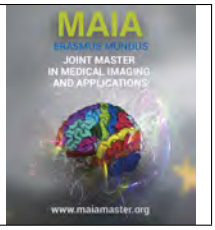
Medical Imaging and Applications

Master Thesis, August 2020



Multiple Sclerosis Lesion Segmentation Using Longitudinal Normalization and Convolutional Recurrent Neural Networks

Sergio Tascon-Morales, Martin Treiber, Stefan Hoffmann, Johannes Gregori

mediri GmbH, Heidelberg, Germany

Abstract

Magnetic resonance imaging (MRI) is the primary clinical tool to examine inflammatory brain lesions in Multiple Sclerosis (MS). Disease progression and inflammatory activities are examined by longitudinal image analysis to support diagnosis and treatment decision. Automated lesion segmentation methods based on deep convolutional neural networks (CNN) have been proposed, but are not yet applied in the clinical setting. Typical CNNs working on cross-sectional single time-point data have several limitations: changes to the image characteristics between single examinations due to scanner and protocol variations have an impact on the segmentation output, while at the same time the additional temporal correlation using pre-examinations is disregarded.

In this work, we investigate approaches to overcome these limitations. Within a CNN architectural design, we propose to use convolutional Long Short-Term Memory (C-LSTM) networks to incorporate the temporal dimension. To reduce scanner- and protocol dependent variations between single MRI exams, we propose a histogram normalization technique as pre-processing step. The ISBI 2015 challenge data were used for cross-validation.

We demonstrate that the combination of the longitudinal normalization and CNN architecture can increase the performance and the inter-time-point stability of the lesion segmentation. The proposed longitudinal architecture produced the highest Dice scores for all the analyzed cases. Furthermore, the combination of the proposed architecture and normalization led to the lowest variation for the Dice score, denoting a higher consistency of the results. The proposed methods can therefore be used to increase the performance and stability of fully automated lesion segmentation applications in the clinical routine or in clinical trials.

Keywords: Segmentation, multiple sclerosis, magnetic resonance imaging (MRI), deep learning, convolutional neural networks, recurrent neural networks, longitudinal normalization

1. Introduction

Multiple sclerosis (MS) is a chronic inflammatory disease of the central nervous system (CNS) that produces demyelination and axonal/neuronal damage (Cohen and Rae-Grant, 2012). The demyelinating process is associated with persistent inflammation throughout the CNS and, as a result, the demyelinated lesion, also known as plaque, is the main pathological feature of MS (Arnon and Miller, 2016; Compston et al., 2005). In terms of location of the lesions, there is a predilection for the periventricular white matter, optic nerves, brainstem, cerebellum and spinal cord (Lucchinetti and Parisi, 2006). Although the etiology of MS remains unknown, the disease appears to be determined by both genetic and environmental factors (Pryse-Phillips and Sloka, 2006). An autoimmune etiology has also been suggested, but remains unproven (Rinker II et al., 2006).

The course of MS can be described in terms of relapses, remissions and chronic progression either from onset or after a period of remissions (Compston et al., 2005). Relapses (attacks) are considered to represent the clinical correlate of recurrent episodes of inflammation and demyelination, often with axonal injury, in the CNS. Remissions are probably due to remyelination and resolution of inflammation and progression is believed to reflect a combination of ongoing demyelination, gliosis and axonal loss (Lucchinetti and Parisi, 2006). Four categories are commonly used to classify

the disease course: Relapsing-remitting, secondary progressive, primary progressive and progressive-relapsing (Compston et al., 2005). Relapsing-remitting MS is characterized by recurrent CNS inflammation with stable clinical manifestations between episodes, whereas in secondary progressive MS there is a gradual neurological deterioration, which occurs with or without superimposed relapses. Both primary progressive and progressive-relapsing MS exhibit gradual neurological deterioration from onset as main feature, but in the case of progressive-relapsing there are also superimposed relapses (Cohen and Rae-Grant, 2012).

According to the MS Atlas, which is a study carried out in 2008 and updated in 2013 by the World Health Organization (WHO) and the MS International Federation (MSIF), there were about 2.1 million people in the world with the disease in 2008 and 2.3 million in 2013. A more recent study reports about 2.2 million people with MS in the world and 18.932 deaths due to MS in 2016 (Wallin et al., 2019). The study also reports greater age-standardized prevalence in North America and some northern European countries (more than 120 cases per 100.000 population), moderate in some countries of Europe and Australasia (60-120 cases per 100.000 population) and lowest in North Africa and the Middle East, Latin America, Asia, Oceania, the Caribbean, and sub-Saharan Africa (<60 cases per 100.000 population).

Besides the uneven geographical distribution, MS has particular incidence and prevalence depending on sex and age. There is a female predominance of approximately 2.5 to 1 (Cohen and Rae-Grant, 2012). Regarding the onset age, although the disease can occur at virtually any age, the incidence of MS is low in childhood, with onset younger than age 10 occurring in about 0.3% of cases. After the age of 18 the incidence increases, reaching a peak between 25 and 35 and then declining. For this reason, MS is the most common non-traumatic neurological disease in young adults. Onset of the disease after the age of 50 is considered rare (Birenbaum and Greenspan, 2016; Lladó et al., 2012; Miller, 2006). The prevalence of MS is similar for boys and girls among preteen children. Divergence appears during adolescence, with higher prevalence among girls as compared to boys. This pattern continues until around the end of the sixth decade of life, when the sex ratio is about 2 to 1 in favor of women. For older people prevalence shows a continued increase for women, while for men there is a slow attenuation (Wallin et al., 2019).

Diagnosis of MS can involve several techniques or approaches that include physical examination, Magnetic Resonance Imaging (MRI), cerebrospinal fluid (CSF) analysis and evoked potentials (Cohen and Rae-Grant, 2012). MRI is widely used for diagnosis and monitoring of MS, due to the high sensitivity that it has for depicting white matter lesions, particularly in terms of dissemination in time and space, which is an important diagnostic criteria (Salem et al., 2019). Depending on the modality or sequence being examined, lesions may appear as hyperintensities, like in the case of T2-weighted (T2w), Proton Density weighted (PDw) and Fluid Attenuated Inversion Recovery (FLAIR), or as hypointensities, like in the case of T1-weighted (T1w) images (Brosch et al., 2016b). Imaging biomarkers such as lesion load and lesion count, which are based on delineation of lesions, are important for determining the progression and treatment effects of MS (Brosch et al., 2016a). Although feasible in practice and considered as the gold standard, manual lesion segmentation from 3D scans is tedious, time-consuming and prone to errors caused by inter- and intra-rater variability (Andermatt et al., 2018; Roy et al., 2018; Valverde et al., 2017). For this reason, automated strategies have been proposed based on traditional machine learning and atlas based techniques (Lladó et al., 2012). More recently, deep neural networks have attracted interest, specially convolutional neural networks (CNN) by proving their effectiveness in tissue segmentation and also brain tumor segmentation (Salem et al., 2019).

From the perspective of how the data is used to train a model, MS lesion segmentation algorithms can be classified as either longitudinal of cross-sectional. Longitudinal approaches make use of the temporal information provided by subsequent scans (known as time-points or visits) of the same patient. In cross-sectional approaches, all scans, even if belonging to the same patient, are treated as independent scans and no time information is considered. Most of the automated methods for MS lesion segmentation found in the literature treat the data as cross-sectional even in the cases in which the images have been acquired in a longitudinal manner.

This master thesis focused on the impact of longitudinal approaches from two perspectives: (1) architecture and (2) data normalization. Both perspectives aim to exploit the temporal information of longitudinal data to produce more consistent segmentation of the MS lesions.

This document is organized as follows. Section 2 presents different normalization methods that have been proposed or adopted for longitudinal MRI in presence of MS lesions. It also presents the state of the art as a comparison between cross-sectional and longitudinal approaches, focusing on deep learning approaches. In Section 3 the methodology and materials are described in detail. The results and their corresponding analysis are presented in Sections 4 and 5, respectively. Finally, conclusions are provided in Section 6.

2. State of the art

2.1. Longitudinal Normalization

One important issue when using longitudinal data is the normalization across time-points, or longitudinal

normalization. The goal is to increase the similarity, in terms of image intensity regarding tissue classes, of the different time-points, without modifying the structures whose changes are due to pathological conditions. MS lesions are an example of those structures, as they can persist, change or disappear in time (Roy et al., 2013). A statistical normalization method is proposed by Shinohara et al. (2014), in which all histograms are centered using statistical measures obtained from the white matter voxels. Sweeney et al. (2013) followed a very similar approach by expressing intensities as a departure from the white matter mean. Other methods based on matching of histograms use landmarks from a reference template to increase similarity through a piecewise linear transformation (Nyúl et al., 2000). This type of approaches can cause, however, undesired mappings, altering key anatomical structures (Roy et al., 2013).

Another longitudinal normalization method was introduced by Roy et al. (2013). In this case voxel changes in time are modeled mathematically depending on the behaviour and the lesion priors are used for keeping the lesion voxels unchanged.

