes of the foreground a weight of 0.308 was applied.

The FCN was trained for 1000 epochs with Adaptive Momentum Estimation (ADAM) optimizer [40], using an initial learning rate of 0.01 with a decay of 2% after every five epochs. The hyperparameters β1 and β2, used to control the exponential decay rates of the gradient, were set to 0.9 and 0.999, respectively. Online data augmentation [41] was implemented using gamma correction, blurring, and random rotation in [-15°, 15°] range. Data augmentation was used to allow the network to better generalize and prevent overfitting. After each epoch, the network performance was evaluated on the validation set. The best model among epochs was selected as the one with the highest Sørensen-Dice coefficient ($Dice$) score on the validation set. The $Dice$ calculates the overlap between foreground regions, and is defined as:

$$Dice = \frac{2 \sum p_{h,w} t_{h,w}}{\sum p_{h,w} + \sum t_{h,w}} \quad (2)$$

where h is the number of classes of the foreground, w is the number of pixels, t is the ground truth label per location and p is prediction of the sigmoid function.

A postprocessing step was also applied to all the trained networks, to obtain more accurate segmentations: in those cases, in which the FCN segmentation resulted in different regions with the same label, only the region with the largest area was preserved, while the remaining regions were classified as background.

All the analyses were performed using Tensorflow on a NVIDIA® GeForce® GTX 1080 (8 GB GDDR5 dedicated) on an Intel® Core™ i7-8700 CPU (3.20 GHz) computer with 16 GB RAM.

Codes are made publicly available at https://github.com/Monzino/Ventricle_segmentation.

2.5 Experimental protocol and statistical analysis

The datasets in DB1 were randomly split at patient level into training (70%), validation (15%) and testing sets (15%), preserving the proportion of patients for each pathology. The performance of the RV endocardial and LV endo- and epicardial segmentation by FCN was measured using sensitivity (*Se*), *Dice* and the Hausdorff distance (*HD*) [27] compared to the ground-truth mask. Results obtained using the standard and modified skip connections were compared (paired t-test for normally distributed variables, and Wilcoxon signed rank test for non-normally distributed variables).

To verify the influence of the basal slice definition criteria on the FCN performance, the analysis was performed training the proposed U-Net using: A) the basal slice as defined in agreement with [32], thus also considering those slices where the myocardium surrounded the LV for at least 50%; B) only the slices where the myocardium completely surrounded the LV, thus excluding some challenging slices.

To consider a practical use case scenario in which the cardiologist uses the proposed segmentation algorithm, the prediction on the test set was reviewed by an independent expert physician, by showing the resulting LV and RV contours overlapped to the original image, and allowing manual correction when needed, using a custom made software implemented in Matlab. The impact of these corrections was computed by comparing the corrected segmentations to the original automated results in respect to the ground-truth, both in term of *Se, Dice* and *HD*. In addition, also a comparison in terms of actual LV and RV

volumes (in ml) and LV mass (in g), compared to the gold standard values by correlation and Bland-Altman analyses, was performed.

To measure the inter-observer variability between the corrected segmentations and the ground truth, for each parameter a coefficient of variation between these two measurements was computed as CV(%)= 100*SD/MEAN and compared with the same result obtained by the FCN contours.

To evaluate the possible dependence of segmentation performance from the morphological characteristics associated to the different pathologies, results in terms of *Dice* and *HD* were grouped according to HCM, DCM and VA , and one-way ANOVA or Kruskal-Wallis test with Bonferroni correction was applied.

To assess the generalization power of the proposed approach on different datasets, the FCN trained on DB1 with criterion A was applied to the datasets in DB2, and compared with the performance obtained on the testing set of DB1 by Mann–Whitney U test.

Normally distributed data were reported as mean ± SD, otherwise as median (25^{th}-75^{th} percentile). Statistical analyses were conducted with IBM SPSS 25 (SPSS Inc, Chicago, IL), with p values < 0.05 to be considered as significant.

3. Results

From the ablation analysis, the proposed architecture consistently outperformed a shallower network.

Specifically, for LV, RV and Myo, respectively, the shallower network achieved a mean Dice of 0.789, 0.687 and 0.462 versus 0.944, 0.908 and 0.851 of the proposed FCN. Even compared to the network with a reduced number of filters, the proposed FCN reached higher performance for both LV (0.944 vs 0.905), RV (0.908 vs 0.858) and Myo (0.851 vs 0.818).

	Left Ventricle (ED)			Left Ventricle (ES)		
	Se	Dice	HD [mm]	Se	Dice	HD [mm]
U-Net	0.967 (0.020)	0.961 (0.009)	6.4 (4.4-9.1)	0.929 (0.056)	0.927 (0.034)	8.8 (4.4-11.9)
U-Net MOD	0.966 (0.021)	0.961 (0.010)	6.0 (4.0-8.3)	0.931 (0.052)	0.927 (0.036)	8.4 (4.8-12.0)
	Right Ventricle (ED)			**Right Ventricle (ES)**		
	Se	Dice	HD [mm]	Se	Dice	HD [mm]
U-Net	0.910 (0.039)	0.932 (0.022)	8.5 (4.9-10.0)	0.836 (0.076)	0.871 (0.054)	8.9 (5.8-12.0)
U-Net MOD	0.920 (0.033)*	0.938 (0.020)*	6.6 (5.0-9.3)	0.842 (0.071)*	0.878 (0.048)*	8.1 (6.0-12.9)
	Myocardium (ED)			**Myocardium (ES)**		
	Se	Dice	HD [mm]	Se	Dice	HD [mm]
U-Net	0.801 (0.082)	0.832 (0.038)	5.9 (4.2-7.2)	0.840 (0.083)	0.867 (0.043)	6.1 (4.1-8.8)
U-Net MOD	0.808 (0.078)*	0.833 (0.035)	5.9 (4.0-7.2)	0.849 (0.080)*	0.870 (0.040)	5.9 (4.4-8.5)

Table 2. Segmentation performance of the proposed U-Net (U-net MOD) compared to the U-Net with standard skip connections (U-Net), at end-diastolic (ED) and end-systolic (ES) phases in DB1, measured as Sensitivity (Se), Dice coefficient, and Hausdorff distance (HD). Se and Dice are reported as mean (SD), while HD is reported as median (25^{th}-75^{th} percentile). *: $p < 0.05$, U-Net vs U-Net MOD (paired t-test or Wilcoxon rank sum).

Table 2 compares the performances obtained with the proposed U-Net versus the same architecture without filters along the skip connections, including in the analysis slices in which the criterion A (i.e., >=50% of myocardium present around the LV) was verified: the implemented skip connections outperformed the standard U-net in particular for the RV, where significant differences in *Se* and *Dice*, in both ED and ES, were highlighted. Also, for the Myo, the redesigned connections performed better, with a significant improvement in *Se* for both ED and ES, while no differences were observed for the LV. Based on these results, and on the improved performance obtained using the redesigned skip connections, further analyses were conducted using this implementation only.

Table 3 reports the performance, in terms of *Se*, *Dice* and *HD*, of the FCN segmentation for LV and RV endocardium and LV myocardium in the ED and ES frames for DB1, both when criteria A or B were applied to define the most basal slice. Comparing the performance between these two criteria (i.e. A and B), results considering cumulatively all the slices showed no significant differences. As regards the performance in different cardiac phases, the model achieved higher Dice and Se in ED for both LV and RV, while for the Myo a significantly higher Dice and Se were obtained in ES.

A. Basal slices with >=50% myocardium around the LV					
Left Ventricle (ED)			Left Ventricle (ES)		
Se	Dice	HD [mm]	Se	Dice	HD [mm]
0.966 (0.021)¥	0.961 (0.010)¥	6.0 (4.0-8.3)	0.931 (0.052)	0.927 (0.036)	8.4 (4.8-12.0)
Right Ventricle (ED)			Right Ventricle (ES)		
Se	Dice	HD [mm]	Se	Dice	HD [mm]
0.920 (0.033)¥	0.938 (0.020)¥	6.6 (5.0-9.3)	0.842 (0.071)	0.878 (0.048)	8.1 (6.0-12.9)
Myocardium (ED)			Myocardium (ES)		
Se	Dice	HD [mm]	Se	Dice	HD [mm]
0.808 (0.078)¥	0.833 (0.035)¥	5.9 (4.0-7.2)	0.849 (0.080)	0.870 (0.040)	5.9 (4.4-8.5)
B. Only basal slices with 100% myocardium around the LV					
Left Ventricle (ED)			Left Ventricle (ES)		
Se	Dice	HD [mm]	Se	Dice	HD [mm]
0.967 (0.020)¥	0.962 (0.010)¥	5.9 (3.9-8.3)	0.933 (0.050)	0.930 (0.035)	6.5 (3.6-11.9)
Right Ventricle (ED)			Right Ventricle (ES)		
Se	Dice	HD [mm]	Se	Dice	HD [mm]
0.920 (0.033)¥	0.939 (0.020)¥	6.6 (5.0-9.3)	0.844 (0.069)	0.879 (0.046)	7.7 (6.0-12.9)
Myocardium (ED)			Myocardium (ES)		
Se	Dice	HD [mm]	Se	Dice	HD[mm]
0.809 (0.079)¥	0.834 (0.036)¥	5.8 (4.0-7.0)	0.853 (0.078)	0.873 (0.038)	5.0 (4.0-6.7)

Table 3. Segmentation performance of the proposed FCN computed cumulatively considering all slices at end-diastolic (ED) and end-systolic (ES) phases in DB1, measured as Sensitivity (Se), Dice coefficient, and Hausdorff distance (HD). Criterion A refers to the FCN trained and tested considering also the slices at LV basal level in which the myocardium was present for at least 50% around the LV cavity, while criterion B refers to the FCN trained and tested considering only the slices at LV basal level in which the myocardium was visible all around the cavity. Se and Dice are reported as mean (SD), while HD is reported as median (25^{th}-75^{th} percentile). *: $p<0.05$, A vs B (unpaired t-test or Mann–Whitney U test); ¥: $p<0.05$, ED vs ES (unpaired t-test or Mann–Whitney U test)

In Figure 2, some examples of the segmented contours at the most basal level, defined by criteria A and B previously described, are shown, paired with the corresponding gold standard manual tracings. In left columns, when the LV myocardium did not completely surround the LV, due to the presence of the aortic root, the automated segmentation sometimes was failing by including it in the LV. For the RV, when a portion of the right atrium was present, sometimes it was included in the RV. In the right columns, where these more problematic slices were excluded from both training and testing, as the morphology of the LV become more similar to the other slices and the RV was more visible and recognizable, these differences with the GS were still present but reduced. Despite these potential errors, cumulative performance computed considering all slices resulted not affected by applying criteria A or B.

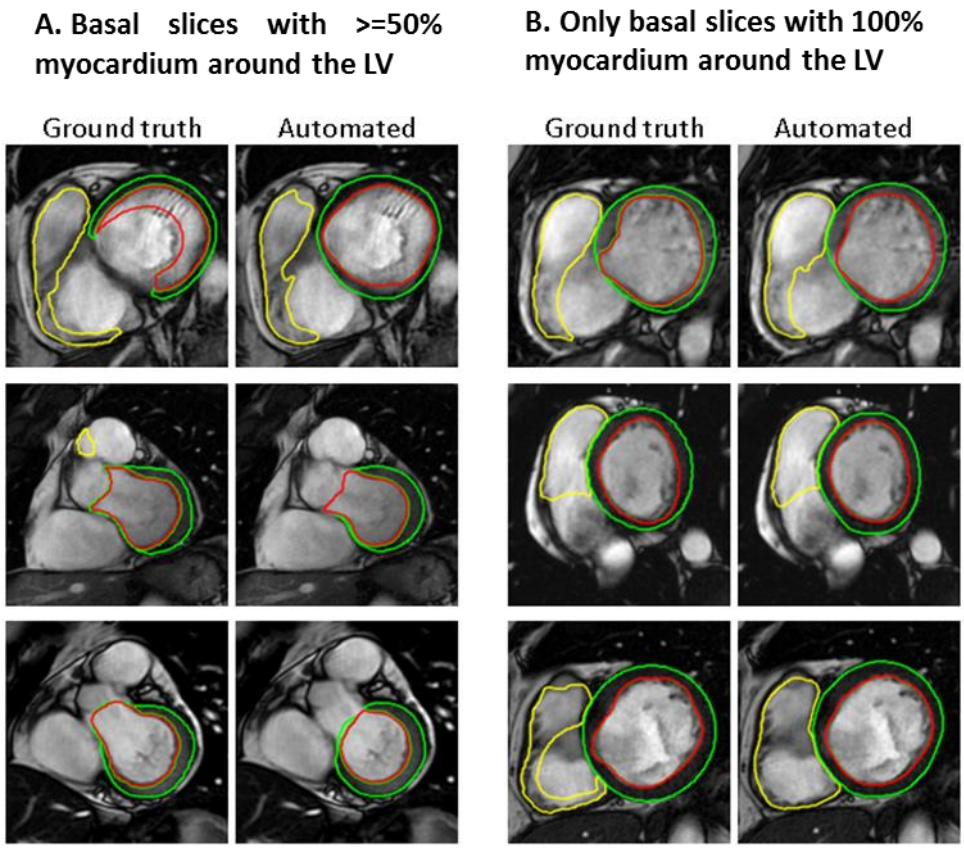


Fig. 2. Example of segmentation results of basal slices obtained using the proposed FCN compared to the manually traced gold standard (RV endocardium: yellow; LV endocardium: red; LV epicardium: green), by considering in the definition of the basal slice criterion A (presence of myocardium around the LV >=50% - left column) or criterion B (only the slices where the myocardium completely surrounds the LV - right column).

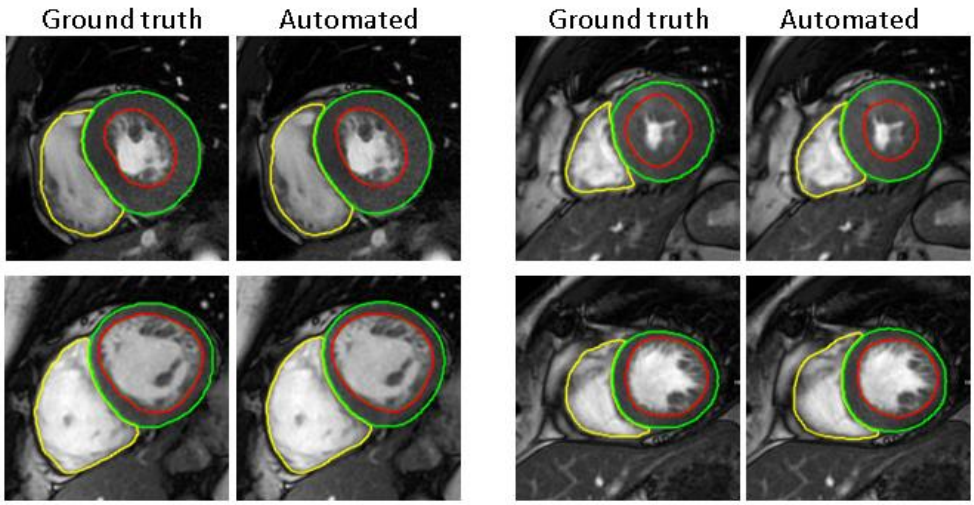


Fig. 3. Example of segmentation results of mid-ventricular level obtained using the proposed FCN compared to the manually traced gold standard (RV endocardium: yellow; LV endocardium: red; LV epicardium: green).

Figure 3 and Figure 4 show examples of the segmented contours at the LV mid and apical levels, respectively. It is possible to appreciate how the automated contours at mid level were able to deal with the presence of papillary muscles, by including them into the LV as traced by the gold standard, and at apical level, despite the small size of the LV (and RV when present), how the FCN was able to detect the correct contours.

