Exploring ensemble applications for multi-sequence myocardial pathology segmentation

Abstract. We tested different loss functions and hyper-parameters using a 2D U-Net architecture (resnet34 backbone) with five-fold cross-validation on the training data. Pathology specific sequence data (e.g. LGE for scar and T_2 for edema) was used as a sole input for training and in combination with all sequences. We wanted to address the question whether for limited training data it is beneficial to incorporate prior knowledge by predicting classes with their appropriate sequence or if a neural network is able to infer these relationships from a multi-sequence dataset. In addition, we aimed to create a model zoo, combining predictions from models with high performance on individual classes. Images were cropped to the central 256x256 region as this contained the region of interest in all cases. To improve robustness and learn more general features extensive data augmentation was used, including both MR artifacts (motion, noise) and standard image transformations (zoom, rotation, brightness, contrast). Variations of training data, loss functions and hyper-parameters led to 21 models trained. The multi-sequence model was trained using all image sequences input via color channels producing pixel-level segmentation for all six classes (background, left ventricle, right ventricle, myocardium, edema, and scar). Cross-entropy as a loss function performed best (metric: dice) for non-pathologic tissue, while pathology weighted focal-loss (0.35 for both scar and edema) had best mean performance on scar and edema. The mean dice scores over all cross validation folds using the multi-sequence model were: $dice_{LV}$=0.853, $dice_{RV}$=0.787, $dice_{MY}$=0.695, $dice_{s\text{-}car}$=0.438, $dice_{edema}$=0.227. The specific models reached dice scores of: $dice_{s\text{-}car}$=0.479, $dice_{edema}$=0.276

These results indicate that the employed neural network is capable of learning multi-sequence segmentation end-to-end. Combining different outputs from a model zoo based on objective criteria proofed difficult.

Keywords: Deep Learning, U-Net, Ensemble, Segmentation, Cardiac MRI

1 Introduction

Cardiac magnetic resonance (CMR) imaging applies methods to investigate cardiac function and pathologies non-invasively. Different measurement sequences are used to produce images with different contrast, enabling diagnosis of varying pathologic tissue alterations. It is common practice to segment the left ventricle and the myocardium to assess clinically relevant parameters like ejection fraction, stroke volume and myocardial mass as well as wall motion. Scar volume, as a result of acute myocardial infarction, has significant prognostic value for outcome prediction and treatment, thus, increasing the importance of accurate pathology segmentation. In clinical practice, such segmentations are commonly done semi-automatically. Fully automatic segmentation

algorithms have been proposed using different methods, including artificial neural networks [1, 2]. However, these networks are usually trained on a single sequence and a subset of tissue/pathology classes. For the prediction of multiple pathologic tissue alterations in parallel, it might be beneficial to train segmentation networks, which combine information from multiple sequences. In the MyoPS 2020 challenge, three different sequences (bSSFP, LGE, SPAIR) were measured for each of 45 patients, providing ground truth segmentation for left ventricular (LV) blood pool, right ventricular (RV) blood pool, LV myocardium (MY), edema, and scar for 25 patients. All data was provided aligned (MvMM method [3, 4]) in a common space with identical spatial resolution by the organizers. The aim of the challenge was to create an algorithm for pixel-wise segmentation of the pathology classes edema and scar. In this study, we employed variations of individual neural networks as well as a model ensemble, combining models with high performance on individual morphologic classes.

2 Material & Methods

2.1 (Hardware and) Software

The experiments and parameter search were done in Google Colab GPU instances. For the final training and prediction, we used our local HPC i. 8x Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz ii. 512 GB of memory iii. 1x NVIDIA Tesla K80 with 12 GB of memory.

We built our model using open source software including python 3.7.7, pytorch 1.5.1 [5], fastai2 0.0.17 [6], torchio 0.15.5 [7], MONAI 0.2.0, nibabel 3.2.1 [8] as well as their dependencies. Our model and code is openly available on GitHub and zenodo (links will be included after anonymous review).

2.2 Processing pipeline and architecture

We converted all images from nifti to png format saving each slice as one image with sequences combined as color channels. Additionally, each sequence was saved independently as a grey-scale image. We tried normalization of the LGE and T_2 images using contrast limited adaptive histogram equalization (CLAHE) [9]. In this step, in addition to the original images, transformed images with simulated MR artifacts (motion and noise) were produced using torchio [7]. These images were used to train U-Nets [10] with a resnet34 [11] backbone (initialized using ImageNet [12] weights) with further augmentations (rotation, brightness, contrast) with fastai2 [6, 13]. Performance of different hyper-parameter settings were evaluated using dice scores from five-fold cross-validation. The same split was used for all experiments and every data set was part of the validation set at least once.