2.2. MS Lesion Segmentation

Automated MS lesion segmentation is not a trivial task due to the fact that lesions vary in size, shape, intensity and location within the brain (Brosch et al., 2016a). A wide variety of algorithms have been proposed in the past to address this problem. We can distinguish between traditional machine learning approaches and deep-learning-based approaches. In the traditional machine learning group, algorithms based on both supervised and unsupervised learning can be found. The supervised learning subgroup includes algorithms based on probabilistic atlases and algorithms that are trained with manual segmentation masks. In the unsupervised learning subgroup, methods can either focus on segmenting brain tissue and detecting lesions as outliers, or they can directly focus on segmenting the lesions (Lladó et al., 2012).

With some recent exceptions such as the one proposed by Wang et al. (2020), in which a Bayesian model is built using Markov and Gibbs random field theorems, the vast majority of modern approaches are based on deep learning, to the point that deep learning approaches outnumber the approaches based on traditional machine learning. More specifically, deep convolutional neural networks (CNN) eliminate the need for handcrafted features or prior guidance and have shown, as mentioned before, outstanding performance in different brain imaging tasks. Furthermore, CNN-based approaches are now in the top of the rankings of international MS lesion segmentation challenges (Salem et al., 2019).

2.2.1. Cross-sectional MS lesion segmentation

Most of the deep learning approaches for MS lesion segmentation are cross-sectional, as shown in Fig. 1. Leading the entry of deep learning into the MS lesion segmentation field, Yoo et al. (2014) used a patch-based deep neural network to extract features that could then be used by a random forests classifier. Shortly thereafter, Vaidya et al. (2015) and Ghafoorian and Bram (2015) used 2D and 3D patch-based CNNs, respectively, not only for extracting features, but also for performing voxel classification using fully connected layers. Brosch et al. (2015) proposed an encoder-decoder called Convolutional Encoder Network (CEN) architecture without skip connections that used whole slices instead of patches. In an attempt to combine the advantages of the CEN with the classic U-Net (Ronneberger et al., 2015) architecture, the same authors added skip connections to the CEN model and used deconvolution instead of upsampling (Brosch et al., 2016a). Following the encoder-decoder architectures, McKinley et al. (2016) proposed to use several networks, one for each orientation (axial, sagittal and coronal) with only one skip connection at the top level. This multi-view style was also exploited by Aslani et al. (2018) using skip connections for all levels, and then by Aslani et al. (2019), who used three parallel independent encoders based on residual blocks, to generate features from three different modalities. These features were then combined and upsampled with one single decoder. The tendency towards the encoder-decoder architectures can be observed in several other approaches (Brugnara et al., 2020; Duong et al., 2019; Gabr et al., 2019; Narayana et al., 2020). As an alternative to this encoder-decoder architectures, a method based on convolutional recurrent neural networks was proposed by Andermatt et al. (2018), but the sequential power of the recurrent networks was not used for considering the time dimension, but rather for treating the spatial dimensions as sequential data. Their method is based on multi-dimensional gated recurrent units (GRU) and considers a filtered version of the images as an additional channel, under the assumption that the filtered images announce changes before they actually occur.

One of the problems faced when segmenting MS lesions is the number of false positive lesions that can be generated by an automated algorithm, due to the high class imbalance (Salehi et al., 2017). To address this problem, Valverde et al. (2017) proposed a special type of CNN architecture. It is a cascaded 3D CNN in which a first network is trained to have high sensitivity so that candidate lesions can be detected, and a second network is trained to reduce the amount of false positives (FP). One advantage of this architecture is that it allows domain adaptation, meaning that after being trained on a certain dataset, it does not have to be completely re-trained for evaluation on another dataset. Fur-

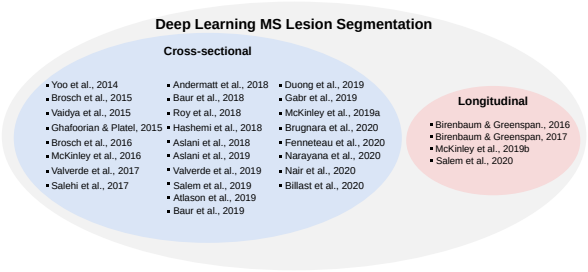


Figure 1: Overview of deep learning methods for MS lesion segmentation according to categories cross-sectional and longitudinal. For years 2014 to 2019 publications with at least 3 citations are considered, whereas for 2020 all found publications are included.

thermore, only some of the fully connected layers have to be re-trained with few new examples (Valverde et al., 2019). Instead of focusing on the architecture itself, Salehi et al. (2017) and Hashemi et al. (2019) used an asymmetric loss function based on the Tversky index. This loss function, which is a generalization of the Dice coefficient and the F_β scores, allows to give more importance to sensitivity or to precision, as determined by two parameters α and β.

Although supervised learning is the predominant type of learning for MS lesion segmentation, other types of learning such as unsupervised (Atlason et al., 2019; Baur et al., 2019b), semi-supervised (Baur et al., 2019a) and self-supervised (Fenneteau et al., 2020) have been explored too.

2.2.2. Longitudinal MS Lesion Segmentation

Only few deep learning approaches can be found in the literature that tackle the problem of MS lesion segmentation in a longitudinal manner. The first CNN-based longitudinal method found in the literature was pro-posed by Birenbaum and Greenspan (2016, 2017). Although their CNN is used only for classifying candidates extracted using intensity and atlas information, the method employs different time-points to perform the task.

McKinley et al. (2020) proposed a method to detect the lesion load change using CNNs. To achieve this, an architecture known as DeepSCAN was used as basis, which is a hybrid between the U-Net and the Dense-Net (Huang et al., 2017). A special type of loss function allows to output, for each voxel and tissue class, the probability that a voxel contains the tissue class, as well as the probability that the predicted label does not match the label of the ground truth. These probabilities and the mask provided by the model are then used to obtain information about lesion change. Following this idea of detecting changes, Salem et al. (2020) proposed an architecture that consists of a first block based on Voxelmorph (Balakrishnan et al., 2019) to learn deformation fields and register baseline image to the follow-up images, and a second block to perform the segmentation of new lesions using the results of the first block.

3. Materials and methods

3.1. Dataset and Pre-processing

One of the most popular datasets for MS lesion segmentation is the one provided by the longitudinal MS lesion segmentation challenge, which was part of the International Symposium on Biomedical Imaging (ISBI) in 2015 and continues to be publicly available. The dataset, acquired with a 3T scanner, is subdivided into training (5 subjects) and testing (14 subjects) sets. Only the training set contains lesion segmentation masks generated by two different expert raters. These masks will be referred to as *mask1* and *mask2* in this document. Each subject contains between 4 and 6 time-points, each of which consists of a T1-weighted (T1-w) magnetization prepared rapid gradient echo (MPRAGE) with TR = 10.3 ms, TE = 6 ms, flip angle = 8°, $0.82 \times 0.82 \times 1.17$ mm^3 voxel size; a double spin echo (DSE) which produces the PD-w and T2-w images with TR = 4177 ms, TE1 = 12.31 ms, TE2 = 80 ms, $0.82 \times 0.82 \times 2.2$ mm^3 voxel size; and Fluid Attenuated Inversion Recovery (FLAIR) with TI = 835 ms, TE = 68 ms, $0.82 \times 0.82 \times 2.2$ mm^3 voxel size. The average time between subsequent time-points is 1 year (Carass et al., 2017; IACL, 2018).

Both the original and the pre-processed images are available for use. The pre-processing steps for each subject are the following: First, the baseline (first time-point) MPRAGE image is corrected using the N4 algorithm, then it is skull-stripped, then dura stripped. After this a second N4 correction takes place and, finally, it is registered to a 1 mm isotropic MNI template. This pre-processed baseline MPRAGE image is then used as target for remaining images of the current patient, which are N4 corrected and then rigidly registered to the baseline MPRAGE image. The masks obtained from skull and dura stripping the baseline image are used on the remaining images (IACL, 2018). For this work, the pre-processed images were used.

3.2. Longitudinal Normalization

A simple yet effective MRI longitudinal normalization based on the Chi-Square metric χ^2 is proposed. The Chi-Square test is commonly used for analyzing the difference between observed and expected distributions (Weaver et al., 2017), but in this case only the metric is used to measure and maximize the similarity between the histograms of volumes of the same modality for different patients and different time-points. Eq. 1 shows the Chi-Square metric as a means of comparison of two histograms H_a and H_b, for voxel intensities I.

