Exploring DeepMedic for the purpose of segmenting white matter hyperintensity lesions

Fiona Lippert[a], Bastian Cheng[b], Amir Golsari[b], Florian Weiler[a], Johannes Gregori[c], Götz Thomalla[b], and Jan Klein[a]

[a]Fraunhofer MEVIS, Center for Medical Image Computing, Bremen, Germany
[b]University Hospital Hamburg-Eppendorf, Germany
[c]mediri GmbH, Germany

ABSTRACT

DeepMedic, an open source software library based on a multi-channel multi-resolution 3D convolutional neural network, has recently been made publicly available for brain lesion segmentations. It has already been shown that segmentation tasks on MRI data of patients having traumatic brain injuries, brain tumors, and ischemic stroke lesions can be performed very well. In this paper we describe how it can efficiently be used for the purpose of detecting and segmenting white matter hyperintensity lesions. We examined if it can be applied to single-channel routine 2D FLAIR data. For evaluation, we annotated 197 datasets with different numbers and sizes of white matter hyperintensity lesions. Our experiments have shown that substantial results with respect to the segmentation quality can be achieved. Compared to the original parametrization of the DeepMedic neural network, the timings for training can be drastically reduced if adjusting corresponding training parameters, while at the same time the Dice coefficients remain nearly unchanged. This enables for performing a whole training process within a single day utilizing a NVIDIA GeForce GTX 580 graphics board which makes this library also very interesting for research purposes on low-end GPU hardware.

Keywords: Machine learning, including deep learning, segmentation, applications: brain

1. INTRODUCTION

The detection and quantification of white matter lesions (WML), e.g., as an expression of the severity of cerebral micro-angiopathy, are of great clinical and scientific importance. An early identification and systematic measurement is the prerequisite for effective prevention and further investigation of the influence of microangiopathy on cognition, dementia risk and stroke risk in observation and therapy studies. The segmentation of WML is the basis for this.

However, as of now there is no established reliable method for clinical use. The heterogeneous, often inhomogeneous signal behavior of WML as well as the interference with other structural changes make automatic segmentation approaches difficult. At the same time, a method for automatic segmentation should provide comparable and reproducible results independent of the scanner, field strength and data quality.

By comparison, non-monitored methods can be used without the use of a training database. Segmentation is carried out here rather by modeling of previous knowledge about position, size, intensity distribution, etc. of lesions. The methods are based, for example, on multi-spectral thresholding,[1] fuzzy clustering[2,3] or fuzzy interference systems.[4] A further approach is to classify healthy tissue, and to define lesions as outliers.[5,6] A detailed overview of published approaches can be found in.[7]

Recently, first approaches based on the concepts of deep learning[8] have been proposed. First publications on the classification of WML[9,10] also show promising results. DeepMedic[11] is an efficient multiscale 3D convolutional neural network (CNN) for the segmentation of brain lesions. It consists of two pathways with 11 layers each,

Note that this is an extended version.

Further author information: (Send correspondence to Jan Klein)
Jan Klein: E-mail: jan.klein@mevis.fraunhofer.de, Telephone: 49 421 218 59239

