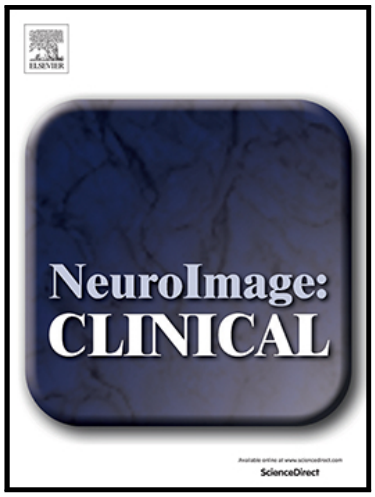
Journal Pre-proof



Automatic detection of lesion load change in Multiple Sclerosis using convolutional neural networks with segmentation confidence

Richard McKinley , Rik Wepfer, Lorenz Grunder,
Fabian Aschwanden, Tim Fischer, Christoph Friedli, Raphaela Muri,
Christian Rummel, Rajeev Verma, Christian Weisstanner,
Benedikt Wiestler, Christoph Berger, Paul Eichinger, Mark Muehlau,
Mauricio Reyes, Anke Salmen, Andrew Chan, Roland Wiest,
Franca Wagner

PII: S2213-1582(19)30451-6
DOI: https://doi.org/10.1016/j.nicl.2019.102104
Reference: YNICL 102104

To appear in: *NeuroImage: Clinical*

Received date: 8 April 2019
Revised date: 27 September 2019
Accepted date: 18 November 2019

Please cite this article as: Richard McKinley , Rik Wepfer, Lorenz Grunder, Fabian Aschwanden, Tim Fischer, Christoph Friedli, Raphaela Muri, Christian Rummel, Rajeev Verma, Christian Weisstanner, Benedikt Wiestler, Christoph Berger, Paul Eichinger, Mark Muehlau, Mauricio Reyes, Anke Salmen, Andrew Chan, Roland Wiest, Franca Wagner, Automatic detection of lesion load change in Multiple Sclerosis using convolutional neural networks with segmentation confidence, *NeuroImage: Clinical* (2019), doi: https://doi.org/10.1016/j.nicl.2019.102104

Highlights

- We introduce a novel method, based on a neural network (DeepSCAN), for detecting lesion change in longitudinal MRI imaging in multiple sclerosis.
- The method had a sensitivity of 1.00 and a positive predictive value of 0.59 for detecting lesion load change, and an AUC of 0.999, when assessed on twenty-six MS patients with longitudinal imaging.
- This compares to a sensitivity of 0.69 and a PPV of 0.18 when lesion volume, as measured by our classifier, was used to assess lesion load change (AUC = 0.71)
- Change in lesion count (another commonly used metric) had a sensitivity of 0.57 and a PPV of 0.38 (AUC = 0.51).

Automatic detection of lesion load change in Multiple Sclerosis using convolutional neural networks with segmentation confidence

Richard McKinley[a], Rik Wepfer[a], Lorenz Grunder[a], Fabian Aschwanden[a], Tim Fischer[f], Christoph Friedli[d], Raphaela Muri[a], Christian Rummel[a], Rajeev Verma[b], Christian Weisstanner[c], Benedikt Wiestler[g], Christoph Berger[i], Paul Eichinger[g], Mark Muehlau[h], Mauricio Reyes[e], Anke Salmen[d], Andrew Chan[d], Roland Wiest[a], Franca Wagner[a]

[a]*Support Center for Advanced Neuroimaging, University Institute for Diagnostic and Interventional Neuroradiology, Inselspital, Bern University Hospital, University of Bern, Switzerland*
[b]*Department of Neuroradiology, Spital Tiefenau, Switzerland*
[c]*Medizinisch Radiologischen Institut, Zurich, Switzerland*
[d]*Univeristy Clinic for Neurology, Inselspital, Bern University Hospital, University of Bern, Switzerland*
[e]*Insel Data Science Centre, Inselspital, Bern University Hospital, University of Bern, Switzerland*
[f]*Universitätsklinik Balgrist, Zurich, Switzerland*
[g]*Diagnostic and Interventional Neuroradiology, Klinikum rechts der Isar der TU München, Munich, Germany*
[h]*Department of Neurology, Klinikum rechts der Isar der TU München, Munich, Germany*
[i]*Center for Translational Cancer Research (TranslaTUM), TU München, Munich, Germany*

Abstract

The detection of new or enlarged white-matter lesions is a vital task in the monitoring of patients undergoing disease-modifying treatment for multiple sclerosis. However, the definition of 'new or enlarged' is not fixed, and it is known that lesion-counting is highly subjective, with high degree of inter- and intra-rater variability. Automated methods for lesion quantification, if accurate enough, hold the potential to make the detection of new and enlarged lesions consistent and repeatable. However, the majority of lesion segmentation algorithms are not evaluated for their ability to separate radiologically progressive from radiologically stable patients, despite this being a pressing clinical use-case. In this paper, we explore the ability of a deep learning segmentation classifier to

Email address: richard.mckinley@insel.ch (Richard McKinley)

separate stable from progressive patients by lesion volume and lesion count, and find that neither measure provides a good separation.. Instead, we propose a method for identifying lesion changes of high certainty, and establish on an internal dataset of longitudinal multiple sclerosis cases that this method is able to separate progressive from stable time-points with a very high level of discrimination (AUC = 0.999), while changes in lesion volume are much less able to perform this separation (AUC = 0.71). Validation of the method on two external datasets confirms that the method is able to generalize beyond the setting in which it was trained, achieving an accuracies of 75 % and 85 % in separating stable and progressive time-points.

Keywords: Deep Learning, Multiple Sclerosis, MRI, Longitudinal Imaging

1. Introduction

Magnetic resonance imaging is the most important imaging method for diagnosis and monitoring of multiple sclerosis. The 2017 revised Mcdonald diagnostic criteria for the diagnosis of multiple sclerosis require the dissemination of lesions in both space and time. Lesion load change is also crucial for the assessment of disease activity, since patients who are assigned with disease modifying therapies and no evidence of disease activity (NEDA) harbor a better prognosis [1, 2, 3, 4]. Radiological progression can be separated into new or enlarged lesions in T2 weighted imaging, and new enhancing lesions on T1 weighted imaging with Gadolinium-based contrast agents (GBCA). While standard imaging protocols for multiple sclerosis have included GBCA, there is increasing evidence that high resolution 3D unenhanced MRI is sufficient to detect the presence of new or enlarged lesions [5].

Detection of new and enlarged lesions in multiple sclerosis imaging by human raters is time-consuming and limited by inter- and intra-rater variability [6]. As a consequence, manual lesion volumetry and lesion counting has limited sensitivity for new lesion detection. Delineation of new and enlarged lesions can be improved by working on subtraction MRI, but this still requires substantial

human user interaction and judgement, as well as manual intensity normalization. A recent study showed that FLAIR subtraction MRI had a sensitivity of 80% for detecting new or enlarged lesions. [7]. Registration errors, flow artifacts and lesion signal intensity differences can result in the detection of false-positive lesions on subtraction images [8].

