

Article

Deep Learning Based Cardiac MRI Segmentation: Do We Need Experts?

Youssef Skandarani ^{1,2,*}, Pierre-Marc Jodoin ³ and Alain Lalande ^{1,4} 

¹ Laboratoire ImVIA, University of Bourgogne Franche-Comte, 21000 Dijon, France; Alain.Lalande@u-bourgogne.fr

² CASIS Inc., 21800 Quetigny, France

³ Department of Computer Science, University of Sherbrooke, Sherbrooke, QC J1K 2R1, Canada; pierre-marc.jodoin@usherbrooke.ca

⁴ Department of Radiology, University Hospital of Dijon, 21000 Dijon, France

* Correspondence: youssef_skandarani@etu.u-bourgogne.fr

Abstract: Deep learning methods are the de facto solutions to a multitude of medical image analysis tasks. Cardiac MRI segmentation is one such application, which, like many others, requires a large number of annotated data so that a trained network can generalize well. Unfortunately, the process of having a large number of manually curated images by medical experts is both slow and utterly expensive. In this paper, we set out to explore whether expert knowledge is a strict requirement for the creation of annotated data sets on which machine learning can successfully be trained. To do so, we gauged the performance of three segmentation models, namely U-Net, Attention U-Net, and ENet, trained with different loss functions on expert and non-expert ground truth for cardiac cine-MRI segmentation. Evaluation was done with classic segmentation metrics (Dice index and Hausdorff distance) as well as clinical measurements, such as the ventricular ejection fractions and the myocardial mass. The results reveal that generalization performances of a segmentation neural network trained on non-expert ground truth data is, to all practical purposes, as good as that trained on expert ground truth data, particularly when the non-expert receives a decent level of training, highlighting an opportunity for the efficient and cost-effective creation of annotations for cardiac data sets.

Keywords: deep learning; MRI; heart; segmentation; annotated data set



Citation: Skandarani, Y.; Jodoin, P.-M.; Lalande, A. Deep Learning Based Cardiac MRI Segmentation: Do We Need Experts? *Algorithms* **2021**, *14*, 212. <https://doi.org/10.3390/a14070212>

Academic Editors: Christian Mata Miquel and Frank Werner

Received: 8 June 2021

Accepted: 13 July 2021

Published: 14 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Deep neural networks (more specifically, *convolutional neural networks*) have deeply percolated through healthcare R&D, addressing various problems, such as survival prediction, disease diagnostics, image registration, anomaly detection, and the segmentation of images, be it magnetic resonance images (MRI), computed tomography (CT) or ultrasound (US), to name a few [1]. The roaring success of deep learning methods is rightly attributed to the unprecedented amount of annotated data across domains. However, ironically, while solutions to decade-long medical problems are at hand [2], the use of neural networks in day-to-day practice is still pending. This can be explained in part by the following two observations. First, while being accurate *on average*, neural networks can nonetheless be sometimes wrong [3], as they provide no strict clinical guarantees. In other words, any neural network within the intra-expert variability is excellent *on average* but not immune to sparse erroneous (yet degenerated) results, which is problematic in clinical practice [2]. Second, machine learning methods are known to suffer from domain adaptation problems, one of the most glaring medical imaging issue of our times [4]. As such, clinically accurate machine learning methods trained on a specific set of data almost always see their performance drop when tested on a data set acquired by following a different protocol. These problems are derived in good part from the fact that current data sets are still relatively

small. According to Maier-Hein et al. [5] most medical imaging challenges organized so far contain less than 100 training and testing cases. This shows that medical applications cannot rely on a *very* large medical data set encompassing tens of thousands of annotated data acquired in various conditions, with machines of various vendors showing clinical conditions and anatomical configurations of all kinds.

This is unlike non-medical computer vision problems, which have had access for a long time to large and varied data sets, such as ImageNet, Coco, PascalVOC, ADE20k, and Youtube-8M, to name a few [6]. The annotation of these data sets relies on non-experts, often through online services, such as Mechanical Turk [7]. Unfortunately, obtaining similarly large annotated data sets in medical imaging is difficult. The challenge stems from the nature of the data, which is sensitive and requires navigating a complicated regulatory framework and privacy safeguards. Furthermore, labeling medical data sets is quite resource intensive and prohibitively costly, as it requires a domain expertise.

For these reasons, the medical imaging literature has had an increasing number of publications, whose goal is to compensate for the lack of expert annotations [8]. While some methods leverage partly-annotated data sets [9], others use domain adaptation strategies to compensate for small training data sets [10]. Some other approaches artificially increase the number of annotated data with generative adversarial networks (GANs) [11,12], while others use third-party neural networks to help experts annotate images more rapidly [13].

While these methods have been shown to be effective for their specific test cases, it is widely accepted that large manually-annotated data sets bring indisputable benefits [14]. In this work, we depart from trying to improve the segmentation methods and focus on the data sets as we challenge the idea that medical data, cardiac cine MRI specifically, needs to be labeled by experts only, and explore the consequences of using non-expert annotations on the generalization capabilities of a neural network. Non-expert here refers to a non-physician who cannot be regarded as a reference in the field. While non-expert annotations are easier and cheaper to obtain, they could be used to build larger data sets faster and at a reduced cost.

This idea was tested on cardiac cine-MRI segmentation. To this end, we had two non-experts labeling cardiac cine-MRI images and compared the performance of neural networks trained on non-expert and expert data. The evaluation between both approaches was conducted with geometric metrics (Dice index and Hausdorff distance) as well as clinical parameters, namely, the ejection fraction for the left and right ventricles and the myocardial mass.