Let s, t and m represent subject, time-point and modality, respectively. For each modality m a reference volume $V^{(m)}_{\hat{s}\hat{t}}$ is selected to normalize the other volumes

$V_{st}^{(m)}$ of that modality, with $s \neq \hat{s}, t \neq \hat{t}$. For each $V_{st}^{(m)}$ an optimal scalar $\theta_{st}^{(m)}$ is found using Eq. 2, where $H_{\hat{s}\hat{t}}^{(m)}$ and $H_{st}^{(m)}$ are the histograms of the normal appearing white matter (NAWM) of $V_{\hat{s}\hat{t}}^{(m)}$ and $V_{st}^{(m)}$, respectively. The normalized images are the result of the product $\theta_{st}^{(m)} V_{st}^{(m)}$. To obtain the NAWM masks, the Computational Anatomy Toolbox (CAT12) applied within the Statistical Parametric Mapping (SPM12) toolkit was used (Gaser, C., Dahnke, 2016; Penny et al., 2011).

$$dist_{\chi^2}(H_a, H_b) = \sum_I \frac{[H_a(I) - H_b(I)]^2}{H_a(I)} \quad (1)$$

$$\theta_{st}^{(m)} = \underset{\theta}{\operatorname{argmin}} \sum_I \frac{[H_{\hat{s}\hat{t}}^{(m)}(I) - H_{st}^{(m)}(\theta \cdot I)]^2}{H_{\hat{s}\hat{t}}^{(m)}(I)} \quad (2)$$

3.3. Patch sampling

Training a patch-based model that considers multiple time-points and multiple MRI modalities requires a proper temporal sampling strategy. In this work, patches with dimensions (T, M, H, W, D) are used, where T and M represent the number of selected time-points to process for each sample and the number of modalities, respectively. H, W and D represent the spatial dimensions, i.e. the height, width and depth of the patches in each volume.

We can subdivide the sampling process into spatial sampling, modality sampling and time sampling. Spatial sampling determines how the 3D patches are selected within each 3D volume. Sub-patches with size (H, W, D) are extracted for each subject in an uniform way from the brain only, using a brain mask generated as the non-zero voxels of the FLAIR image of the first time-point. Because of the multi-modal and longitudinal character, the selected sub-patches are also extracted across modalities and across time-points, as determined by the modality and time sampling.

Modality sampling refers to which modalities are used for generating the patches. Regarding modality sampling, considering all four modalities has been shown to bring the best performances in MS lesion segmentation as compared to using only some of them (Narayana et al., 2020). For this reason, all four available modalities are used, therefore $M = 4$ in all cases.

Finally, sampling in time refers to how the patches are selected across time-points, as determined by the desired number of time-points to be analyzed in each sample (parameter T). This sampling is made by slicing a window of size T through all available time-points. Choosing an odd value for T becomes convenient, so that the segmentation can be provided for the time-point in the middle, which is possible thanks to the bi-directional implementation of the C-LSTMs, as explained later on. This, however, raises the question about how to segment the $\lfloor T/2 \rfloor$ first and last time-points. This was solved by applying time padding, i.e. by repeating the first and last $\lfloor T/2 \rfloor$ time-points. Fig. 2 shows this strategy for the case when $T = 3$. The first and last time-points are copied for all modalities, which is indicated with blue arrows in the figure. It is also shown which time-point is selected in the ground truth, which, as mentioned before, is chosen to be the one in the middle of the window. Thus, this padding allows to generate samples in the positions where the sliding window would not have information available.

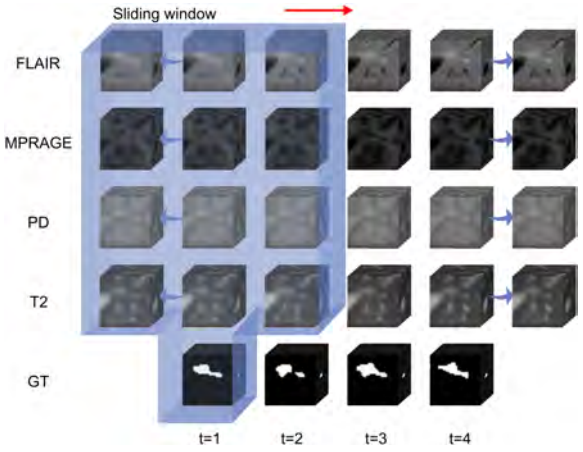


Figure 2: Time padding strategy when $T = 3$. First and last time-points are repeated for all modalities (blue arrows). The red arrow indicates how the temporal sliding window moves.

3.4. Architecture

In order to exploit temporal information, a 3D extension of the architecture presented by Novikov et al. (2019) is proposed for the segmentation of MS lesions from longitudinal multi-modal brain images. This architecture is a hybrid between the well known U-Net and a variant of the Convolutional Long Short-Term Memory (C-LSTM) network (Xingjian et al., 2015). Figure 4 shows the architecture. Just like in the U-Net, there is an encoder for extracting hierarchical features, but these are extracted for each time-point separately. These features are then combined, at the deepest level and for all input time-points, by a bidirectional C-LSTM. After the features are processed by the first C-LSTM, a decoder upsamples them so that the input dimensions can be reached again. At the output of the decoder, a second bidirectional C-LSTM combines the features of the different time-points again. Finally, the feature maps corresponding to a specific time-point (e.g. the one in the middle, if T is odd) are selected and a last convolution takes place to reduce the number of maps to 2, one for each class, lesion or non-lesion. The selection of a time-point after the second C-LSTM implies that when training the network, the ground truth mask of the same time-point must be used, as indicated previously in Fig. 2.

The inner structure of the units or cells that compose each bidirectional C-LSTM block is shown in Fig. 3, where C and h correspond to the cell and hidden states, respectively. Contrary to traditional LSTM networks used in other fields and although not visible in the figure, the C-LSTM uses convolutions (Xingjian et al., 2015), as determined by Eq. 3 to 8, where σ corresponds to the sigmoid function, tanh is the hyperbolic tangent function, $*$ denotes convolution and $\circ$ represents the Hadamard product.

Figure 5 shows how the C-LSTM blocks are built for the case when $T = 3$. The bidirectional nature is achieved by processing the sequences in both possible directions of the time dimension and then adding the outputs. This allows to better capture the temporal behaviour of the lesions and also makes possible to take advantage of using the time-point in the middle during training.

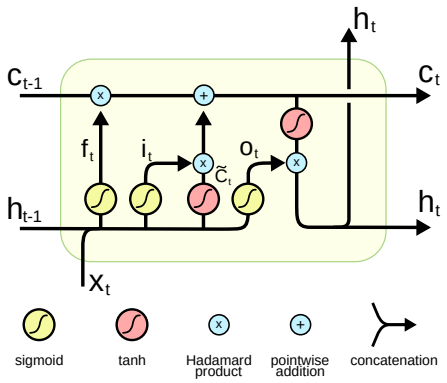


Figure 3: C-LSTM basic unit. Convolutions described by Eq. 3 to 8 are not shown in the figure. Diagram based on Phi (2018).

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i) \tag{3}$$

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f) \tag{4}$$

$$o_t = \sigma(W_o * [h_{t-1}, x_t] + b_o) \tag{5}$$

$$\tilde{C}_t = \tanh(W_C * [h_{t-1}, x_t] + b_C) \tag{6}$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t \tag{7}$$

$$h_t = o_t \circ \tanh(C_t) \tag{8}$$

3.5. Post-Processing

After a segmentation is produced, a post-processing step is performed to exclude potential false-positive (FP) detected lesions. This is achieved by imposing a minimal lesion size of 3 mm^3, as it has been found to improve the performance of MS lesion segmentation methods (Fartaria et al., 2018).

3.6. Experimental Setup

3.6.1. Normalization Configuration

One important step in the proposed longitudinal normalization is the white matter segmentation, which was performed on each volume using CAT12 with the default parameters. For finding the values of $\theta_{st}^{(m)}$, the first time-point of subject 01 was selected as reference for each modality. The Nelder–Mead Simplex method (Dennis Jr and Woods, 1985) was employed for the minimization of the distance function.

For comparison purposes, the min-max normalization (Eq. 9) and the standardization (Eq. 10) are also considered, since they are widely used in MS lesion segmentation. In min-max normalization the intensity values are mapped to the interval $[0, 1]$, whereas in standardization the goal is to have zero mean and standard deviation one. In Eq. 9, I_{orig}, I_{min} and I_{max} represent the original, minimum and maximum intensities of a volume, respectively, and I_{norm} is the assigned intensity. In Eq. 10 the term μ corresponds to the mean of the intensities and σ is the standard deviation. I_{orig} and I_{norm} have the same meaning explained for Eq. 9.

$$I_{norm} = \frac{I_{orig} - I_{min}}{I_{max} - I_{min}} \tag{9}$$

$$I_{norm} = \frac{I_{orig} - \mu}{\sigma} \tag{10}$$

3.7. Training and Cross-validation

After having normalized the pre-processed images, a leave-one-out (subject-wise) cross-validation was performed on the training set. For each fold, from the 4 subjects not used for testing, one was used for validation and 3 for training. The model was trained using 32×32×32 spatial patches with step size 16×16×16. All four modalities were used ($M = 4$) and three time-points were considered for each training sample ($T = 3$). This means the size of each sample patch is (3,4,32,32,32). Training was performed using the Adam optimizer (Kingma and Ba, 2015) for a maximum of 200 epochs with an early stopping condition of 20 epochs, and a batch size of 16. To reduce the effect of the class imbalance (more normal tissue as compared to lesion tissue), the dice loss function (Milletari et al., 2016) was used, as defined in Eq. 11, where p_i and g_i denote the predicted binary segmentation and ground truth binary volume, respectively, and N is the total number of voxels. All models were separately trained for both available masks, and the subjects were assigned for each subject according to Table 1, where the validation subjects were randomly chosen once and then set to be the same for all experiments. No data augmentation was performed in order to increase the comparability between different experiments.