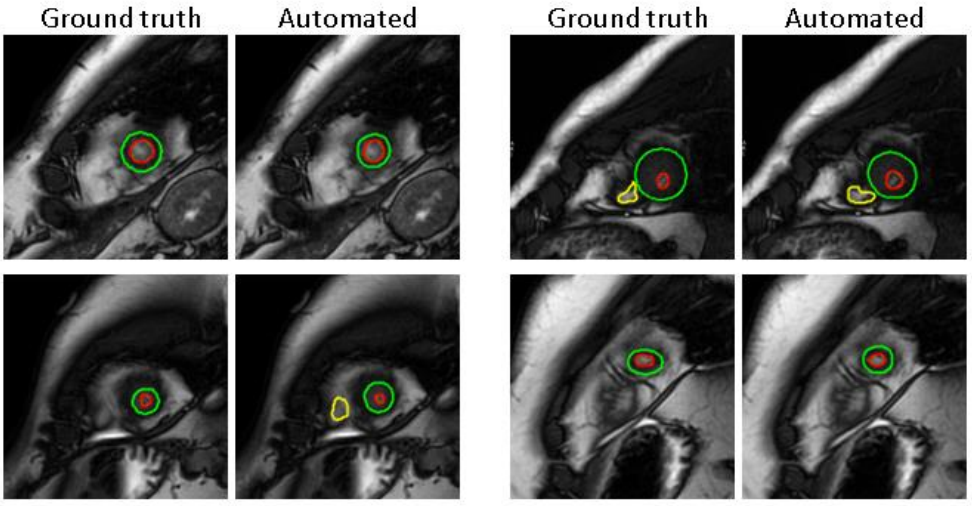


Fig. 4. Example of segmentation results of apical slices obtained using the proposed FCN compared to the manually traced gold standard (RV endocardium: yellow; LV endocardium: red; LV epicardium: green).

	A. Basal (Myo mass around the LV more than 50% but less than 100%)					
	LV			Myo		
	Se	Dice	HD [mm]	Se	Dice	HD [mm]
FCN+corr.	0.958 (0.831-0.983)#	0.927 (0.815-0.949)#	7.5 (5.2-12.2)#	0.768 (0.614-0.852)#	0.811 (0.690-0.847)#	13.5 (9.3-20.8)#
FCN	0.952 (0.832-0.981)	0.909 (0.829-0.951)#	9.7 (5.3-16.2)#	0.768 (0.612-0.836)#	0.808 (0.682-0.838)#	13.0 (9.3-18.3)#
	B. Basal (Myo mass surrounds completely the LV)					
	LV			Myo		
	Se	Dice	HD [mm]	Se	Dice	HD [mm]
FCN+corr.	0.979 (0.953-0.993)*	0.958 (0.949-0.969)	4.7 (2.9-7.1)	0.849 (0.753-0.905)*	0.874 (0.823-0.902)	4.5 (3.6-5.6)
FCN	0.976 (0.945-0.990)	0.959 (0.949-0.970)	5.0 (2.8-7.5)	0.851 (0.756-0.907)	0.873 (0.812-0.900)	4.5 (3.6-5.9)
	Apical					
	LV			Myo		
	Se	Dice	HD [mm]	Se	Dice	HD [mm]
FCN+corr.	0.890 (0.658-0.976)*	0.844 (0.715-0.907)	4.1 (2.8-6.2)	0.707 (0.455-0.849)	0.694 (0.557-0.815)	5.7 (3.6-6.9)
FCN	0.862 (0.592-0.963)	0.867 (0.709-0.900)	3.6 (2.4-5.2)	0.699 (0.435-0.855)	0.682 (0.532-0.811)	5.2 (3.6-6.7)

Table 4. Segmentation performance of the proposed FCN specifically at the basal (criteria A and B) and apical level, before (FCN) and after manual correction (FCN+corr) considering both end-diastole (ED) and end-systole (ES) , measured as Sensitivity (Se), Dice coefficient, and Hausdorff distance (HD) reported as median (25th-75th percentile),.

*: $p < 0.05$, FCN vs FCN+corr (Wilcoxon rank sum); #: $p<0.05$, A vs B (Mann–Whitney U test)

In Table 4, the accuracy of the automated segmentation specifically for the basal (defined as based on A or B criteria) and apical levels, together with the results following-manual correction of the automated segmentation by an expert physician, are shown in term of Se, Dice and HD with respect to the ground-truth. At the LV apex, the improvement relevant to manual correction was visible only for Se, but not for the Myo, for both ED and ES. At the basal region, when criterion B (i.e., with 100% of Myo around the LV) was applied, the effect of manually adjusted segmentation resulted in a higher Se for the LV for both ED and ES, but in slightly lower Se for the Myo ($p < 0.05$) in ES, with respect to the gold standard. When criterion A (i.e., >50% myocardium around the LV) was applied, respective performance in terms of Se, Dice and HD decreased (as expected) in both automated and corrected contours, where manual correction unexpectedly did not result in significant improvement in respect to automated results, thus highlighting a higher uncertainty and inter-observer variability in tracing the contours at this level.

In Table 5 the results of linear correlation (r^2) and Bland-Altman (Bias ± 2σ) analyses for the automated and manually adjusted segmentation, compared to the gold standard in terms of clinical parameters, are reported. For LV ED and ES volumes, as well as for RV ED, the FCN resulted in a strong correlation with no bias and narrow limits of agreement, both with and without correction, while a slight overestimation was measured for RV ES compared to the gold standard. For Myo mass, still a very good correlation was reported, with a slight overestimation ($p<0.05$) of both the FCN and the following correction compared to the gold standard, and a wider (but still acceptable) limits of agreement.

In addition, the CV% for each parameter is reported: it is worth noting that, considering the result of corrected contours by the expert observer, values of CV% can be interpreted as interobserver variability (between this observer and the gold standard). The CV% of the FCN contours compared to the gold standard showed values comparable to those of the expert observer, thus mimicking the performance in terms of interobserver variability of a human expert.

	LV_{EDV}			LV_{ESV}		
	r^2	Bias ± 2σ [ml]	CV (%)	r^2	Bias ± 2σ [ml]	CV (%)
FCN+corr	0.99	-2 ± 13	2.40	0.98	-2 ± 14	5.34
FCN	0.99	-1 ± 12	2.22	0.98	-1 ± 14	4.98
	RV_{EDV}			RV_{ESV}		
	r^2	Bias ± σ [ml]	CV (%)	r^2	Bias ± 2σ [ml]	CV (%)
FCN+corr	0.98	4 ± 14	3.03	0.96	7 ± 12*	8.02
FCN	0.99	6 ± 12	3.26	0.95	6 ± 14*	7.65
	$Myo_{ED\ Mass}$			$Myo_{ES\ Mass}$		
	r^2	Bias ± σ [g]	CV (%)	r^2	Bias ± 2σ [g]	CV (%)
FCN+corr	0.92	7 ± 26*	9.44	0.91	6 ± 28*	8.27
FCN	0.92	6 ± 26*	9.13	0.92	5 ± 26*	8.07

Table 5. Clinical metrics before (FCN) and after manual correction (FCN+corr) reported in terms of correlation (r^2), bias ± 2*standard deviation (Bias ± 2σ) and Coefficient of variation (CV), separately for LV, RV and Myo, on basal slices defined as criterion A (>=50% myocardium around the LV). EDV: End diastolic volume; ESV: End systolic volume. *: p<0.05 vs null (paired t-test)

In Figure 5, the boxplots for Dice and HD of each segmented structure in each patient group (i.e. HCM, DCM, and VA) are reported for comparison. It is possible to notice that the FCN performed similarly in all groups, except for the LV at ES phase in HCM patients, where a significant lower Dice was reported.

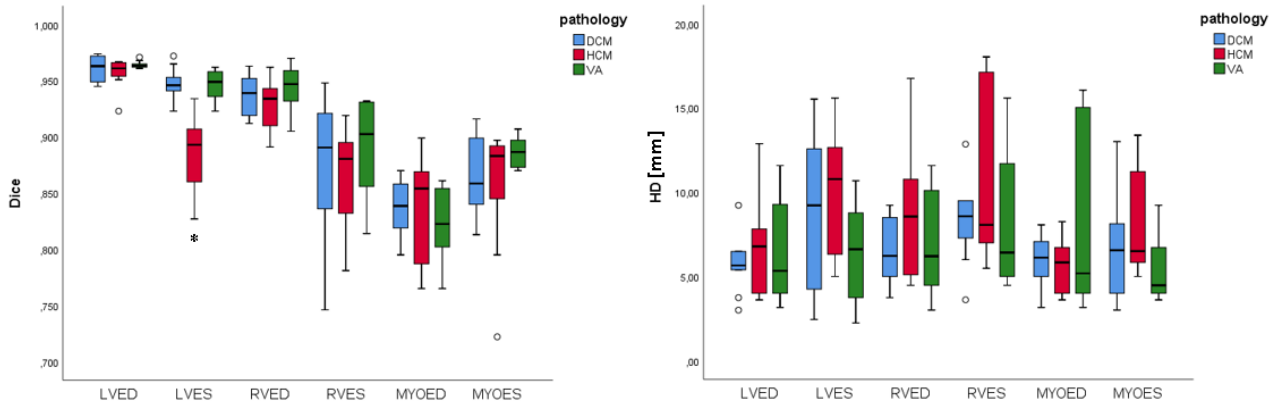


Fig. 5. Boxplots comparing Dice index and Hausdorff distance (HD) of each segmented cardiac structure at end-diastole (ED) and end-systole (ES), grouped by patient's pathology (DCM: dilated cardiomyopathy; HCM: hypertrophic cardiomyopathy; VA: ventricular arrhythmias). LV: left ventricular, RV: right ventricular; MYO: myocardium; ED: end-diastolic; ES: end-systolic. *: p<0.05 vs HCM (ANOVA or Kruskal-Wallis test with Bonferroni correction)

As regards generalizability of the results, Table 6 reports the cumulative performance for all slices of the segmentation for LV and RV endocardium and LV myocardium in the ED and ES frames for DB2, when the proposed FCN, trained on images of DB1 with criteria A to define the most basal slice, was applied.

Compared to the corresponding results obtained in DB1 (see Table 3), the obtained performance appears very similar, with a slight worsening in Se for LV both at ED and ES, in Dice for RV at ED, but a slight improvement in Se for the myocardium both at ED and ES.

Left Ventricle (ED)			Left Ventricle (ES)		
Se	Dice	HD [mm]	Se	Dice	HD [mm]
0.935 (0.011)*¥	0.957 (0.005)¥	7.8 (5.7-8.6)¥	0.893 (0.019)*	0.922 (0.011)	12.6 (9.1-13.3)*
Right Ventricle (ED)			**Right Ventricle (ES)**		
Se	Dice	HD [mm]	Se	Dice	HD [mm]
0.898 (0.042)¥	0.901 (0.034)*¥	6.4 (4.8-7.9)	0.838 (0.052)	0.858 (0.036)	6.6 (5.2-8.1)
Myocardium (ED)			**Myocardium (ES)**		
Se	Dice	HD [mm]	Se	Dice	HD[mm]
0.884 (0.028)*¥	0.834 (0.025)¥	4.2 (3.7-5.9)¥	0.929 (0.025)*	0.878 (0.013)	5.6 (4.3-17.3)

Table 6. Segmentation performance of the proposed FCN applied on DB2 at end-diastolic (ED) and end-systolic (ES) phases, measured as Sensitivity (Se), Dice coefficient, and Hausdorff distance (HD). Se and Dice are reported as mean (SD), while HD is reported as median (25^{th}-75^{th} percentile). *: $p<0.05$, DT2 vs DT1 (unpaired t-test or Mann–Whitney U test); ¥: $p<0.05$, ED vs ES (unpaired t-test or Mann–Whitney U test).

When the performance of the automated segmentation of DB2 specifically for the basal slice defined as criterion A is compared with that obtained for DB1, the only statistically significant difference was an improvement in LV HD, that decreased from 6.0 (4-8.3) to 5.0 (3.9-8.5).

4. Discussion and Conclusion

In this paper, a fully automated method based on FCN for accurate multi-structure segmentation on CMR, capable of reliably producing high quality LV and RV contour tracings, was proposed. To this aim, we used a FCN derived from the consolidated U-Net architecture as a common reference network [20], opportunely modified with a new scheme for skip connections. Especially in medical image, where precise segmentation is fundamental for accurate diagnosis, skip connections showed to be effective in recovering fine details lost in down-sampling path [20,42]. Skip connections between the corresponding layers of down-sampling and up-sampling path preserve the spatial information lost during the pooling operation and alleviate the vanishing-gradient problem [43]. In our specific implementation, compared to the traditional U-net, where high resolution feature maps are directly propagated from the down-sampling to the up-sampling path possibly degrading the segmentation, we took advantage from dense skip connections, that alleviate the

semantic gap generated when higher level features are merged with lower level features, with benefits in term of both network optimization [23,44] and consequently on its segmentation performance. Similar to ResNets [45], where identity shortcut connections are introduced, dense block tends to maximize the information flow, because each layer has a direct access to the other and to the input feature maps, with benefits in term of gradients propagation, and vanishing problem [37,44]. Differently from ResNets, in our implementation of dense blocks, the output feature maps were concatenated with the incoming ones and, from the down-sampling to the up-sampling path corresponding layer, the feature maps were propagated through Dense block. This had a direct consequence on the reduction in the number of parameters to learn, exploiting the potential of network through a feature reuse. This feature propagation improved the training phase, leading an implicit deep supervision, because each layer received an additional supervision from the loss function, which made the training phase easier.

An ablation study was performed to investigate the performance of the network at different depths and number of filters. A shallower architecture showed a drastic reduction in performance due to the limited capacity of the network to catch high-level features, as well as decreasing the number of filters which limited the number of extracted feature maps.

To validate the performance of the proposed FCN in line with current CMR guidelines [32], all the slices from the base to the apex of the heart where included. In addition, the analysis was repeated to evaluate the effect of the applied inclusion criteria for the LV basal slice, which represents the most variable and difficult image to be traced. Furthermore, CMR datasets from three different pathological groups were included, to avoid polarization of results for a very specific group of patients.

CMR dataset							
Name	Training Samples	Testing samples	LV	RV	Myo	Pathology	Basal slice [32]
Our Dataset	180	30	✓	✓	✓	✓	✓
TWINS-UK [16]	58	10	✓	–	–	–	–
Sunnybrook [15]	45		✓	–	–	✓	–
Preterm [15]	214	20	✓	–	–	✓	–
York [29]	33		✓	–	–	✓	–
STACOM [30]	100	100	✓	–	✓	–	–
MICCAI RV [31]	16	32	–	✓	–	–	–
KAGGLE [27]	500	200	–	–	–	–	–
ACDC [27]	100	50	✓	✓	✓	✓	–

Table 7. Summary of available CMR short-axis cine image datasets and relevant use for training and testing FCN-basedd methods made available in different challenges. LV/RV/Myo: ground truth segmentation available; Pathology: different clinical conditions included; Basal slice: basal slice defined in agreement with guidelines [32].

Table 7 shows a summary of the available CMR datasets based on SAX cine images used for validating FCN-based methods. Savioli et al [16] recruited 68 female subjects from the TWINS-UK cohort for LV segmentation, with each acquisition covering both ventricles from the atrioventricular ring to the apex. Poudel et al [15] utilized two datasets, the Sunnybrook Cardiac Data, made available during the MICCAI 2009 challenge on automated LV segmentation, consisting 45 patients with different pathologies (heart failure, hypertrophy, and heathy subjects), and the Preterm dataset, including 234 subjects with and without premature birth, for LV segmentation only. Nasr-Esfahani [29] focused on LV segmentation using a publicly available dataset of the York University including 33 patients of different pathologies such as cardiomyopathy, aortic regurgitation, enlarged ventricles, and ischemia. In 2011, as part of the STACOM challenge on automated LV myocardium segmentation, another dataset consisting of 200 patients was made available [30]. To focus on RV endocardium and epicardium segmentation, in 2012 MICCAI challenge provided a dataset with 48 cases [31]. In 2015, the KAGGLE challenge organized a competition with the goal to automatically measure LV ED and ES volumes, but no ground truth was provided [27]. In 2017, the MICCAI challenge provided the ACDC dataset for LV and RV segmentation with 150 patients of different pathologies, specifically myocardium infarction, DCM, HCM, abnormal RV and healthy subjects, but basal

slices according to CMR guidelines [32] and apical slices were not available. In comparison with these images, our dataset allows a more realistic evaluation of the proposed FCN for cardiac ventricular analysis in clinical practice, by considering both LV and RV, different pathologies, and including basal and apical slices according to CMR guidelines [32].