2.3 Hyper-parameter search

Preliminary experiments. In preliminary experiments the effect of contrast enhancements using CLAHE as well as cropping vs resizing to 256x256 pixels were tested.

Systematic experiments. For the general multi-channel/multi-class networks, different losses were tested. Cross-entropy loss (ce) was compared to differently weighted focal loss [14]. We experimented with some classes receiving higher weights, while the other classes received balanced weights:

- all classes with equal weights (balanced)
- myocardium (0.2, 0.3), edema (0.2, 0.3) and scar (0.2, 0.3), label: multi_pathoMyo
- edema (0.2, 0.35, 0.49) and scar (0.2, 0.35, 0.49), label: multi_patho
- edema (0.2, 0.4, 0.6, 0.8, 0.99), label: multi_edema
- scar (0.2, 0.4, 0.6, 0.8, 0.99), label: multi_scar

Additionally, pathology specific networks (t2_edemaOnly, lge_scarOnly) were trained on their corresponding sequence only (edema with T_2 and scar with LGE) using two different weightings of the focal loss (0.5 and 0.8). In total 21 networks were trained this way for 30 epochs (10 epochs frozen, 20 epochs unfrozen) and a base learning rate of 10^{-3}.

Targeted experiments. The best performing networks from the systematic experiments were selected based on mean dice score over all cross validations. For LV, RV and myocardium only the network with the highest dice score was selected. For the pathology classes first the network with highest mean dice over both classes was selected, then for each class the two remaining networks with highest individual dice scores in the respective class were selected. This way a total of six networks were selected. These networks were trained for 60 epochs (20 frozen, 40 unfrozen) in order to assess benefits of prolonged training duration.

Final training. For the evaluation on the test set, the six networks from the targeted experiments were trained from scratch using all 25 data sets for training and no validation set. Training was done for 60 epochs (20 frozen, 40 unfrozen), since average performance was increased with prolonged training duration.

2.4 Ensemble method

The networks were trained with different foci, which led to different strengths and weaknesses. Therefore, we combined predictions from the different networks in a bagging approach. This combination included predictions from all six networks from the final training. Class probabilities were averaged over all networks, taking into account that the specialized networks only returned predictions for their respective pathology class. The final prediction for each pixel was the argmax of these averages (**Fig. 1**).