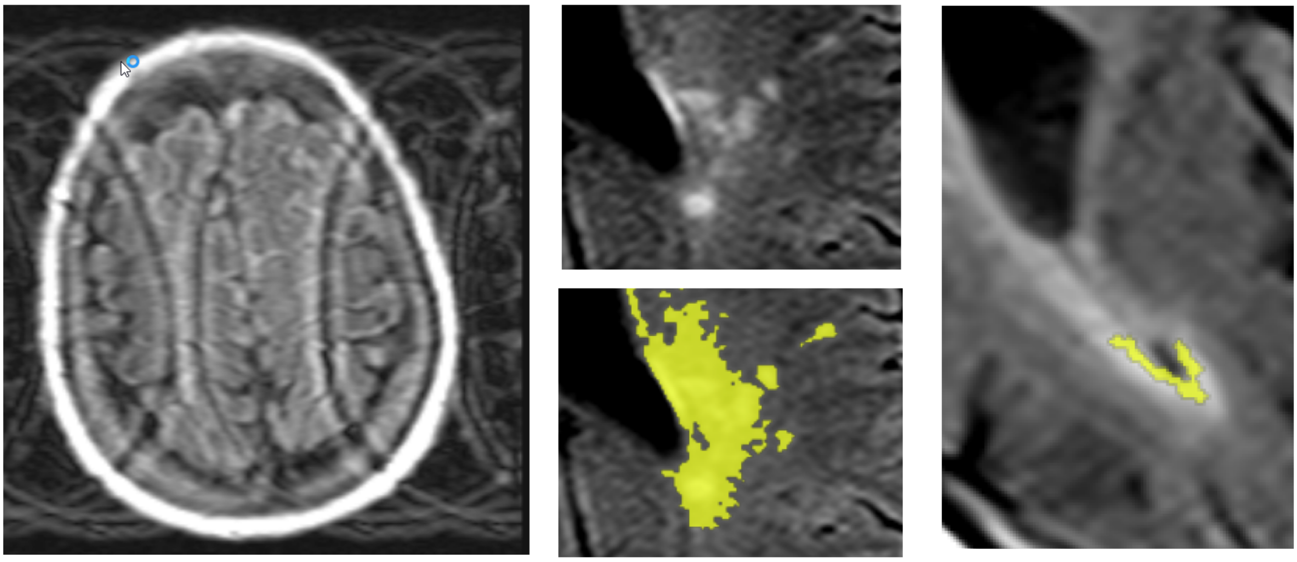

Figure 1. Examples of excluded data. From left to right: imaging artifact, over-segmentation, under-segmentation.

in which the input image is processed in parallel with two different resolutions. The quality of these procedures depends on the amount and quality of the available training data as well as on the data to be segmented.

The aim of this paper is to explore the usefulness of DeepMedic as a basis for development of software applications for automated segmentation using clinical routine MRI data. For this study, data from the clinical routine has been used where no extra time was available, neither for high-resolution 3D FLAIR data, nor for multi-modal MRI acquisitions which can improve the segmentation quality.

2. METHODS

For this work, we used a datasets consisting of 2D FLAIR images acquired within the I-KNOW study[12] and the RETIS study.[13] Acquisition details are given in,[12,13] Tab. 1 summarizes some basic numbers.

2.1 Annotation of MRI data

The presence and extent of hyperintensities of the white matter can qualitatively be determined in the context of neuroradiological findings. Such a qualitative assessment, however, sets a systematic analysis, particularly in the comparison between different persons and in the course of time, narrow methodological boundaries and is, moreover, naturally subjective.

The use of standardized assessment scales such as the Fazekas and Schmidt scale[14] allows for a standardized visual assessment and estimation of the extent of the lesions in the white matter.

Thus, we have developed a semi-automatic method for creating reference annotations. For the determination of the WML, we individually adapted the brightness and contrast of the FLAIR data from the acute dataset of each patient so that the contrast between healthy brain tissue and leukoaraiosis was visually optimal. Afterwards, the area of the WML was manually defined with a certain safety margin. In this marked area, the threshold value was individually adjusted under visual control, thus ensuring an optimal demarcation of the WML against the healthy brain tissue. Artifacts and cortex were excluded. If necessary, in a last step, non-plausible voxels were manually removed from the selection. In the FLAIR sequence, already visible acute stroke lesions were separated from the WML by comparison with the DWI or ADC map. It is important to note that an individual threshold has been defined for each patient.

For quality assurance, all lesion masks have been critically examined with regard to anatomical distribution and have been compared with a visual scoring system (Fazekas and Schmidt scale); a ”good” comparability was established.

	Study1 (104 patients)	**Study2 (93 patients)**
Resolution (mm^3)	$0.4 \times 0.4 \times 6.5$	$0.7 \times 0.7 \times 6.6$
Overall	104	93
SelectedData1	87	74
SelectedData2	70	47

Table 1. During/after the annotations, cases with strong artifacts or missing images were excluded (number of remaining data sets: SelectedData1). In a second refinement step datasets with only moderate annotations were removed (number of remaining data: SelectedData2).

Afterwards, low quality datasets have been removed as shown in Table 1. Additionally, a second selection on this cleaned data has been performed by removing datasets with still ambiguous segmentations.

2.2 Preprocessing

DeepMedic uses ROI masks to limit the learning process and the prediction to the respective relevant image area. Therefore, brain masks are created for all patients. DeepMedic also expects uniform voxel sizes for all records used (both training and test data) so that resampling was applied to all records (Gaussian nearest neighbor filter with resulting voxel size is $0.449 \times 0.449 \times 6.5mm^3$). The image intensities are normalized so that they have zero-mean and unit variance.[11]