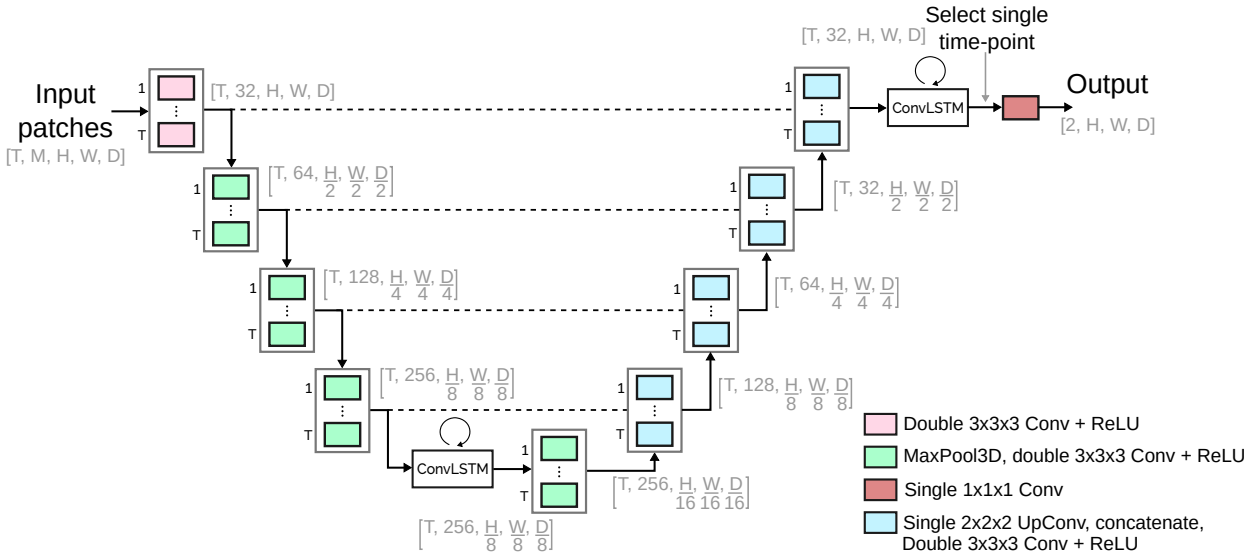


Figure 4: U-Net ConvLSTM architecture. Patch dimensions are included in gray text, where T and M denote the number of selected time-points and number of modalities, respectively. H, W and D denote the spatial dimensions of the patches in the volumes. Horizontal dashed lines denote skip connections by copying and concatenation.

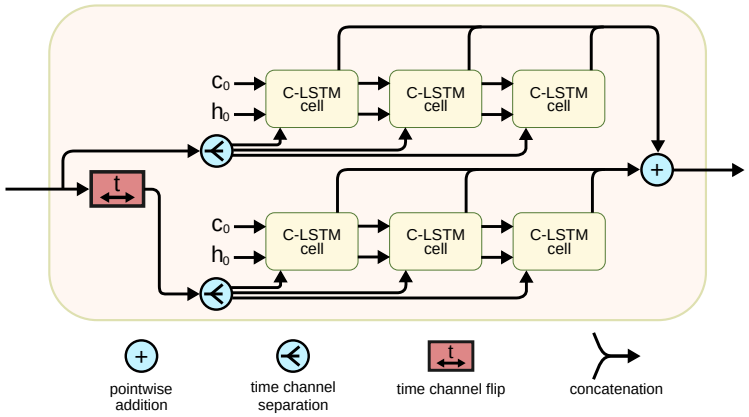


Figure 5: Bidirectional C-LSTM block for the case when $T = 3$. Both the cell and hidden states C_0 and h_0 are initialized to zero for the first unit.

$$\mathcal{L}_{dice} = 1 - \frac{2\sum_{i}^{N} p_i g_i}{\sum_{i}^{N} p_i^2 + \sum_{i}^{N} g_i^2} \tag{11}$$

Table 1: Cross-validation subject selection

Fold	Train	Validation	Test
1	02, 04, 05	03	01
2	01, 04, 05	03	02
3	01, 02, 05	04	03
4	01, 02, 03	05	04
5	01, 03, 04	02	05

A cross-sectional version of the model, was also trained under the same parameters described before. This cross-sectional model, which is shown in Fig. 6, is the version resulting from the proposed model when the C-LSTM blocks are removed and the time-dimension is not included in the samples.

Both models (U-Net and U-Net ConvLSTM) were trained using cross-validation for the three described normalization methods (min-max, standardization and the proposed one), and for both available segmentation masks (mask1 and mask2). This means a total of 12 experiments were carried out to determine the advantages of the proposed normalization method as well as the advantages of incorporating time information to the U-Net with the bi-directional C-LSTM blocks.

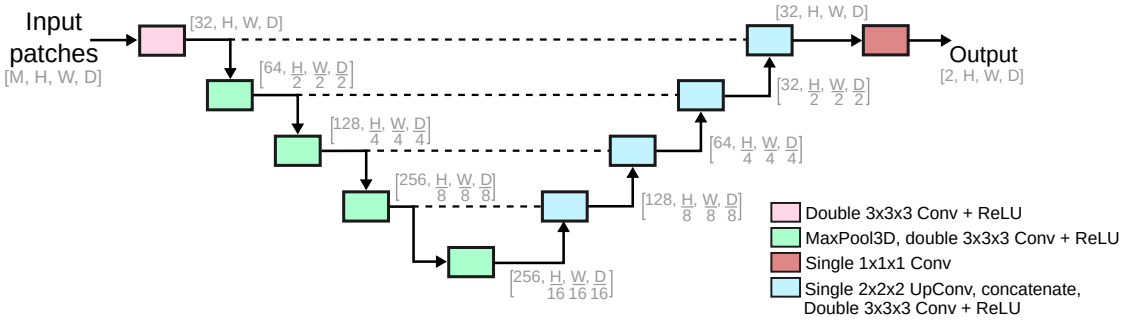


Figure 6: U-Net architecture. Patch dimensions are included in gray text, where M denotes number of modalities, respectively. H, W and D denote the spatial dimensions of the patches in the volumes. Dashed lines denote skip connections implemented by copying and concatenation.

3.8. Evaluation Metrics

To evaluate the performance of the longitudinal method, the Dice score (DSC), lesion-wise false positive rate (LFPR) and lesion-wise true positive rate (LTPR) were used. The DSC is computed according to Eq. 12, where TP, FP and FN denote number of true positive, false positive, and false negative voxels, respectively. The LFPR (Eq. 13) is the number of lesions in the produced segmentation that do not overlap with a lesion in the ground truth, divided by the total number of lesions in the produced segmentation. The LTPR (Eq. 14) is computed as the number of lesions in the ground truth that overlap with a lesion in the produced segmentation, divided by the total number of lesions in the ground truth (Aslani et al., 2018).

$$DSC = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{12}$$

$$LFPR = \frac{LFP}{\#PL} \tag{13}$$

$$LTPR = \frac{LTP}{\#RL} \tag{14}$$

3.9. Implementation

The models were implemented in PyTorch, using a GPU NVIDIA Tesla T4.

4. Results

The histograms after normalization are presented in Fig. 7 for all pre-processed training subjects of the dataset. Background voxels are ignored for the computation of the histograms.

Tables 2 to 5 show the results of the cross-validation process for all four possible combinations: training and evaluation with mask1 (Table 2), training with mask1 and evaluation with mask2 (Table 3), training with mask 2 and evaluation with mask1 (Table 4), and training and evaluation with mask2 (Table 5). The metrics are computed as global averages for all time-points of all subjects and standard deviations are shown in parentheses. When the same ground truth is used for both training and evaluation (Tables 2 and 5), the proposed pipeline produces the best DSC (0.711) in the case of mask1 and the second best (0.676) in the case of mask2, compared to the other evaluated methods. In both situations, however, the proposed pipeline leads to the lowest standard deviation of the DSC. When different masks are used for training and evaluation (Tables 3 and 4), the proposed pipeline leads to the highest DSC and also the lowest standard deviations. The best results could be achieved when training and evaluating the network on mask1, as well as when training on mask2 and evaluating on mask1 (Tables 2 and 4), with a DSC of > 0.71 and standard deviation of $\leqslant 0.085$.

To demonstrate how the standard deviation changes for the different evaluated methods, Fig. 8 and 9 show scatter plots of the DSC metric for the cases in which both training and evaluation are performed using the same ground truth. Particularly, with respect to the simple cross-sectional model with min-max normalization (leftmost), the proposed pipeline (rightmost) reduces the standard deviation of the DSC by 56.2% and 33.8% for mask1 and mask2, respectively. Especially, the amount of results with low DSC is reduced.

Resulting lesion segmentation examples are shown as overlay on the FLAIR images in Fig. 10 and 11 for one slice of specific subjects.

