

Automatic Segmentation of Lymphatic Perfusion in Patients with Congenital Single Ventricle Defects

Marietta Stegmaier¹, Johanna P. Müller¹, Christian Schröder², Thomas Day³,
Michela Cuomo², Oliver Dewald², Sven Dittrich², Bernhard Kainz^{1,4}

¹FAU Erlangen-Nürnberg, DE

²University Hospital Erlangen, DE

³King's College London, UK

⁴Imperial College London, UK

johanna.paula.mueller@fau.de

Abstract. The Fontan circulation is the surgical end-point for a variety of single-ventricle congenital heart lesions. While recent decades have witnessed substantial improvements in survival rates, the associated physiology remains susceptible to severe complications such as protein-losing enteropathy and plastic bronchitis. These complications are often indicative of abnormal congestion in the lymphatic system, underscoring their significance as harbingers of impending comorbidities. An accurate assessment of congestion severity requires the detailed annotation of lymphatic perfusion patterns in each volumetric scan slice. The manual labelling of such intricate data is time-consuming and demands a high level of expertise, rendering it unfeasible within the confines of standard clinical protocols. We use a curated database consisting of manually annotated *T2*-weighted Magnetic Resonance Imaging (MRI) scans from 71 Fontan patients post-surgery. Following the current state-of-the-art method for biomedical image segmentation, we evaluate its performance on multiple independent test sets regarding the degree of severity and imaging quality. Incorporating the best-performing model, we have developed a user-friendly interface for the automatic segmentation of lymphatic malformations, which will be published before the conference starts.

1 Introduction

Single ventricle malformations are relatively rare diagnoses ($\sim 1.25\%$ of infants) manifesting a genetic component [1], and when left untreated, they can be fatal in the neonatal period. To address this, the majority of univentricular abnormalities undergo a multi-stage surgical protocol, ultimately resulting in the establishment of the Fontan circulation. While there have been significant improvements in the long-term survival rates for Fontan patients over recent decades, their unique hemodynamics can lead to severe complications, such as protein-losing enteropathy and plastic bronchitis. Detecting early signs of these comorbidities is crucial. They are often manifested as abnormalities in lymphatic perfusion patterns, resulting from congestion in the venous system. To facilitate the recognition of these complications, early diagnosis, and the categorisation of lymphatic malformations into different degrees of severity is imperative. *T2*-weighted MRI of the thorax and abdomen has proven to be a valuable tool in this classification process, as it portrays lymphatic fluid with high intensity [2]. We established a curated

dataset comprising Fontan patients who received treatment at the Erlangen University Hospital between the years 2007 and 2020. As part of their routine follow-up examinations, conducted six months after Fontan Completion, these patients underwent a standard MRI protocol. At the time of our study, our database encompassed records from 71 patients. Collaborating with a pediatric cardiologist, we manually annotated lymphatic perfusion patterns using MRI-specific labelling software. Manual labelling introduces fuzzy labels into the learning pipeline, e.g., induced by different levels of annotator expertise, specific ambiguity inherited in the data, varying granularity and complexity of annotations, and annotator bias, represented by specific uncertainty and ambiguity. In our specific case, semantic annotation required a pixel-wise binary classification of lymphatic and non-lymphatic structures. Given the complexity of lymphatic perfusion patterns, we had to consider several factors to distinguish between foreground and background. The defining characteristic of lymphatic fluid is its heightened intensity in $T2$ -weighted MR images, which proves insufficient due to the presence of numerous structures exhibiting similar brightness. Thus, anatomical expertise played a pivotal role in excluding gastrointestinal areas and blood vessels. We utilised quality ratings assigned by the pediatric cardiologist and focused on developing models capable of making accurate predictions with a limited amount of labelled data through data augmentation, cross-validation, and distinctions based on image quality, severity levels, and anatomical regions. So far, the authors could only identify one similar study [3] with fewer patients ($n = 48$) for measuring lymphatic burden, and a related one [4] with fewer patients ($n = 22$) for quantifying thoracic lymphatic flow patterns using dynamic contrast-enhanced MR lymphangiography.

Contributions.

1. We manually annotated lymphatic perfusion abnormalities in $T2$ -weighted MRI samples of Fontan patients six months after surgery completion. We curated the collection of annotated samples to a novel database consisting of 71 samples.
2. We test and evaluate the performance of state-of-the-art segmentation models on ill-defined labels in few-shot scenarios. Through 5-fold cross-validation and ensemble prediction, we show the generalisability of our trained models in challenging cases.
3. To complete our contributions, we offer a user-friendly interface that facilitates the automatic prediction of lymphatic perfusion patterns in input data, also with probability maps for prediction.