A joint analysis of Table 3 and 4 shows that the effect of the basal slice inclusion criteria on the outcome of the automated segmentation becomes visible only if the performance is reported separately for the basal slices, while it does not impact cumulative results where all slices are considered, due to the larger number of not basal slices. This fact suggests that, to allow better comparison of performances in CMR segmentation algorithms, the criteria of definition of basal slice should be always reported, and the corresponding results should be separately given, thus avoiding reporting only cumulative results. In fact, as expected, the automated segmentation of the LV and RV contours in the basal slice is arduous, due to the presence of the pulmonary valve, aortic valve, atrium and outflow tract, thus resulting in higher inter-observer variability, thus impacting on the segmentation accuracy. For example, including or not the portion of outflow tract in the manual segmentation is strictly dependent on the physician's decision on that particular patient, with the support of visualizing also the long axis views, not available to the FCN [46]. This variability generates instability during the training step, with the network continuously learning to adapt to different patterns. Despite these limitations, the performance of the proposed FCN could be considered still good also in those slices, in particular for LV and RV endocardium.

As regards the apical slices, the model was almost always able to localize the LV and RV, despite their small size, but with accuracy lower than that at basal level. Compared to [6], which suggests to focus on apical slices as a possible way to improve segmentation accuracy, our opinion is that also the algorithm performance should be separately reported for this slice, but more attention should be directed towards the basal contours, due to their major impact on clinical volumetric parameters due to the relevant chamber size, and to the higher variability in the cardiac structures.

The performance for the RV was slightly lower than for the LV, which it is also highlighted by the reduced correlation and greater bias (Table 5). The segmentation of the RV is more challenging, as demonstrated by

the lower Dice scores and higher HD. This is probably due to the more complicated geometry, especially at ES phase, when at maximum contraction the papillary muscles and trabeculations are more compacted. In addition, a higher variability from the basal to the apical slice, and a lower contrast with the surrounding structures, makes RV contour detection more complicated than the LV [46], limiting the accuracy of the segmentation process. It is important noticing that trabeculations and papillary muscles were included in the cavity, also in agreement with the current guidelines. Even the Myo was well segmented, even if with more difficulties when compared to the LV and RV, probably related to the detection of two contours (i.e. endocardium and epicardium) instead than a single one [27].

The analysis of possible challenges in the segmentation process relevant to a specific pathological condition did not show a systematic bias in performance, thus demonstrating the generalizability and accuracy of the proposed FCN algorithm.

The adjusted segmentation of an external physician returned the fidelity of the automated contours, demonstrating the possibilities and the potential offered by the proposed FCN in term of ventricular segmentations. As mentioned in [47], Neural Networks are emerging as gold standard in many medical fields outperforming expert physicians, but the object of the research would be to understand if these models together with a physician are better than the same physician alone. At this purpose our approach, comparing automated and corrected segmentation, showed a very similar performance, and inter-observer variability when compared to the gold standard, with the advantage of drastically reducing the manual segmentation time by only reviewing the detected boundaries instead of tracing them.

The analysis of the generalizability of the results, when the FCN was applied to a different database (DB2) obtained with a 3T scanner and different acquisition settings, showed very similar performances despite a slight worsening in Se for the LV, that might be due to the different intensity distributions and the lower resolution of these images, not included in the training phase, and even a slight improvement in HD when only basal slices were considered.

One possible limitation of our study is that the gold standard was not based on independent multiple tracings but on a single expert annotation, and only in cases with uncertainty in the image interpretation, the contours were reviewed through group consensus. In addition, the model was trained only on ED and ES phases, because manual contours were available for these two phases only; however, we expect the proposed FCN to be able to perform well even on the other frames in the cardiac cycle.

This work may be extended in different ways, by evaluating the performance of the proposed FCN model on other CMR datasets covering different pathologies. Furthermore, spatio-temporal features [48] could be also evaluated, with possible improvement in segmentation accuracy.

In conclusion, a deep learning approach for semantic segmentation on both right and left ventricle in CMR images with Dense skip connection was proposed, demonstrating its utility and potential in clinical practice, reaching high correlations with ground truth and accurate ventricular volumes estimation, and good generalizability of the obtained performances. The additional manual correction to the automatically obtained contours revealed how our supervised learning-based method was able to manage the real-world variability of functional and anatomical characteristics across different pathological conditions, with results in the range of expert inter-observer variability. The importance of clearly defining the inclusion criteria of basal slices in the testing dataset was underlined, as well as the need to separately report the algorithm performances in the most difficult slices (i.e., basal and apical). Such results reveal the potential of deep learning approach for fast, accurate and reliable automated cardiac segmentation in real clinical applications.

ACKNOWLEDGEMENTS

None.

CONFLICT OF INTEREST

There are no conflicts of interest that could inappropriately influence this research work.

ETHICAL APPROVAL

This study was approved by the Ethics Committee of the "Centro Cardiologico Monzino IRCCS", Milan, Italy.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

AUTHORSHIP RESPONSIBILITY

- The material in this manuscript is original and contains no matter libelous or otherwise unlawful.
- The manuscript represents valid work and that neither this manuscript nor any other with substantially similar content under my authorship has been published or is being considered for publication elsewhere.
- All authors have participated sufficiently in the work to take public responsibility for all its content.

REFERENCES

[1] Nichols, M., Townsend, N., Scarborough, P. and Rayner, M., 2014. Cardiovascular disease in Europe 2014: epidemiological update. European heart journal, 35(42), pp.2950-2959. https://doi.org/10.1093/eurheartj/ehu299.

[2] Kang, D., Woo, J., Kuo, C.J., Slomka, P.J., Dey, D. and Germano, G., 2012. Heart chambers and whole heart segmentation techniques. Journal of Electronic Imaging, 21(1), p.010901. https://doi.org/10.1117/1.JEI.21.1.010901.

[3] Suinesiaputra, A., Cowan, B.R., Al-Agamy, A.O., Elattar, M.A., Ayache, N., Fahmy, A.S., Khalifa, A.M., Medrano-Gracia, P., Jolly, M.P., Kadish, A.H. and Lee, D.C., 2014. A collaborative resource to build consensus for automated left ventricular segmentation of cardiac MR images. Medical image analysis, 18(1), pp.50-62. https://doi.org/10.1016/j.media.2013.09.001.

[4] Tavakoli, V. and Amini, A.A., 2013. A survey of shaped-based registration and segmentation techniques for cardiac images. Computer Vision and Image Understanding, 117(9), pp.966-989. https://doi.org/10.1016/j.cviu.2012.11.017.

[5] Peng, P., Lekadir, K., Gooya, A., Shao, L., Petersen, S.E. and Frangi, A.F., 2016. A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging. Magnetic Resonance Materials in Physics, Biology and Medicine, 29(2), pp.155-195. https://doi.org/10.1007/s10334-015-0521-4.

[6] Petitjean, C. and Dacher, J.N., 2011. A review of segmentation methods in short axis cardiac MR images. Medical image analysis, 15(2), pp.169-184. https://doi.org/10.1016/j.media.2010.12.004.

[7] Bai, W., Shi, W., Ledig, C. and Rueckert, D., 2015. Multi-atlas segmentation with augmented features for cardiac MR images. Medical image analysis, 19(1), pp.98-109. https://doi.org/10.1016/j.media.2014.09.005.

[8] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B. and Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical image analysis, 42, pp.60-88. https://doi.org/10.1016/j.media.2017.07.005.

[9] Li, Z., Dong, M., Wen, S., Hu, X., Zhou, P. and Zeng, Z., 2019. CLU-CNNs: Object detection for medical images. Neurocomputing, 350, pp.53-59. https://doi.org/10.1016/j.neucom.2019.04.028

[10] Pereira, S., Pinto, A., Alves, V. and Silva, C.A., 2016. Brain tumor segmentation using convolutional neural networks in MRI images. IEEE transactions on medical imaging, 35(5), pp.1240-1251. https://doi.org/10.1109/TMI.2016.2538465.

[11] Yadav, S.S. and Jadhav, S.M., 2019. Deep convolutional neural network based medical image classification for disease diagnosis. Journal of Big Data, 6(1), p.113. https://doi.org/10.1186/s40537-019-0276-2.

[12] Colleoni, E., Moccia, S., Du, X., De Momi, E. and Stoyanov, D., 2019. Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers. IEEE Robotics and Automation Letters, 4(3), pp.2714-2721. . https://doi.org/10.1109/LRA.2019.2917163.

[13] Moccia, S., Banali, R., Martini, C., Muscogiuri, G., Pontone, G., Pepi, M. and Caiani, E.G., 2019. Development and testing of a deep learning-based strategy for scar segmentation on CMR-LGE images. Magnetic Resonance Materials in Physics, Biology and Medicine, 32(2), pp.187-195. https://doi.org/10.1007/s10334-018-0718-4.

[14] Isensee, F., Jaeger, P.F., Full, P.M., Wolf, I., Engelhardt, S. and Maier-Hein, K.H., 2017. Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features. In International workshop on statistical atlases and computational models of the heart , pp. 120-129. https://doi.org/10.1007/978-3-319-75541-0_13.

[15] Poudel, R.P., Lamata, P. and Montana, G., 2016. Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation. In Reconstruction, segmentation, and analysis of medical images, pp.83-94. https://doi.org/10.1007/978-3-319-52280-7_8.

[16] Savioli, N., Vieira, M.S., Lamata, P. and Montana, G., 2018, October. Automated segmentation on the entire cardiac cycle using a deep learning work-flow. In 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS),pp.153-158. https://doi.org/10.1109/SNAMS.2018.8554962.

[17] Zhen, X., Wang, Z., Islam, A., Bhaduri, M., Chan, I. and Li, S., 2016. Multi-scale deep networks and regression forests for direct bi-ventricular volume estimation. Medical image analysis, 30, pp.120-129. https://doi.org/10.1016/j.media.2015.07.003.

[18] Yang, X., Bian, C., Yu, L., Ni, D. and Heng, P.A., 2017, September. Class-balanced deep neural network for automatic ventricular structure segmentation. In International workshop on statistical atlases and computational models of the heart (pp. 152-160). Springer, Cham. https://doi.org/10.1007/978-3-319-75541-0_16.

[19] Baumgartner, C.F., Koch, L.M., Pollefeys, M. and Konukoglu, E., 2017, September. An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation. In International Workshop on Statistical Atlases and Computational Models of the Heart, pp.111-119. Springer, Cham. https://doi.org/10.1007/978-3-319-75541-0_12.

[20] Ronneberger, O., Fischer, P. and Brox, T., 2015, October. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pp.234-241. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28.

[21] Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S. and Pal, C., 2016. The importance of skip connections in biomedical image segmentation. In Deep Learning and Data Labeling for Medical Applications, pp. 179-187. https://doi.org/10.1007/978-3-319-46976-8_19.

[22] Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N. and Liang, J., 2019. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging, 39(6), pp.1856-1867. https://doi.org/10.1109/TMI.2019.2959609.

[23] Ibtehaz, N. and Rahman, M.S., 2020. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural Networks, 121, pp.74-87. https://doi.org/10.1016/j.neunet.2019.08.025

[24] Xue, W., Lum, A., Mercado, A., Landis, M., Warrington, J. and Li, S., 2017, September. Full quantification of left ventricle via deep multitask learning network respecting intra-and inter-task relatedness. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp.276-284. Springer, Cham. https://doi.org/10.1007/978-3-319-66179-7_32.

[25] Zange, L., Muehlberg, F., Blaszczyk, E., Schwenke, S., Traber, J., Funk, S. and Schulz-Menger, J., 2019. Quantification in cardiovascular magnetic resonance: agreement of software from three different vendors on assessment of left ventricular function, 2D flow and parametric mapping. Journal of Cardiovascular Magnetic Resonance, 21(1), pp.12. https://doi.org/10.1186/s12968-019-0522-y.

[26] Bai, W., Sinclair, M., Tarroni, G., Oktay, O., Rajchl, M., Vaillant, G., Lee, A.M., Aung, N., Lukaschuk, E., Sanghvi, M.M. and Zemrak, F., 2018. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. Journal of Cardiovascular Magnetic Resonance, 20(1), pp.65. https://doi.org/10.1186/s12968-018-0471-x.

[27] Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G. and Sanroma, G., 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?. IEEE transactions on medical imaging, 37(11), pp.2514-2525. https://doi.org/10.1109/TMI.2018.2837502

[28] Petitjean, C., Zuluaga, M.A., Bai, W., Dacher, J.N., Grosgeorge, D., Caudron, J., Ruan, S., Ayed, I.B., Cardoso, M.J., Chen, H.C. and Jimenez-Carretero, D., 2015. Right ventricle segmentation from cardiac MRI: a collation study. Medical image analysis, 19(1), pp.187-202. https://doi.org/10.1016/j.media.2014.10.004.

[29] Nasr-Esfahani, M., Mohrekesh, M., Akbari, M., Soroushmehr, S.R., Nasr-Esfahani, E., Karimi, N., Samavi, S. and Najarian, K., Left ventricle segmentation in cardiac MR images using fully convolutional network. In

2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 1275-1278). IEEE..

[30] Queirós, S., Barbosa, D., Heyde, B., Morais, P., Vilaça, J.L., Friboulet, D., Bernard, O. and D'hooge, J. Fast automatic myocardial segmentation in 4D cine CMR datasets. Medical image analysis, 18(7), pp.1115-1131. https://doi.org/10.1016/j.media.2014.06.001.

[31] Petitjean, C., Zuluaga, M.A., Bai, W., Dacher, J.N., Grosgeorge, D., Caudron, J., Ruan, S., Ayed, I.B., Cardoso, M.J., Chen, H.C. and Jimenez-Carretero, D. Right ventricle segmentation from cardiac MRI: a collation study. Medical image analysis, 19(1), pp.187-202. https://doi.org/10.1016/j.media.2014.10.004

[32] Schulz-Menger, J., Bluemke, D.A., Bremerich, J., Flamm, S.D., Fogel, M.A., Friedrich, M.G., Kim, R.J., von Knobelsdorff-Brenkenhoff, F., Kramer, C.M., Pennell, D.J. and Plein, S., 2013. Standardized image interpretation and post processing in cardiovascular magnetic resonance: Society for Cardiovascular Magnetic Resonance (SCMR) board of trustees task force on standardized post processing. Journal of Cardiovascular Magnetic Resonance, 15(1), pp.1-19. https://doi.org/10.1186/1532-429X-15-35.

[33] Caiani, E.G., Landreani, F., Costantini, L., Mulder, E., Gerlach, D., Vaada, P. and Migeotte, P.F., 2018. Effectiveness of high-intensity jump training countermeasure on mitral and aortic flow after 58-days head-down bed-rest assessed by phase-contrast MRI. In International Astronautical Congress (Vol. 2018, pp. 1-6).

[34] Cherabit, N., Chelali, F.Z. and Djeradi, A., 2012. Circular hough transform for iris localization. Science and Technology, 2(5), pp.114-121. https://doi.org/10.5923/j.scit.20120205.02.

[35] Santurkar, S., Tsipras, D., Ilyas, A. and Madry, A., 2018. How does batch normalization help optimization?. In Advances in Neural Information Processing Systems, pp.2483-2493.

[36] Nair, V. and Hinton, G.E., 2010, January. Rectified linear units improve restricted boltzmann machines. In ICML.

[37] Zeiler, M.D., Krishnan, D., Taylor, G.W. and Fergus, R., 2010, June. Deconvolutional networks. In 2010 IEEE Computer Society Conference on computer vision and pattern recognition, pp. 2528-2535. https://doi.org/10.1109/CVPR.2010.5539957.

[38] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., 2017. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp.2261-2269. https://doi.org/ 10.1109/CVPR.2017.243.

[39] He, K., Zhang, X., Ren, S. and Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision, pp.1026-1034. . https://doi.org/10.1109/ICCV.2015.123.

[40] Dogo, E.M., Afolabi, O.J., Nwulu, N.I., Twala, B. and Aigbavboa, C.O., 2018, December. A comparative analysis of gradient descent-based optimization algorithms on convolutional neural networks. In 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), pp.92-99. IEEE. https://doi.org /10.1109/CTEMS.2018.8769211.

[41] Shorten, C. and Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), p.60. https://doi.org/10.1186/s40537-019-0197-0

[42] Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp.3431-3440. . https://doi.org/10.1109/CVPR.2015.7298965.

[43] Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S. and Pal, C., 2016. The importance of skip connections in biomedical image segmentation. In Deep Learning and Data Labeling for Medical Applications, pp.179-187. https://doi.org/10.1007/978-3-319-46976-8_19.