2. Methods and Data

As mentioned before, medical data annotation requires rightful expertise so that the labeling can be used with full confidence. Expert annotators are typically medical doctors or medical specialists whose training and experience are reliable sources of truth for the problem at hand. These experts often have close collaborators working daily with medical data, typically computer scientists, technicians, biophysicist, etc. While their understanding of the data is real, these non-experts are typically not considered a reliable source of truth. Non-experts are thus considered as people who can manually label data but whose annotations are biased and/or noisy and thus unreliable.

In this study, two non-experts were asked to label 1902 cardiac images. We defined a non-expert as someone with no professional expertise on cardiac anatomy nor on cine-MR images. Non-Expert 1 is a technician in biotechnology who received 30 min of training by a medical expert on how to recognize and outline cardiac structures. The training was done on a few examples where the expert showed what the regions of interest in the image look like and where their boundaries lie. Training also came with an introduction to the cardiac anatomy and its temporal dynamics. Non-Expert 2 is a computer scientist with 4 years of active research in cardiac cine-MRI with several months of training. In the case of Non-Expert 2, the training spanned several months, where directions about the imaging modality as well as the anatomy and pathologies were thoroughly explained. In addition,

fine delineation guidelines were provided to disambiguate good from poor annotations. In this study, we both gauge the effect of training a neural network on non-expert data and also verify how the level of training of the non-experts impact the overall results.

The non-experts were asked to delineate three structures of the heart, namely, the left ventricular cavity (endocardial border), the left ventricle myocardium (epicardial border) and the endocardial border of the right ventricle. No further quality control was done to validate the non-expert annotations. Segmentations were used as is for the subsequent tasks.

We used the gold standard for medical image segmentation U-Net [15] as the baseline network. In addition, the well-known Attention U-Net [16] and ENet [17] networks were trained in order to ensure that the results are affected by the differences in annotations and not the network architecture. We first trained the the segmentation models (U-Net, Attention U-Net and ENet) on the original ACDC data set (Automated Cardiac Diagnosis Challenge) [2] with its associated ground truth, using a classical supervised training scheme, with a combined cross-entropy and Dice loss:

$$L_{ce} = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^3 y_{ki} \log \hat{y}_{ki} \quad (1)$$

$$L_{dice} = 1 - \frac{1}{N} \left[\sum_{i=1}^N \frac{2 \times \sum_k \hat{y}_{ki} y_{ki}}{\sum_k \hat{y}_{ki} + \sum_k y_{ki}} \right] \quad (2)$$

where \hat{y}_{ki} is the probabilistic output for image $i \in N$ (N is the number of images in the batch) and class $k \in \{1, 2, 3\}$ (3 is the number of classes). \hat{y} is the predicted output of the network, y is a one-hot encoding of the ground truth segmentation map.

We then re-trained the neural networks with the non-expert labels, using the same training configuration. Furthermore, considering that the non-expert annotations can be seen as noisy versions of the true annotation (i.e., $y' = y + \epsilon$ where y' is the non-expert annotation, y is the ground truth and ϵ a random variable), we also trained the networks with a mean absolute error loss which, as shown by Ghosh et al. [18], has the solve property of compensating for labeling inaccuracies.

3. Experimental Setup

To test whether non-expert annotated data sets hold any value for cardiac MRI segmentation, the following two cardiac cine MRI data sets were used:

- Automated Cardiac Diagnosis Challenge (ACDC) data set [2]: This data set comprises 150 exams acquired at the University Hospital of Dijon (all from different patients). It is divided into 5 evenly distributed subgroups (4 pathological plus 1 healthy subject groups) and split into 100 exams for training, and 50 are held out set for testing. The exams were acquired using two MRI scanners with different magnetic strengths (1.5 T and 3 T). The pixel spacing varies from 0.7 mm to 1.9 mm with a slice spacing varying between 5 mm and 10 mm. An example of images with the different expert and non-expert annotations is shown in Figure 1.
- Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Image Segmentation (M&M) data set [19]: This data set consists of 375 cases from 3 different countries (Spain, Germany and Canada) totaling 6 different centers with 4 different MRI manufacturers (Siemens, General Electric, Philips and Canon). The cohort is composed of patients with hypertrophic and dilated cardiomyopathies as well as healthy subjects. The cine MR images were annotated by experienced clinicians from the respective centers.

We trained the segmentation models on the 100 ACDC training subjects on either the expert or non-expert ground truth data. Training was done with a fixed set of hyper-parameters, chosen through a cross-validated hyper-parameters search to best fit the 3 annotators, without tuning it further. The networks were trained three times in order to reduce the effect of the stochastic nature of the training process on the results.