2 Materials and Methods

Timely classification of patients into different severity types based on lymphatic malformations is crucial due to potential complications. Manual annotation, while time-consuming and impractical for routine clinical use, can be significantly streamlined by automated segmentation tools, taking advantage of recent deep learning advancements. To facilitate supervised learning, a dataset of $T2$ -weighted MR images from Fontan patients with lymphatic malformations has been annotated.

Dataset curation. In Erlangen University Hospital, all patients with univentricular heart defects undergo a standardised protocol which was first introduced in 2007. Apart from numerous other steps, the plan includes a retrospective analysis of *T2*-weighted MR images with fat saturation six months after the Fontan Completion to assess potential lymphatic malformations. In the years 2007 – 2020, a total of 71 children received this scan creating the basis for a dataset that can be extended in the future. The dataset properties and patient distribution are presented in the underlying medical study [2], which compiles abnormal abdominal lymphatic perfusion patterns and investigates their impact on serum protein readings. Reference annotations were created for all patients with the assistance of a specialised pediatric cardiologist. The semantic annotation involved a binary pixel-wise classification into lymphatic and non-lymphatic structures. Due to the complexity of the perfusion patterns, several aspects had to be considered to differentiate between the fore- and background. Considering only the high intensity of lymphatic fluid in *T2*-weighted MR images, this criterion is not sufficient because multiple other structures are depicted in a similar brightness. Hence, anatomical knowledge was vital to rule out gastrointestinal areas or blood vessels. Apart from that, previous research about common locations of lymphatic abnormalities was considered during the decision-making process [2, 5]. To factor these disparities into the evaluation, the scans

Tab. 1. Data Folds for Model Training.

	Model	1	2	3	4	5	6	7.1	7.2	7.3	Dataset
	Samples	81	81	81	60	45	80	12	46	20	105
Region	Abdominal	28	28	28	19	17	27	3	17	8	36
	Whole-body	25	25	25	22	11	26	6	12	4	33
	Thoracic	28	28	28	19	17	27	3	17	8	36
	0	3	3	3	3	1	3	0	0	0	4
Quality (n)	1	12	12	12	12	6	14	12	0	0	18
	2	46	46	46	32	20	42	0	46	0	52
	3	20	20	20	12	14	20	0	0	18	26
	4	0	0	0	1	4	1	0	0	2	5
Severity (n)	1	2	2	2	0	7	5	1	1	0	7
	2	28	28	28	0	38	29	3	15	10	38
	3	50	50	50	58	0	45	8	29	10	58
	4	1	1	1	2	0	1	0	1	0	2

were rated concerning their quality by the pediatric cardiologist. Given the large differences in intensity between *BLADE* and *SPACE* MRI sequences, the latter were exempt from the general quality assessment. The remaining images were assigned ranks from one to four, with one representing the best and four representing the worst quality, see quality distribution in Tab. 1. Reasons for low quality were large slice thickness, motion artefacts and imaging resolution. As a final step of the data curation, the cases were assigned a severity type describing the degree of lymphatic malformations. All patients in the database had been previously classified by two independent experts according to the system introduced by [2, 5]. The para-aortic, portal-venous, and thoracic regions are considered separately for this evaluation, resulting in three ranks per patient. For

thoracic-only scans, the thoracic ranks by both raters were averaged (see, e.g., Fig. 1); abdominal-only images were assigned the mean of the para-aortic and portal-venous ratings by both raters. The severity types mapped to whole-body scans comprised the ranks of all three regions.

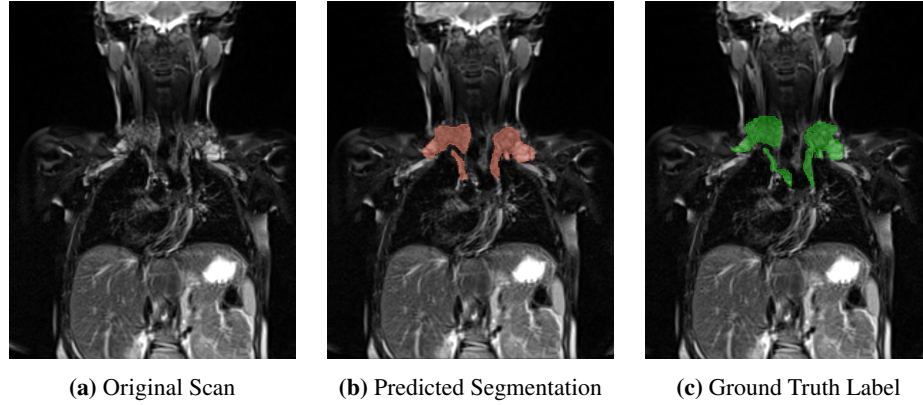


Fig. 1. Example of a high-performing MRI scan with characteristics: Region "Thoracic-only", Severity type "2", Quality rank "2".