[44] Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N. and Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp.3-11. https://doi.org/10.1007/978-3-030-00889-5_1.

[45] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp.770-778. https://doi.org/10.1109/CVPR.2016.90.

[46] Marchesseau, S., Ho, J.X. and Totman, J.J., 2016. Influence of the short-axis cine acquisition protocol on the cardiac function evaluation: a reproducibility study. European journal of radiology open, 3, pp.60-66. https://doi.org/10.1016/j.ejro.2016.03.003.

[47] Baselli, G., Codari, M. and Sardanelli, F., 2020. Opening the black box of machine learning in radiology: can the proximity of annotated cases be a way?. European Radiology Experimental, 4, pp.1-7. https://doi.org/10.1186/s41747-020-00159-0.

[48] Moccia, S., Migliorelli, L., Carnielli, V. and Frontoni, E., 2019. Preterm infants' pose estimation with spatio-temporal features. IEEE Transactions on Biomedical Engineering. https://doi.org/10.1109/tbme.2019.2961448.

APPENDIX

CLAIM: Checklist for Artificial Intelligence in Medical Imaging

Section / Topic	No.	Item	
TITLE / ABSTRACT			
	1	Identification as a study of AI methodology, specifying the category of technology used (e.g., deep learning)	**Pag 1, title** **Pag 2, row 6**
	2	Structured summary of study design, methods, results, and conclusions	**Pag 2-3**
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the AI approach	**Pag 4-5**
	4	Study objectives and hypotheses	**Pag 6, row 6-22**
METHODS			
Study Design	**5**	Prospective or retrospective study	**Pag 7, row 1**
	6	Study goal, such as model creation, exploratory study, feasibility study, non-inferiority trial	**Pag 4, row 16-26** **Pag 6, row 6-22**
Data	**7**	Data sources	**Pag 7, row 1-4, 21-26; page 8, row 1-5**
	8	Eligibility criteria: how, where, and when potentially eligible participants or studies were identified (e.g., symptoms, results from previous tests, inclusion in registry, patient-care setting, location, dates)	**Pag 7-8, sec 2.1**
	9	Data pre-processing steps	**Pag 8-9, sec 2.2** **Pag 11, sec 2.4**
	10	Selection of data subsets, if applicable	**Pag 8-9, sec 2.2**
	11	Definitions of data elements, with references to Common Data Elements	**Pag 11-13, sec 2.5**
	12	De-identification methods	**N/A**
	13	How missing data were handled	**N/A**
Ground Truth	**14**	Definition of ground truth reference standard, in sufficient detail to allow replication	**Pag 7, row 8-26; pag 8, row 4-5**
	15	Rationale for choosing the reference standard (if alternatives exist)	**N/A**
	16	Source of ground-truth annotations; qualifications and preparation of annotators	**Pag 7, sec 2.1, row 9-10** **Pag 8, row 5-**

			3
	17	Annotation tools	**Pag 7, row 9-10**
	18	Measurement of inter- and intrarater variability; methods to mitigate variability and/or resolve discrepancies	**Pag 7, row 15-16; pag 13, row 3-5**
Data Partitions	**19**	Intended sample size and how it was determined	**Pag 7, sec 2.1** **Pag 7, row 26, pag 8, row 1-2**
	20	How data were assigned to partitions; specify proportions	**Pag 12, sec 2.5 row 10-11**
	21	Level at which partitions are disjoint (e.g., image, study, patient, institution)	**Pag 12, sec 2.5, row 10**
Model	**22**	Detailed description of model, including inputs, outputs, all intermediate layers and connections	**Pag 9-10, sec 2.3**
	23	Software libraries, frameworks, and packages	**Pag 12, row 8**
	24	Initialization of model parameters (e.g., randomization, transfer learning)	**Pag 11, sec 2.4 row 4**
Training	**25**	Details of training approach, including data augmentation, hyperparameters, number of models trained	**Pag 11-12, sec 2.4**
	26	Method of selecting the final model	**Pag 11-12, sec 2.4**
	27	Ensembling techniques, if applicable	**N/A**
Evaluation	**28**	Metrics of model performance	**Pag 12, sec 2.5 row 11-15**
	29	Statistical measures of significance and uncertainty (e.g., confidence intervals)	**Pag 13, sec 2.5 row 1-2** **Pag 13 row 12-14**
	30	Robustness or sensitivity analysis	**Pag 13, row 12-14**
	31	Methods for explainability or interpretability (e.g., saliency maps), and how they were validated	**N/A**
	32	Validation or testing on external data	**Pag 13, row 9-11**
RESULTS			
Data	**33**	Flow of participants or cases, using a diagram to indicate inclusion and exclusion	**Pag 7, sec 2.1**
	34	Demographic and clinical characteristics of cases in each partition	**Pag 8 table1**
Model performance	**35**	Performance metrics for optimal model(s) on all data partitions	**Pag 13-14, section 3**
	36	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	**Pag 18,**
	37	Failure analysis of incorrectly classified cases	**Fig 2,3,4**
DISCUSSION			
	38	Study limitations, including potential bias, statistical uncertainty, and	**Pag 25, row**

		generalizability	**1-8**
	39	Implications for practice, including the intended use and/or clinical role	**Pag 25, row 9-18**
OTHER INFORMATION			
	40	Registration number and name of registry	**N/A**
	41	Where the full study protocol can be accessed	**N/A**
	42	Sources of funding and other support; role of funders	**Pag 26**

Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020; 2(2):e200029. https://doi.org/10.1148/ryai.2020200029

Revised Manuscript clean version (Without Track changes)

Automated Left and Right Ventricular Chamber Segmentation in Cardiac Magnetic Resonance Images Using Dense Fully Convolutional Neural Network

Marco Penso[a], Sara Moccia[b,c], Stefano Scafuri[a], Giuseppe Muscogiuri[a], Gianluca Pontone[a], Mauro Pepi[a], Enrico Gianluca Caiani[d,e]

[a] Department of Cardiovascular Imaging, Centro Cardiologico Monzino IRCCS, Milan, Italy.
(Marco.Penso@cardiologicomonzino.it, scafuristefano90@gmail.com, Giuseppe.Muscogiuri@cardiologicomonzino.it, gianluca.pontone@cardiologicomonzino.it, Mauro.Pepi@cardiologicomonzino.it)

[b] Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy.
(s.moccia@staff.univpm.it)

[c] Department of Advanced Robotics, Istituto Italiano di Tecnologia, Genoa, Italy.

[d] Department of Electronics, Information and Biomedical engineering, Politecnico di Milano, Milan, Italy.
(enrico.caiani@polimi.it)

[e] Consiglio Nazionale delle Ricerche, Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni, Milan, Italy.

Corresponding author: Prof. Enrico Gianluca Caiani

Politecnico di Milano

Department of Electronics, Information and Biomedical engineering

P.zza L. da Vinci 32

20133 Milan, Italy

Office: +39.02.2399.3390

Abstract

Background and Objective: Segmentation of the left ventricular (LV) myocardium (Myo) and RV endocardium on cine cardiac magnetic resonance (CMR) images represents an essential step for cardiac-function evaluation and diagnosis. In order to have a common reference for comparing segmentation algorithms, several CMR image datasets were made available, but in general they do not include the most apical and basal slices, and/or gold standard tracing is limited to only one of the two ventricles, thus not fully corresponding to real clinical practice. Our aim was to develop a deep learning (DL) approach for automated segmentation of both RV and LV chambers from short-axis (SAX) CMR images, reporting separately the performance for basal slices, together with the applied criterion of choice.

Method: A retrospectively selected database (DB1) of 210 cine sequences (3 pathology groups) was considered: images (GE, 1.5 T) were acquired at Centro Cardiologico Monzino (Milan, Italy), and end-diastolic (ED) and end-systolic frames (ES) were manually segmented (gold standard, GS). Automatic ED and ES RV and LV segmentation were performed with a U-Net inspired architecture, where skip connections were redesigned introducing dense blocks to alleviate the semantic gap between the U-Net encoder and decoder. The proposed architecture was trained including: A) the basal slices where the Myo surrounded the LV for at least the 50% and all the other slice; B) all the slices where the Myo completely surrounded the LV. To evaluate the clinical relevance of the proposed architecture in a practical use case scenario, a graphical user interface was developed to allow clinicians to revise, and correct when needed, the automatic segmentation. Additionally, to assess generalizability, analysis of CMR images obtained in 12 healthy volunteers (DB2) with different equipment (Siemens, 3T) and settings was performed.

Results: The proposed architecture outperformed the original U-Net. Comparing the performance on DB1 between the two criteria, no significant differences were measured when considering all slices together, but were present when only basal slices were examined. Automatic and manually-adjusted segmentation performed similarly compared to the GS (bias±95%LoA): LVEDV -1±12 ml, LVESV -1±14 ml, RVEDV 6±12 ml, RVESV 6±14 ml, ED LV mass 6±26 g) , ES LV mass 5±26 g. Also, generalizability showed very similar

performance, with Dice scores of 0.944 (LV), 0.908 (RV) and 0.852 (Myo) on DB1, and 0.940 (LV), 0.880 (RV), and 0.856 (Myo) on DB2.

Conclusions: Our results support the potential of DL methods for accurate LV and RV contours segmentation and the potential of dense skip connections to alleviate the semantic gap generated when high level features are concatenated with lower level feature. The evaluation on our dataset, considering separately the performance on basal and apical slices, reveals the potential of DL approaches for fast, accurate and reliable automated cardiac segmentation in a real clinical setting.

Keywords: Convolutional Neural Networks, Cardiac Segmentation, Cine Cardiac Magnetic Resonance, Dense skip connections

1. Introduction:

Cardiovascular disease represents one of the main causes of death in Europe [1]. Cardiac imaging is used to obtain structural and functional cardiac parameters for both diagnosis and disease evaluation [2]. In particular, Cardiac Magnetic Resonance (CMR) is the reference technique for estimating parameters such as left (LV) and right (RV) ventricular end-diastolic (EDV) and end-systolic (ESV) volumes, stroke volume (SV), ejection fraction (EF), and myocardial mass, as a result of an endocardial and epicardial contour segmentation. Although this operation can be performed using software for automated segmentation, visual inspection and manual correction are still requested, in particular for the right ventricular endocardium and for the LV basal and apical slices [3]. In the last decade, several model-based approaches to automatize the segmentation process were developed and tested, such as deformable models, cardiac atlas-based registration, level sets, statistical shape and active appearance models [4-7]. More recently, machine learning (ML) algorithms have been introduced to deal with the huge variability of clinical data. Specifically, Convolutional Neural Networks (CNNs) are emerging as the new state of the art for medical-image segmentation [8-12]. One of the advantages of the CNNs is the ability to learn more robust features through a hierarchical structure composed of learnable filters, without relying on hand-crafted features.

CNNs in different configurations have been applied to CMR images for segmenting both myocardial fibrosis and cardiac chambers. For example, Moccia et al. [13] obtained promising performance in LV scar tissue segmentation on LGE images using an encoder-decoder fully convolutional network (FCN). On cine images, Isensee et al. [14] implemented an ensemble of 2D and 3D U-Net architectures for LV and RV segmentation, while Poudel et al. [15] proposed a recurrent neural network for LV segmentation in multi-slice CMR cine images to capture global anatomical structure of the heart; similarly, Savioli et al. [16] adopted a Gated Recurrent Unit to exploit information of consecutive CMR cine frames, that might affect the backpropagation through time, limiting the number of consecutive frames that can be analyzed. Differently, Zhen et al. [17] developed a multi-scale CNN followed by random forests for a direct estimation of cardiac biventricular volumes, while Yang et al. [18] implemented a 3D U-Net with residual connections for ventricular volumetric segmentation in cine images. However, using a 3D network, the amount of

training images is drastically reduced and a consistent downsampling of the data is required due to memory constraints, leading to a potential loss of information [19]. The majority of this approaches are typically based on U-Net [20], an architecture made of a down-sampling (encoder) and up-sampling (decoder) path, and its variations. The success of U-Net may be attributed to the introduction of the skip connection for combining coarse-grained with fine-grained feature maps, recovering lost details of the target structures to be segmented [21]. A possible limitation of the original skip connections presented in [20] can be seen in the direct fusion of feature maps that are semantically different. Currently, this feature fusion is not supported by a solid theory which guarantees it to be the best solution [22]. Hence, feature maps of the up-sampling path, which are computed from the deepest layers of the network, are supposed to encode high-level information, while the maps of the down-sampling path refer to more general, low-level features. This generates a semantic gap, which could affect the segmentation accuracy [23].

CNNs achieved remarkable success as described in the literature, but an accurate segmentation of the myocardium in cine CMR images remains a challenging task in the present clinical routine. This is due to the presence of papillary muscles and trabeculations, and to the shape variability of ventricles across slices, cardiac phases, and pathologies [24]. This is particularly true at the ventricular basal and apical level: the former, with the presence of different anatomical structures such as the aortic and tricuspid valves and atrial chambers, and the latter due to the reduced size of the ventricles. Even if commercial software for automated ventricular boundary detection and computation of clinical parameters are available [25], the accuracy of the detected contours still needs to be improved to reach that of manual segmentation in all slices, and in particular for the basal and apical ones, usually more prone to discrepancies compared to manual tracings. As an example, in [26] the correspondence between manual and automated contours evaluated by two different observers in a subsample of 250 patients resulted as 85% and 92% in the mid-ventricular slices, while decreased down to 40% and 33% for the basal slices, and down to 44% and 81% for apical slices, respectively for the first and second observer.

To possibly develop and compare the performance of deep learning (DL) algorithms, over the past years different CMR image datasets were made available through several challenges [27]. However, they

generally did not include the most apical and basal slices, reference contours were limited to only one of the two ventricles, or the patient population was relevant to a specific clinical conditions, thus resulting in a non-optimal reference for comparison when referred to real clinical practice [15,28-31]. At the same time, the variability encoded in such datasets is limited when compared to the real clinical scenario, posing issues to the development of DL algorithms.

In this work, we hypothesized that CNNs could achieve segmentation accuracy comparable to that of expert physicians in the tracing of LV and RV endocardium and epicardium even when including the more basal and apical slices. Accordingly, our contributions can be summarized as follows:

1) Developing an FCN architecture based on U-Net [20], trained end-to-end for semantic segmentation of LV, RV and LV myocardium (Myo) on short-axis (SAX) cine CMR images. To attenuate the semantic gap that occurs in the original implementation of skip connections and improve segmentation accuracy, a novel skip connections with dense block is proposed.

2) Performing an experimental analysis on a large clinical dataset including both LV and RV chambers of patients with different pathologies and evaluating how the performance of the proposed FCN changes according to the definition criterion of the basal slices (i.e., following current recommendations of the Society for Cardiovascular Magnetic Resonance (SCMR) [32], or adopting a simplification criteria as in publicly-available datasets).

3) Evaluating the extent of manual correction needed after automated segmentation for a real use case scenario, and the relevant clinical impact on the obtained measurements of LV and RV volumes and LV mass.

4) Assessing the generalizability of the proposed approach by applying it to a different CMR image database obtained using different settings and equipment.

2. Materials and Methods:

2.1 Patient population

A retrospectively selected database (DB1) consisting of SAX cine (steady-state free-precession) CMR images (size 512x512 pixels, in-plane pixel resolution from 1.2 to 2.0 mm, acquired (GE, 1.5 T) in 210 patients referred to CMR evaluation at IRCCS Centro Cardiologico Monzino (Milan, Italy) along a 6 months period (September 2019-February 2020) was analyzed. The dataset included patients with the following diagnoses: 70 patients with hypertrophic cardiomyopathy (HCM), 70 with dilated cardiomyopathy (DCM) and 70 with ventricular arrhythmias (VA). Each stack of SAX images was acquired during breath hold, including both the LV and RV from base to apex, with a slice thickness of 8 mm, no inter-slice gap, and a field of view ranging from 300 to 360 mm^2. For each patient, images were acquired with 30 phases/cardiac cycle. The manual tracings of the LV endocardium and epicardium and RV endocardium (cvi software, ver. 5.1.1, Circle Cardiovascular Imaging Inc., Calgary, Canada), performed as part of the clinical evaluation at the time of the exam by one of the experienced cardiologists (10 years of experience) in the team at IRCCS Centro Cardiologico Monzino (Milan, Italy), were used as gold standard (GS). To determine where to trace the contours, the expert cardiologist usually examines all the acquired datasets, including other views (apical 2-, 3-, 4-chambers), to appreciate the dynamic motion and contiguity in space of structures visible in adjacent slices. In case of doubts in the interpretation and tracings, the expert cardiologist could count on the second opinion of additional experts in the group. For each patient, the manual tracing was provided only at the end-diastolic (ED) and at the end-systolic (ES) frames, which were chosen as the image with the largest and smallest LV blood volume, respectively, by examining the full image stack. The LV papillary muscles and trabeculae were included as part of LV and RV blood cavity, in agreement with what indicated in the guidelines of the SCMR [32].