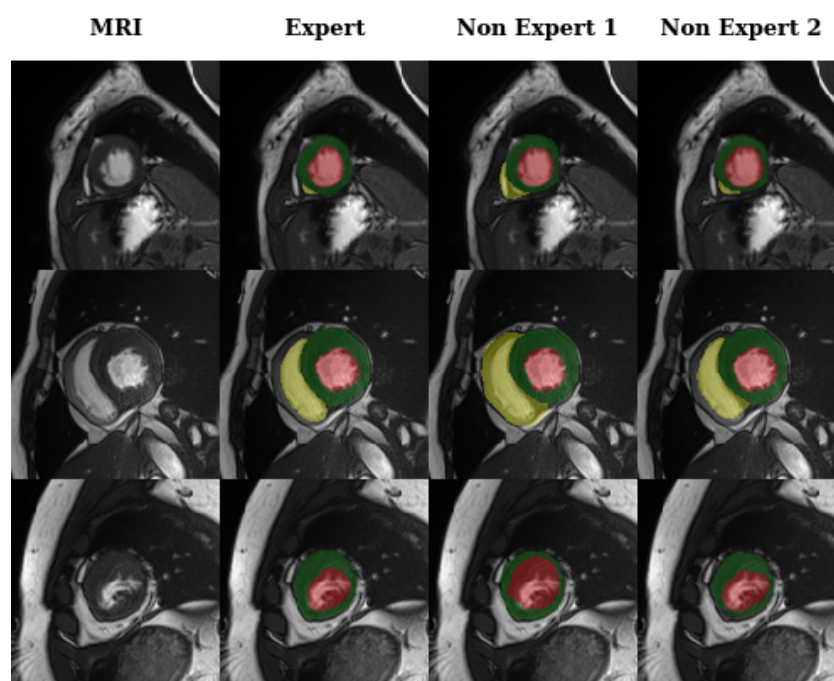


Figure 1. ACDC Annotation differences between Expert, Non-Expert 1 and Non-Expert 2.

As mentioned before, we first trained the neural networks on non-expert data with exactly the same setup as for the expert annotations. Then, we retrained from scratch the neural networks (U-Net, Attention U-Net and ENet) with a L1 loss, which was shown to be robust to noisy labels [18].

We then tested in turn on the 50 ACDC test subjects and the 150 M&Ms training data. The M&Ms data set constitutes data with ground truth that are not biased toward either of the annotators of the training set, be it the expert or the non-expert. Moreover, testing on different data sets provided an inter-expert variability range as well as a domain generalization setup.

4. Results and Discussion

The first set of results are laid out in Table 1. It corresponds to standard geometrical metrics, i.e., the Dice score and the Hausdorff distance (HD) for the left ventricular (LV) cavity (Table 1), the myocardium (MYO) (considering the endocardial and epicardial of the left ventricle) (Table 2) and the cavity of the right ventricle (RV) (Table 3). It also contains the end-diastolic volume (EDV) as well as the ejection fraction (EF) for the LV and the RV and the myocardial mass error. For all three tables, the networks (U-Net, Attention U-Net and ENet) were trained on the ACDC training set and tested on the ACDC testing set and the M&Ms training set.

Results for the ACDC testing set reveal that for the LV (Table 1) the networks trained on the non-expert annotations (Non-Expert 1 as well as Non-Expert 2) manage to achieve performance that is statistically indistinguishable from that of the expert. This is true regardless of the training loss (CE+Dice vs MAE+Dice) and the metric (Dice, HD, and EF). The only exception is for the EF error for Non-Expert 1 with loss function MAE+Dice.

The situation, however, is more fuzzy for the MYO and the RV. In both cases, we can see that results for the Non-Expert 1 are almost always worse than that of the expert, especially for the CE+Dice loss. For example, there is a Dice score drop of 12% on the myocardium. In addition, the clinical results on the RV (Table 3) show a clear gap between the Non-Expert 1 and the other annotators. However, we can appreciate how the MAE+Dice loss improves results for both non-experts. Overall, for the MYO and the RV, results for the Non-Expert 2 are very close (if not better) than that of the expert. This is obvious when considering the average myocardial mass error in Table 2. Although, one recurrent result

from our experiments is the hit-and-miss performance of all the evaluated networks on the M&Ms data set, where in a number of cases, the output segmentation is completely degenerated as shown in Figure 2. Moreover, the difference in segmentation performance between all the annotators is similar regardless of the segmentation network (U-Net, Attention U-Net or ENet) used, although Attention U-Net shows the best performance overall, which is to be expected, given its larger capacity.

Table 1. Dice score (DSC), Hausdorff distance (HD), average ejection fraction (EF) error, and end diastolic volume (EDV) of left ventricular cavity for U-Net, Attention U-Net and E-Net with expert and non-expert annotations with data augmentation.