Segmentation Framework. To create a standardised and robust segmentation tool, we chose to use the nnU-Net [6] as a state-of-the-art segmentation model. This self-configuring model preserves the overall structure of conventional U-Nets, yet systematizes the re-design of the pipeline for individual tasks. One key advantage of the network is its ability to learn efficiently from minimal training data, which is particularly important in our project since the available training cases of Fontan patients with lymphatic malformations are very limited. Another benefit of the nnU-Net is its accessibility without demanding expert knowledge or advanced computing resources to provide fast and accurate results, especially, for future clinical deployment.

3 Results

In Table 2, the 2D-, 3D- and Ensemble U-Nets yield similar results. A more detailed evaluation reveals that the superiority of specific configurations depends on the anatomical region of the scans. Abdominal-only images perform better with the 2D network, while thoracic-only scans tend to favour the 3D U-Net, and no clear preference between configurations is observed for whole-body cases. Concerning the different training schemes for the models, one can note that the performance of the networks improves substantially the more data augmentation functions are included. This observation also applies to mirroring, comparing Model 3 with the default augmentation and Model 2 without mirroring. Independent of the specific configurations and data folds, the anatomical region of a scan largely dictates the results of the segmentation as depicted in Table

Tab. 2. Evaluation scores for predictions with the best checkpoint (best scores of each configuration shown in bold) (DSC: Dice Similarity Coefficient; IoU: Intersection over Union).

		2D Configuration		3D Configuration		Ensemble of 2 D and 3 D	
		DSC	IoU	DSC	IoU	DSC	IoU
Model	1	0.39 $\pm$ 0.21	0.24 $\pm$ 0.16	0.38 $\pm$ 0.22	0.23 $\pm$ 0.18	0.39 $\pm$ 0.22	0.24 $\pm$ 0.18
	2	0.43 $\pm$ 0.20	0.27 $\pm$ 0.16	0.43 $\pm$ 0.19	0.27 $\pm$ 0.16	0.43 $\pm$ 0.20	0.27 $\pm$ 0.17
	3	0.45 $\pm$ 0.20	0.29 $\pm$ 0.17	0.44 $\pm$ 0.20	0.28 $\pm$ 0.19	0.45 $\pm$ 0.22	0.29 $\pm$ 0.20
	4	0.41 $\pm$ 0.18	0.26 $\pm$ 0.14	0.38 $\pm$ 0.16	0.23 $\pm$ 0.12	0.40 $\pm$ 0.17	0.25 $\pm$ 0.14
	5	0.40 $\pm$ 0.19	0.25 $\pm$ 0.15	0.37 $\pm$ 0.18	0.23 $\pm$ 0.14	0.38 $\pm$ 0.19	0.23 $\pm$ 0.15
	6	0.45 $\pm$ 0.17	0.29 $\pm$ 0.16	0.46 $\pm$ 0.17	0.30 $\pm$ 0.15	0.46 $\pm$ 0.18	0.30 $\pm$ 0.17
	7	0.38 $\pm$ 0.23	0.23 $\pm$ 0.20	0.39 $\pm$ 0.22	0.24 $\pm$ 0.21	0.40 $\pm$ 0.24	0.25 $\pm$ 0.22

Tab. 3. Average scores for anatomical regions, qualities, and severities (best scores of each category shown in bold) (DSC: Dice Similarity Coefficient; IoU: Intersection over Union; NSD: Normalized Surface Distance).