Several groups have proposed automated methods for multiple sclerosis lesion segmentation, mostly validated in a cross-sectional fashion. [9, 10, 11, 12] Even where longitudinal data was used to assess the performance of classifiers, consistency of segmentations over time, or the ability to detect new lesions were not investigated [13]. Since MR contrast will differ between time-points, even on the same scanner, and since the borders of MS lesions are often not well defined, automated methods will typically show small differences in the boundaries of lesions at different time-points, even if no lesion growth has taken place. Since even the best automated methods also make false positive and false negative lesion identifications, lesion counts may also not be reliable in a longitudinal setting. Several researchers have proposed methods to harmonize segmentations across two or more time-points. Jain et al propose a joint expectation-maximization (EM) framework for two time-point white matter (WM) lesion segmentation, and the Lesion Segmentation Toolkit, a tool integrated in SPM, has a longitudinal pipeline which adapts existing segmentations across multiple time-points [14, 15]. Meanwhile, Salem et al proposed a logistic regression classifier for detected new and enlarged lesions showing "considerable growth" using features derived from subtraction imaging and deformation fields derived from registration of two time-points.[16]

In a companion paper, we have introduced a novel method (DeepSCAN MS) based on convolutional neural networks (CNNs), for multiple sclerosis lesion segmentation, which we demonstrated to outperform previous methods.[17] In this paper, we demonstrate that changes in lesion count and volume change, estimated using our method, do not perform well as a method for separating stable and progressive MS cases. Simultaneous lesion growth and lesion resolution may occur at a single time-point, which will not be apparent from simply observing

volume changes. Further, variations in image contrast between acquisitions can lead to substantial volumetric changes in automated lesion delineation, even when using 'state-of-the-art' classification methods. Lesion counts are also only approximate measures of activity, since lesions may be missed or undersegmented, false positives may give the impression of lesion growth where none exists, and lesions may become confluent, leading to an increase in lesion tissue but a decrease in lesion count.

As a potential solution to this issue, we instead propose to identify new and missing lesion tissue by using the *confidence* of an automated classifier in its own segmentation. Measures of segmentation uncertainty have previously been proposed as a method of rejecting false positive MS lesion identifications. [18] To our knowledge, our method is the first to leverage segmentation confidence in the detection of longitudinal change. Our recently introduced MS lesion classifier, DeepSCAN, produces for each tissue map a 'label-flip probability', which is a measure of uncertainty derived from the training data. We use the segmentation of the classifier and the label-flip map to distinguish between patients with no new or enlarged lesions (those satisfying that component of the NEDA criteria) and those with genuinely new or enlarged lesions. We identify as new lesion tissue only those voxels that were confidently not present at time-point t=0 but that are confidently lesion tissue at time-point t=1. The method requires T1, FLAIR and T2 imaging adhering to modern best-practice imaging standards in MS (specifically, a 3D FLAIR and 3D T1 acquisition), such as those specified in the OFSEP minimal MRI protocol. [19].

2. Methods

In this paper, we study the ability of a previously trained deep learning classifier to detect longitudinal changes in T2 lesion load, by several means: lesion counting, overall lesion volume, detecting voxel-by-voxel change using coregistration, detecting voxel-by-voxel *confident* change using a method which incorporates classifier confidence. We describe the patient cohorts, the deep

learning method, and the methods for detecting lesion growth. We utilise data from three sources. The first are MRI datasets of patients with remitting-relapsing multiple sclerosis that were identified from the MS cohort databank of the University of Bern. Use of data for this study was approved by the local ethics committee (Cantonal Ethics Commission Bern, Switzerland 'MS segmentation disease monitoring', approval number 2016-02035) and all patients gave general consent for data storage and analysis of their MRI datasets. This data was from the same centre and scanner as that used for the training of our fully convolutional deep learning classifier (DeepSCAN).

Additional anonymized datasets were provided by Radiology Center Bethanien, (which we subsequently refer to as the Zurich dataset), and from the Klinikum Rechts der Isar, Munich, Germany (which we subsequently refer to as the Munich dataset).

2.1. Patient cohorts and MR imaging

Patients from the Bernese MS cohort were included in the Bern dataset if they had at least three consecutive MRI datasets, and were not among the 50 casesused in training of the DeepSCAN classifier. [17] All patients fulfilled the revised McDonald criteria of 2010 for relapsing-remitting multiple sclerosis.[20]

MR images from the Bern dataset were acquired on a 3T MRI (Siemens Verio, Siemens, Erlangen, Germany) . The protocol settings were i) T1 weighted MP-RAGE pre- and post gadobutrol i.v. (TR 2530 ms, TE 2.96 ms, averages 1, FoV read 250 mm, FoV phase 87.5 % voxel size 1.0 x 1.0 x 1.0 mm, flip angle 7, acquisition time 4:30 min. slices per slab 160, slice thickness of 1.0 mm) ii) T2- weighted imaging (TR 6580 ms, TE 85 ms, averages 2, FoV read 220 mm, FoV phase 87.5 %, voxel size 0.7 x 0.4 x 3.0 mm, flip angle 150, acquisition time 6:03 min, 42 parallel images were acquired with a slice thickness of 3.0 mm,) iii) 3D FLAIR imaging (TR 5000 ms, TE 395 ms, averages 1, FoV read 250 mm, FoV phase 100 %, voxel size 1.0 x 1.0 x 1.0 mm, acquisition time 6:27 min. A total of 176 parallel images were acquired with a slice thickness of 1.0 mm). All patients received Gadobutrol (Gadovist) 0.1 ml/kg bodyweight immediately

after the acquisition of the unenhanced T1w sequence.

MR images from the Zurich dataset were acquired using a standardized acquisition protocol on a 3T MRI (Siemens Skyra, Siemens, Erlangen, Germany), including: i) T1 weighted MP-RAGE precontrast (TR 2300 ms, TE 2.9 ms, TI 900 ms, averages 1, FoV read 250 mm, FoV phase 93.75 % voxel size 1.0 x 1.0 x 1.0 mm, flip angle 9, acquisition time 05:12 min.) ii) T2- weighted imaging (TR 4790 ms, TE 100 ms, averages 1, FoV read 220 mm, FoV phase 100 %, voxel size 0.7 x 0.4 x 3.0 mm, flip angle 150, acquisition time 02:16 min iii) 3D FLAIR imaging (TR 5000 ms, TE 398 ms, TI 1800 ms, averages 1, FoV read 250 mm, FoV phase 100 %, voxel size 1.0 x 1.0 x 1.0 mm, flip angle 120, acquisition time 04:17 min.).

MR images from the Munich dataset were acquired with a 3T MRI (Achieva; Philips Healthcare, Best, the Netherlands) including: i) 3D T1 gradient-echo imaging, performed before and at least 5 minutes after administration of 0.1 mmol/kg gadolinium-based contrast material : voxel size 1.0 x 1.0 x 1.0 mm; acquisition time, 6 minutes ii) a three-dimensional fluid-attenuated inversion-recovery (FLAIR) sequence, voxel size, 1.03 x 1.03 x 1.5 mm3; acquisition time, 5 minutes iii) T2-weighted imaging: voxel size, 1.03 1.03 1.5 mm; TR 40006000 ms (variable); TE 35 ms; acquisition time 5 min.