U-Net							
Test Set	ACDC Trainset	Loss	Avg. DSC	Avg. HD	Avg. EF (%)	Avg EF Err.	Avg. EDV (mL)
ACDC	Expert	CE+Dice	0.92 ± 0.08	11.84 ± 11.43	47.57 ± 19.62	3.59 ± 3.17	175.70 ± 69.03
	Non-Expert 1	CE+Dice	0.92 ± 0.08	12.15 ± 10.63	50.27 ± 20.00	4.37 ± 3.69	173.05 ± 67.06
		MAE+Dice	0.91 ± 0.08	12.80 ± 10.48	50.37 ± 20.14	4.52 * ± 4.32	173.57 ± 70.29
	Non-Expert 2	CE+Dice	0.93 ± 0.09	11.68 ± 12.36	49.73 ± 20.70	3.35 ± 3.21	182.98 ± 72.51
		MAE+Dice	0.92 ± 0.09	11.76 ± 12.34	49.45 ± 20.68	3.69 ± 3.59	181.19 ± 70.93
	M&Ms	Expert	CE+Dice	0.86 ± 0.11	15.03 ± 8.07	54.22 ± 15.11	7.47 ± 5.79
Non-Expert 1		CE+Dice	0.86 ± 0.10	11.76 ± 6.04	56.81 * ± 15.37	5.53 * ± 5.22	149.75 * ± 57.33
		MAE+Dice	0.86 ± 0.10	11.52 ± 5.55	57.14 * ± 15.23	5.39 * ± 5.38	151.08 * ± 58.08
Non-Expert 2		CE+Dice	0.88 ± 0.09	11.58 ± 6.93	56.70 * ± 15.33	5.54 * ± 4.57	164.95 ± 62.00
		MAE+Dice	0.88 ± 0.10	11.84 ± 7.93	56.74 * ± 15.45	6.17 * ± 5.11	165.78 ± 62.00
Attention U-Net							
Test Set	ACDC Trainset	Loss	Avg. DSC	Avg. HD	Avg. EF (%)	Avg EF Err.	Avg. EDV
ACDC	Expert	CE+Dice	0.92 ± 0.08	11.78 ± 11.14	46.83 ± 19.50	3.81 ± 3.63	177.77 ± 71.08
	Non-Expert 1	CE+Dice	0.92 ± 0.08	12.07 ± 11.20	50.27 ± 20.07	4.28 ± 3.94	170.86 ± 67.86
		MAE+Dice	0.92 ± 0.09	13.42 ± 12.62	50.35 ± 20.28	4.74 * ± 4.15	173.43 ± 67.61
	Non-Expert 2	CE+Dice	0.93 ± 0.07	11.05 ± 11.45	50.56 ± 21.24	3.50 ± 3.86	180.28 ± 70.50
		MAE+Dice	0.93 ± 0.08	12.52 ± 11.62	49.78 ± 21.10	3.93 ± 3.96	178.82 ± 68.16
	M&Ms	Expert	CE+Dice	0.88 ± 0.09	11.90 ± 7.38	55.77 ± 15.23	5.59 ± 4.73
Non-Expert 1		CE+Dice	0.86 ± 0.12	11.90 ± 7.73	57.67 ± 17.34	6.64 * ± 6.45	148.25 * ± 57.84
		MAE+Dice	0.86 ± 0.10	12.40 ± 6.79	57.40 ± 17.44	6.23 ± 6.18	148.22 * ± 57.37
Non-Expert 2		CE+Dice	0.87 ± 0.10	11.10 ± 6.59	58.51 * ± 16.33	5.99 ± 5.01	161.20 ± 61.40
		MAE+Dice	0.87 ± 0.09	11.14 ± 6.43	58.20 * ± 16.34	6.29 * ± 5.36	161.29 ± 62.34
E-Net							
Test Set	ACDC Trainset	Loss	Avg. DSC	Avg. HD	Avg. EF (%)	Avg EF Err.	Avg. EDV
ACDC	Expert	CE+Dice	0.92 ± 0.09	12.28 ± 11.38	46.68 ± 19.06	3.94 ± 3.86	178.48 ± 70.68
	Non-Expert 1	CE+Dice	0.91 ± 0.09	12.72 ± 12.28	48.64 ± 20.33	4.31 ± 4.99	172.60 ± 67.69
		MAE+Dice	0.91 ± 0.10	13.07 ± 15.00	50.10 ± 19.99	4.15 ± 3.65	171.53 ± 67.29
	Non-Expert 2	CE+Dice	0.92 ± 0.09	12.18 ± 14.01	49.05 ± 20.52	3.25 ± 3.09	180.77 ± 71.20
		MAE+Dice	0.92 ± 0.08	11.69 ± 11.16	48.83 ± 20.57	3.45 ± 3.53	181.07 ± 70.45
	M&Ms	Expert	CE+Dice	0.86 ± 0.10	14.46 ± 7.54	53.91 ± 16.04	7.19 ± 5.98
Non-Expert 1		CE+Dice	0.86 ± 0.12	13.64 ± 10.60	53.10 ± 21.96	7.41 ± 12.67	152.59 * ± 59.87
		MAE+Dice	0.86 ± 0.12	11.90 ± 8.59	55.90 † ± 18.26	5.81 * † ± 7.64	150.79 * ± 57.79
Non-Expert 2		CE+Dice	0.87 ± 0.10	12.39 ± 7.63	53.53 ± 17.49	6.39 ± 7.25	165.05 ± 61.73
		MAE+Dice	0.87 ± 0.10	12.72 ± 8.41	54.99 ± 16.29	6.76 ± 5.56	165.36 ± 60.71

* p -value < 0.05 non-expert vs. expert; † p -value < 0.05 MAE vs. CE.

Table 2. Dice score, Hausdorff distance and mass error for the **myocardium** for U-Net, Attention U-Net and E-Net with expert and non-expert annotations with data augmentation.