5. Discussion

5.1. Longitudinal Normalization

A longitudinal pipeline for the segmentation of MS lesions has been presented in this document. The first step in the pipeline is the longitudinal normalization, which is based on the optimization of the Chi-Square metric, and which is performed not only across time-points for every single subject, but also across all subjects. This allows to increase the homogeneity of the whole dataset while preserving the contrast characteristics of the lesions and other structures. As shown in Fig 7, the alignment of the histograms is higher for the

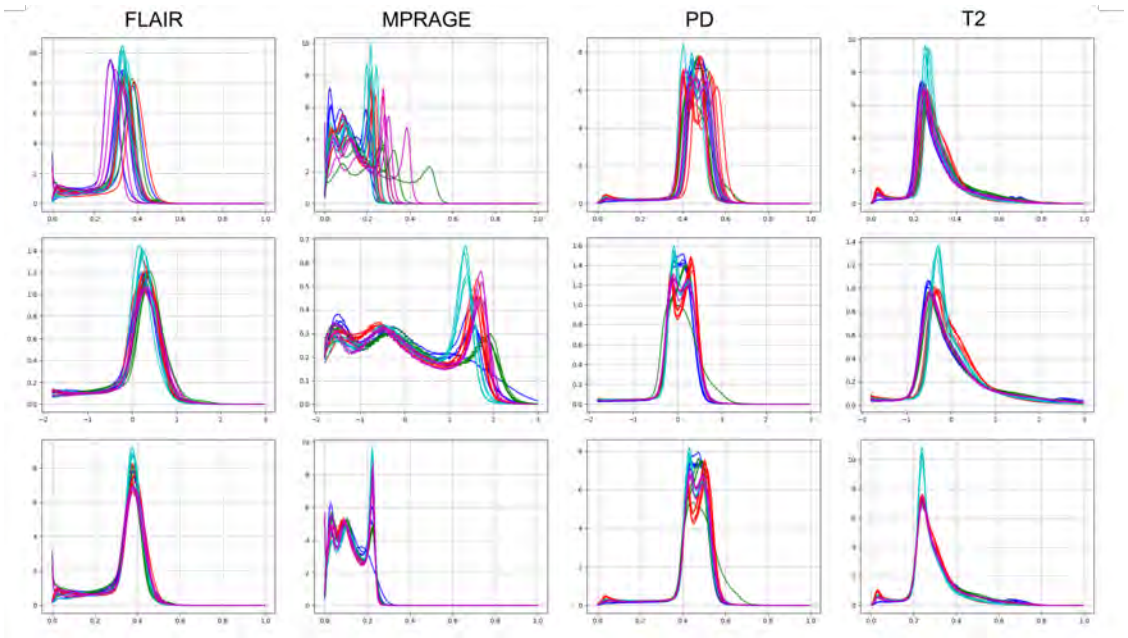


Figure 7: Histograms of the training set images for all four modalities after min-max normalization (top row), after standardization (middle row) and after the Chi-Square based normalization (bottom row).

Table 2: Segmentation results for different models and normalization methods. For the cross-validation mask1 was used for both training and evaluation. Metrics are computed as the averages for all time-points of all subjects.

Method	Normalization	Mean DSC	Mean LFPR	Mean LTPR
U-Net	min-max	0.636 (0.194)	0.396 (0.143)	0.616 (0.189)
	standardization	0.685 (0.159)	0.348 (0.163)	0.662 (0.199)
	proposed	0.651 (0.148)	0.453 (0.240)	0.664 (0.198)
Proposed	min-max	0.646 (0.179)	0.407 (0.170)	0.645 (0.160)
	standardization	0.684 (0.143)	0.371 (0.178)	0.656 (0.174)
	proposed	0.711 (0.085)	0.398 (0.134)	0.667 (0.171)

Table 3: Segmentation results for different models and normalization methods. For the cross-validation mask1 was used for training and mask2 for evaluation. Metrics are computed as the averages for all time-points of all subjects.

Architecture	Normalization	Mean DSC	Mean LFPR	Mean LTPR
U-Net	min-max	0.608 (0.185)	0.360 (0.165)	0.445 (0.182)
	standardization	0.635 (0.165)	0.338 (0.190)	0.455 (0.149)
	proposed	0.605 (0.148)	0.411 (0.265)	0.485 (0.158)
Proposed	min-max	0.605 (0.169)	0.392 (0.176)	0.458 (0.143)
	standardization	0.625 (0.144)	0.355 (0.193)	0.465 (0.144)
	proposed	0.658 (0.085)	0.377 (0.189)	0.479 (0.139)

proposed method in comparison to the classic standardization, in which the overall alignment is not always achieved.

In comparison to other normalization methods that require a reference such as histogram matching, the proposed method allows to preserve the basic shape of the histograms, which prevents from loosing key intensity information about the lesions. The optimization of the similarity metric reduces problems that peak/landmark based methods can exhibit when the histograms differ too much before normalization, especially in MPRAGE and PD images, where several peaks can be observed in the histograms. We chose an approach using a pre-segmented WM mask, assuming that normalizing the

Table 4: Segmentation results for different models and normalization methods. For the cross-validation mask2 was used for training and mask1 for evaluation. Metrics are computed as the averages for all time-points of all subjects.

Architecture	Normalization	Mean DSC	Mean LFPR	Mean LTPR
U-Net	min-max	0.670 (0.129)	0.420 (0.187)	0.678 (0.162)
	standardization	0.695 (0.167)	0.441 (0.161)	0.740 (0.135)
	proposed	0.659 (0.129)	0.544 (0.118)	0.750 (0.138)
Proposed	min-max	0.680 (0.124)	0.406 (0.179)	0.694 (0.130)
	standardization	0.712 (0.127)	0.446 (0.111)	0.750 (0.136)
	proposed	0.713 (0.080)	0.455 (0.134)	0.720 (0.118)

Table 5: Segmentation results for different models and normalization methods. For the cross-validation mask2 was used for both training and evaluation. Metrics are computed as the averages for all time-points of all subjects.

Architecture	Normalization	Mean DSC	Mean LFPR	Mean LTPR
U-Net	min-max	0.664 (0.142)	0.359 (0.180)	0.505 (0.145)
	standardization	0.663 (0171)	0.375 (0.139)	0.561 (0.089)
	proposed	0.638 (0.135)	0.481 (0.163)	0.580 (0.122)
Proposed	min-max	0.673 (0.137)	0.329 (0.156)	0.542 (0.135)
	standardization	0.685 (0.140)	0.385 (0.116)	0.550 (0.108)
	proposed	0.676 (0.094)	0.392 (0.191)	0.534 (0.099)

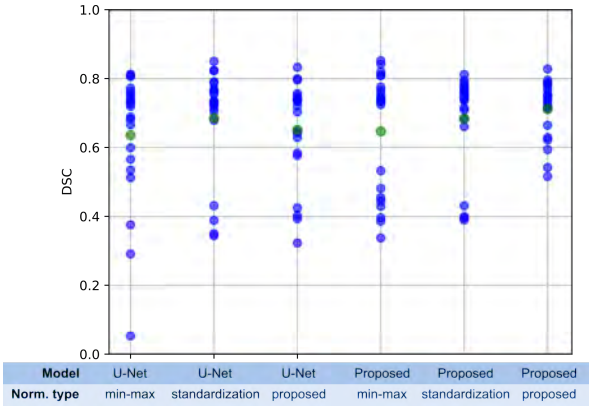


Figure 8: Scatter plot of DSC metric for models trained and evaluated with mask1 using cross-validation, for different normalization methods. Blue circles represent the value of the metric for all subjects and time-points, and green circles represent the average value.

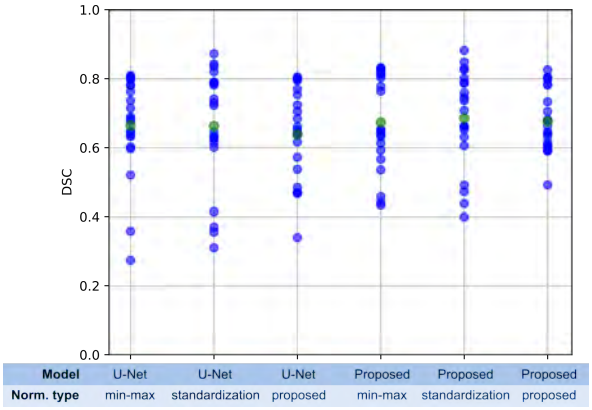


Figure 9: Scatter plot of DSC metric for models trained and evaluated with mask2 using cross-validation, for different normalization methods. Blue circles represent the value of the metric for all subjects and time-points, and green circles represent the average value.

surrounding tissue value of white matter lesions optimally supports the detection of the pathological lesions. This approach relies on a rough segmentation of the white matter before applying the CNN. The WM segmentation can be affected by the presence of lesions, but the influence in the final displacement of the histograms was found to be very small. However, when a WM mask is not available, the normalization can also be applied on the original histograms, at the cost of a higher influence of the lesion volume in the quality of the normalization, but still allowing the alignment of the histograms.

One disadvantage of the proposed normalization is the fact that it requires a reference time-point, which is used for all remaining images. Even though the effect of the selection of this reference or the generation of a synthetic reference was not studied in this work, it is expected to have an impact in the performance of the subsequent steps of the pipeline.