CNN name	N_{Train}	N_{Segm}	E	SE	**Number of feature maps in $L1$**	**Number of feature maps in $L2$**
Original	62	1000	35	20	[30,30,40,40,40,40,50,50]	[30,30,40,40,40,40,50,50]
Original(masked)	62	1000	35	20	[30,30,40,40,40,40,50,50]	[30,30,40,40,40,40,50,50]
Orig.(w/o masks)	62	1000	35	20	[30,30,40,40,40,40,50,50]	[30,30,40,40,40,40,50,50]
Smaller	25	1000	20*	15*	[20,20,30,30,30,30,40,40]*	[30,30,40,40,40,40,50,50]
Smaller2	10	1000	15*	10*	[20,20,30,30,30,30,40,40]*	[20,20,30,30,30,30,40,40]*
LessFMs	62	1000	35	10*	[20,20,30,30,30,30,40,40]*	[15,15,25,25,25,25,35,35]*
LessData(masked)	10	1000	35	20	[30,30,40,40,40,40,50,50]	[30,30,40,40,40,40,50,50]
LessLoaded(masked)	62	500*	35	20	[30,30,40,40,40,40,50,50]	[30,30,40,40,40,40,50,50]
CNN3(masked)	46	750*	30*	20	[30,30,40,40,40,40,50,50]	[20,20,30,30,30,30,40,40]*

Table 2. CNN parameters. The name 'Original' refers to the original DeepMedic architecture. The asterisk(*) marks parameters that are different from the original. The extension '(masked)' in the CNN name means that the masked image was used during training. '(w/o masks)' means that the CNN has been trained completely without specifying ROI masks.

2.3 Neural networks and Training

In the following, the CNN proposed by Kamnitsas[11] is referred as 'Original'. Different derived configurations of this CNN are given in Table 2.

The training consists of several iterations which are called "epochs" (E) which are divided into "sub-epochs" (SE). The number of training samples is denoated as N_{Train}. The CNN is trained batchwise. For this purpose, a certain number N_{Segm} of segments are extracted from the training data. The given DeepMedic CNN consists of eight convolutional layers in case of the normal pathway ($L1$) and consists of two fully connected layers and a classification layer in case of the low-resolution pathway ($L2$). The number of feature maps (FM) in the convolutional layers has been varied.

During the first training cycles, the brain masks were used. For predictions (validation / tests), however, the total image data seemed to be taken into account since the predicted segmentation was partially outside the given ROI. The brain masks were previously used for normalization and passed to the algorithm for extracting training segments, but are obviously not applied to the records during the entire algorithm.

CNN name	**Per Epoch**	**Training incl. validation**	**Application on patient datasets** ($n = 25$)	**DSC**
Original	1 h 18 min	50 h 36 min	0 h 54 min	0.6341
Original(masked)	1 h 18 min	45 h 33 min	0 h 54 min	0.6492
Orig.(w/o masks)	1 h 18 min	54 h 13 min	1 h 36 min	0.6027
Smaller	0 h 49 min	17 h 42 min	0 h 40 min	0.5750
Smaller2	0 h 30 min	08 h 02 min	0 h 37 min	0.5420
LessFMs	0 h 57 min	36 h 04 min	0 h 37 min	0.6169
LessData(masked)	1 h 18 min	50 h 18 min	0 h 54 min	0.5982
LessLoaded(masked)	0 h 49 min	33 h 30 min	0 h 54 min	0.6094
CNN3(masked)	0 h 42 min	21 h 01 min	0 h 52 min	0.6437

Table 3. Performance of the CNNs configured as shown in Table 2. The Dice Similarity Coefficient (DSC) was used for validation.

3. RESULTS

3.1 Speed vs quality

Since the training with the original CNN takes very long, the first goal was to reduce the computing time while keeping the quality of the results as good as possible. The different CNNs were compared on 25 patients from the Study1. Initially, only patients from this dataset were used for the training and the validation. As selection, SelectedData1 was used.

The CNN3 was also trained with data from these 25 patients. Therefore, all patient data from Study1 (SelectedData2) was used for testing CNN3, except the training data.

As expected, the adjustments in the CNN architecture, which allow a reduction of the computation time, result in poorer segmentation results. Nevertheless, the deviations of the DSC values from the DeepMedic original are below 0.1. The CNN3 even reaches a comparable average DSC while reducing the computation time to less than half the original.

For applying the trained networks to Study2 datasets, the selection SelectedData2 was used since the quality of many MRI recordings is significantly worse than in Study1 and there are very strong over- and undersegmentations. This would lead to highly falsified results in validation. The results for the segmentation on patients of Study2 datasets are, as expected, significantly worse than for the Study1 datasets used for the training process. The rules learned from this dataset are obviously not directly transferable to Study2. Relevant differences could be lower resolution, larger lesions, and deviating imaging parameters. In the case of larger lesions, there is a tendency to undersegment.

Also, it is noticeable that both the CNN LessData and LessDataLoaded produce better results on average than the original network, although less data was used for the training using the same architecture. One reason for this is might be that less incomplete groundtruth segmentation has been used for learning.