U-Net					
Test Set	ACDC Trainset	Loss	Avg. DSC	Avg. HD	Avg. Mass Err.
ACDC	Expert	CE+Dice	0.88 ± 0.03	11.02 ± 6.25	16.88 ± 9.66
	Non-Expert 1	CE+Dice	0.82 ± 0.03	11.45 ± 4.88	36.40 * ± 13.37
		MAE+Dice	0.86 ± 0.03	11.69 ± 5.18	15.93 † ± 8.95
	Non-Expert 2	CE+Dice	0.87 ± 0.03	10.13 ± 5.02	11.68 * ± 7.74
		MAE+Dice	0.89 ± 0.03	10.43 ± 5.00	6.58 * † ± 5.93
	M&Ms	Expert	CE+Dice	0.80 ± 0.06	17.57 ± 7.65
Non-Expert 1		CE+Dice	0.76 ± 0.08	14.91 ± 8.60	22.14 * ± 19.71
		MAE+Dice	0.78 ± 0.09	14.32 ± 7.95	18.78 † ± 17.80
Non-Expert 2		CE+Dice	0.80 ± 0.07	14.29 ± 7.53	16.36 ± 13.69
		MAE+Dice	0.80 ± 0.07	13.76 ± 6.79	19.44 * † ± 16.79
Attention U-Net					
Test Set	ACDC Trainset	Loss	Avg. DSC	Avg. HD	Avg. Mass Err.
ACDC	Expert	CE+Dice	0.88 ± 0.03	11.57 ± 7.13	17.95 ± 10.28
	Non-Expert 1	CE+Dice	0.84 ± 0.03	10.27 ± 5.25	33.99 * ± 13.17
		MAE+Dice	0.86 ± 0.03	13.11 ± 8.65	20.12 † ± 8.82
	Non-Expert 2	CE+Dice	0.87 ± 0.03	10.86 ± 6.17	13.76 * ± 8.99
		MAE+Dice	0.88 ± 0.03	11.53 ± 6.62	9.68 * † ± 7.32
	M&Ms	Expert	CE+Dice	0.82 ± 0.05	15.85 ± 9.84
Non-Expert 1		CE+Dice	0.78 ± 0.08	13.95 ± 8.04	19.44 * ± 17.61
		MAE+Dice	0.77 ± 0.09	16.19 ± 9.78	20.80 * ± 20.08
Non-Expert 2		CE+Dice	0.80 ± 0.07	13.43 ± 7.85	16.17 ± 13.58
		MAE+Dice	0.78 ± 0.08	13.75 ± 7.56	26.89 * † ± 21.43
E-Net					
Test Set	ACDC Trainset	Loss	Avg. DSC	Avg. HD	Avg. Mass Err.
ACDC	Expert	CE+Dice	0.87 ± 0.03	10.99 ± 5.61	17.14 ± 9.05
	Non-Expert 1	CE+Dice	0.82 ± 0.04	11.91 ± 7.80	39.81 * ± 15.58
		MAE+Dice	0.86 ± 0.04	11.51 ± 6.03	21.26 * † ± 9.49
	Non-Expert 2	CE+Dice	0.85 ± 0.04	10.96 ± 6.11	15.64 ± 9.59
		MAE+Dice	0.87 ± 0.03	10.49 ± 5.16	7.83 * † ± 6.24
	M&Ms	Expert	CE+Dice	0.80 ± 0.07	15.81 ± 7.55
Non-Expert 1		CE+Dice	0.77 ± 0.10	16.12 ± 12.23	28.40 * ± 23.91
		MAE+Dice	0.78 ± 0.09	14.51 ± 9.26	18.35 † ± 17.10
Non-Expert 2		CE+Dice	0.80 ± 0.06	15.18 ± 8.41	16.08 ± 14.98
		MAE+Dice	0.79 ± 0.08	15.69 ± 8.39	18.14 † ± 15.82

* p -value < 0.05 non-expert vs. expert; † p -value < 0.05 MAE vs. CE.

Further analysis of the segmentation performance on the different sections of the heart (Figure 3), namely the base, the middle and the apex, show that the differences between the non-experts and the expert annotations lie heavily on the two ends of the heart. The performance gap is more pronounced on the apex for the three anatomical structures. In parallel, when we look at the performance from the disease groups (Figure 4), we can distinguish a relative similarity in the Dice score between the different annotators and disease groups.

Our experiments also reveal some interesting results on the M&Ms data set, a data set with different acquisition settings than ACDC. In that case, we see that the gap in performance between the expert and the non-expert decreases substantially. For example, when comparing the Non-Expert 1 results with MAE+Dice loss and those from the expert annotation, we see that the Dice difference for the RV decreases from a 6% on the ACDC

data set to a mere 4% on the M&Ms data set; overall, the results of Non-Expert 2 are similar (and sometimes better) than those of the expert.

Table 3. Dice score, Hausdorff distance, average ejection fraction (EF) error, and end diastolic volume (EDV) for the **right ventricular cavity** for U-Net, Attention U-Net and E-Net with expert and non-expert annotations with data augmentation.