5.2. Lesion Segmentation

The second step in the pipeline is the lesion segmentation, for which an improvement in the DSC is observed when the mask1 is used for both training and evaluation, whereas for training and evaluation performed with mask2 the standardization produced, to-

gether with the proposed architecture, a better DSC as compared to the other cases. When training and evaluation is performed with different masks, the proposed pipeline produced the highest DSC. Furthermore, in all 4 possible combination of masks for training an evaluation the proposed architecture produced the best results in terms of DSC, either with standardization or with the proposed normalization procedure. This contributes to the validness of the initial hypothesis that considering time information can help produce better segmentation results, as suggested previously by Birenbaum and Greenspan (2016, 2017).

Table 6 shows a summary and comparison of the results, with respect to previous reported deep learning approaches in the literature for the same general cross-validation procedure followed in this work. The table also includes the metrics computed between both raters. Taking into account that no data augmentation took place for the proposed pipeline, the obtained results are close and even higher to some of the previously proposed methods in some of the metrics, particularly the DSC and the LFPR. However, the table does not reflect the inter- and intra time-point consistency of the results, which is also one important advantage of the proposed pipeline.

In order to characterize MS lesion activity and the temporal change of lesion size correctly, it is crucial that all time-points lead to consistent results compared to a human expert rater. To study this, tables 2 to 5 include the standard deviation of the obtained DSC for the respective method. Outliers lead to higher standard deviations, while results consistent with the human rater should yield low standard deviations of DSC. Thus, the standard deviation of DSC is an important quality metric in this work, to characterize the quality for a segmentation when longitudinal data is used.

Another important fact to be considered in the longitudinal setting is that, for a given subject, the images of subsequent time-points are not expected to have significant or aggressive changes. Instead, they are expected to be relatively similar considering also that the average interval between time-points for the used dataset is one year. This consistency in the volumes implies consistency in the segmentations of the lesions. While the proposed histogram normalization method is expected to reduce outliers induced by time-points with different image contrast, the proposed longitudinal architecture including C-LSTM is expected to improve the temporal consistency of the segmentations.

The cross-sectional approach with min-max normalization can produce segmentations that are highly different for a certain time-point of a subject, as shown in the first column of Fig. 10, where no lesion is detected in the second time-point for the shown slice. This is corrected by implementing a better normalization strategy. However, in order to increase the intra and inter time-point consistency, the use of a longitudinal model together with the proposed normalization led to the most compact ranges for the DSC metric in terms of low standard deviation, as shown in the scatter plots and standard deviations. This does not mean that the model can only detect lesions that appear in all time-points. Fig. 11 shows an example of a lesion that changes in time, and for which the proposed architecture, when combined with the proposed normalization or with standardization, is able to capture the change in the lesion.

Regarding the LFPR and LTPR, it does not seem to be an improvement nor deterioration of the obtained values, or at least a general trend. In some cases the longitudinal model led to higher values, whereas in some other cases the cross-sectional approach caused higher values for both LFPR and LTPR.

In terms of training and inference times, although this aspect was not analyzed thoroughly, the addition of the bidirectional C-LSTM blocks causes additional computation time which is highly dependent on the implementation of these blocks. The training time of the proposed architecture was found to be about 1.5x the time of a normal cross-sectional U-Net. This factor is of course dependent on the implementation of the C-LSTM blocks, which were not optimized for time efficiency in this work.

Diagnosis and treatment decision based on lesion inspection on MRI data is a central aspect in MS. The clinical workflow also contains the comparison to pre-examinations to assess inflammatory activity. This process is tedious when looking at up to above 100 slices in high resolution imaging, at least four modalities and several pre-examinations. Still, common solutions for automated lesion segmentation do not rely on neural networks and are not typically applied in the clinical setting. Thus, the work presented in this thesis is highly relevant as it investigates ways to improve the state-of-the-art regarding the important aspect of longitudinal analysis, in order to make longitudinal lesion segmentation applicable in clinical MS neuroimaging.

6. Conclusions and Future Work

In this study we have proposed a supervised longitudinal pipeline for MS lesion segmentation from multi-modal MR images. The approach combines a whole-volume longitudinal normalization scheme with a patch-based 3D CNN architecture that exploits time information. The method was evaluated on data from the ISBI 2015 challenge, obtaining result that are consistent in time as well as across subjects, allowing also improvements in the segmentation metrics, especially in the DSC. Lesion segmentation consistency in time for each subject should be an important goal of the segmentation algorithms, as it is a natural consequence of the non-sudden variations in the different scans that a subject can have in longitudinal studies.

Table 6: Comparison of different deep learning segmentation methods for leave-one-out cross-validation on the ISBI challenge dataset. The word proposed represents in this case the whole pipeline i.e. the proposed normalization followed by the proposed model.

Method	**mask1**			**mask2**		
	DSC	**LFPR**	**LTPR**	**DSC**	**LFPR**	**LTPR**
Rater 1	-	-	-	0.732	0.174	0.645
Rater 2	0.732	0.355	0.8260	-	-	-
Brosch et al., 2016 (mask1)	0.684	0.546	0.746	0.644	0.529	0.633
Brosch et al., 2016 (mask2)	0.683	0.646	0.783	0.659	0.620	0.693
Aslani et al., 2018 (mask1)	0.698	0.482	0.746	0.651	0.451	0.641
Aslani et al., 2018 (mask2)	0.694	0.497	0.784	0.664	0.442	0.695
Aslani et al., 2019 (mask1)	0.765	0.120	0.670	0.699	0.123	0.536
Aslani et al., 2019 (mask2)	0.765	0.202	0.700	0.713	0.190	0.572
Proposed (mask1)	0.711	0.398	0.667	0.658	0.377	0.479
Proposed (mask2)	0.713	0.455	0.720	0.676	0.392	0.534

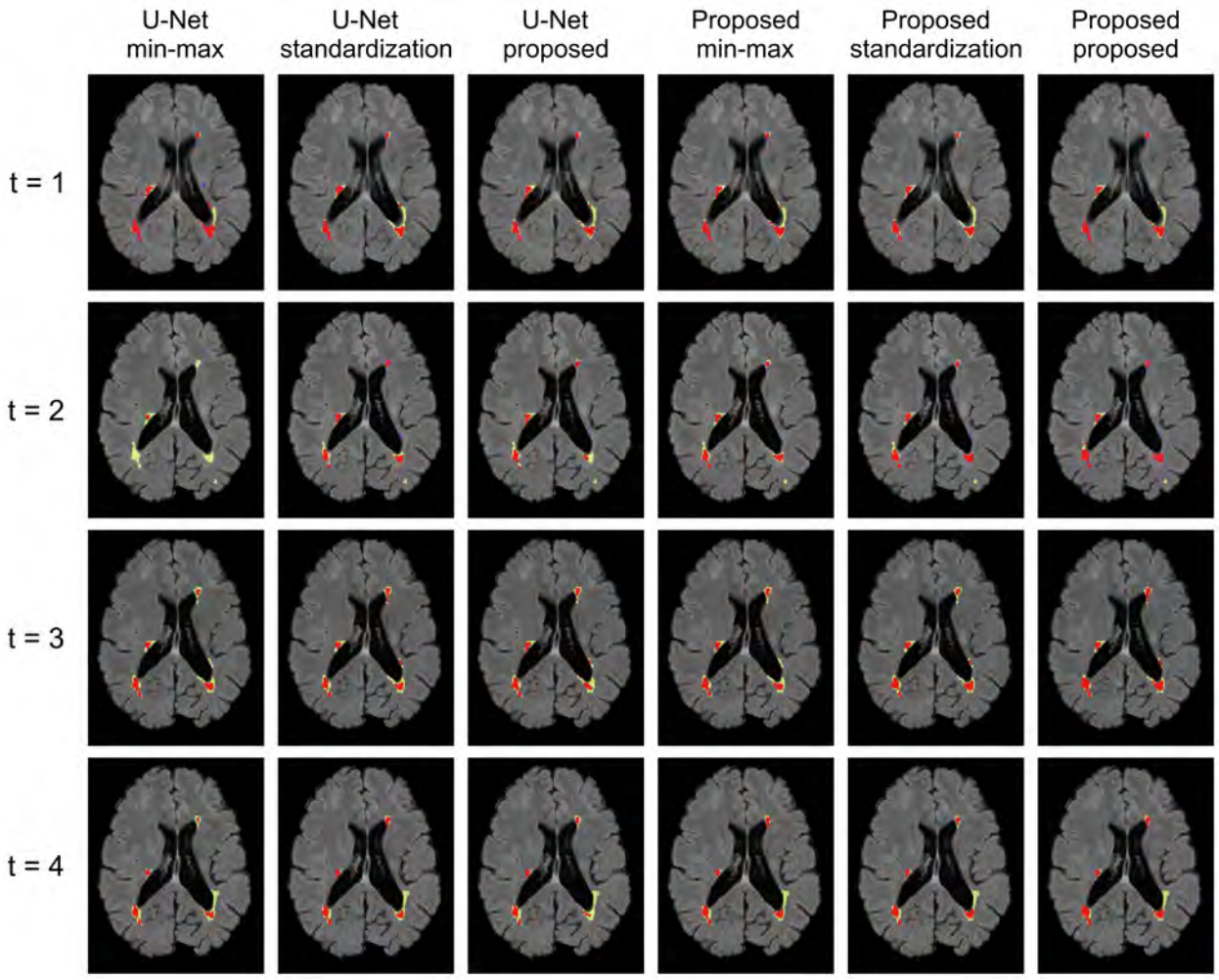


Figure 10: Example of resulting segmentation masks from cross-validation experiment for patient 01, slice 89 from the ISBI training dataset for cross-sectional and longitudinal models, and for three types of normalization. *t* denotes time-point index. Pixel colors correspond to true positives (red), false negatives (yellow) and false positives (blue), using *mask1* as reference. The proposed pipeline (rightmost column) produces the highest DSC score.