2.2. The DeepSCAN MS lesion classifier

In a previous paper on brain tumor segmentation [21] , we proposed a hybrid of U-net [22] and Densenet [23], in which the bottleneck layer of the Unet is a single dense block, and in which some of the pooling and upscaling is replaced by dilated convolutions. In a subsequent paper, we introduced a new loss function (label-flip loss), in which the probability that classification output differs from the ground truth used for supervision is used to anneal gradients coming uncertain datapoints, and demonstrated that this loss function leads to improved results in brain segmentation.[24]. In a companion paper to this paper, we trained a classifier, which we call DeepSCAN MS, on fifty cases from the Bernese MS cohort databank [17]. In this section, we first summarize the proce-

dure for training the DeepSCAN MS classifier, and then describe its application in detecting longitudinal changes in MS.

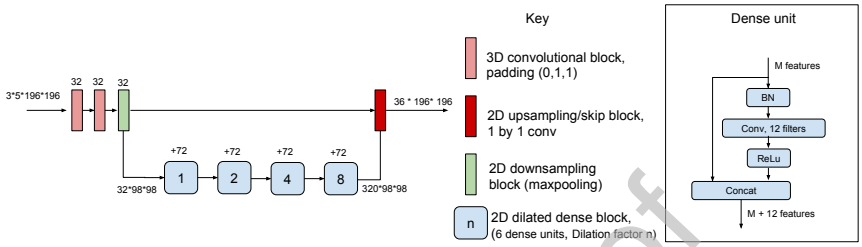


Figure 1: The DeepSCAN architecture used in this paper for lesion and brain-structure segmentation

The DeepSCAN MS classifier is shown in Figure 1: it is a fully-convolutional neural network trained on fifty cases from the Bernese MS cohort databank, which provides segmentations of white-matter lesions, together with segmentations of the cerebellum, subcortical grey matter structures, and cortical grey and white matter, in MS patients. (In this study we only use the lesion segmentations produced by the classifier.) The network was trained using a combination of *focal loss* and our previously defined *label-flip loss*, on lesion labels provided by manual raters, and brain anatomy labels provided by Freesurfer. In label flip loss, for each voxel, and tissue class, the network outputs two probabilities: the probability p that voxel contains the tissue class, and the probability q that the label predicted does not correspond to the label in the ground-truth annotation (i.e., the probability of a 'label flip'). IF BCE stands for the standard binary cross-entropy loss, and y is the target label, then the label-uncertainty loss is:

$$BCE(p, (1-x) * q + x * (1-q)) + BCE(q, z) \tag{1}$$

where

$$z = (p > 0.5) * (1-x) + (p < 0.5) * x \tag{2}$$

If q is close to zero, and the label is correct, the first term is approximately

the ordinary BCE loss: if q is close to 0.5 (representing total uncertainty as to the correct label) the first term tends to zero. This loss therefore attenuates loss in areas of high uncertainty (i.e., where the network is likely to disagree with the ground truth) during training, and indicates areas where segmentation reliability may be poor when applied to new data.

On an internal dataset of 32 patients, the DeepSCAN classifier achieved a mean Dice coefficient of 0.60 versus a manual consensus ground truth for the task of segmenting MS lesions, compared to a mean Dice coefficient of 0.58 between two independent manual raters. This result was sustained when we examined external data from the MSSEG challenge [25]. This dataset consists of fifteen cases, from two centres and three scanners, each rated by seven independent manual raters. Imaging quality is of a similar standard to that used in the Bernese MS cohort. [19]. Versus the independent raters, mean Dice coefficient with the output of DeepSCAN (without retraining on the external data) ranged between 0.56 and 0.61. For comparison, the mean Dice coefficient between the MSSEG raters on the training data ranged between 0.54 and 0.75.

As we have already discussed, manual segmentations of MS lesions have large inter- and intra-rater variability, and so we must accept that this 'ground-truth' may, for lesion segmentation, contain many inconsistencies: missed or under-segmented lesions, and false identifications or over-segmented lesions. For example, a retrospective analysis of the 32 manual lesion segmentations used to validate the DeepSCAN classifier found an average of 18 false positive lesions and 4 missed lesions per subject.

For full details of the training and validation of the DeepSCAN MS classifier, please see McKinley et. al [17].

2.3. Dichotomization of imaging data: progressive vs stable

For each patient and each time-point, a decision was made by an experienced neuroradiologist if that time-point represented, from an imaging standpoint, progressive disease (PD, if any new FLAIR- or contrast-enhancing lesions was detected) or stable disease (SD, if the number of lesions remained stable or

reduced over time) ,based on visual analysis by one of the authors (LG for cases from Bern, CW for cases from Zurich, PE for cases from Munich). In each case, the full clinical sequence (including T1 post-contrast for all sites, and Double Inversion Recovery for Munich) was included in the analysis.

2.4. Automated Segmentation by DeepSCAN convolutional neural network

For each patient and time-point we used the DeepSCAN classifier to generate lesion masks and label-flip maps for MS lesions lesions, using the T1-weighted, T2-weighted, and T2 FLAIR imaging as input. To aid in comparison between time-points, these maps were resampled to $1mm^3$ isotropic resolution. The classifier also returns a $1mm^3$ isotropic skull-stripped FLAIR image in the same space as the lesion and label-flip maps.

2.5. Coregistration

In order to compare cases across time-points, it was necessary to register all imaging for each patient to a common space. To avoid biases inherent in registering to a particular time-point, we applied a robust registration technique (the Robust Template method from Freesurfer) to the skull-stripped FLAIR images produced by our CNN tool, in which all time-points are registered to a common patient-specific template . [26] After construction of the template, lesion masks and lesion confidence maps were rigidly registered to the template space using the transforms output by the robust template method.

2.6. Lesion change detection by classification uncertainty

We describe here the decision procedure for labelling a voxel as 'new lesion', given lesion mask and label-flip maps at time-points A and B in a common, coregistered space, and a threshold q determining acceptable confidence. For each time-point, a voxel is labelled as 'confident lesion' if it is in the lesion mask, and if the label-flip probability is less than q. A voxel is labelled 'confident non-lesion' if it is not in the lesion mask, and if the label-flip probability is less than q. A voxel is labelled as 'new lesion' at time-point B, if it is labelled as

'confident non-lesion' at time-point A, and 'confident lesion' at time-point B. It is labelled 'missing lesion' at time-point B, if it is labelled as 'confident lesion' at time-point A, and 'confident non-lesion' at time-point B. Finally, connected components of the 'new lesion' and 'missing lesion' maps were calculated.

We subsequently identified all connected components of "new lesion" tissue. To improve robustness to coregistration artifacts, all connected components of the new lesion map containing fewer than 12 voxels were deleted.

For the purposes of our initial investigation, we set the value of q to be 0.05: i.e., we determine a voxel to be classified with confidence if the model predicts a 5% or lower chance of the predicted label disagreeing with the manual rater.