U-Net							
Test Set	ACDC Train set	Loss	Avg. DSC	Avg. HD	Avg. EF (%)	Avg EF Err.	Avg. EDV (mL)
ACDC	Expert	CE+Dice	0.90 ± 0.06	14.72 ± 6.12	42.77 ± 15.25	6.69 ± 6.21	183.73 ± 70.68
	Non-Expert 1	CE+Dice	0.78 ± 0.11	19.59 ± 8.07	32.22 * ± 13.49	14.77 * ± 9.88	216.16 * ± 69.96
		MAE+Dice	0.84 ± 0.09	16.78 ± 7.23	35.51 *,† ± 14.64	13.52 * ± 8.88	181.61 † ± 61.85
	Non-Expert 2	CE+Dice	0.86 ± 0.07	15.25 ± 6.00	40.10 ± 14.42	7.98 ± 7.34	200.67 * ± 73.35
		MAE+Dice	0.89 ± 0.07	14.65 ± 6.34	41.61 ± 14.85	7.38 ± 7.27	186.59 ± 71.56
	M&Ms	Expert	CE+Dice	0.82 ± 0.15	16.91 ± 12.65	52.90 ± 19.79	9.12 ± 12.73
Non-Expert 1		CE+Dice	0.74 ± 0.15	22.51 ± 15.82	37.68 * ± 21.06	17.88 * ± 18.47	173.34 * ± 70.88
		MAE+Dice	0.78 ± 0.15	16.93 ± 10.95	38.75 * ± 36.59	18.12 * ± 33.08	144.36 † ± 60.57
Non-Expert 2		CE+Dice	0.79 ± 0.16	20.31 ± 17.56	44.11 * ± 29.89	12.86 * ± 26.88	168.76 * ± 66.89
		MAE+Dice	0.81 ± 0.17	18.05 ± 15.60	39.17 * ± 91.83	20.72 * ± 88.32	149.17 † ± 66.41
Attention U-Net							
Test Set	ACDC Train set	Loss	Avg. DSC	Avg. HD	Avg. EF (%)	Avg EF Err.	Avg. EDV (mL)
ACDC	Expert	CE+Dice	0.89 ± 0.07	16.62 ± 7.71	38.73 ± 15.58	8.71 ± 7.46	194.05 ± 71.01
	Non-Expert 1	CE+Dice	0.79 ± 0.10	20.60 ± 9.19	30.12 * ± 13.11	15.96 * ± 9.27	218.32 * ± 69.08
		MAE+Dice	0.83 ± 0.09	20.08 ± 11.27	33.00 * ± 14.62	13.57 *,† ± 9.68	198.36 † ± 68.13
	Non-Expert 2	CE+Dice	0.86 ± 0.07	16.33 ± 7.92	37.96 ± 15.19	9.20 ± 7.15	200.98 ± 71.70
		MAE+Dice	0.88 ± 0.09	16.30 ± 9.18	39.65 ± 15.58	8.31 ± 8.28	185.30 ± 70.51
	M&Ms	Expert	CE+Dice	0.82 ± 0.15	17.94 ± 12.54	50.23 ± 15.46	9.39 ± 8.21
Non-Expert 1		CE+Dice	0.76 ± 0.13	20.96 ± 11.07	37.69 * ± 19.13	17.86 * ± 15.87	179.85 * ± 67.54
		MAE+Dice	0.77 ± 0.14	24.24 ± 14.61	37.94 * ± 18.12	17.67 * ± 15.47	167.74 *,† ± 60.49
Non-Expert 2		CE+Dice	0.79 ± 0.15	18.09 ± 13.02	47.28 * ± 19.16	12.20 * ± 12.84	159.95 ± 64.95
		MAE+Dice	0.80 ± 0.17	20.72 ± 15.26	44.72 * ± 27.57	13.22 * ± 23.62	151.40 † ± 63.41
E-Net							
Test Set	ACDC Train set	Loss	Avg. DSC	Avg. HD	Avg. EF (%)	Avg EF Err.	Avg. EDV (mL)
ACDC	Expert	CE+Dice	0.88 ± 0.07	15.76 ± 6.38	38.65 ± 15.27	9.11 ± 7.17	192.65 ± 70.02
	Non-Expert 1	CE+Dice	0.77 ± 0.11	22.14 ± 9.01	30.12 * ± 13.12	16.29 * ± 10.02	223.07 * ± 71.29
		MAE+Dice	0.83 ± 0.10	17.47 ± 8.29	34.48 *,† ± 13.68	13.00 *,† ± 9.11	189.00 † ± 65.27
	Non-Expert 2	CE+Dice	0.84 ± 0.09	16.19 ± 6.29	36.25 ± 14.13	10.46 ± 7.91	203.49 ± 71.15
		MAE+Dice	0.87 ± 0.08	16.46 ± 7.34	36.87 ± 13.84	9.90 ± 7.91	194.54 ± 69.42
	M&Ms	Expert	CE+Dice	0.79 ± 0.18	19.57 ± 16.32	48.86 ± 20.66	12.06 ± 13.44
Non-Expert 1		CE+Dice	0.72 ± 0.16	28.38 ± 17.94	34.20 * ± 30.51	22.27 * ± 26.54	178.67 * ± 72.98
		MAE+Dice	0.78 ± 0.14	19.30 ± 12.58	41.63 *,† ± 19.67	14.83 *,† ± 15.06	156.65 † ± 61.99
Non-Expert 2		CE+Dice	0.79 ± 0.14	19.13 ± 12.09	41.88 * ± 18.65	14.17 * ± 13.61	166.83 * ± 66.72
		MAE+Dice	0.80 ± 0.14	20.73 ± 13.26	42.94 * ± 16.39	13.35 ± 11.36	169.43 * ± 63.54

* p -value < 0.05 non-expert vs. expert; † p -value < 0.05 MAE vs. CE.

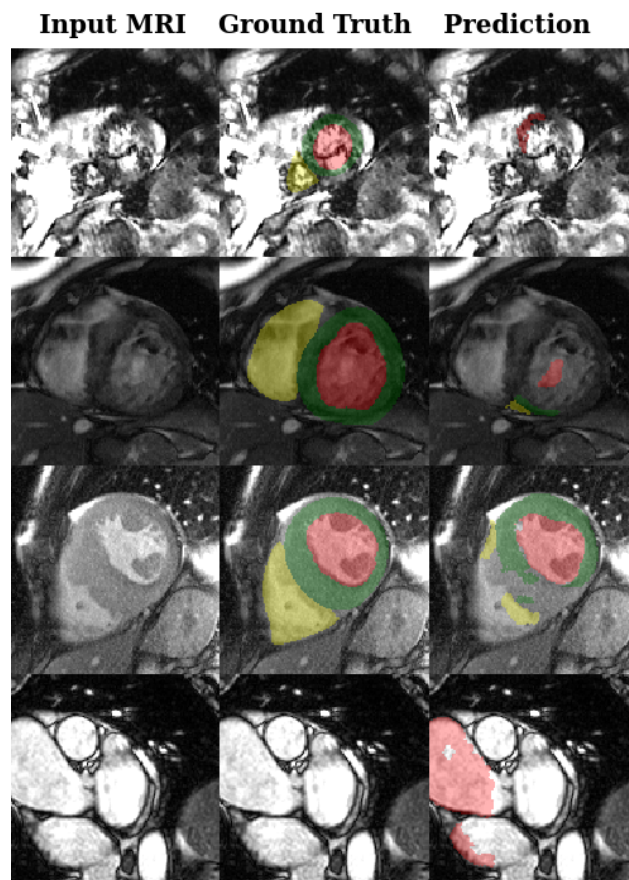


Figure 2. Examples of bad output segmentation on the M&Ms data set for the left ventricle, right ventricle and myocardium.