3.2 Impact of annotation quality

Complete coverage of all lesions

As first step, we examined if all areas of increased intensity which point to white matter were correctly segmented by our technique described in Sec. 2.1. In particular, problems occurred in distinguishing between WML and other types of lesions or imaging-related artifacts, e.g., at the boundaries of ventricles. After a manual improvement of those problems, the CNN3 was trained with these improved ground truth masks. As before, 46 patients from Study1 dataset were used.

For both studies (Study1 and Study2), the average DSC could be slightly increased by the CNN trained with improved masks, but this improvement is negligibly small especially for the Study1 dataset. As can be seen in Fig. 3, it is striking that the change in the DSC varies greatly among different patients as expected because

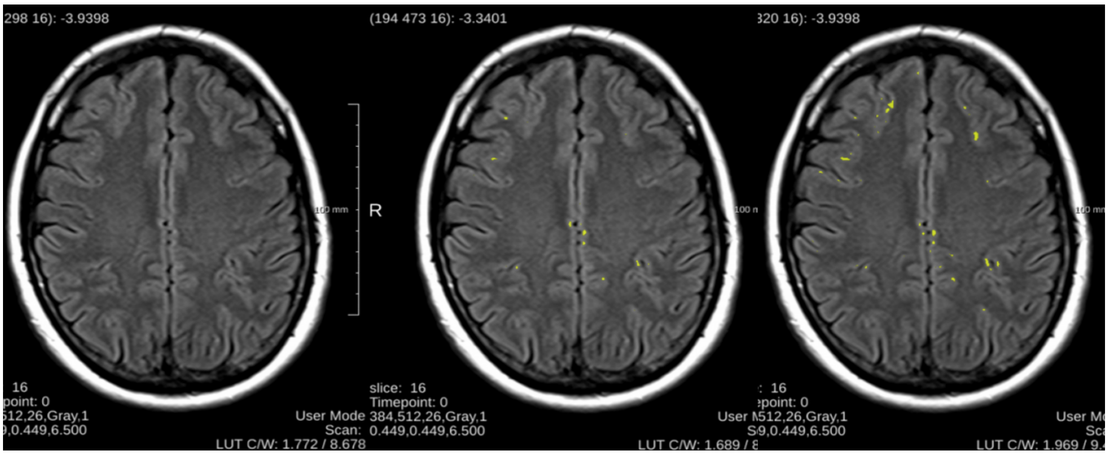

Figure 2. False positives by improvedSeg. From left to right: reference, originalSeg, improvedSeg.

different adjustments were made depending on the quality of the original segmentation mask. The relatively high deviations in the DSC among many patients in the Study2 dataset show how important the reference segmentations are in the assessment of the results and how strongly the assessment can be influenced by them.

On closer examination of the predicted segmentations, it is also becomes clear that the proportion of false-positive segmentations was increased by the adapted groundtruth masks. Particularly locations with increased intensity in areas of the gray matter and brightness caused by noise are segmented by the CNN, see Fig. 2.

	Reference (Training)	**Reference (Validation)**	**DSC**
Study1 (SelectedData1, w/o training data)	originalSeg	originalSeg	0.5756
	originalSeg	improvedSeg	0.5847
	improvedSeg	improvedSeg	0.5910
	improvedSeg2	improvedSeg2	0.6051
Study2 (SelectedData2)	originalSeg	originalSeg	0.4309
	originalSeg	improvedSeg	0.4884
	improvedSeg	improvedSeg	0.5060
	improvedSeg2	improvedSeg2	0.5137

Table 4. Overview of the results of CNN3. originalSeg" refers to the original manual segmentation.

Remove all artifacts

As a consequence, we examined a second improvement: particular care was taken to limit the segmentation to relevant lesions and thus to discard very small areas of increased intensity. This should reduce the segmentation of noise and facilitate the learning of specific characteristics of distinct lesions. In addition, attempts were made to segment lesions in the larger connected regions in order to obtain this property more strongly in the predictions. In addition, the segmentation at the borders of ventricles was also improved. The test segmentations and validations were performed again on the new adapted ground truth masks. An overview of the dice scores when using CNN3 is shown in Tab. 4.

3.3 Different acquisition protocols

The CNN3 was trained on the previously used 46 patients from the Study1 dataset and on additional 20 patients from the Study2 dataset. The improved GroundTruth masks "improvedSeg2" were used for training and validation. When validating the Study1 data, it is noticeable that both the DSC values and the visual inspection of the segmentation results have not changed significantly, see Tab. 5.