2.7. Lesion change detection by threshold margin

A more simplistic methodology for labelling lesions as confidently or uncertainly classified is to set a margin around the ordinary decision threshold, 0.5, and to label all voxels outside of this margin as 'confident'. This method has the advantage that it may be applied to classifiers which do not output a label-flip probability: however, in general the output of modern neural networks is not well calibrated: the scores output by deep networks do not correspond to observed probabilities and are typically overconfident [27].

Concretely, we set a margin $0 < m < 0.5$, and classify every voxel with $p < 0.5 - m$ as confident nonlesion, while every voxel with $p > m + 0.5$ is classified as confident lesion. The measure of new lesion tissue is then as above: a voxel is new lesion if it is labelled as 'confident lesion' at time-point A, and 'confident non-lesion' at time-point B. As above, connected components below 12 voxels were deleted.

For the purposes of our initial investigation, we set the value of m to be 0.45: i.e., we determine a voxel to be classified as confident lesion if the model predicts a score of .95 or greater and to be classified as confident non-lesion if the model predicts a score of 0.05 or less.

2.8. Evaluation

We compared our proposed methods to four other methods on our internal (Bernese) test set: absolute change in lesion volume, relative change in lesion volume, change in lesion count, and total new lesion volume (equivalent to our method with $q = 0.5$). To test the power of these measures to separate progressive and stable time-points, we plotted the receiver-operating characteristic (ROC) curves for each of the above methods. While ROC-AUC analysis gauges the ability of a metric to separate positive and negative examples across all operating thresholds, clinical applicability required that a particular threshold is chosen. .We therefore tested the performance of our metrics at an operating threshold corresponding to 'no lesion change' (i.e. lesion count > 0, lesion volume change > 0, and new lesion volume > 0).

We assessed the sensitivity of our method to its parameters, by comparing the ROC curves of the method at different values of uncertainty threshold q, margin m, and small-growth threshold.

3. Results

Twenty-six patients from the Bernese MS databank satisfied the inclusion criteria, of which 16 were judged from radiological reports to have no lesion changes in any of the time-points, and so were labelled as having stable disease (SD). The remaining 10 cases were judged to have progressive disease (PD). The mean number of time-points per patient was 4.4 for the progressive patients, and 4.9 for the stable patients. Among the ten progressive patients, there were a total of 13 time-points where the radiological reports indicated progression, meaning that approximately 30% of the time-points in those patients showed lesion progression. Mean time between examinations for 223 days, with a standard deviation of 98 days.

3.1. ROC-AUC analysis

For each proposed method, we computed the area under the receiver-operating characteristic for the bernese dataset: see Figure 2. Lesion counting performed

	TN	FP	FN	TP	Accuracy	Sensitivity	PPV	FPR
Confidence method> 0	74	9	0	13	0.91	1.00	0.59	0.11
Margin method > 0	83	0	6	7	0.94	0.54	1.00	0.00
New lesion volume > 0	8	75	0	13	0.22	1.00	0.15	0.90
Volume change> 0	41	42	4	9	0.52	0.69	0.18	0.51
Lesion count change > 0	50	33	8	5	0.57	0.38	0.13	0.40

Table 1: Ability to distinguish progressive vs stable MS at thresholds corresponding to no lesion change, on internal test set, showing the number of true negatives (TN), false positives (FP), false negatives (FN) and true positives (TP), together with accuracy, positive predictive value and recall. Metrics are shown for the label-flip method (Confidence method) and the margin-based method (Margin method), together with new lesion volume, lesion volume change and lesion count change.

worst, with a ROC-AUC of 0.51, while absolute and relative volume change performed comparably, with ROC-AUCs of 0.70 and 0.71 respectively. The proposed method using score margins had an AUC of 0.77. Meanwhile, the proposed method using network-derived uncertainty had a ROC-AUC of 0.999.

3.2. Performance at meaningful thresholds

Results of this analysis are shown in Table 1.

For lesion counting, this metric leads to a total of 33 time-points being identified as progressive, when in fact they were stable according to radiological reports. For lesion volumetry, 42 time-points were falsely identified as being stable. For the proposed method, nine stable time-points were labelled as progressive. Meanwhile, the proposed method based on uncertainty successfully identified all progressive time-points. By comparison, the lesion volume metric failed to find four of the progressive time-points, and lesion counting failed to find eight progressive time-points. The proposed method based on a margin around the decision boundary made no false positive identifications, but failed to find six of the progressive time-points.

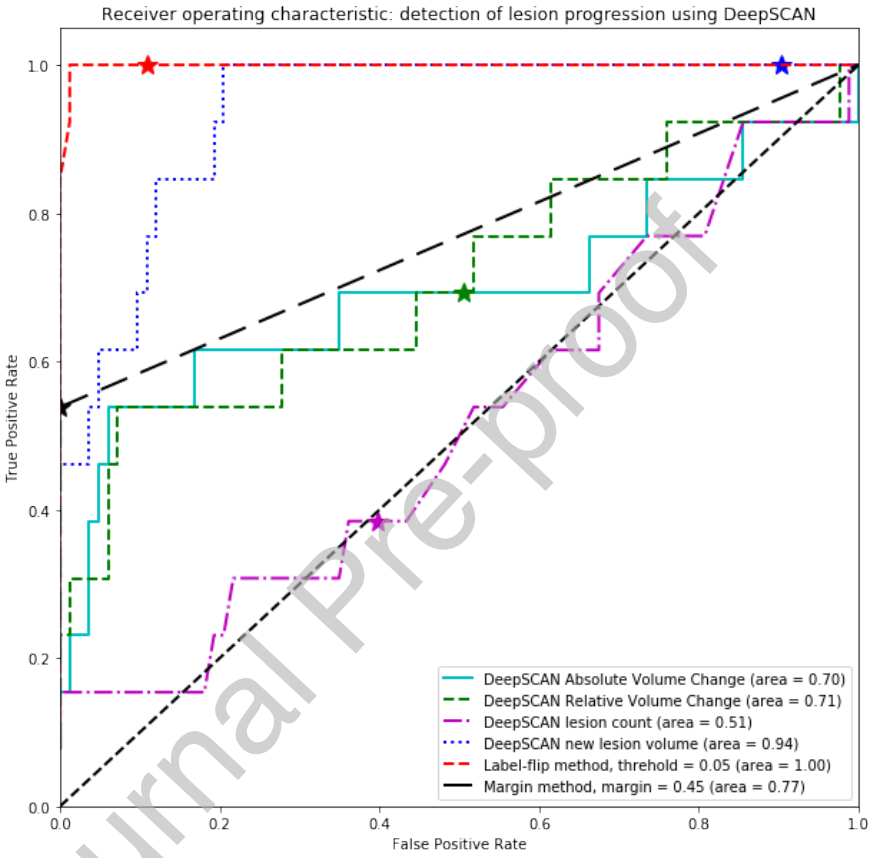


Figure 2: Receiver operating curves for the detection of lesion progression using DeepSCAN, on our internal validation set, via absolute lesion volume change (AUC=0.70), relative volume change (AUC = 0.71), lesion count change (AUC = 0.51), the proposed method using a score margin of .45 (AUC=0.77) and the proposed method using an uncertainty threshold of 0.05 (AUC $\approx$ 1). The star on each curve represents a cutoff where the patient is labelled as stable if the considered metric is less than or equal to zero.