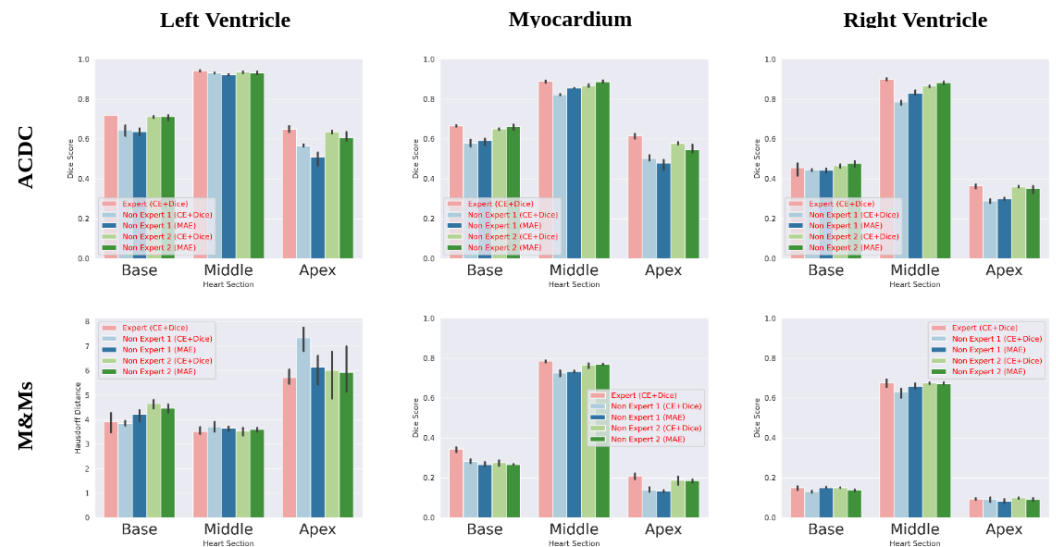


Figure 3. Dice score on ACDC and M&Ms data set per anatomical structure and per slice location.

Throughout our experiments, the performance of the three neural networks (U-Net, Attention U-Net and ENet) trained on the Non-Expert 2 annotations with MAE+Dice loss has been roughly on par, if not better, with those trained on the expert annotations. This is especially true for the LV. For Non-Expert 1, most likely due to a lack of proper training, the results on both test sets and most MYO and RV metrics are worse than those of the expert. In fact, a statistical test reveals that the results from Non-Expert 1 are almost always statistically different than those of the expert. We also evaluated the statistical difference

between the CE+Dice and the MAE+Dice losses and observed that the MAE+Dice loss provides overall better results for both non experts.

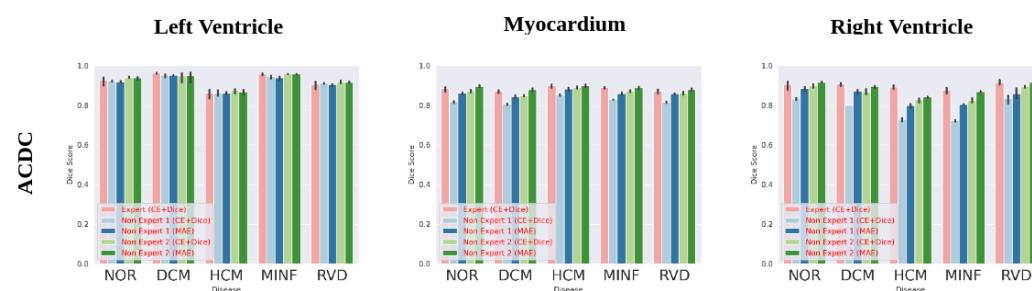


Figure 4. Dice score on ACDC per disease. NOR: Normal, DCM: Dilated cardiomyopathy, HCM: Hypertrophic cardiomyopathy, MINF: Myocardial infarction, RVD: Right ventricle disease.

Overall on M&Ms, while the expert obtained a better MYO mass error and a better RV EF error, the MYO HD was lower for Non-Expert 2 and the Dice score and the RV HD of Non-Expert 2 are statistically similar. These results underline the idea that well-trained non-expert and expert annotations could be used interchangeably to build reliable annotated cardiac data sets. In contrast, the number of non-experts we evaluated might be considered a limitation of our study; however, this still provides encouraging results for settings where experts are not readily available to annotate whole data sets, but can provide training to a non-expert to effectively annotate in their stead. We leave the investigation on more data sets to future works that could transpose the setup to more difficult problems and a larger number of non-experts. Our work supplements previous endeavors that rely on non-experts to annotate medical data sets; Heim et al. [20] showcased the ability of crowd-sourced expertise to reliably annotate the liver data set, although their approach proposes initial segmentations to the non-expert, which might have biased their decision. Likewise, Ganz et al. [21] proposed to make use of non-experts as a crowd-sourced error detection framework. In contrast, our approach evaluated the effectiveness of non-expert knowledge without any prior input. This further reinforces the idea that crowd-sourced medical annotations are a viable solution for the lack of data.