Test data	Training	DSC
Study1 (SelectedData1, without training data)	ImprovedSeg2, only Study1	0.6051
	ImprovedSeg2, Study1 and Study2	0.5961
Study2 (SelectedData2, without training data)	ImprovedSeg2, only Study1	0.5008
	ImprovedSeg2, Study1 and Study2	0.5525

Table 5. DSC comparison between the training with only one and both datasets (Study1 and Study2).

On the other hand, the results for the Study2 patients were significantly improved, see Tab. 5 and Fig. 4. The average DSC has risen significantly compared to the changes resulting from the previous adjustments. In addition, a clear improvement can also be perceived visually, see Fig. 6.

Further results are shown in Fig. 3 and Fig. 4 where the ground truth segmentation has been further improved and where both studies are used for training. A omparison between the CNN3 and the original DeepMedic architecture is given in Fig. 6.

Test data	CNN	DSC
Study1 (SelectedData1, without training data)	CNN3	0.5962
	Original DeepMedic	0.6197
Study2 (SelectedData2, without training data)	CNN3	0.5525
	Original DeepMedic	0.5645

Table 6. DSC comparison between the CNN3 and the original DeepMedic architecture.

4. DISCUSSION AND CONCLUSIONS

We explored the usage of DeepMedic on single-modality, low-quality FLAIR data. In addition, we have examined how to modify the neural networks so that the training process can be done on low-end GPUs with nearly the same quality.

The heterogeneous, frequently inhomogeneous signal behavior of hyperintensities of the white matter as well as the interference with other structural changes makes automatic segmentation approaches more difficult. At the same time, a method for automatic segmentation should work independently of the model, of the scanner used, the field strength and largely independent of the quality of the images used and provide comparable and reproducible results. To the best of our knowledge, such an approach is not yet available. However, DeepMedic is an excellent library to achieve results of substantial quality,[15] also on low-budget GPU hardware. Our results have shown that this can also be achieved when having only clinical routine 2D FLAIR data available and not, as demonstrated in several other publications, only by utilizing multi-modal high quality MRI data. One open problem that has not been addressed by DeepMedic until now is possibility of differentiating between two or more different pathological manifestations like stroke lesions and WML. This issue, however, seems to be solvable by a subsequent classification task or by using a u-shaped residual network (uResNet) architecture.[16]

REFERENCES

1. C. R. Jack, P. C. O'Brien, D. W. Rettman, M. M. Shiung, Y. Xu, R. Muthupillai, A. Manduca, R. Avula, and B. J. Erickson, "Flair histogram segmentation for measurement of leukoaraiosis volume," *Journal of Magnetic Resonance Imaging* **14**(6), pp. 668–676, 2001.
2. M. L. Seghier, A. Ramlackhansingh, J. Crinion, A. P. Leff, and C. J. Price, "Lesion identification using unified segmentation-normalisation models and fuzzy clustering," *NeuroImage* **41**(4), pp. 1253 – 1266, 2008.
3. M. Anitha, T. S. Perumal, and V. Palanisamy, "Wml detection of brain images using fuzzy and possibilistic approach in feature space," **11**, pp. 180–189, 06 2012.
4. F. Admiraal-Behloul, D. van den Heuvel, H. Olofsen, M. van Osch, J. van der Grond, M. van Buchem, and J. Reiber, "Fully automatic segmentation of white matter hyperintensities in mr images of the elderly," *NeuroImage* **28**(3), pp. 607 – 617, 2005.