.

3.3. Sensitivity to uncertainty threshold, score-margin and small-growth threshold

The best-performing method according to area under the ROC curve, according to our initial analysis, was achieved using our uncertainty-based method with an uncertainty threshold of 0.05: i.e. voxels which had a flip-probability greater than 0.05 at either time-point are not used to calculate lesion change. At a fixed operating threshold, meanwhile, our two proposed methods performed similarly in terms of accuracy, but the method derived from label-flip confidence had perfect sensitivity and lower PPV, while the method derived from a margin around the threshold had perfect PPV and lower sensitivity.

Both of these methods rely on a parameter which can be varied, with an effect on the performance. In this section we investigate the effect of changing those parameters.

3.3.1. Effect of changing uncertainty threshold

For uncertainty threshold values lower than the one we initially selected (0.0005, 0.001 and 0.01), the AUC was slightly reduced, at 0.92. At larger uncertainty thresholds than initially selected, the AUC was also slightly lower: a threshold of 0.1 gave an AUC of 0.99, and a threshold of 0.2 gave an AUC of 0.96.

3.3.2. Effect of changing classification margin

The effect of changing the classification margin was much more drastic. By setting a narrower classification margin (0.15), we were able to achieve an AUC close to the performance of the uncertainty-based method (AUC = 0.998). A slightly larger margin of 0.2 gave worse performance (AUC = 0.96), while a slightly narrower margin of 0.1 led to a smaller decrease in performance (AUC = 0.996).

3.3.3. Effect of changing threshold for growth

In the method as described, areas of growth below 12 voxels do not count towards lesion growth. The method is reasonably robust to changes in this

lesion-growth threshold. A larger threshold of 24 voxels led to an AUC of 0.96, while a smaller threshold of 6 voxels led to an AUC of 0.997. Not applying a threshold yielded an AUC of 0.98.

3.4. Performance on external data

Several authors have reported difficulties of automated methods for MS lesion segmentation to perform on out-of-sample data.[25, 11] In our previous paper, we already validated that performance of the DeepSCAN MS classifier is not substantially degraded when applied to data adhering to similar protocol standards from different centres[17]. In this section, we report the ability of the uncertainty-based method, as described above to identify progressive time-points in external data. The method was applied to data from eight patients, each having four consecutive time-points (thirty-two datsets, twenty-four after baseline) from the Zurich dataset. This data was supplied full anonymized. In a second test of generalization, the full lesion segmentation algorithm and uncertainty-based method was containerized using Docker, and provided to the co-authors from Munich (BW, CB, PE, MM), who applied the classifier to cases from their centre.

The Zurich dataset consisted of four consecutive time-points (thirty-two datasets, twenty-four after baseline imaging). Of the twenty-four follow-up time-points, five were judged by the rater (CW) to have new or enlarged lesions. The proposed method successfully identified three of the five progressive time-points (sensitivity of 60%) and labelled an additional three incorrectly as being progressive. (PPV of 84 %), Overall accuracy on this dataset was 75 %

The Munich dataset consisted of 53 pairs of baseline and followup image, of which 24 were judged progressive, and 29 judged stable. The method successfully identified 16 of the 22 progressive time-points (Sensitivity of 72%) and correctly identified all of the stable time-points. (PPV of 100 %) Overall accuracy on this dataset was 85%

	Accuracy	Sensitivity	PPV
Zurich	0.75	0.60	0.84
Munich	0.85	0.72	1.00
Bern	0.91	1.00	0.59

Table 2: Performance of the confidence-based method on the three datasets studied in this paper, showing Accuracy, Sensitivity, and Positive Predicative Value (PPV)

4. Discussion

MRI is the method of choice to determine lesion load evolution in patients with multiple sclerosis. The accurate detection of new or enlarged white-matter lesions in multiple sclerosis patients is a pivotal task of the disease monitoring process in patients who receive disease-modifying treatment. However, the definition of 'new or enlarged' remains ill-defined, and lesion counting remains subjective with a considerable degree of inter- and intra-rater variability depending on the level of experience of the reporting expert. Automated methods for lesion quantification, if accurate, hold the potential to make the detection of new and grown lesions consistent and repeatable. Until now, the majority of lesion segmentation algorithms are not well evaluated for their ability to accurately separate radiologically progressive disease course from radiologically stable patients during follow-up. Despite this being the pressing clinical use-case and information for the clinicians with impact on further treatment regime selection for the MS patients. We demonstrate that measures of new lesion load derived from label-flip uncertainty outperform lesion counting as well as absolute and relative volume change detection in the longitudinal analysis of MS lesions. The major advantage of the proposed approach is to identify the time-point during follow-up where lesion progression was evident with a very high accuracy and positive predictive value. The method is fully automated, and therefore offers the benefit of being objective and independent from user bias, thus leading to more trustful longitudinal evaluations.

The method developed relies on a minimum standard of MR imaging cor-

responding to a modern MRI protocol for imaging of demyelinating disease: in particular a 3D T1 and 3D FLAIR acquisition (with approximately $1mm^3$ or better voxels). The recommended protocol is in keeping with the 2016 Consortium of MS Centers Task Force recommendations and can be executed in approximately 20 minutes. In particular, the method does not rely on the availability of a post-contrast T1 sequence: recent research suggests that modern 3D imaging at 3T can reduce or eliminate the need for contrast-enhanced sequences. [5, 7]

The method in this paper proposes to track changes in lesion load by leveraging measures of uncertainty in the location of lesion boundaries, based on the predictions of a deep learning convolutional neural network classifier, DeepSCAN. This method has already been shown to perform well at lesion segmentation in a cross-sectional setting: the classifier was more than twice as effective in lesion detection as both previous generations of CNN-based segmentation tools and freely-available lesion segmentation SPM toolboxes. [17] In this paper, we sought to demonstrate the same classifiers ability to detect lesion change: by considering as new lesion tissue only those voxels which are classified confidently by DeepSCAN, progressive time-points were detected with an accuracy of 0.91 and a recall of 1.0, when applied to data from the same centre as those used to train the classifier. By comparison with standard metrics, such as lesion count progression or volume changes, no progressive time-points were falsely identified as stable, and the risk of false positive results decreased by more than a factor of three, in comparison with lesion counting, and a factor of eight compared to simply counting new lesion tissue voxels. An alternative method, relying on a margin around the decision boundary rather than uncertainty, performed similarly to the label-flip confidence method, but only after the correct margin was found. We therefore tend to prefer the uncertainty-based method.