5. Conclusions

In this work, we studied the usefulness of training deep learning models with non-expert annotations for the segmentation of cardiac MR images. The need for medical experts was probed in a comparative study with non-physician sourced labels. Through framing the problem of relying on non-expert annotations as noisy data, we managed to obtain good performance on two public data sets, one of which was used to emulate an out-of-distribution data set. We found that training a deep neural network, regardless of its capacity (U-Net, Attention U-Net or ENet), with data labeled by a well-trained non-expert achieved comparable performance than on expert data. Moreover, the performance gap between the networks with non-expert and expert annotations on the out-of-distribution data set was less pronounced than the gap on the training data set. Future endeavors could focus on crowd sourcing large-scale medical data sets and tailoring approaches that take their noisiness into account.

Author Contributions: Conceptualization, P.-M.J. and A.L.; methodology, Y.S.; software, Y.S.; validation, P.-M.J. and A.L.; formal analysis, P.-M.J. and A.L.; investigation, Y.S. and P.-M.J.; resources, A.L.; data curation, Y.S.; writing—original draft, Y.S.; writing—review and editing, P.-M.J. and A.L.; supervision, P.-M.J. and A.L.; funding acquisition, A.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We would like to acknowledge Ivan Porcherot for the tremendous work in annotating the data sets.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
2. Bernard, O.; Lalande, A.; Zotti, C.; Cervenansky, F.; Yang, X.; Heng, P.A.; Cetin, I.; Lekadir, K.; Camara, O.; Ballester, M.A.G.; et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Trans. Med. Imaging* **2018**, *37*, 2514–2525. [CrossRef] [PubMed]
3. Painchaud, N.; Skandarani, Y.; Judge, T.; Bernard, O.; Jodoin, A.P.M. Cardiac Segmentation with Strong Anatomical Guarantees. *IEEE Trans. Med. Imaging* **2020**, *39*, 3703–3713. [CrossRef] [PubMed]
4. Venkataramani, R.; Ravishankar, H.; Anamandra, S. Towards Continuous Domain Adaptation for Medical Imaging. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging, Venice, Italy, 8–11 April 2019; pp. 443–446.
5. Maier-Hein, L.; Eisenmann, M.; Reinke, A.; Onogur, S.; Stankovic, M.; Scholz, P.; Arbel, T.; Bogunović, H.; Bradley, A.; Carass, A.; et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. Commun.* **2018**, *9*, 5217. [CrossRef] [PubMed]
6. Dataset List—A List of the Biggest Machine Learning Datasets. 2021. Available online: <https://www.datasetlist.com/> (accessed on 14 July 2021).
7. Amazon Mechanical Turk. 2021. Available online: <https://www.mturk.com/> (accessed on 14 July 2021).
8. Karimi, D.; Dou, H.; Warfield, S.K.; Gholipour, A. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Med. Image Anal.* **2020**, *65*, 101759. [CrossRef] [PubMed]
9. Can, Y.B.; Chaitanya, K.; Mustafa, B.; Koch, L.M.; Konukoglu, E.; Baumgartner, C.F. Learning to Segment Medical Images with Scribble-Supervision Alone. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer International Publishing: Cham, Switzerland, 2018; pp. 236–244.
10. Choudhary, A.; Tong, L.; Zhu, Y.; Wang, M. Advancing Medical Imaging Informatics by Deep Learning-Based Domain Adaptation. *Yearb. Med. Inf.* **2020**, *29*, 129–138. [CrossRef] [PubMed]
11. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *2*, 2672–2680.
12. Skandarani, Y.; Painchaud, N.; Jodoin, P.M.; Lalande, A. On the effectiveness of GAN generated cardiac MRIs for segmentation. *arXiv* **2020**, arXiv:2005.09026.
13. Girum, K.B.; Créhange, G.; Hussain, R.; Lalande, A. Fast interactive medical image segmentation with weakly supervised deep learning method. *Int. J. Comput. Assist. Radiol. Surg.* **2020**, *15*, 1437–1444. [CrossRef] [PubMed]
14. Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 843–852.
15. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2015; pp. 234–241.
16. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.J.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.G.; Hammerla, N.; Kainz, B.; et al. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv* **2018**, arXiv:1804.03999.
17. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv* **2016**, arXiv:1606.02147.
18. Ghosh, A.; Kumar, H.; Sastry, P. Robust loss functions under label noise for deep neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–10 February 2017; Volume 31.
19. Campello, V.M.; Gkontra, P.; Izquierdo, C.; Martín-Isla, C.; Sojoudi, A.; Full, P.M.; Maier-Hein, K.; Zhang, Y.; He, Z.; Ma, J.; et al. Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge. *IEEE Trans. Med. Imaging* **2021**. [CrossRef]
20. Heim, E.; Roß, T.; Seitel, A.; März, K.; Stieltjes, B.; Eisenmann, M.; Lebert, J.; Metzger, J.; Sommer, G. Large-scale medical image annotation with crowd-powered algorithms. *J. Med. Imaging* **2018**, *5*, 1. [CrossRef] [PubMed]
21. Ganz, M.; Kondermann, D.; Andrusis, J.; Knudsen, G.M.; Maier-Hein, L. Crowdsourcing for error detection in cortical surface delineations. *Int. J. Comput. Assist. Radiol. Surg.* **2016**, *12*, 161–166. [CrossRef] [PubMed]