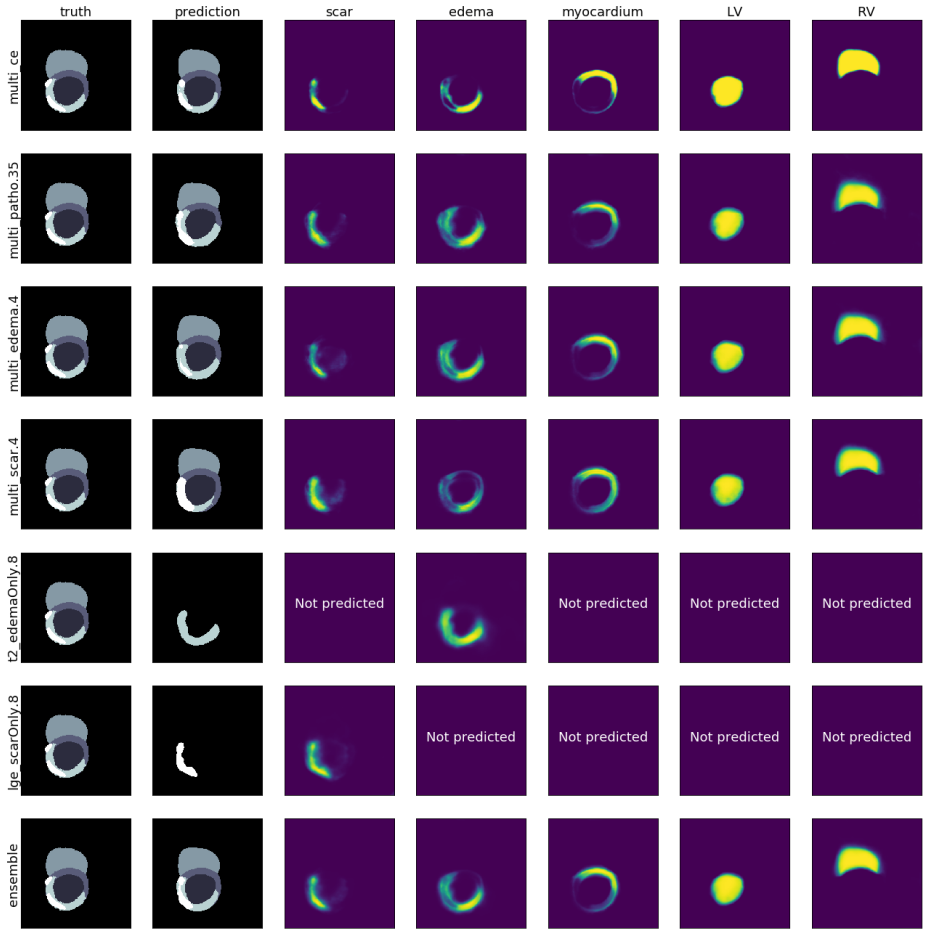

Fig. 1. Probability maps for all classes and derived prediction (second column) for the six networks and the ensemble, compared to the ground truth (first column) for a single slice of the training data. The result of the ensemble method (bottom row) is the mean over the probability maps of the six separate networks above.

3 Results

3.1 Cross-validation results on training set

Preliminary experiments indicated that not using CLAHE and cropping to 256x256 pixels yields better results than normalization or resizing. Thus, only cropping was used in the systematic experiments. In the systematic experiments, the network with cross-entropy loss reached the best results for LV, RV and myocardium segmentation with

mean values of $dice_{LV}$=0.855, $dice_{RV}$=0.783 and $dice_{MY}$=0.696. The best mean performance on both pathology classes: mean($dice_{edema}$, $dice_{scar}$)=0.345 was achieved using the multi-channel network (multi_patho) with weights of 0.35 for both pathology classes. Of the remaining networks the highest dice on scar was reached by the multi-channel network (multi_scar, weight: 0.4) and by the specialized LGE network (lge_scarOnly, weight 0.8) (**Fig. 2**), while the specialized T_2 network (t2_edemaOnly, weight: 0.8) and the multi-channel network (multi_edema, weight: 0.4) reached the highest dice scores for edema (**Fig. 2**). Longer training improved dice scores for almost all classes and networks (**Fig. 3**, **Table 1**).

Table 1. Mean performance of the targeted experiment networks over the five-fold cross-validation after 60 epochs of training. Highest dice for each class in bold.

Network	$dice_{LV}$	$dice_{MY}$	$dice_{RV}$	$dice_{edema}$	$dice_{scar}$	mean dice edema, scar
multi_patho.35	0.829	0.649	0.773	0.246	0.441	**0.343**
multi_scar.4	0.850	0.690	0.779	0.202	**0.479**	0.341
multi_ce	**0.853**	**0.695**	**0.787**	0.227	0.438	0.333
multi_edema.4	0.843	0.664	0.781	0.261	0.400	0.330
lge_scarOnly.8	-	-	-	-	0.467	-
t2_edemaOnly.8	-	-	-	**0.276**	-	-