5. K. V. Leemput, F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens, "Automated segmentation of multiple sclerosis lesions by model outlier detection," *IEEE Transactions on Medical Imaging* **20**, pp. 677–688, Aug 2001.
6. F. Rousseau, F. Blanc, J. de Seze, L. Rumbach, and J. P. Armspach, "An a contrario approach for outliers segmentation: Application to multiple sclerosis in mri," in *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 9–12, May 2008.
7. M. E. Caligiuri, P. Perrotta, A. Augimeri, F. Rocca, A. Quattrone, and A. Cherubini, "Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: A review," *Neuroinformatics* **13**, pp. 261–276, Jul 2015.
8. J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks* **61**, pp. 85 – 117, 2015.
9. T. Brosch, *Efficient deep learning of 3D structural brain MRIs for manifold learning and lesion segmentation with application to multiple sclerosis.* PhD thesis, University of British Columbia, 2016.
10. M. Ghafoorian, N. Karssemeijer, I. W. M. van Uden, F.-E. de Leeuw, T. Heskes, E. Marchiori, and B. Platel, "Automated detection of white matter hyperintensities of all sizes in cerebral small vessel disease," *Medical Physics* **43**(12), pp. 6246–6258, 2016.
11. K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *Medical Image Analysis* **36**, pp. 61 – 78, 2017.
12. B. Cheng, N. D. Forkert, M. Zavaglia, C. C. Hilgetag, A. Golsari, S. Siemonsen, J. Fiehler, S. Pedraza, J. Puig, T.-H. Cho, J. Alawneh, J.-C. Baron, L. Ostergaard, C. Gerloff, and G. Thomalla, "Influence of stroke infarct location on functional outcome measured by the modified rankin scale," *Stroke* **45**(6), pp. 1695–1702, 2014.
13. A. Golsari, D. Bittersohl, B. Cheng, P. Griem, C. Beck, A. Hassenstein, M. Nedelmann, T. Magnus, J. Fiehler, C. Gerloff, and G. Thomalla, "Silent brain infarctions and leukoaraiosis in patients with retinal ischemia," *Stroke* **48**(5), pp. 1392–1396, 2017.
14. P. Kapeller, R. Barber, R. Vermeulen, H. Adèr, P. Scheltens, W. Freidl, O. Almkvist, M. Moretti, T. del Ser, P. Vaghfeldt, C. Enzinger, F. Barkhof, D. Inzitari, T. Erkinjunti, R. Schmidt, and F. Fazekas, "Visual rating of age-related white matter changes on magnetic resonance imaging," *Stroke* **34**(2), pp. 441–445, 2003.
15. J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics* **33**(1), pp. 159–174, 1977.
16. R. Guerrero, C. Qin, O. Oktay, C. Bowles, L. Chen, R. Joules, R. Wolz, M. Valds-Hernndez, D. Dickie, J. Wardlaw, and D. Rueckert, "White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks," *NeuroImage: Clinical* **17**, pp. 918 – 934, 2018.

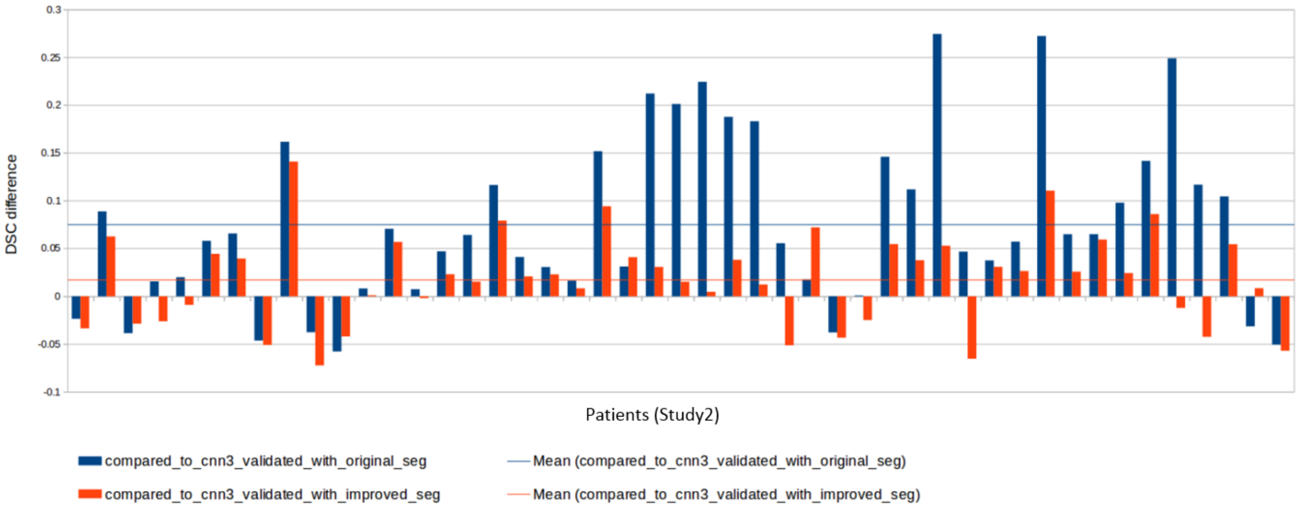

Figure 3. Modification of the DSC values by the use of the improved ground-truth masks during the training of the CNN3. The patients from Study2 / SelectedData2 are considered.