Furthermore, our method (trained on fifty cases from a single institution) also performs well when applied to two datasets from external centres. While detection of progression was perfect on the internal validation set, the method failed to identify progression at two time-points in the Zurich dataset and eight

time-points from the Munich data set. This was caused by small new lesions which were correctly identified, but too small to be identified confidently. For example, the two cases mislabelled as stable in the Zurich dataset each had a single, small new lesion. In the first case this was a small faint lesion in deep white matter, and in the second it was a small periventricular lesion. In both cases these lesions were correctly segmented by DeepSCAN, but not at a sufficient level of confidence to deem them confident new lesion tissue. Representative slices from these two cases are shown in Figures 3 and 4. A representative slice from a further case from the external dataset, showing two correctly identified instances of lesion growth, is shown in Figure 5. We can hope that detection of missed lesions can be improved by training on larger, more diverse datasets, or by the inclusion of more sensitive sequences. In the case of the Munich dataset, a Double Inversion Recovery sequence was used by human raters in addition to FLAIR to identify lesions. Detection of lesions on FLAIR only was shown in a recent study to miss 27.6 % of new or grown lesions, compared to DIR.[5] It is therefore perhaps not surprising that some time-points labelled as stable were judged as progressive by the human raters, as the new lesions may not have been visible in the FLAIR sequence. This suggests that it would be worthwhile to extend our approach to incorporate DIR imaging. This would, however, limit the applicability of the technique in clinical practice. Alternatively, the proposed method could be used by a reader, in conjunction with segmentations from the separate time-points, to streamline semi-automatic detection of new lesions.

Semi-automated methods for MS lesion segmentation provide a simple method to assess the change in lesion load of an MS patient. Simple FLAIR image subtraction methods or background subtractions of binarized image have been used to manually identify new lesion tissue with high accuracy and low error rates. Other methods included graph cuts, i.e. graph-based segmentation techniques that employ seed points set by the user and a cost function or active contouring using prior information. These methods still require a degree of human interaction, are time consuming and require an expert-in-the-loop. Currently,

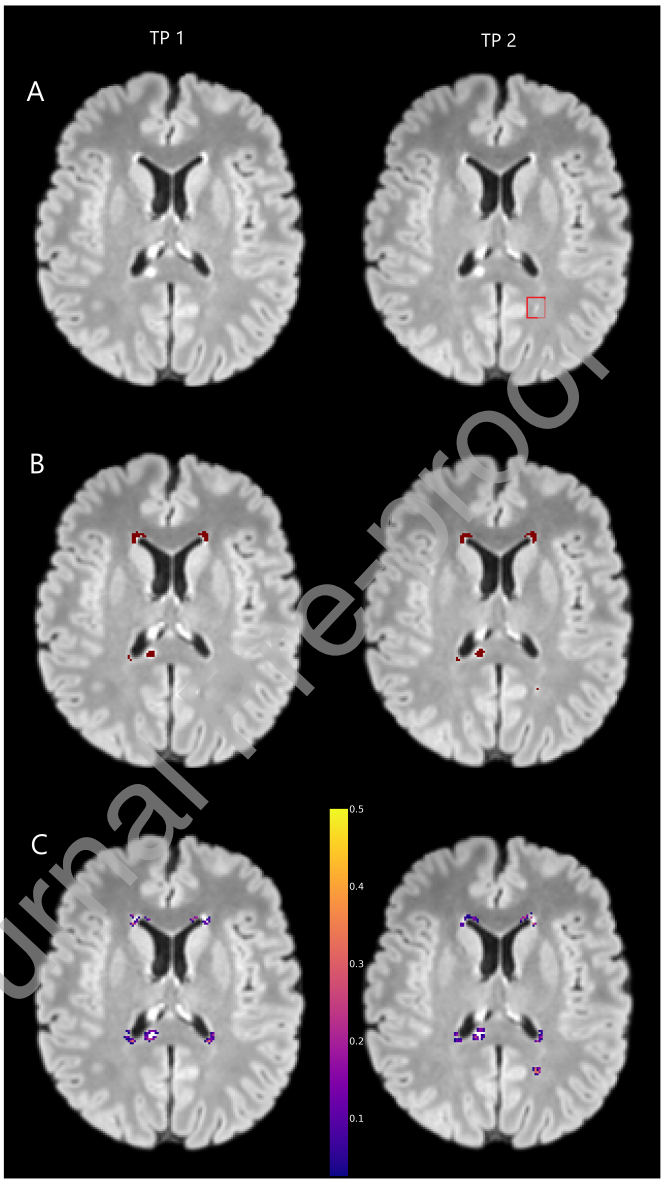


Figure 3: Two time-points from the external dataset, showing a missed new lesion. (A) coregistered FLAIR, (B) lesion segmentations, (C) Label-flip maps. New lesion is correctly detected by DeepSCAN at TP2 , but not labelled as confident new lesion. Small, faint lesions are more likely to be labelled as uncertain than large, clear lesions.

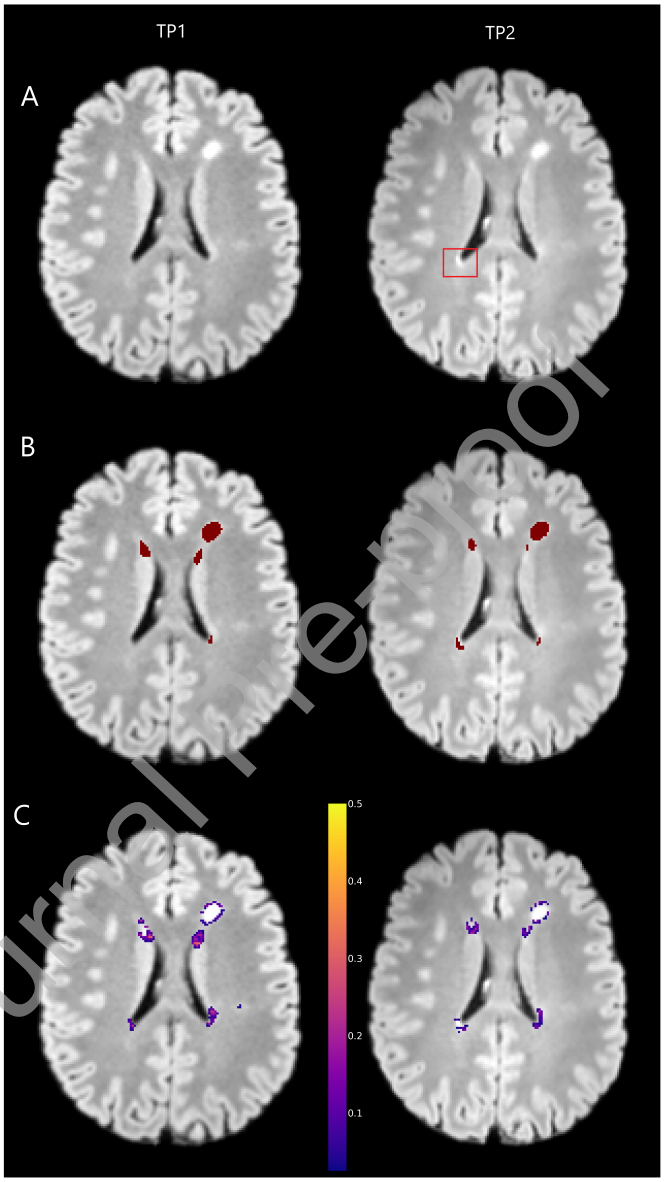


Figure 4: Two time-points from the external dataset, showing a missed new periventricular lesion. (A) coregistered FLAIR, (B) lesion segmentations, (C) Label-flip maps. Lesion is detected by DeepSCAN at TP2, but location of new lesion is uncertain at TP1. Owing to the similar appearance of periventricular lesions and subependymal gliosis, label confidence is typically low in this region.