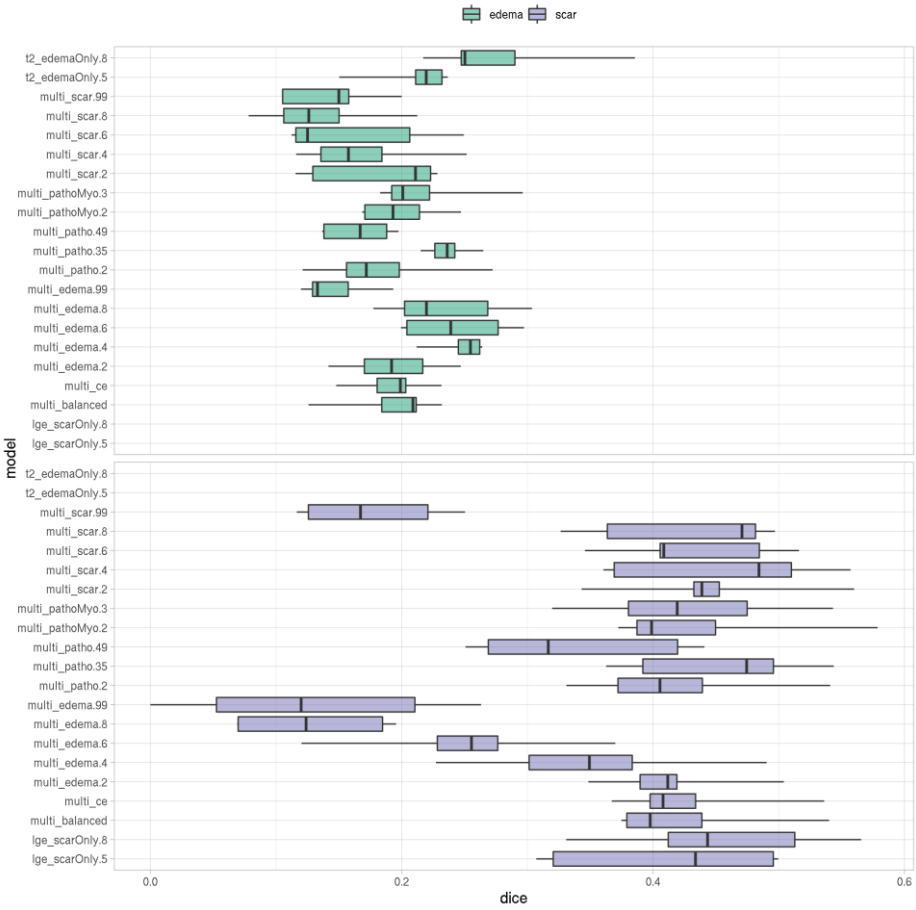

Fig. 2. Dice score for edema (top panel) and scar (bottom panel) over the five cross-validation folds of each of the 21 networks from the systematic parameter search. Naming of models: input channels (multi, lge, t2), focused classes (scar, edema, patho: scar+edema, pathoMyo: scar+edema+MY, ce for cross-entropy and balanced have identical weight for all classes) and weight for those classes as suffix.

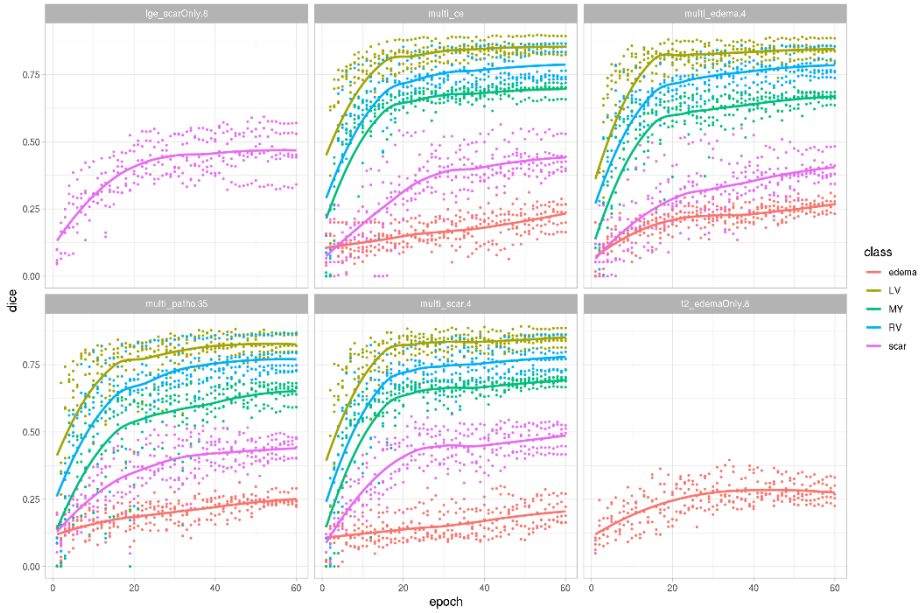

Fig. 3. Dice scores throughout training of the six networks from the targeted experiments. Data for all five cross-validation folds is shown with loess-smoothed lines for each class. The first 20 epochs were trained with frozen weights, the remaining 40 with unfrozen weights.

3.2 Performance on test set

Evaluation results on the test set were provided by the challenge organizers for two models, the multi_patho.35 network and the ensemble method. The ensemble reached better performance with mean ± standard deviation of $dice_{scar}$=0.620±0.240 and $dice_{edema+scar}$=0.665±0.137 compared to $dice_{scar}$=0.593±0.232 and $dice_{edema+scar}$=0.611±0.111 for the single network. For all but one patient dice scores for scar were greater than 0 indicating at least some overlap between truth and prediction.