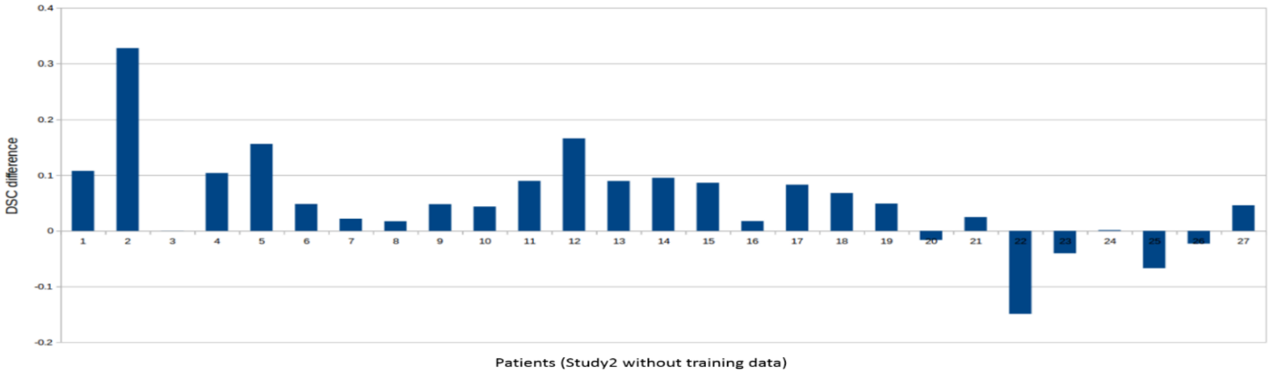

Figure 4. Datasets from both studies are used for training. The plot shows the change of the DSC values for the patients used for validation from the Study2 dataset (positive values mean an improvement by using both datasets).

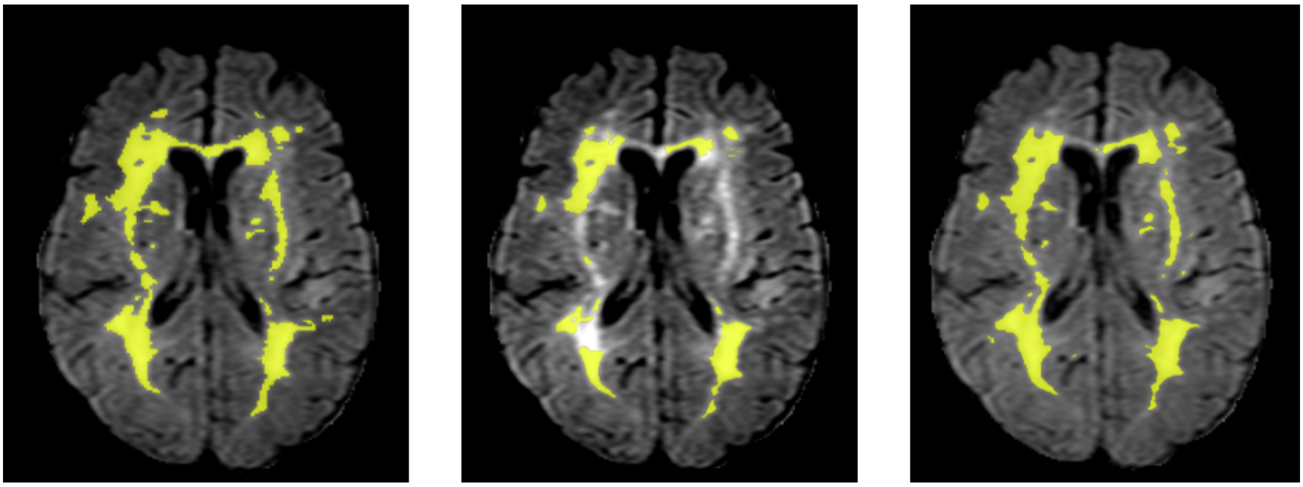

Figure 5. Results on Study 2. From left to right: reference annotation, Original, CNN3.

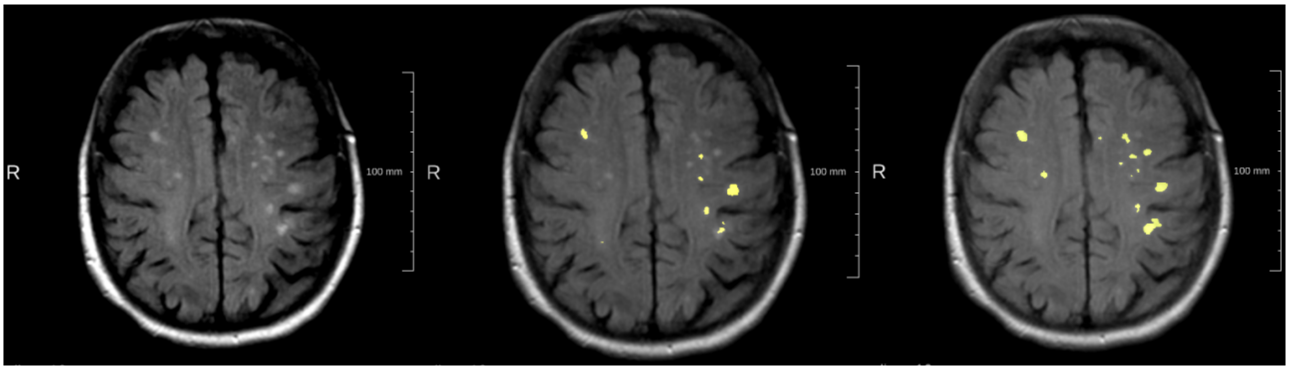

Figure 6. Left: MRI scan without segmentation, center: result of CNN3 trained on Study1 and improvedSeg2, result of CNN3 trained on both datasets and improvedSeg2.