Additionally, to investigate the generalization power of the proposed network, a second database (DB2) in our availability, acquired with different settings and equipment, was used as additional testing platform. DB2 consisted of SAX cine CMR images (size from 208x162 to 208x210 pixels, in-plane pixel resolution 1.63 mm, slice thickness 7 mm, 25 frames/cardiac cycle) acquired during breath hold, including both the LV and

RV from base to apex, using a Biograph mMR 3-Tesla scanner (Siemens, Erlangen, Germany) at the :envihab facility of the Institute of Aerospace Medicine at the German Aerospace Center (DLR, Koln, Germany) in 12 healthy volunteers in the context of a head-down bed rest study organized by the European Space Agency [33]. Also for this database, GS manual contours of the LV endocardium and epicardium and RV endocardium traced by an experienced cardiologist for the ED and ES phases were available.

The population demographic, together with the GS measurements for DB1 (considering both all 210 patients altogether and separately in the three HCM, VA and DCM groups) and DB2 are presented in Table 1.

Characteristics	**DB 1**				**DB 2**
	All (n = 210)	VA (n = 70)	DCM (n = 70)	HCM (n = 70)	All (n = 12)
Age, years	56 (42-65)	43 (31-59)	60 (50-69)	57 (44-64)	35 (28-43)
Female, n (%)	77 (37%)	33 (47%)	25 (36%)	19 (27%)	6 (50%)
Body surface area, m^2	1.9 (1.7-2.0)	1.8 (1.7-2.0)	1.9 (1.7-2.0)	1.9 (1.8-2.0)	1.9 (1.9-2.0)
LVEDV index (mL/m^2)	89 (77-113)	90 (74-100)	108 (83-133)	85 (76-92)	92 (81-101)
LVESV index (mL/m^2)	42 (33-56)	40 (32-49)	58 (42-89)	36 (29-41)	32 (28-35)
LVSV (mL/m^2)	47 (40-52)	48 (44-55)	45 (37-52)	48 (43-52)	60(53-67)
LVEF (%)	53.0 (44.0-60.0)	55.0 (50.5-60.0)	41.0 (32.8-53.0)	58.0 (52.0-63.0)	65.1 (64.7-66.0)
RVEDV index (mL/m^2)	81 (69-94)	86 (73-92)	80 (67-101)	76 (69-91)	104 (88-120)
RVESV index (mL/m^2)	36 (28-45)	36 (30-42)	39 (28-51)	32 (27-40)	46 (33-53)
RVSV (mL/m^2)	46 (40-51)	48 (41-53)	44 (36-50)	46 (41-50)	59 (53-62)
RVEF (%)	56.0 (51.0-61.0)	56.0 (51.5-62.0)	54.0 (49.8-59.0)	58.0 (53.3-62.0)	58.1 (52.4-61.5)

Table 1. Values are reported as median (25th-75th percentile), or n (%). LV: left ventricular; RV: right ventricular; EDV: end diastolic volume; ESV: end systolic volume; SV: stroke volume; EF: ejection fraction; HCM: hypertrophic cardiomyopathy; DCM: dilated cardiomyopathy; VA: ventricular arrhythmias.

2.2 Pre-processing

Data imported as DICOM images underwent a pre-processing step, consisting in cropping a 2D region of interest (ROI) centered with respect to the heart, normalizing voxel intensities to [0, 1] and a down-sampling operation for DT1 or an up-sampling operation for DT2 to obtain a ROI of 176x176 pixels size, to comply with memory constraints.

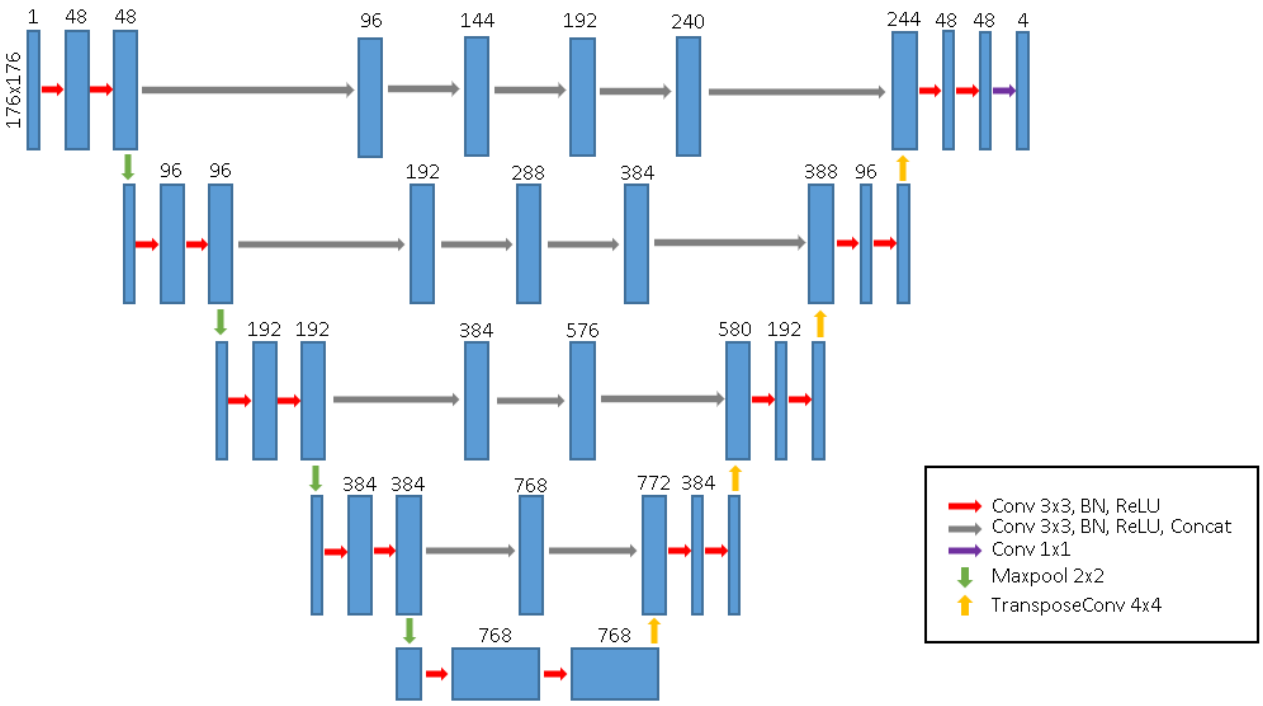


Fig. 1. Architecture of the proposed segmentation network for a 176x176 pixels input image. Each blue box corresponds to a multi-channel feature map. The number of channels is reported on top of each box.

Image cropping to obtain a squared 2D ROI was fully automatic, and was performed on a mid-ventricular slice for each patient, using the circular Hough Transform for the identification of the LV cavity centroid, and then transposed to all other slices by centering the crop area with the detected LV center. Selecting a ROI was useful in alleviating the class-imbalance problem between the analyzed structures (i.e. LV, RV and Myo) and background pixels [34]. The cine CMR slices were rotated and converted to the Right-Anterior-Head reference frame.

2.3 Network Architecture

The proposed FCN architecture takes inspiration from U-Net [20], which has symmetric encoder (down-sampling) and (up-sampling) decoder paths, with skip connections between them. With the down-sampling path, U-Net encodes context information that are transferred with skip connections to the up-sampling path, which is used to restore feature-map resolution and to produce the segmentation maps. Each step of the down-sampling path consists of two 3x3 convolutions with batch normalization [35] and activations with the rectified linear unit (ReLU) [36]. The sequence of convolution, batch normalization and ReLU is

called convolutional block (conv block), where each block is followed by a max-pooling operation with size 2x2 and stride of 2. At each downsampling step, the number of feature maps is doubled, and this process is repeated four times.

The bottom layer consists of a conv block linking the down-sampling to the up-sampling path. In this work, transposed convolution was implemented in the up-sampling path instead of up-convolution (used in the original U-Net implementation), as suggested in the literature [37]. A series of 4x4 transposed convolutions with stride 2 and conv block are repeated four times, similar to the down-sampling path. A final 1x1 convolution layer with softmax activation function, that maps the features maps to 4 classes, is used to generate the final segmentation maps (one for each target structure). The resulting network has a different number of filters in each conv block, from 48 (first layer) to 768 (bottom layer).

In this work, the original skip connections of U-Net were modified incorporating dense blocks [38] to reduce the semantic gap between the up- and down-sampling path [22,23]. Specifically, each layer of the dense block consisted of 3x3 convolutions with ReLU and batch normalization. Each layer concatenated its own feature maps with the feature maps of all the preceding layers, and transferred its own feature maps to all the following ones, thus adding and preserving information. The number of filters in each convolutional layer in a dense block was the same as in the corresponding convolutional layer in the down-sampling path. The number of layers used in each dense block were 4, 3, 2, and 1, respectively, along the four skip connections.

A further modification compared to the original U-Net was the reduction of the number of filters in the transposed convolutional layer, with a number of feature maps equal to the number of classes. As reported in [19], the full complexity of the up-sampling path is not necessary, with benefits in terms of reduction of learnable parameters. A schematic illustration of the proposed architecture is given in Figure 1.

To evaluate the accuracy and computations trade-off of the proposed FCN, an ablation study with two experiments was carried out: in the former, the deeper layer of the network was removed, while in the

latter the number of filters in each conv block was changed, from 32 in the first layer to 512 in the bottom layer.

2.4 Parameter setting and data augmentation

The weights of each convolution were initialized following the standard *He normal* initialization [39]. To deal with the class imbalance between the background and the foreground pixels, weighted cross-entropy (WCE) was adopted as loss function [20]: WCE measures the pixel-wise error probability between the predicted output and target class and accumulates summing the errors across all the pixels. The WCE is defined as:

$$WCE = -\sum_{x} w_{t(x)} \log(p(x, t(x))) \quad (1)$$

where w is the weight map to balance the class frequencies, and $p(x, t(x))$ is the probability prediction for each class at each pixel x after the softmax function. The targets class label at location x is denoted by $t(x)$. Class weights were set to be inversely proportional to their frequencies in the training data and normalized, so that for the background class a weight of 0.076 was used, whereas for all the classes of the foreground a weight of 0.308 was applied.

The FCN was trained for 1000 epochs with Adaptive Momentum Estimation (ADAM) optimizer [40], using an initial learning rate of 0.01 with a decay of 2% after every five epochs. The hyperparameters β1 and β2, used to control the exponential decay rates of the gradient, were set to 0.9 and 0.999, respectively. Online data augmentation [41] was implemented using gamma correction, blurring, and random rotation in [-15°, 15°] range. Data augmentation was used to allow the network to better generalize and prevent overfitting. After each epoch, the network performance was evaluated on the validation set. The best model among epochs was selected as the one with the highest Sørensen-Dice coefficient ($Dice$) score on the validation set. The $Dice$ calculates the overlap between foreground regions, and is defined as:

$$Dice = \frac{2 \sum p_{h,w} t_{h,w}}{\sum p_{h,w} + \sum t_{h,w}} \quad (2)$$

where h is the number of classes of the foreground, w is the number of pixels, t is the ground truth label per location and p is prediction of the sigmoid function.

A postprocessing step was also applied to all the trained networks, to obtain more accurate segmentations: in those cases, in which the FCN segmentation resulted in different regions with the same label, only the region with the largest area was preserved, while the remaining regions were classified as background.

All the analyses were performed using Tensorflow on a NVIDIA® GeForce® GTX 1080 (8 GB GDDR5 dedicated) on an Intel® Core™ i7-8700 CPU (3.20 GHz) computer with 16 GB RAM.

Codes are made publicly available at https://github.com/Monzino/Ventricle_segmentation.

2.5 Experimental protocol and statistical analysis

The datasets in DB1 were randomly split at patient level into training (70%), validation (15%) and testing sets (15%), preserving the proportion of patients for each pathology. The performance of the RV endocardial and LV endo- and epicardial segmentation by FCN was measured using sensitivity (*Se*), *Dice* and the Hausdorff distance (*HD*) [27] compared to the ground-truth mask. Results obtained using the standard and modified skip connections were compared (paired t-test for normally distributed variables, and Wilcoxon signed rank test for non-normally distributed variables).

To verify the influence of the basal slice definition criteria on the FCN performance, the analysis was performed training the proposed U-Net using: A) the basal slice as defined in agreement with [32], thus also considering those slices where the myocardium surrounded the LV for at least 50%; B) only the slices where the myocardium completely surrounded the LV, thus excluding some challenging slices.

To consider a practical use case scenario in which the cardiologist uses the proposed segmentation algorithm, the prediction on the test set was reviewed by an independent expert physician, by showing the resulting LV and RV contours overlapped to the original image, and allowing manual correction when needed, using a custom made software implemented in Matlab. The impact of these corrections was computed by comparing the corrected segmentations to the original automated results in respect to the ground-truth, both in term of *Se, Dice* and *HD*. In addition, also a comparison in terms of actual LV and RV

volumes (in ml) and LV mass (in g), compared to the gold standard values by correlation and Bland-Altman analyses, was performed.

To measure the inter-observer variability between the corrected segmentations and the ground truth, for each parameter a coefficient of variation between these two measurements was computed as CV(%)= 100*SD/MEAN and compared with the same result obtained by the FCN contours.

To evaluate the possible dependence of segmentation performance from the morphological characteristics associated to the different pathologies, results in terms of *Dice* and *HD* were grouped according to HCM, DCM and VA , and one-way ANOVA or Kruskal-Wallis test with Bonferroni correction was applied.

To assess the generalization power of the proposed approach on different datasets, the FCN trained on DB1 with criterion A was applied to the datasets in DB2, and compared with the performance obtained on the testing set of DB1 by Mann–Whitney U test.

Normally distributed data were reported as mean ± SD, otherwise as median (25^{th}-75^{th} percentile). Statistical analyses were conducted with IBM SPSS 25 (SPSS Inc, Chicago, IL), with p values < 0.05 to be considered as significant.

3. Results

From the ablation analysis, the proposed architecture consistently outperformed a shallower network.

Specifically, for LV, RV and Myo, respectively, the shallower network achieved a mean Dice of 0.789, 0.687 and 0.462 versus 0.944, 0.908 and 0.851 of the proposed FCN. Even compared to the network with a reduced number of filters, the proposed FCN reached higher performance for both LV (0.944 vs 0.905), RV (0.908 vs 0.858) and Myo (0.851 vs 0.818).

	Left Ventricle (ED)			Left Ventricle (ES)		
	Se	Dice	HD [mm]	Se	Dice	HD [mm]
U-Net	0.967 (0.020)	0.961 (0.009)	6.4 (4.4-9.1)	0.929 (0.056)	0.927 (0.034)	8.8 (4.4-11.9)
U-Net MOD	0.966 (0.021)	0.961 (0.010)	6.0 (4.0-8.3)	0.931 (0.052)	0.927 (0.036)	8.4 (4.8-12.0)
	Right Ventricle (ED)			**Right Ventricle (ES)**		
	Se	Dice	HD [mm]	Se	Dice	HD [mm]
U-Net	0.910 (0.039)	0.932 (0.022)	8.5 (4.9-10.0)	0.836 (0.076)	0.871 (0.054)	8.9 (5.8-12.0)
U-Net MOD	0.920 (0.033)*	0.938 (0.020)*	6.6 (5.0-9.3)	0.842 (0.071)*	0.878 (0.048)*	8.1 (6.0-12.9)
	Myocardium (ED)			**Myocardium (ES)**		
	Se	Dice	HD [mm]	Se	Dice	HD [mm]
U-Net	0.801 (0.082)	0.832 (0.038)	5.9 (4.2-7.2)	0.840 (0.083)	0.867 (0.043)	6.1 (4.1-8.8)
U-Net MOD	0.808 (0.078)*	0.833 (0.035)	5.9 (4.0-7.2)	0.849 (0.080)*	0.870 (0.040)	5.9 (4.4-8.5)

Table 2. Segmentation performance of the proposed U-Net (U-net MOD) compared to the U-Net with standard skip connections (U-Net), at end-diastolic (ED) and end-systolic (ES) phases in DB1, measured as Sensitivity (Se), Dice coefficient, and Hausdorff distance (HD). Se and Dice are reported as mean (SD), while HD is reported as median (25^{th}-75^{th} percentile). *: $p < 0.05$, U-Net vs U-Net MOD (paired t-test or Wilcoxon rank sum).