4 Discussion

It is possible to train neural networks both on separate sequences and on multiple sequences with good performance. For scar the reported dice score is higher than that achieved by individual observers reported as 0.524±0.158 [4]. Segmentation quality can be further improved by training a model zoo with focus on different classes and combining their predictions using a bagging ensemble method. We showed that it is even possible to combine predictions from networks that were trained on different input data (channels) with a different set of output channels using averaging. While these

results are promising, further experiments are needed to optimize the hyper-parameters for this challenging task. Additionally more and diverse training data is needed to train an algorithm with good performance and to reliably estimate its performance on unseen data.

References

1. Bai, W., Sinclair, M., Tarroni, G., Oktay, O., Rajchl, M., Vaillant, G., Lee, A.M., Aung, N., Lukaschuk, E., Sanghvi, M.M., Zemrak, F., Fung, K., Paiva, J.M., Carapella, V., Kim, Y.J., Suzuki, H., Kainz, B., Matthews, P.M., Petersen, S.E., Piechnik, S.K., Neubauer, S., Glocker, B., Rueckert, D.: Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. Journal of Cardiovascular Magnetic Resonance 20, (2018)
2. Baumgartner, C.F., Koch, L.M., Pollefeys, M., Konukoglu, E.: An Exploration of 2D and 3D Deep Learning Techniques for Cardiac MR Image Segmentation. pp. 111-119. Springer International Publishing, Cham (Year)
3. Zhuang, X.: Multivariate Mixture Model for Cardiac Segmentation from Multi-Sequence MRI. pp. 581-588. Springer International Publishing, (Year)
4. Zhuang, X.: Multivariate Mixture Model for Myocardial Segmentation Combining Multi-Source Images. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 2933-2946 (2019)
5. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. 8024--8035 (2019)
6. Howard, J., Gugger, S.: fastai: A Layered API for Deep Learning. Information 11, 108 (2020)
7. Pérez-García, F., Sparks, R., Ourselin, S.: TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. arXiv:2003.04696 [cs, eess, stat] (2020)
8. Brett, M., Markiewicz, C.J., Hanke, M., Côté, M.-A., Cipollini, B., McCarthy, P., Jarecka, D., Cheng, C.P., Halchenko, Y.O., Cottaar, M., Ghosh, S., Larson, E., Wassermann, D., Gerhard, S., Lee, G.R., Wang, H.-T., Kastman, E., Kaczmarzyk, J., Guidotti, R., Duek, O., Rokem, A., Madison, C., Morency, F.C., Moloney, B., Goncalves, M., Markello, R., Riddell, C., Burns, C., Millman, J., Gramfort, A., Leppäkangas, J., Sólon, A., van den Bosch, J.J.F., Vincent, R.D., Braun, H., Subramaniam, K., Gorgolewski, K.J., Raamana, P.R., Nichols, B.N., Baker, E.M., Hayashi, S., Pinsard, B., Haselgrove, C., Hymers, M., Esteban, O., Koudoro, S., Oosterhof, N.N., Amirbekian, B., Nimmo-Smith, I., Nguyen, L., Reddigari, S., St-Jean, S., Panfilov, E., Garyfallidis, E., Varoquaux, G., Legarreta, J.H., Hahn, K.S., Hinds, O.P., Fauber, B., Poline, J.-B., Stutters, J., Jordan, K., Cieslak, M., Moreno, M.E., Haenel, V., Schwartz, Y., Baratz, Z., Darwin, B.C., Thirion, B., Papadopoulos Orfanos, D., Pérez-García, F., Solovey, I., Gonzalez, I., Palasubramaniam, J., Lecher, J., Leinweber, K., Raktivan, K., Fischer, P., Gervais, P., Gadde, S., Ballinger, T., Roos, T., Reddam, V.R., freec84: nibabel. vol. 10.5281/zenodo.591597. Zenodo (2020)

9. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., Terhaarromeny, B., Zimmerman, J.B., Zuiderveld, K.: Adaptive Histogram Equalization and Its Variations. Comput Vision Graph 39, 355-368 (1987)
10.Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. pp. 234-241. Springer International Publishing, Cham (Year)
11.He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778. (Year)
12.Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 115, 211-252 (2015)
13.Ankenbrand, M.J., Lohr, D., Schlötelburg, W., Reiter, T., Wech, T., Schreiber, L.M.: A Deep Learning Based Cardiac Cine Segmentation Framework for Clinicians - Transfer Learning Application to 7T. medRxiv 2020.2006.2015.20131656 (2020)
14.Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal Loss for Dense Object Detection. arXiv:1708.02002 [cs] (2017)