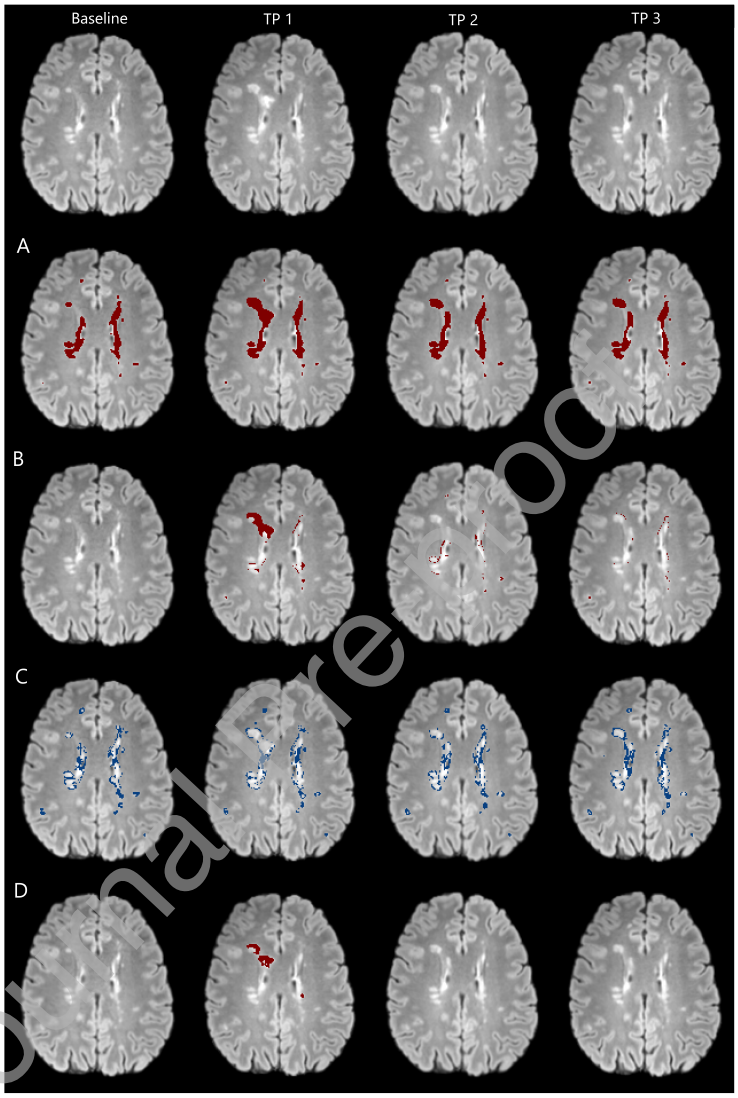


Figure 5: A case from the Zurich dataset. Top Row: FLAIR imaging at baseline and three subsequent time-points. A: FLAIR images with lesion masks as provided by the DeepSCAN classifier. B: FLAIR images with masks indicating naive lesion change (lesion is absent at previous time-point but present at current time-point). time-points 3 and 4 show new lesion tissue due to differences in imaging, rather than genuine lesion growth. C: Regions where DeepSCAN flip probability > 0.05 highlighted in blue. D: Confident new lesion tissue maps as provided by the method, showing correctly detected new lesion tissue at time-point 2, and no change at time-points 3 and 4

substantial effort is being invested in the development of fully-automated lesion annotation methods, and results indicate that advances in model architecture and training techniques, together with increasing availability of expert-labelled data, have brought us close to, or even allow us to exceed, the performance of expert human raters [25, 17]. However, in the study at hand, we could demonstrate that despite the effectiveness of automated lesion segmentation, automatically detected *changes* in lesion volume in MS patients alone is not a sufficient method for performing separation between radiologically progressive course from radiologically stable patients. Instead, we propose a method for identifying lesion changes of high certainty. We conclude that, while solitary lesion volume or total lesion load - together with clinical disease course / EDSS of MS patients - are strong predictors of disease course across a reference MS population, in the individual MS patient changes in these measures are not an adequate means to clear differentiate progressive disease course from no disease activity.

We believe that the performance shown by our method will encourage the MS community to investigate its use in different clinical settings. The benefits of automated methods lie not only in terms of the accuracy in differentiation of progressive versus stable disease course on MR imaging but also in the related reductions in time and economic costs derived from manual lesion labelling. While there is an increasing level of evidence that CNNs are comparable to human rater's performance in cross-sectional studies, only longitudinal clinical follow-up studies will demonstrate the utility of these methods for identifying patients who remain stable under DMT.

Acknowledgements

This research was supported by the Swiss Multiple Sclerosis Society and a grant from the Novartis Forschungsstiftung.

References

References

[1] D. L. Arnold, P. A. Calabresi, B. C. Kieseier, S. I. Sheikh, A. Deykin, Y. Zhu, S. Liu, X. You, B. Sperling, S. Hung, Effect of peginterferon beta-1a on MRI measures and achieving no evidence of disease activity: results from a randomized controlled trial in relapsing-remitting multiple sclerosis, BMC Neurology 14 (2014) 240.

[2] E. Havrdova, S. Galetta, M. Hutchinson, D. Stefoski, D. Bates, C. H. Polman, P. W. O'Connor, G. Giovannoni, J. T. Phillips, F. D. Lublin, A. Pace, R. Kim, R. Hyde, Effect of natalizumab on clinical and radiological disease activity in multiple sclerosis: a retrospective analysis of the natalizumab safety and efficacy in relapsing-remitting multiple sclerosis (AFFIRM) study, The Lancet Neurology 8 (2009) 254 – 260.

[3] E. Havrdova, G. Giovannoni, D. Stefoski, S. Forster, K. Umans, L. Mehta, S. Greenberg, J. Elkins, Disease-activity-free status in patients with relapsingremitting multiple sclerosis treated with daclizumab high-yield process in the select study, Multiple Sclerosis Journal 20 (2014) 464–470. PMID: 24022270.

[4] R. Nixon, N. Bergvall, D. Tomic, N. Sfikas, G. Cutter, G. Giovannoni, no evidence of disease activity: Indirect comparisons of oral therapies for the treatment of relapsing-remitting multiple sclerosis (2014).

[5] P. Eichinger, S. Schn, V. Pongratz, H. Wiestler, H. Zhang, M. Bussas, M.-M. Hoshi, J. Kirschke, A. Berthele, C. Zimmer, B. Hemmer, M. Mhlau, B. Wiestler, Accuracy of unenhanced MRI in the detection of new brain lesions in multiple sclerosis, Radiology 291 (2019) 429–435. PMID: 30860448.

[6] E. Erbayat Altay, E. Fisher, S. Jones, C. Hara-Cleaver, J.-C. Lee, R. Rudick, Reliability of classifying multiple sclerosis disease activity using

magnetic resonance imaging in a multiple sclerosis clinic, JAMA Neurol. 70 (2013) 338–44.