Table 2 compares the performances obtained with the proposed U-Net versus the same architecture without filters along the skip connections, including in the analysis slices in which the criterion A (i.e., >=50% of myocardium present around the LV) was verified: the implemented skip connections outperformed the standard U-net in particular for the RV, where significant differences in *Se* and *Dice*, in both ED and ES, were highlighted. Also, for the Myo, the redesigned connections performed better, with a significant improvement in *Se* for both ED and ES, while no differences were observed for the LV. Based on these results, and on the improved performance obtained using the redesigned skip connections, further analyses were conducted using this implementation only.

Table 3 reports the performance, in terms of *Se*, *Dice* and *HD*, of the FCN segmentation for LV and RV endocardium and LV myocardium in the ED and ES frames for DB1, both when criteria A or B were applied to define the most basal slice. Comparing the performance between these two criteria (i.e. A and B), results considering cumulatively all the slices showed no significant differences. As regards the performance in different cardiac phases, the model achieved higher Dice and Se in ED for both LV and RV, while for the Myo a significantly higher Dice and Se were obtained in ES.

A. Basal slices with >=50% myocardium around the LV					
Left Ventricle (ED)			**Left Ventricle (ES)**		
Se	Dice	HD [mm]	Se	Dice	HD [mm]
0.966 (0.021)¥	0.961 (0.010)¥	6.0 (4.0-8.3)	0.931 (0.052)	0.927 (0.036)	8.4 (4.8-12.0)
Right Ventricle (ED)			**Right Ventricle (ES)**		
Se	Dice	HD [mm]	Se	Dice	HD [mm]
0.920 (0.033)¥	0.938 (0.020)¥	6.6 (5.0-9.3)	0.842 (0.071)	0.878 (0.048)	8.1 (6.0-12.9)
Myocardium (ED)			**Myocardium (ES)**		
Se	Dice	HD [mm]	Se	Dice	HD [mm]
0.808 (0.078)¥	0.833 (0.035)¥	5.9 (4.0-7.2)	0.849 (0.080)	0.870 (0.040)	5.9 (4.4-8.5)
B. Only basal slices with 100% myocardium around the LV					
Left Ventricle (ED)			**Left Ventricle (ES)**		
Se	Dice	HD [mm]	Se	Dice	HD [mm]
0.967 (0.020)¥	0.962 (0.010)¥	5.9 (3.9-8.3)	0.933 (0.050)	0.930 (0.035)	6.5 (3.6-11.9)
Right Ventricle (ED)			**Right Ventricle (ES)**		
Se	Dice	HD [mm]	Se	Dice	HD [mm]
0.920 (0.033)¥	0.939 (0.020)¥	6.6 (5.0-9.3)	0.844 (0.069)	0.879 (0.046)	7.7 (6.0-12.9)
Myocardium (ED)			**Myocardium (ES)**		
Se	Dice	HD [mm]	Se	Dice	HD[mm]
0.809 (0.079)¥	0.834 (0.036)¥	5.8 (4.0-7.0)	0.853 (0.078)	0.873 (0.038)	5.0 (4.0-6.7)

Table 3. Segmentation performance of the proposed FCN computed cumulatively considering all slices at end-diastolic (ED) and end-systolic (ES) phases in DB1, measured as Sensitivity (Se), Dice coefficient, and Hausdorff distance (HD). Criterion A refers to the FCN trained and tested considering also the slices at LV basal level in which the myocardium was present for at least 50% around the LV cavity, while criterion B refers to the FCN trained and tested considering only the slices at LV basal level in which the myocardium was visible all around the cavity. Se and Dice are reported as mean (SD), while HD is reported as median (25^{th}-75^{th} percentile). *: $p<0.05$, A vs B (unpaired t-test or Mann–Whitney U test); ¥: $p<0.05$, ED vs ES (unpaired t-test or Mann–Whitney U test)

In Figure 2, some examples of the segmented contours at the most basal level, defined by criteria A and B previously described, are shown, paired with the corresponding gold standard manual tracings. In left columns, when the LV myocardium did not completely surround the LV, due to the presence of the aortic root, the automated segmentation sometimes was failing by including it in the LV. For the RV, when a portion of the right atrium was present, sometimes it was included in the RV. In the right columns, where these more problematic slices were excluded from both training and testing, as the morphology of the LV become more similar to the other slices and the RV was more visible and recognizable, these differences with the GS were still present but reduced. Despite these potential errors, cumulative performance computed considering all slices resulted not affected by applying criteria A or B.

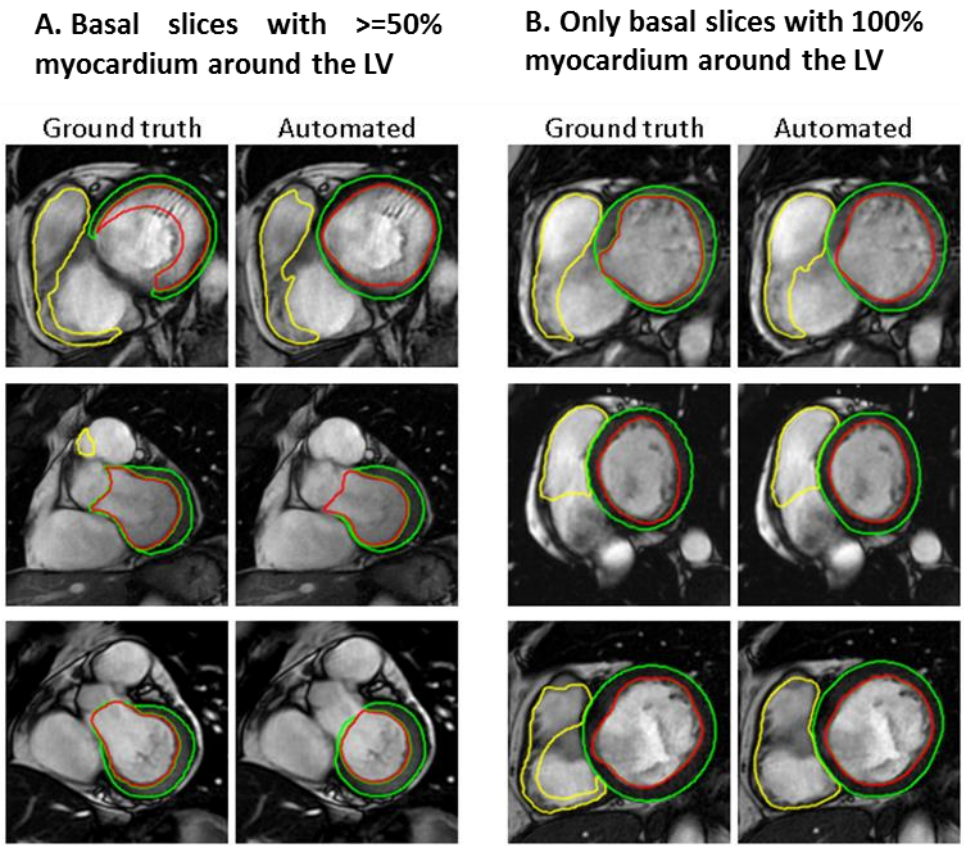


Fig. 2. Example of segmentation results of basal slices obtained using the proposed FCN compared to the manually traced gold standard (RV endocardium: yellow; LV endocardium: red; LV epicardium: green), by considering in the definition of the basal slice criterion A (presence of myocardium around the LV >=50% - left column) or criterion B (only the slices where the myocardium completely surrounds the LV - right column).

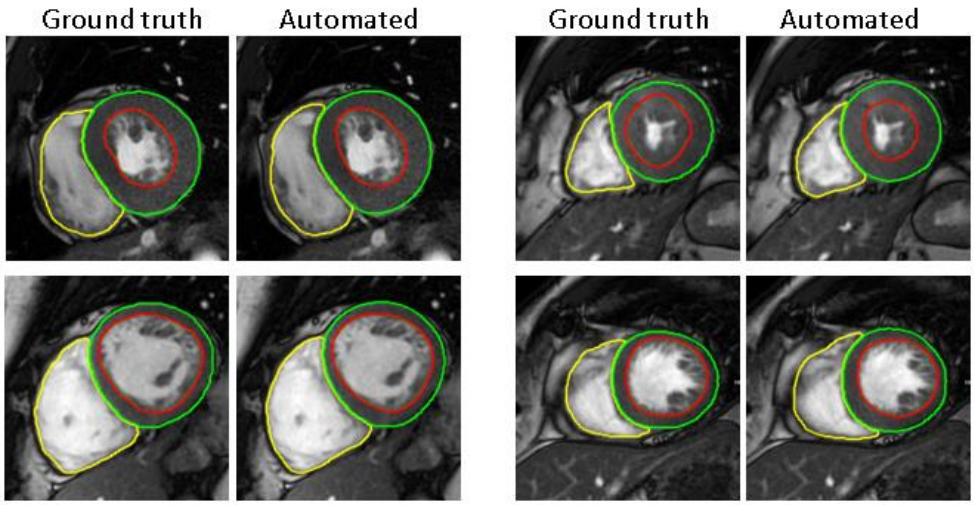


Fig. 3. Example of segmentation results of mid-ventricular level obtained using the proposed FCN compared to the manually traced gold standard (RV endocardium: yellow; LV endocardium: red; LV epicardium: green).

Figure 3 and Figure 4 show examples of the segmented contours at the LV mid and apical levels, respectively. It is possible to appreciate how the automated contours at mid level were able to deal with the presence of papillary muscles, by including them into the LV as traced by the gold standard, and at apical level, despite the small size of the LV (and RV when present), how the FCN was able to detect the correct contours.

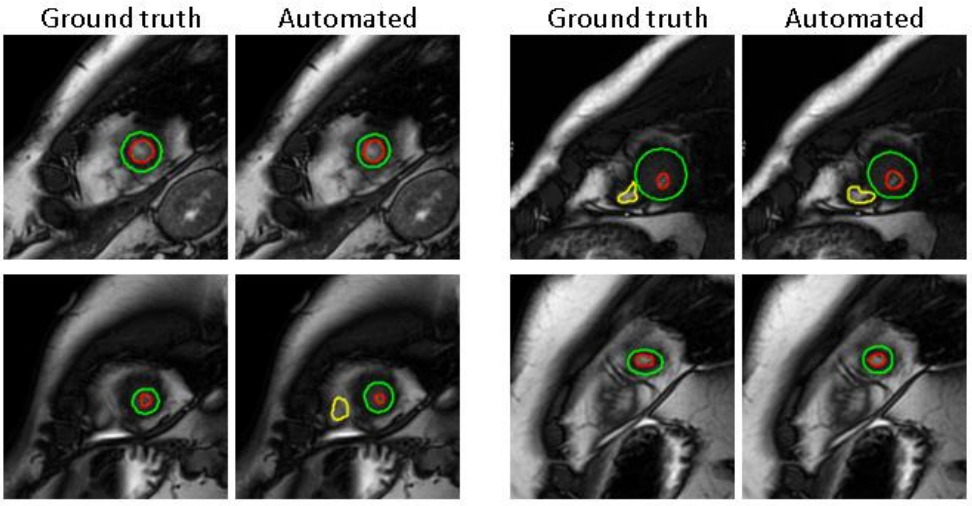


Fig. 4. Example of segmentation results of apical slices obtained using the proposed FCN compared to the manually traced gold standard (RV endocardium: yellow; LV endocardium: red; LV epicardium: green).

A. Basal (Myo mass around the LV more than 50% but less than 100%)						
	LV			Myo		
	Se	Dice	HD [mm]	Se	Dice	HD [mm]
FCN+corr.	0.958 (0.831-0.983)#	0.927 (0.815-0.949)#	7.5 (5.2-12.2)#	0.768 (0.614-0.852)#	0.811 (0.690-0.847)#	13.5 (9.3-20.8)#
FCN	0.952 (0.832-0.981)	0.909 (0.829-0.951)#	9.7 (5.3-16.2)#	0.768 (0.612-0.836)#	0.808 (0.682-0.838)#	13.0 (9.3-18.3)#
B. Basal (Myo mass surrounds completely the LV)						
	LV			Myo		
	Se	Dice	HD [mm]	Se	Dice	HD [mm]
FCN+corr.	0.979 (0.953-0.993)*	0.958 (0.949-0.969)	4.7 (2.9-7.1)	0.849 (0.753-0.905)*	0.874 (0.823-0.902)	4.5 (3.6-5.6)
FCN	0.976 (0.945-0.990)	0.959 (0.949-0.970)	5.0 (2.8-7.5)	0.851 (0.756-0.907)	0.873 (0.812-0.900)	4.5 (3.6-5.9)
Apical						
	LV			Myo		
	Se	Dice	HD [mm]	Se	Dice	HD [mm]
FCN+corr.	0.890 (0.658-0.976)*	0.844 (0.715-0.907)	4.1 (2.8-6.2)	0.707 (0.455-0.849)	0.694 (0.557-0.815)	5.7 (3.6-6.9)
FCN	0.862 (0.592-0.963)	0.867 (0.709-0.900)	3.6 (2.4-5.2)	0.699 (0.435-0.855)	0.682 (0.532-0.811)	5.2 (3.6-6.7)

Table 4. Segmentation performance of the proposed FCN specifically at the basal (criteria A and B) and apical level, before (FCN) and after manual correction (FCN+corr) considering both end-diastole (ED) and end-systole (ES) , measured as Sensitivity (Se), Dice coefficient, and Hausdorff distance (HD) reported as median (25th-75th percentile),.

*: $p < 0.05$, FCN vs FCN+corr (Wilcoxon rank sum); #: $p<0.05$, A vs B (Mann–Whitney U test)

In Table 4, the accuracy of the automated segmentation specifically for the basal (defined as based on A or B criteria) and apical levels, together with the results following-manual correction of the automated segmentation by an expert physician, are shown in term of Se, Dice and HD with respect to the ground-truth. At the LV apex, the improvement relevant to manual correction was visible only for Se, but not for the Myo, for both ED and ES. At the basal region, when criterion B (i.e., with 100% of Myo around the LV) was applied, the effect of manually adjusted segmentation resulted in a higher Se for the LV for both ED and ES, but in slightly lower Se for the Myo ($p < 0.05$) in ES, with respect to the gold standard. When criterion A (i.e., >50% myocardium around the LV) was applied, respective performance in terms of Se, Dice and HD decreased (as expected) in both automated and corrected contours, where manual correction unexpectedly did not result in significant improvement in respect to automated results, thus highlighting a higher uncertainty and inter-observer variability in tracing the contours at this level.

In Table 5 the results of linear correlation (r^2) and Bland-Altman (Bias ± 2σ) analyses for the automated and manually adjusted segmentation, compared to the gold standard in terms of clinical parameters, are reported. For LV ED and ES volumes, as well as for RV ED, the FCN resulted in a strong correlation with no bias and narrow limits of agreement, both with and without correction, while a slight overestimation was measured for RV ES compared to the gold standard. For Myo mass, still a very good correlation was reported, with a slight overestimation ($p<0.05$) of both the FCN and the following correction compared to the gold standard, and a wider (but still acceptable) limits of agreement.

In addition, the CV% for each parameter is reported: it is worth noting that, considering the result of corrected contours by the expert observer, values of CV% can be interpreted as interobserver variability (between this observer and the gold standard). The CV% of the FCN contours compared to the gold standard showed values comparable to those of the expert observer, thus mimicking the performance in terms of interobserver variability of a human expert.