The longitudinal normalization pre-processing method increased the robustness of a trained network

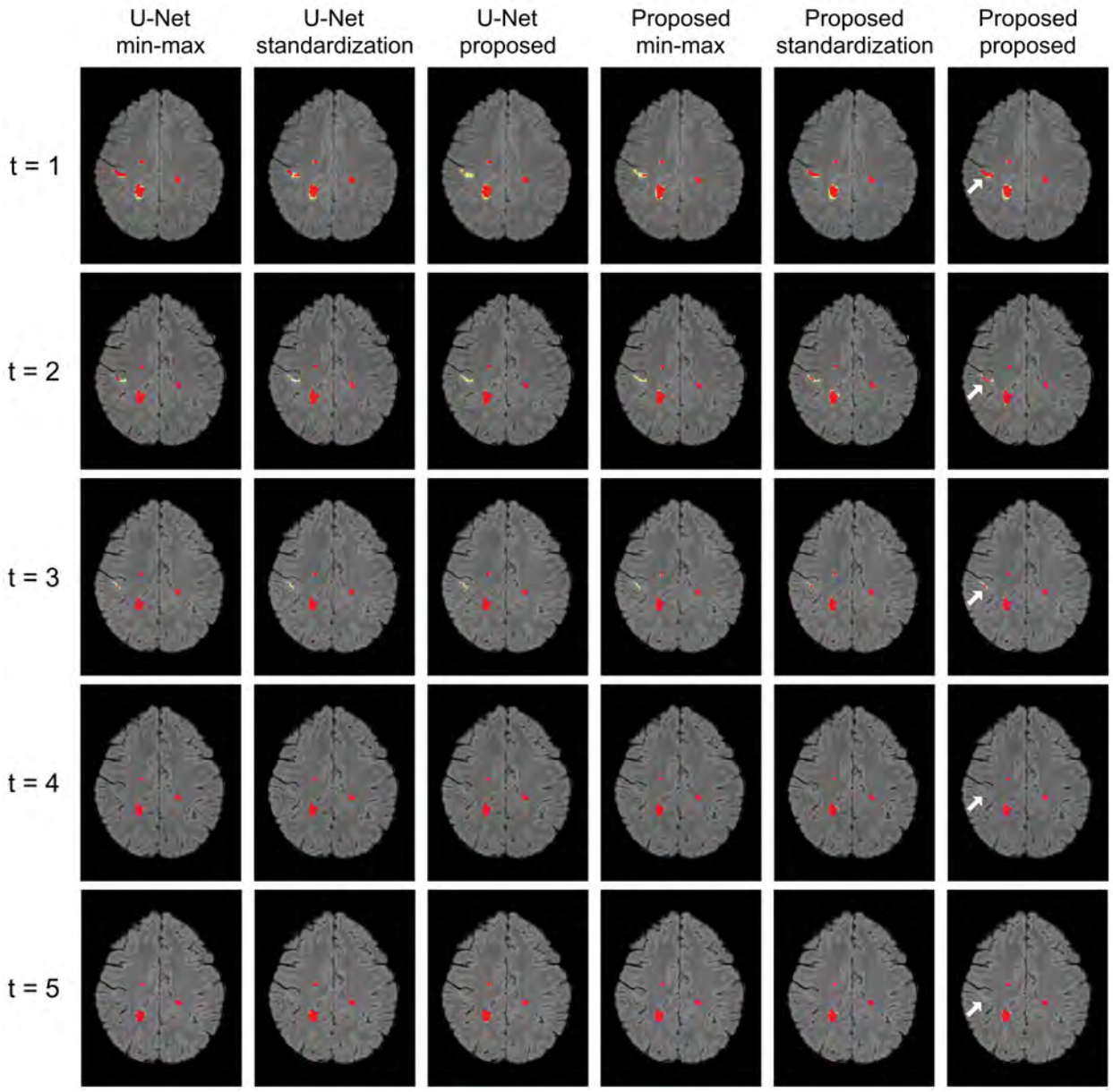


Figure 11: Example of resulting segmentation masks from cross-validation experiment for patient 03, slice 109 from the ISBI training dataset for cross-sectional and longitudinal models, and for three types of normalization. *t* denotes time-point index. Pixel colors correspond to true positives (red), false negatives (yellow) and false positives (blue), using *mask1* as reference. The white arrow points (for simplicity only in the last column) to a lesion that disappears in time, and whose change can be properly detected by the longitudinal pipeline.

in respect to the histogram variations of the input data, which were present in the ISBI 2015 training data. Thus, it is a promising technique to be applied also on MRI data from various sources, e.g. in the context of multi-center trials. Future work of our group will therefore include the validation of the algorithm on heterogeneous data from clinical studies and the evaluation of diagnostic relevance together with clinical partners. Future improvements also include the automatic selection of the reference images for the normalization process or eventually the generation of a synthetic template.

7. Acknowledgments

We would like to thank Prof. Dr. Matthias Günther and Annika Hänsch for their support whenever questions related with their fields of expertise arose. Additionally, special thanks to Daniel Mensing for his feedback to some parts of this document.

References

Andermatt, S., Pezold, S., Cattin, P.C., 2018. Automated segmentation of multiple sclerosis lesions using multi-dimensional gated recurrent units. BrainLes 2017 10670, 31–42. doi:10.1007/978-3-319-75238-9_3.

Arnon, R., Miller, A., 2016. Translational neuroImmunology in multiple sclerosis. 1st ed., Academic Press, Inc., London.

Aslani, S., Dayan, M., Murino, V., Sona, D., 2018. Deep 2D encoder-decoder convolutional neural network for multiple sclerosis lesion segmentation in brain MRI, in: International MICCAI Brainlesion Workshop, pp. 132–141. doi:10.1007/978-3-030-11723-8_13.

Aslani, S., Dayan, M., Storelli, L., Filippi, M., Murino, V., Rocca, M.A., Sona, D., 2019. Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. NeuroImage 196, 1–15. doi:10.1016/j.neuroimage.2019.03.068, arXiv:1811.02942.

Atlason, H.E., Love, A., Sigurdsson, S., Gudnason, V., 2019. Unsupervised brain lesion segmentation from MRI using a convolutional autoencoder, in: Medical Imaging 2019: Image Processing, International Society for Optics and Photonics. p. 109491H. arXiv:arXiv:1811.09655v1.

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2019. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. IEEE Transactions on Medical Imaging 38, 1788–1800. doi:10.1109/TMI.2019.2897538, arXiv:1809.05231.

Baur, C., De Benedikt Wiestler, C.B., Albarqouni, S., Navab, N., 2019a. Fusing Unsupervised and Supervised Deep Learning for White Matter Lesion Segmentation. Proceedings of Machine Learning Research 102, 63–72. URL: https://openreview.net/pdf?id=ryxNhZGlxV.

Baur, C., Wiestler, B., Albarqouni, S., Navab, N., 2019b. Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11383 LNCS, 161–169. doi:10.1007/978-3-030-11723-8_16, arXiv:1804.04488.

Billast, M., Meyer, M.I., Sima, D.M., Robben, D., 2020. Improved inter-scanner ms lesion segmentation by adversarial training on longitudinal data. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11992 LNCS, 98–107. doi:10.1007/978-3-030-46640-4_10, arXiv:2002.00952.

Birenbaum, A., Greenspan, H., 2016. Longitudinal multiple sclerosis lesion segmentation using multi-view convolutional neural networks 10008, 58–67. URL: http://link.springer.com/10.1007/978-3-319-46976-8, doi:10.1007/978-3-319-46976-8_7.

Birenbaum, A., Greenspan, H., 2017. Multi-view longitudinal CNN for multiple sclerosis lesion segmentation. Engineering Applications of Artificial Intelligence 65, 111–118. URL: http://dx.doi.org/10.1016/j.engappai.2017.06.006, doi:10.1016/j.engappai.2017.06.006.

Brosch, T., Tang, L.Y., Yoo, Y., Li, D.K., Traboulsee, A., Tam, R., 2016a. Deep 3D Convolutional Encoder Networks With Shortcuts for Multiscale Feature Integration Applied to Multiple Sclerosis Lesion Segmentation. IEEE Transactions on Medical Imaging 35, 1229–1239. doi:10.1109/TMI.2016.2528821.

Brosch, T., Yoo, Y., Tang, L., Tam, R., 2016b. Deep learning of brain images and its application to multiple sclerosis, in: Wu, G., Shen, D., Sabuncu, M.R. (Eds.), Machine learning and medical imaging. 1st ed.. Academic Press, Inc., London. chapter Deep learn, pp. 69–97.