[7] J. D. Rudie, R. R. Mattay, M. Schindler, S. Steingall, T. S. Cook, L. A. Loevner, M. D. Schnall, A. C. Mamourian, M. Bilello, An initiative to reduce unnecessary gadolinium-based contrast in multiple sclerosis patients, Journal of the American College of Radiology 16 (2019) 1158 – 1164.

[8] B. Moraal, D. S. Meier, P. A. Poppe, J. J. G. Geurts, H. Vrenken, W. M. A. Jonker, D. L. Knol, R. A. van Schijndel, P. J. W. Pouwels, C. Pohl, L. Bauer, R. Sandbrink, C. R. G. Guttmann, F. Barkhof, Subtraction MR images in a multiple sclerosis multicenter clinical trial setting., Radiology 250 2 (2009) 506–14.

[9] R. McKinley, R. Wepfer, T. Gundersen, F. Wagner, A. Chan, R. Wiest, M. Reyes, Nabla-net: A deep dag-like convolutional architecture for biomedical image segmentation, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer International Publishing, Cham, 2016, pp. 119–128.

[10] S. Valverde, M. Cabezas, E. Roura, S. Gonzalez-Villa, D. Pareto, J. C. Vilanova, L. Ramio-Torrenta, A. Rovira, A. Oliver, X. Llado, Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach, Neuroimage 155 (2017) 159–168.

[11] S. Valverde, M. Salem, M. Cabezas, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, A. Rovira, J. Salvi, A. Oliver, X. Lladó, One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks, NeuroImage. Clinical (2018) 101638.

[12] M. Fartaria, A. Todea, T. Kober, K. O'brien, G. Krueger, R. Meuli, C. Granziera, A. Roche, M. Bach Cuadra, Partial volume-aware assessment of multiple sclerosis lesions, NeuroImage : Clinical 18 (2018) 245–253.

[13] A. Carass, S. Roy, A. Jog, J. L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C. H. Sudre, M. J. Cardoso, N. Cawley, O. Ciccarelli, C. A. Wheeler-Kingshott, S. Ourselin, L. Catanese, H. Deshpande, P. Maurel, O. Commowick, C. Barillot, X. Tomas-Fernandez, S. K. Warfield, S. Vaidya, A. Chunduru, R. Muthuganapathy, G. Krishnamurthi, A. Jesson, T. Arbel, O. Maier, H. Handels, L. O. Iheme, D. Unay, S. Jain, D. M. Sima, D. Smeets, M. Ghafoorian, B. Platel, A. Birenbaum, H. Greenspan, P.-L. Bazin, P. A. Calabresi, C. M. Crainiceanu, L. M. Ellingsen, D. S. Reich, J. L. Prince, D. L. Pham, Longitudinal multiple sclerosis lesion segmentation: Resource and challenge, NeuroImage 148 (2017) 77 – 102.

[14] S. Jain, A. Ribbens, D. M. Sima, M. Cambron, J. De Keyser, C. Wang, M. H. Barnett, S. Van Huffel, F. Maes, D. Smeets, Two time point ms lesion segmentation in brain MRI: An expectation-maximization framework, Frontiers in Neuroscience 10 (2016) 576.

[15] P. Schmidt, C. Gaser, M. Arsic, D. Buck, A. Förschler, A. Berthele, M. Hoshi, R. Ilg, V. J. Schmid, C. Zimmer, B. Hemmer, M. Mühlau, An automated tool for detection of flair-hyperintense white-matter lesions in multiple sclerosis, NeuroImage 59 (2012) 3774–3783.

[16] M. Salem, M. Cabezas, S. Valverde, D. Pareto, A. Oliver, J. Salvi, lex Rovira, X. Llad, A supervised framework with intensity subtraction and deformation field features for the detection of new t2-w lesions in multiple sclerosis, NeuroImage: Clinical 17 (2018) 607 – 615.

[17] R. McKinley, R. Wepfer, F. Aschwanden, L. Grunder, R. Muri, C. Rummel, R. Verma, C. Weisstanner, M. Reyes, A. Salmen, A. Chan, F. Wagner, R. Wiest, Simultaneous lesion and neuroanatomy segmentation in multiple sclerosis using deep neural network, eprint, available at https://arxiv.org/abs/1901.07419 (2019).

[18] T. Nair, D. Precup, D. L. Arnold, T. Arbel, Exploring uncertainty measures

in deep networks for multiple sclerosis lesion detection and segmentation, in: Proc. MICCAI 2018, 2018.

[19] F. Cotton, S. Kremer, S. Hannoun, S. Vukusic, V. Dousset, OFSEP, a nationwide cohort of people with multiple sclerosis: Consensus minimal MRI protocol, Journal of Neuroradiology 42 (2015) 133 – 140.

[20] C. H. Polman, S. C. Reingold, B. Banwell, M. Clanet, J. A. Cohen, M. Filippi, K. Fujihara, E. Havrdova, M. Hutchinson, L. Kappos, F. D. Lublin, X. Montalban, P. O'Connor, M. Sandberg-Wollheim, A. J. Thompson, E. Waubant, B. Weinshenker, J. S. Wolinsky, Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria, Annals of Neurology 69 (2011) 292–302.

[21] R. McKinley, R. Maier, R. Wiest, Ensembles of densely-connected cnns with label-uncertainty for brain tumor segmentation, in: A. Crimi, S. Bakas, H. Kuijf, B. Menze, M. Reyes (Eds.), Proc Brainles 2018, Springer International Publishing, Cham, 2019, pp. 456–465.

[22] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (2015) 234–241.

[23] G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 2017.

[24] R. McKinley, M. Rebsamen, R. Meier, M. Reyes, C. Rummel, R. Wiest, Few-shot brain segmentation from weakly labeled data with deep heteroscedastic multi-task networks, arXiv e-print, available at https://arxiv.org/abs/1904.02436 (2019).

[25] O. Commowick, A. Istace, M. Kain, B. Laurent, F. Leray, M. Simon, S. Pop, P. Girard, R. Améli, J. Ferré, A. Kerbrat, T. Tourdias, F. Cervenansky, T. Glatard, J. Beaumont, S. Doyle, F. Forbes, J. Knight, A. Khademi,

A. Mahbod, C. Wang, R. McKinley, F. Wagner, J. Muschelli, E. Sweeney, E. Roura, X. Lladó, M. Santos, W. Santos, A. Silva-Filho, X. Tomas-Fernandez, H. Urien, I. Bloch, S. Valverde, M. Cabezas, F. Vera-Olmos, N. Malpica, C. Guttmann, S. Vukusic, G. Edan, M. Dojat, M. Styner, S. Warfield, F. Cotton, C. Barillot, Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure, Scientific Reports 8 (2018).

[26] M. Reuter, N. J. Schmansky, H. D. Rosas, B. Fischl, Within-subject template estimation for unbiased longitudinal image analysis, NeuroImage 61 (2012) 1402 – 1418.

[27] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: Proceedings, ICML, 2017, 2017.

List of changes