		DSC	IoU	Box-IoU	NSD
Region	Abdominal	0.42 $\pm$ 0.19	0.27 $\pm$ 0.16	0.29 $\pm$ 0.10	0.70 $\pm$ 0.13
	Whole-body	0.26 $\pm$ 0.16	0.15 $\pm$ 0.11	0.22 $\pm$ 0.08	0.60 $\pm$ 0.23
	Thoracic	0.49 $\pm$ 0.18	0.32 $\pm$ 0.18	0.34 $\pm$ 0.12	0.74 $\pm$ 0.26
Quality (n)	0	0.19 $\pm$ 0.17	0.10 $\pm$ 0.12	0.18 $\pm$ 0.11	0.55 $\pm$ 0.25
	1	0.33 $\pm$ 0.19	0.20 $\pm$ 0.14	0.27 $\pm$ 0.10	0.69 $\pm$ 0.22
	2	0.41 $\pm$ 0.18	0.26 $\pm$ 0.15	0.29 $\pm$ 0.13	0.71 $\pm$ 0.28
	3	0.41 $\pm$ 0.17	0.26 $\pm$ 0.13	0.30 $\pm$ 0.11	0.78 $\pm$ 0.20
	4	0.44 $\pm$ 0.30	0.28 $\pm$ 0.29	0.30 $\pm$ 0.15	0.39 $\pm$ 0.28
Severity (n)	1	0.26 $\pm$ 0.14	0.15 $\pm$ 0.10	0.20 $\pm$ 0.09	0.59 $\pm$ 0.26
	2	0.39 $\pm$ 0.19	0.24 $\pm$ 0.16	0.28 $\pm$ 0.11	0.67 $\pm$ 0.29
	3	0.43 $\pm$ 0.21	0.27 $\pm$ 0.19	0.31 $\pm$ 0.13	0.71 $\pm$ 0.28
	4	0.39 $\pm$ 0.13	0.24 $\pm$ 0.08	0.29 $\pm$ 0.07	0.76 $\pm$ 0.12

3. Thoracic-only images are predicted most accurately, followed by abdominal-only- and, lastly, whole-body cases. Apart from the anatomical region, the severity ranks assigned to the scans seem to have a noticeable impact on the scores, with types 3 and 4 outperforming groups 1 and 2. In contrast to severity ranks, image qualities have a minor impact on the results except from the SPACE sequence (quality 0) with poor performance due to its contrast characteristics.

4 Discussion

The choice of model configuration is crucial for optimising segmentation results in the context of lymphatic abnormalities, particularly when considering different anatomical regions. The 2D and 3D U-Net models and their ensemble do not show a significant trend in performance. While the 2D model works well for abdominal-only scans, the 3D model excels in segmenting thoracic-only images due to the need for three-dimensional context. The ensemble model does not consistently outperform the others. The comparison

of quality distributions reveals that the quality of training images does not significantly impact segmentation performance. In the context of lymphatic perfusion patterns, the rarity of positive foreground pixels in scans can lead to high precision but low sensitivity, which is problematic for clinical diagnosis. Comparing different models with varying foreground-to-background ratios in training sets, it is evident that the size of the training set has a more significant impact on performance than the fore-to-background ratio, highlighting the importance of having a substantial number of training cases in the dataset. For challenging scans with reduced intensity contrast between lymphatic abnormalities and surrounding structures, e.g., where the gastrointestinal tract's similar intensity and texture further complicate the task, the models tend to under-segment the portal-venous region instead of over-segmenting the gastrointestinal tract, as the latter is a larger structure, and the models favour not assigning a foreground label when uncertain. Although the current model performance may benefit from more extensive training data, the efforts to enhance the database of Fontan patients at Erlangen University Hospital are ongoing. To aid clinical professionals in the annotation process, a user interface has been developed that provides predicted segmentation masks and probability maps.

Acknowledgement. The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) under the NHR projects b143dc and b180dc. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683. Additional support was also received by the ERC - project MIA-NORMAL 101083647, DFG KA 5801/2-1, INST 90/1351-1 and by the state of Bavaria.

References

1. Marin-Garcia J. Post-genomic cardiology. Academic Press, 2011.
2. Schroeder C, Moosmann J, Cesnjevar R, Purbojo A, Rompel O, Dittrich S. A classification of abdominal lymphatic perfusion patterns after Fontan surgery. *European Journal of Cardio-Thoracic Surgery*. 2022;62(4):ezac103.
3. Vaikom House A, David D, Aguet J, Dipchand AI, Honjo O, Jean-St-Michel E et al. Quantification of lymphatic burden in patients with Fontan circulation by T2 MR lymphangiography and associations with adverse Fontan status. *European Heart Journal-Cardiovascular Imaging*. 2023;24(2):241–9.
4. Zheng Q, Itkin M, Fan Y. Quantification of thoracic lymphatic flow patterns using dynamic contrast-enhanced MR lymphangiography. *Radiology*. 2020;296(1):202–7.
5. Biko DM, DeWitt AG, Pinto EM, Morrison RE, Johnstone JA, Griffis H et al. MRI evaluation of lymphatic abnormalities in the neck and thorax after Fontan surgery: relationship with outcome. *Radiology*. 2019;291(3):774–80.
6. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*. 2021;18(2):203–11.