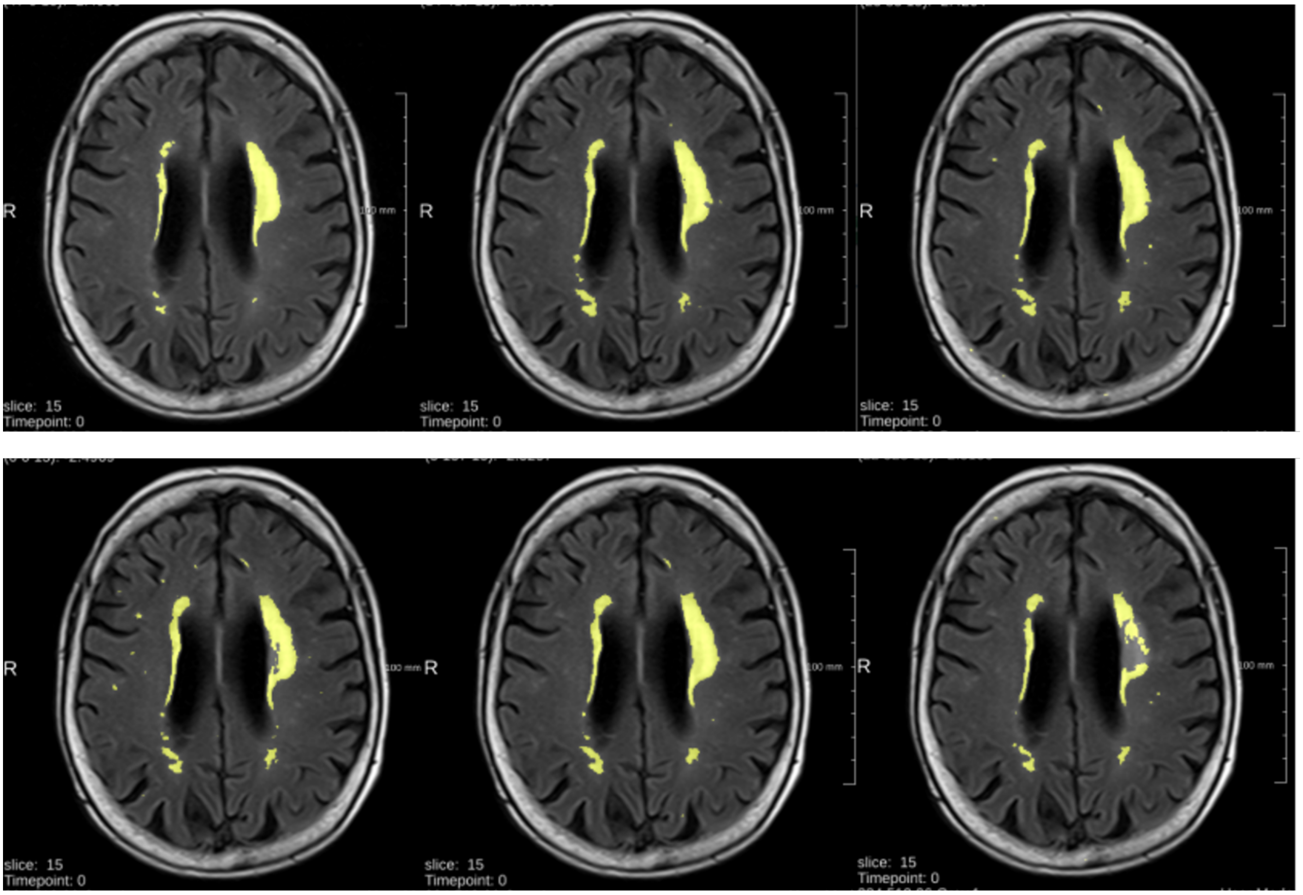

Figure 7. Results on Study1. Upper row, left to right: reference annotation, Original, LessFMs. Bottom, left to right: LessData, LessLoaded, Smaller2.