	LV_{EDV}			LV_{ESV}		
	r^2	Bias ± 2σ [ml]	CV (%)	r^2	Bias ± 2σ [ml]	CV (%)
FCN+corr	0.99	-2 ± 13	2.40	0.98	-2 ± 14	5.34
FCN	0.99	-1 ± 12	2.22	0.98	-1 ± 14	4.98
	RV_{EDV}			RV_{ESV}		
	r^2	Bias ± σ [ml]	CV (%)	r^2	Bias ± 2σ [ml]	CV (%)
FCN+corr	0.98	4 ± 14	3.03	0.96	7 ± 12*	8.02
FCN	0.99	6 ± 12	3.26	0.95	6 ± 14*	7.65
	$Myo_{ED\ Mass}$			$Myo_{ES\ Mass}$		
	r^2	Bias ± σ [g]	CV (%)	r^2	Bias ± 2σ [g]	CV (%)
FCN+corr	0.92	7 ± 26*	9.44	0.91	6 ± 28*	8.27
FCN	0.92	6 ± 26*	9.13	0.92	5 ± 26*	8.07

Table 5. Clinical metrics before (FCN) and after manual correction (FCN+corr) reported in terms of correlation (r^2), bias ± 2*standard deviation (Bias ± 2σ) and Coefficient of variation (CV), separately for LV, RV and Myo, on basal slices defined as criterion A (>=50% myocardium around the LV). EDV: End diastolic volume; ESV: End systolic volume. *: p<0.05 vs null (paired t-test)

In Figure 5, the boxplots for Dice and HD of each segmented structure in each patient group (i.e. HCM, DCM, and VA) are reported for comparison. It is possible to notice that the FCN performed similarly in all groups, except for the LV at ES phase in HCM patients, where a significant lower Dice was reported.

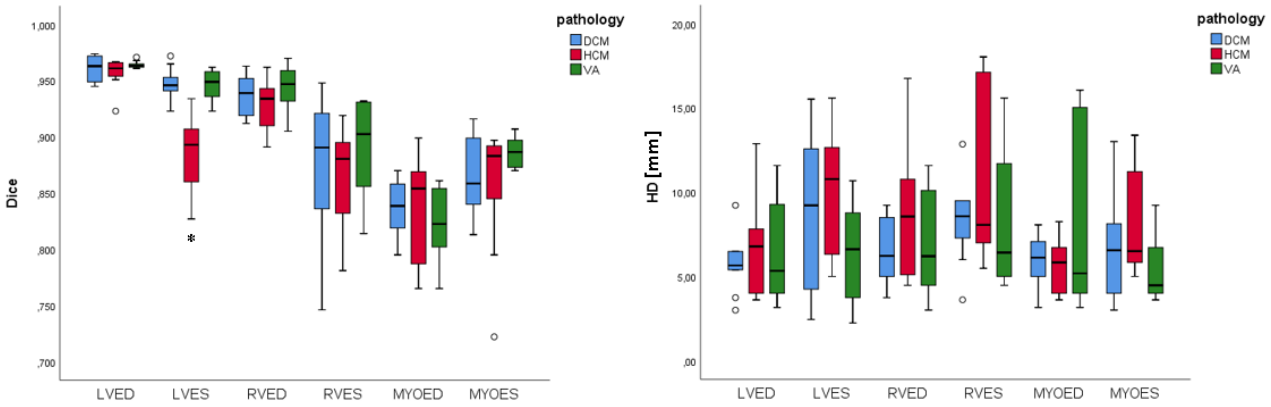


Fig. 5. Boxplots comparing Dice index and Hausdorff distance (HD) of each segmented cardiac structure at end-diastole (ED) and end-systole (ES), grouped by patient's pathology (DCM: dilated cardiomyopathy; HCM: hypertrophic cardiomyopathy; VA: ventricular arrhythmias). LV: left ventricular, RV: right ventricular; MYO: myocardium; ED: end-diastolic; ES: end-systolic. *: p<0.05 vs HCM (ANOVA or Kruskal-Wallis test with Bonferroni correction)

As regards generalizability of the results, Table 6 reports the cumulative performance for all slices of the segmentation for LV and RV endocardium and LV myocardium in the ED and ES frames for DB2, when the proposed FCN, trained on images of DB1 with criteria A to define the most basal slice, was applied.

Compared to the corresponding results obtained in DB1 (see Table 3), the obtained performance appears very similar, with a slight worsening in Se for LV both at ED and ES, in Dice for RV at ED, but a slight improvement in Se for the myocardium both at ED and ES.

Left Ventricle (ED)			**Left Ventricle (ES)**		
Se	Dice	HD [mm]	Se	Dice	HD [mm]
0.935 (0.011)*¥	0.957 (0.005)¥	7.8 (5.7-8.6)¥	0.893 (0.019)*	0.922 (0.011)	12.6 (9.1-13.3)*
Right Ventricle (ED)			**Right Ventricle (ES)**		
Se	Dice	HD [mm]	Se	Dice	HD [mm]
0.898 (0.042)¥	0.901 (0.034)*¥	6.4 (4.8-7.9)	0.838 (0.052)	0.858 (0.036)	6.6 (5.2-8.1)
Myocardium (ED)			**Myocardium (ES)**		
Se	Dice	HD [mm]	Se	Dice	HD[mm]
0.884 (0.028)*¥	0.834 (0.025)¥	4.2 (3.7-5.9)¥	0.929 (0.025)*	0.878 (0.013)	5.6 (4.3-17.3)

Table 6. Segmentation performance of the proposed FCN applied on DB2 at end-diastolic (ED) and end-systolic (ES) phases, measured as Sensitivity (Se), Dice coefficient, and Hausdorff distance (HD). Se and Dice are reported as mean (SD), while HD is reported as median (25^{th}-75^{th} percentile). *: $p<0.05$, DT2 vs DT1 (unpaired t-test or Mann–Whitney U test); ¥: $p<0.05$, ED vs ES (unpaired t-test or Mann–Whitney U test).

When the performance of the automated segmentation of DB2 specifically for the basal slice defined as criterion A is compared with that obtained for DB1, the only statistically significant difference was an improvement in LV HD, that decreased from 6.0 (4-8.3) to 5.0 (3.9-8.5).

4. Discussion and Conclusion

In this paper, a fully automated method based on FCN for accurate multi-structure segmentation on CMR, capable of reliably producing high quality LV and RV contour tracings, was proposed. To this aim, we used a FCN derived from the consolidated U-Net architecture as a common reference network [20], opportunely modified with a new scheme for skip connections. Especially in medical image, where precise segmentation is fundamental for accurate diagnosis, skip connections showed to be effective in recovering fine details lost in down-sampling path [20,42]. Skip connections between the corresponding layers of down-sampling and up-sampling path preserve the spatial information lost during the pooling operation and alleviate the vanishing-gradient problem [43]. In our specific implementation, compared to the traditional U-net, where high resolution feature maps are directly propagated from the down-sampling to the up-sampling path possibly degrading the segmentation, we took advantage from dense skip connections, that alleviate the

semantic gap generated when higher level features are merged with lower level features, with benefits in term of both network optimization [23,44] and consequently on its segmentation performance. Similar to ResNets [45], where identity shortcut connections are introduced, dense block tends to maximize the information flow, because each layer has a direct access to the other and to the input feature maps, with benefits in term of gradients propagation, and vanishing problem [37,44]. Differently from ResNets, in our implementation of dense blocks, the output feature maps were concatenated with the incoming ones and, from the down-sampling to the up-sampling path corresponding layer, the feature maps were propagated through Dense block. This had a direct consequence on the reduction in the number of parameters to learn, exploiting the potential of network through a feature reuse. This feature propagation improved the training phase, leading an implicit deep supervision, because each layer received an additional supervision from the loss function, which made the training phase easier.

An ablation study was performed to investigate the performance of the network at different depths and number of filters. A shallower architecture showed a drastic reduction in performance due to the limited capacity of the network to catch high-level features, as well as decreasing the number of filters which limited the number of extracted feature maps.

To validate the performance of the proposed FCN in line with current CMR guidelines [32], all the slices from the base to the apex of the heart where included. In addition, the analysis was repeated to evaluate the effect of the applied inclusion criteria for the LV basal slice, which represents the most variable and difficult image to be traced. Furthermore, CMR datasets from three different pathological groups were included, to avoid polarization of results for a very specific group of patients.

CMR dataset							
Name	Training Samples	Testing samples	LV	RV	Myo	Pathology	Basal slice [32]
Our Dataset	180	30	✓	✓	✓	✓	✓
TWINS-UK [16]	58	10	✓	–	–	–	–
Sunnybrook [15]	45		✓	–	–	✓	–
Preterm [15]	214	20	✓	–	–	✓	–
York [29]	33		✓	–	–	✓	–
STACOM [30]	100	100	✓	–	✓	–	–
MICCAI RV [31]	16	32	–	✓	–	–	–
KAGGLE [27]	500	200	–	–	–	–	–
ACDC [27]	100	50	✓	✓	✓	✓	–

Table 7. Summary of available CMR short-axis cine image datasets and relevant use for training and testing FCN-basedd methods made available in different challenges. LV/RV/Myo: ground truth segmentation available; Pathology: different clinical conditions included; Basal slice: basal slice defined in agreement with guidelines [32].

Table 7 shows a summary of the available CMR datasets based on SAX cine images used for validating FCN-based methods. Savioli et al [16] recruited 68 female subjects from the TWINS-UK cohort for LV segmentation, with each acquisition covering both ventricles from the atrioventricular ring to the apex. Poudel et al [15] utilized two datasets, the Sunnybrook Cardiac Data, made available during the MICCAI 2009 challenge on automated LV segmentation, consisting 45 patients with different pathologies (heart failure, hypertrophy, and heathy subjects), and the Preterm dataset, including 234 subjects with and without premature birth, for LV segmentation only. Nasr-Esfahani [29] focused on LV segmentation using a publicly available dataset of the York University including 33 patients of different pathologies such as cardiomyopathy, aortic regurgitation, enlarged ventricles, and ischemia. In 2011, as part of the STACOM challenge on automated LV myocardium segmentation, another dataset consisting of 200 patients was made available [30]. To focus on RV endocardium and epicardium segmentation, in 2012 MICCAI challenge provided a dataset with 48 cases [31]. In 2015, the KAGGLE challenge organized a competition with the goal to automatically measure LV ED and ES volumes, but no ground truth was provided [27]. In 2017, the MICCAI challenge provided the ACDC dataset for LV and RV segmentation with 150 patients of different pathologies, specifically myocardium infarction, DCM, HCM, abnormal RV and healthy subjects, but basal

slices according to CMR guidelines [32] and apical slices were not available. In comparison with these images, our dataset allows a more realistic evaluation of the proposed FCN for cardiac ventricular analysis in clinical practice, by considering both LV and RV, different pathologies, and including basal and apical slices according to CMR guidelines [32].

A joint analysis of Table 3 and 4 shows that the effect of the basal slice inclusion criteria on the outcome of the automated segmentation becomes visible only if the performance is reported separately for the basal slices, while it does not impact cumulative results where all slices are considered, due to the larger number of not basal slices. This fact suggests that, to allow better comparison of performances in CMR segmentation algorithms, the criteria of definition of basal slice should be always reported, and the corresponding results should be separately given, thus avoiding reporting only cumulative results. In fact, as expected, the automated segmentation of the LV and RV contours in the basal slice is arduous, due to the presence of the pulmonary valve, aortic valve, atrium and outflow tract, thus resulting in higher inter-observer variability, thus impacting on the segmentation accuracy. For example, including or not the portion of outflow tract in the manual segmentation is strictly dependent on the physician's decision on that particular patient, with the support of visualizing also the long axis views, not available to the FCN [46]. This variability generates instability during the training step, with the network continuously learning to adapt to different patterns. Despite these limitations, the performance of the proposed FCN could be considered still good also in those slices, in particular for LV and RV endocardium.

As regards the apical slices, the model was almost always able to localize the LV and RV, despite their small size, but with accuracy lower than that at basal level. Compared to [6], which suggests to focus on apical slices as a possible way to improve segmentation accuracy, our opinion is that also the algorithm performance should be separately reported for this slice, but more attention should be directed towards the basal contours, due to their major impact on clinical volumetric parameters due to the relevant chamber size, and to the higher variability in the cardiac structures.

The performance for the RV was slightly lower than for the LV, which it is also highlighted by the reduced correlation and greater bias (Table 5). The segmentation of the RV is more challenging, as demonstrated by

the lower Dice scores and higher HD. This is probably due to the more complicated geometry, especially at ES phase, when at maximum contraction the papillary muscles and trabeculations are more compacted. In addition, a higher variability from the basal to the apical slice, and a lower contrast with the surrounding structures, makes RV contour detection more complicated than the LV [46], limiting the accuracy of the segmentation process. It is important noticing that trabeculations and papillary muscles were included in the cavity, also in agreement with the current guidelines. Even the Myo was well segmented, even if with more difficulties when compared to the LV and RV, probably related to the detection of two contours (i.e. endocardium and epicardium) instead than a single one [27].

The analysis of possible challenges in the segmentation process relevant to a specific pathological condition did not show a systematic bias in performance, thus demonstrating the generalizability and accuracy of the proposed FCN algorithm.

The adjusted segmentation of an external physician returned the fidelity of the automated contours, demonstrating the possibilities and the potential offered by the proposed FCN in term of ventricular segmentations. As mentioned in [47], Neural Networks are emerging as gold standard in many medical fields outperforming expert physicians, but the object of the research would be to understand if these models together with a physician are better than the same physician alone. At this purpose our approach, comparing automated and corrected segmentation, showed a very similar performance, and inter-observer variability when compared to the gold standard, with the advantage of drastically reducing the manual segmentation time by only reviewing the detected boundaries instead of tracing them.

The analysis of the generalizability of the results, when the FCN was applied to a different database (DB2) obtained with a 3T scanner and different acquisition settings, showed very similar performances despite a slight worsening in Se for the LV, that might be due to the different intensity distributions and the lower resolution of these images, not included in the training phase, and even a slight improvement in HD when only basal slices were considered.

One possible limitation of our study is that the gold standard was not based on independent multiple tracings but on a single expert annotation, and only in cases with uncertainty in the image interpretation, the contours were reviewed through group consensus. In addition, the model was trained only on ED and ES phases, because manual contours were available for these two phases only; however, we expect the proposed FCN to be able to perform well even on the other frames in the cardiac cycle.

This work may be extended in different ways, by evaluating the performance of the proposed FCN model on other CMR datasets covering different pathologies. Furthermore, spatio-temporal features [48] could be also evaluated, with possible improvement in segmentation accuracy.

In conclusion, a deep learning approach for semantic segmentation on both right and left ventricle in CMR images with Dense skip connection was proposed, demonstrating its utility and potential in clinical practice, reaching high correlations with ground truth and accurate ventricular volumes estimation, and good generalizability of the obtained performances. The additional manual correction to the automatically obtained contours revealed how our supervised learning-based method was able to manage the real-world variability of functional and anatomical characteristics across different pathological conditions, with results in the range of expert inter-observer variability. The importance of clearly defining the inclusion criteria of basal slices in the testing dataset was underlined, as well as the need to separately report the algorithm performances in the most difficult slices (i.e., basal and apical). Such results reveal the potential of deep learning approach for fast, accurate and reliable automated cardiac segmentation in real clinical applications.

ACKNOWLEDGEMENTS

None.

CONFLICT OF INTEREST

There are no conflicts of interest that could inappropriately influence this research work.

ETHICAL APPROVAL

This study was approved by the Ethics Committee of the "Centro Cardiologico Monzino IRCCS", Milan, Italy.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

AUTHORSHIP RESPONSIBILITY

- The material in this manuscript is original and contains no matter libelous or otherwise unlawful.
- The manuscript represents valid work and that neither this manuscript nor any other with substantially similar content under my authorship has been published or is being considered for publication elsewhere.
- All authors have participated sufficiently in the work to take public responsibility for all its content.

REFERENCES

[1] Nichols, M., Townsend, N., Scarborough, P. and Rayner, M., 2014. Cardiovascular disease in Europe 2014: epidemiological update. European heart journal, 35(42), pp.2950-2959. https://doi.org/10.1093/eurheartj/ehu299.

[2] Kang, D., Woo, J., Kuo, C.J., Slomka, P.J., Dey, D. and Germano, G., 2012. Heart chambers and whole heart segmentation techniques. Journal of Electronic Imaging, 21(1), p.010901. https://doi.org/10.1117/1.JEI.21.1.010901.

[3] Suinesiaputra, A., Cowan, B.R., Al-Agamy, A.O., Elattar, M.A., Ayache, N., Fahmy, A.S., Khalifa, A.M., Medrano-Gracia, P., Jolly, M.P., Kadish, A.H. and Lee, D.C., 2014. A collaborative resource to build consensus for automated left ventricular segmentation of cardiac MR images. Medical image analysis, 18(1), pp.50-62. https://doi.org/10.1016/j.media.2013.09.001.

[4] Tavakoli, V. and Amini, A.A., 2013. A survey of shaped-based registration and segmentation techniques for cardiac images. Computer Vision and Image Understanding, 117(9), pp.966-989. https://doi.org/10.1016/j.cviu.2012.11.017.

[5] Peng, P., Lekadir, K., Gooya, A., Shao, L., Petersen, S.E. and Frangi, A.F., 2016. A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging. Magnetic Resonance Materials in Physics, Biology and Medicine, 29(2), pp.155-195. https://doi.org/10.1007/s10334-015-0521-4.

[6] Petitjean, C. and Dacher, J.N., 2011. A review of segmentation methods in short axis cardiac MR images. Medical image analysis, 15(2), pp.169-184. https://doi.org/10.1016/j.media.2010.12.004.

[7] Bai, W., Shi, W., Ledig, C. and Rueckert, D., 2015. Multi-atlas segmentation with augmented features for cardiac MR images. Medical image analysis, 19(1), pp.98-109. https://doi.org/10.1016/j.media.2014.09.005.

[8] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B. and Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical image analysis, 42, pp.60-88. https://doi.org/10.1016/j.media.2017.07.005.

[9] Li, Z., Dong, M., Wen, S., Hu, X., Zhou, P. and Zeng, Z., 2019. CLU-CNNs: Object detection for medical images. Neurocomputing, 350, pp.53-59. https://doi.org/10.1016/j.neucom.2019.04.028

[10] Pereira, S., Pinto, A., Alves, V. and Silva, C.A., 2016. Brain tumor segmentation using convolutional neural networks in MRI images. IEEE transactions on medical imaging, 35(5), pp.1240-1251. https://doi.org/10.1109/TMI.2016.2538465.

[11] Yadav, S.S. and Jadhav, S.M., 2019. Deep convolutional neural network based medical image classification for disease diagnosis. Journal of Big Data, 6(1), p.113. https://doi.org/10.1186/s40537-019-0276-2.

[12] Colleoni, E., Moccia, S., Du, X., De Momi, E. and Stoyanov, D., 2019. Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers. IEEE Robotics and Automation Letters, 4(3), pp.2714-2721. . https://doi.org/10.1109/LRA.2019.2917163.

[13] Moccia, S., Banali, R., Martini, C., Muscogiuri, G., Pontone, G., Pepi, M. and Caiani, E.G., 2019. Development and testing of a deep learning-based strategy for scar segmentation on CMR-LGE images. Magnetic Resonance Materials in Physics, Biology and Medicine, 32(2), pp.187-195. https://doi.org/10.1007/s10334-018-0718-4.

[14] Isensee, F., Jaeger, P.F., Full, P.M., Wolf, I., Engelhardt, S. and Maier-Hein, K.H., 2017. Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features. In International workshop on statistical atlases and computational models of the heart , pp. 120-129. https://doi.org/10.1007/978-3-319-75541-0_13.

[15] Poudel, R.P., Lamata, P. and Montana, G., 2016. Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation. In Reconstruction, segmentation, and analysis of medical images, pp.83-94. https://doi.org/10.1007/978-3-319-52280-7_8.

[16] Savioli, N., Vieira, M.S., Lamata, P. and Montana, G., 2018, October. Automated segmentation on the entire cardiac cycle using a deep learning work-flow. In 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS),pp.153-158. https://doi.org/10.1109/SNAMS.2018.8554962.

[17] Zhen, X., Wang, Z., Islam, A., Bhaduri, M., Chan, I. and Li, S., 2016. Multi-scale deep networks and regression forests for direct bi-ventricular volume estimation. Medical image analysis, 30, pp.120-129. https://doi.org/10.1016/j.media.2015.07.003.

[18] Yang, X., Bian, C., Yu, L., Ni, D. and Heng, P.A., 2017, September. Class-balanced deep neural network for automatic ventricular structure segmentation. In International workshop on statistical atlases and computational models of the heart (pp. 152-160). Springer, Cham. https://doi.org/10.1007/978-3-319-75541-0_16.

[19] Baumgartner, C.F., Koch, L.M., Pollefeys, M. and Konukoglu, E., 2017, September. An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation. In International Workshop on Statistical Atlases and Computational Models of the Heart, pp.111-119. Springer, Cham. https://doi.org/10.1007/978-3-319-75541-0_12.

[20] Ronneberger, O., Fischer, P. and Brox, T., 2015, October. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pp.234-241. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28.

[21] Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S. and Pal, C., 2016. The importance of skip connections in biomedical image segmentation. In Deep Learning and Data Labeling for Medical Applications, pp. 179-187. https://doi.org/10.1007/978-3-319-46976-8_19.

[22] Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N. and Liang, J., 2019. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging, 39(6), pp.1856-1867. https://doi.org/10.1109/TMI.2019.2959609.

[23] Ibtehaz, N. and Rahman, M.S., 2020. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural Networks, 121, pp.74-87. https://doi.org/10.1016/j.neunet.2019.08.025

[24] Xue, W., Lum, A., Mercado, A., Landis, M., Warrington, J. and Li, S., 2017, September. Full quantification of left ventricle via deep multitask learning network respecting intra-and inter-task relatedness. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp.276-284. Springer, Cham. https://doi.org/10.1007/978-3-319-66179-7_32.

[25] Zange, L., Muehlberg, F., Blaszczyk, E., Schwenke, S., Traber, J., Funk, S. and Schulz-Menger, J., 2019. Quantification in cardiovascular magnetic resonance: agreement of software from three different vendors on assessment of left ventricular function, 2D flow and parametric mapping. Journal of Cardiovascular Magnetic Resonance, 21(1), pp.12. https://doi.org/10.1186/s12968-019-0522-y.

[26] Bai, W., Sinclair, M., Tarroni, G., Oktay, O., Rajchl, M., Vaillant, G., Lee, A.M., Aung, N., Lukaschuk, E., Sanghvi, M.M. and Zemrak, F., 2018. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. Journal of Cardiovascular Magnetic Resonance, 20(1), pp.65. https://doi.org/10.1186/s12968-018-0471-x.

[27] Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G. and Sanroma, G., 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?. IEEE transactions on medical imaging, 37(11), pp.2514-2525. https://doi.org/10.1109/TMI.2018.2837502

[28] Petitjean, C., Zuluaga, M.A., Bai, W., Dacher, J.N., Grosgeorge, D., Caudron, J., Ruan, S., Ayed, I.B., Cardoso, M.J., Chen, H.C. and Jimenez-Carretero, D., 2015. Right ventricle segmentation from cardiac MRI: a collation study. Medical image analysis, 19(1), pp.187-202. https://doi.org/10.1016/j.media.2014.10.004.

[29] Nasr-Esfahani, M., Mohrekesh, M., Akbari, M., Soroushmehr, S.R., Nasr-Esfahani, E., Karimi, N., Samavi, S. and Najarian, K., Left ventricle segmentation in cardiac MR images using fully convolutional network. In

2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 1275-1278). IEEE..

[30] Queirós, S., Barbosa, D., Heyde, B., Morais, P., Vilaça, J.L., Friboulet, D., Bernard, O. and D'hooge, J. Fast automatic myocardial segmentation in 4D cine CMR datasets. Medical image analysis, 18(7), pp.1115-1131. https://doi.org/10.1016/j.media.2014.06.001.

[31] Petitjean, C., Zuluaga, M.A., Bai, W., Dacher, J.N., Grosgeorge, D., Caudron, J., Ruan, S., Ayed, I.B., Cardoso, M.J., Chen, H.C. and Jimenez-Carretero, D. Right ventricle segmentation from cardiac MRI: a collation study. Medical image analysis, 19(1), pp.187-202. https://doi.org/10.1016/j.media.2014.10.004

[32] Schulz-Menger, J., Bluemke, D.A., Bremerich, J., Flamm, S.D., Fogel, M.A., Friedrich, M.G., Kim, R.J., von Knobelsdorff-Brenkenhoff, F., Kramer, C.M., Pennell, D.J. and Plein, S., 2013. Standardized image interpretation and post processing in cardiovascular magnetic resonance: Society for Cardiovascular Magnetic Resonance (SCMR) board of trustees task force on standardized post processing. Journal of Cardiovascular Magnetic Resonance, 15(1), pp.1-19. https://doi.org/10.1186/1532-429X-15-35.

[33] Caiani, E.G., Landreani, F., Costantini, L., Mulder, E., Gerlach, D., Vaada, P. and Migeotte, P.F., 2018. Effectiveness of high-intensity jump training countermeasure on mitral and aortic flow after 58-days head-down bed-rest assessed by phase-contrast MRI. In International Astronautical Congress (Vol. 2018, pp. 1-6).

[34] Cherabit, N., Chelali, F.Z. and Djeradi, A., 2012. Circular hough transform for iris localization. Science and Technology, 2(5), pp.114-121. https://doi.org/10.5923/j.scit.20120205.02.

[35] Santurkar, S., Tsipras, D., Ilyas, A. and Madry, A., 2018. How does batch normalization help optimization?. In Advances in Neural Information Processing Systems, pp.2483-2493.

[36] Nair, V. and Hinton, G.E., 2010, January. Rectified linear units improve restricted boltzmann machines. In ICML.

[37] Zeiler, M.D., Krishnan, D., Taylor, G.W. and Fergus, R., 2010, June. Deconvolutional networks. In 2010 IEEE Computer Society Conference on computer vision and pattern recognition, pp. 2528-2535. https://doi.org/10.1109/CVPR.2010.5539957.

[38] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., 2017. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp.2261-2269. https://doi.org/ 10.1109/CVPR.2017.243.

[39] He, K., Zhang, X., Ren, S. and Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision, pp.1026-1034. . https://doi.org/10.1109/ICCV.2015.123.

[40] Dogo, E.M., Afolabi, O.J., Nwulu, N.I., Twala, B. and Aigbavboa, C.O., 2018, December. A comparative analysis of gradient descent-based optimization algorithms on convolutional neural networks. In 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), pp.92-99. IEEE. https://doi.org /10.1109/CTEMS.2018.8769211.

[41] Shorten, C. and Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), p.60. https://doi.org/10.1186/s40537-019-0197-0

[42] Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp.3431-3440. . https://doi.org/10.1109/CVPR.2015.7298965.

[43] Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S. and Pal, C., 2016. The importance of skip connections in biomedical image segmentation. In Deep Learning and Data Labeling for Medical Applications, pp.179-187. https://doi.org/10.1007/978-3-319-46976-8_19.

[44] Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N. and Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp.3-11. https://doi.org/10.1007/978-3-030-00889-5_1.

[45] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp.770-778. https://doi.org/10.1109/CVPR.2016.90.

[46] Marchesseau, S., Ho, J.X. and Totman, J.J., 2016. Influence of the short-axis cine acquisition protocol on the cardiac function evaluation: a reproducibility study. European journal of radiology open, 3, pp.60-66. https://doi.org/10.1016/j.ejro.2016.03.003.

[47] Baselli, G., Codari, M. and Sardanelli, F., 2020. Opening the black box of machine learning in radiology: can the proximity of annotated cases be a way?. European Radiology Experimental, 4, pp.1-7. https://doi.org/10.1186/s41747-020-00159-0.

[48] Moccia, S., Migliorelli, L., Carnielli, V. and Frontoni, E., 2019. Preterm infants' pose estimation with spatio-temporal features. IEEE Transactions on Biomedical Engineering. https://doi.org/10.1109/tbme.2019.2961448.

APPENDIX

CLAIM: Checklist for Artificial Intelligence in Medical Imaging

Section / Topic	No.	Item	
TITLE / ABSTRACT			
	1	Identification as a study of AI methodology, specifying the category of technology used (e.g., deep learning)	**Pag 1, title** **Pag 2, row 6**
	2	Structured summary of study design, methods, results, and conclusions	**Pag 2-3**
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the AI approach	**Pag 4-5**
	4	Study objectives and hypotheses	**Pag 6, row 6-22**
METHODS			
Study Design	**5**	Prospective or retrospective study	**Pag 7, row 1**
	6	Study goal, such as model creation, exploratory study, feasibility study, non-inferiority trial	**Pag 4, row 16-26** **Pag 6, row 6-22**
Data	**7**	Data sources	**Pag 7, row 1-4, 21-26; page 8, row 1-5**
	8	Eligibility criteria: how, where, and when potentially eligible participants or studies were identified (e.g., symptoms, results from previous tests, inclusion in registry, patient-care setting, location, dates)	**Pag 7-8, sec 2.1**
	9	Data pre-processing steps	**Pag 8-9, sec 2.2** **Pag 11, sec 2.4**
	10	Selection of data subsets, if applicable	**Pag 8-9, sec 2.2**
	11	Definitions of data elements, with references to Common Data Elements	**Pag 11-13, sec 2.5**
	12	De-identification methods	**N/A**
	13	How missing data were handled	**N/A**
Ground Truth	**14**	Definition of ground truth reference standard, in sufficient detail to allow replication	**Pag 7, row 8-26; pag 8, row 4-5**
	15	Rationale for choosing the reference standard (if alternatives exist)	**N/A**
	16	Source of ground-truth annotations; qualifications and preparation of annotators	**Pag 7, sec 2.1, row 9-10** **Pag 8, row 5-**

			3
	17	Annotation tools	**Pag 7, row 9-10**
	18	Measurement of inter- and intrarater variability; methods to mitigate variability and/or resolve discrepancies	**Pag 7, row 15-16; pag 13, row 3-5**
Data Partitions	**19**	Intended sample size and how it was determined	**Pag 7, sec 2.1** **Pag 7, row 26, pag 8, row 1-2**
	20	How data were assigned to partitions; specify proportions	**Pag 12, sec 2.5 row 10-11**
	21	Level at which partitions are disjoint (e.g., image, study, patient, institution)	**Pag 12, sec 2.5, row 10**
Model	**22**	Detailed description of model, including inputs, outputs, all intermediate layers and connections	**Pag 9-10, sec 2.3**
	23	Software libraries, frameworks, and packages	**Pag 12, row 8**
	24	Initialization of model parameters (e.g., randomization, transfer learning)	**Pag 11, sec 2.4 row 4**
Training	**25**	Details of training approach, including data augmentation, hyperparameters, number of models trained	**Pag 11-12, sec 2.4**
	26	Method of selecting the final model	**Pag 11-12, sec 2.4**
	27	Ensembling techniques, if applicable	**N/A**
Evaluation	**28**	Metrics of model performance	**Pag 12, sec 2.5 row 11-15**
	29	Statistical measures of significance and uncertainty (e.g., confidence intervals)	**Pag 13, sec 2.5 row 1-2** **Pag 13 row 12-14**
	30	Robustness or sensitivity analysis	**Pag 13, row 12-14**
	31	Methods for explainability or interpretability (e.g., saliency maps), and how they were validated	**N/A**
	32	Validation or testing on external data	**Pag 13, row 9-11**
RESULTS			
Data	**33**	Flow of participants or cases, using a diagram to indicate inclusion and exclusion	**Pag 7, sec 2.1**
	34	Demographic and clinical characteristics of cases in each partition	**Pag 8 table1**
Model performance	**35**	Performance metrics for optimal model(s) on all data partitions	**Pag 13-14, section 3**
	36	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	**Pag 18,**
	37	Failure analysis of incorrectly classified cases	**Fig 2,3,4**
DISCUSSION			
	38	Study limitations, including potential bias, statistical uncertainty, and	**Pag 25, row**

		generalizability	**1-8**
	39	Implications for practice, including the intended use and/or clinical role	**Pag 25, row 9-18**
OTHER INFORMATION			
	40	Registration number and name of registry	**N/A**
	41	Where the full study protocol can be accessed	**N/A**
	42	Sources of funding and other support; role of funders	**Pag 26**

Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020; 2(2):e200029. https://doi.org/10.1148/ryai.2020200029