Brosch, T., Yoo, Y., Tang, L.Y.W., Li, D.K.B., Traboulsee, A., Tam, R., 2015. Deep Convolutional Encoder Networks for Multiple Sclerosis Lesion Segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 3–11. URL: http://link.springer.com/10.1007/978-3-319-24574-4_1, doi:10.1007/978-3-319-24574-4_1.

Brugnara, G., Isensee, F., Neuberger, U., Bonekamp, D., Petersen, J., Diem, R., Wildemann, B., Heiland, S., Wick, W., Bendszus, M., Maier-Hein, K., Kickingereder, P., 2020. Automated volumetric assessment with artificial neural networks might enable a more accurate assessment of disease burden in patients with multiple sclerosis. European Radiology 30, 2356–2364. doi:10.1007/s00330-019-06593-y.

Carass, A., Roy, S., Jog, A., Cuzzocreo, J.L., Magrath, E., Gherman, A., Button, J., Nguyen, J., Bazin, P.L., Calabresi, P.A., Crainiceanu, C.M., Ellingsen, L.M., Reich, D.S., Prince, J.L., Pham, D.L., 2017. Longitudinal multiple sclerosis lesion segmentation data resource. NeuroImage 12, 346–350. URL: http://dx.doi.org/10.1016/j.dib.2017.04.004, doi:10.1016/j.dib.2017.04.004.

Cohen, J.A., Rae-Grant, A., 2012. Handbook of multiple sclerosis. 1st ed., Springer Healthcare, London. doi:10.1007/978-1-907673-50-4.

Compston, A., Confavreux, C., Lassmann, H., McDonald, I., Miller, D., Noseworthy, J., Smith, K., Wekerle, H., 2005. McAlpine's multiple sclerosis. 4th ed., Churchill Livingstone, Elsevier, Inc.

Dennis Jr, J.E., Woods, D.J., 1985. Optimization on microcomputers. The nelder-mead simplex algorithm.

Duong, M.T., Rudie, J.D., Wang, J., Xie, L., Mohan, S., Gee, J.C., Rauschecker, A.M., 2019. Convolutional neural network for automated flair lesion segmentation on clinical brain MR imaging. American Journal of Neuroradiology 40, 1282–1290. doi:10.3174/ajnr.A6138.

Fartaria, M.J., Todea, A., Kober, T., O'brien, K., Krueger, G., Meuli, R., Granziera, C., Roche, A., Bach Cuadra, M., 2018. Partial volume-aware assessment of multiple sclerosis lesions. NeuroImage: Clinical 18, 245–253. URL: https://doi.org/10.1016/j.nicl.2018.01.011, doi:10.1016/j.nicl.2018.01.011.

Fenneteau, A., Bourdon, P., Helbert, D., Habas, C., Guillevin, R., Fenneteau, A., Bourdon, P., Helbert, D., Fernandez-maloigne, C., 2020. Learning a CNN on multiple sclerosis lesion segmentation with self-supervision, in: 3D Measurement and Data Processing, IST Electronic Imaging 2020 Symposium, San Francisco.

Gabr, R.E., Coronado, I., Robinson, M., Sujit, S.J., Datta, S., Sun, X., Allen, W.J., Lublin, F.D., Wolinsky, J.S., Narayana, P.A., 2019. Brain and lesion segmentation in multiple sclerosis using fully convolutional neural networks: A large-scale study. Multiple Sclerosis Journal , 1–10doi:10.1177/1352458519856843.

Gaser, C., Dahnke, R., 2016. CAT-A Computational Anatomy Toolbox for the Analysis of Structural MRI Data. Human Brain Mapping 32, 336–348.

Ghafoorian, M., Bram, P., 2015. Convolutional Neural Networks for MS Lesion Segmentation, Method Description of DIAG team, in: Proceedings of the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge, pp. 1–2.

Hashemi, S.R., Salehi, S.S.M., Erdogmus, D., Prabhu, S.P., Warfield, S.K., Gholipour, A., 2019. Asymmetric Loss Functions and Deep Densely-Connected Networks for Highly-Imbalanced Medical Image Segmentation: Application to Multiple Sclerosis Lesion Detection. IEEE Access 7, 1721–1735. doi:10.1109/ACCESS.2018.2886371, arXiv:1803.11078.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-Janua, 2261–2269. doi:10.1109/CVPR.2017.243, arXiv:1608.06993.

IACL, 2018. The 2015 Longitudinal MS Lesion Segmentation Challenge. URL: http://iacl.ece.jhu.edu/MSChallenge.

Kingma, D.P., Ba, J.L., 2015. Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings , 1–15arXiv:1412.6980.

Lladó, X., Oliver, A., Cabezas, M., Freixenet, J., Vilanova, J.C., Quiles, A., Valls, L., Ramió-Torrent, L., Rovira, L., 2012. Segmentation of multiple sclerosis lesions in brain MRI: A review of automated approaches. Information Sciences 186, 164–185. doi:10.1016/j.ins.2011.10.011.

Lucchinetti, C.F., Parisi, J.E., 2006. Pathology: What may it tell us?, in: Cook, S.D. (Ed.), Handbook of multiple sclerosis. 4th ed.. Taylor & Francis Group, New York. chapter Pathology:, pp. 114–115.

McKinley, R., Gundersen, T., Wagner, F., Chan, A., Wiest, R., Reyes, M., 2016. Nabla-net: a deep dag-like convolutional architecture for biomedical image segmentation: application to white-matter lesion segmentation in multiple sclerosis, in: MSSEG Challenge Proceedings: Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure, pp. 37–43. URL: http://www.hal.inserm.fr/inserm-01397806.

McKinley, R., Wepfer, R., Aschwanden, F., Grunder, L., Muri, R., Rummel, C., Verma, R., Weisstanner, C., Reyes, M., Salmen, A., Chan, A., Wagner, F., Wiest, R., 2019. Simultaneous lesion and neuroanatomy segmentation in multiple sclerosis using deep neural networks. arXiv preprint arXiv:arXiv:1901.07419.

McKinley, R., Wepfer, R., Grunder, L., Aschwanden, F., Fischer, T., Friedli, C., Muri, R., Rummel, C., Verma, R., Weisstanner, C., Wiestler, B., Berger, C., Eichinger, P., Muhlau, M., Reyes, M., Salmen, A., Chan, A., Wiest, R., Wagner, F., 2020. Automatic detection of lesion load change in Multiple Sclerosis using convolutional neural networks with segmentation confidence. NeuroImage: Clinical 25, 102104. URL: https://doi.org/10.1016/j.nicl.2019.102104, doi:10.1016/j.nicl.2019.102104, arXiv:1904.03041.

Miller, A.E., 2006. Clinical features, in: Cook, S.D. (Ed.), Handbook of Multiple Sclerosis. 4th ed.. Taylor & Francis Group, New York. chapter 6, pp. 153–178.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016 , 565–571doi:10.1109/3DV.2016.79, arXiv:1606.04797.

Nair, T., Precup, D., Arnold, D.L., Arbel, T., 2020. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Medical image analysis 59. URL: http://dx.doi.org/10.1007/978-3-030-00928-1_30, doi:10.1007/978-3-030-00928-1, arXiv:1805.10884.

Narayana, P.A., Coronado, I., Sujit, S.J., Sun, X., Wolinsky, J.S., Gabr, R.E., 2020. Are multi-contrast magnetic resonance images necessary for segmenting multiple sclerosis brains? A large cohort study based on deep learning. Magnetic Resonance Imaging 65, 8–14. URL: https://doi.org/10.1016/j.mri.2019.10.003, doi:10.1016/j.mri.2019.10.003.

Novikov, A.A., Major, D., Wimmer, M., Lenis, D., Buhler, K., 2019. Deep sequential segmentation of organs in volumetric medical scans. IEEE Transactions on Medical Imaging 38, 1207–1215. doi:10.1109/TMI.2018.2881678, arXiv:1807.02437.

Nyúl, L.G., Udupa, J.K., Zhang, X., 2000. New variants of a method of MRI scale standardization. IEEE Transactions on Medical Imaging 19, 143–150. doi:10.1109/42.836373.

Penny, W.D., Friston, K.J., Ashburner, J.T., Kiebel, S.J., Nichols, T.E., 2011. Statistical parametric mapping: the analysis of functional brain images. Elsevier.

Phi, M., 2018. Illustrated Guide to LSTM's and GRU's: A step by step explanation. URL: https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21.

Pryse-Phillips, W., Sloka, S., 2006. Etiopathogenesis and epidemiology: clues to etiology, in: Cook, S.D. (Ed.), Handbook of multiple sclerosis. 4th ed.. Taylor & Francis Group, New York. chapter 1, pp. 1–39.

Rinker II, J.R., Naismith, R.T., Cross, A.H., 2006. Multiple sclerosis: an autoimmune disease of the central nervous system?, in: Cook, S.D. (Ed.), Handbook of multiple sclerosis. 4th ed.. Taylor & Francis Group, New York. chapter 4, pp. 95–112.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, pp. 234–241. doi:10.1007/978-3-319-24574-4_28, arXiv:1505.04597.

Roy, S., Butman, J.A., Reich, D.S., Calabresi, P.A., Pham, D.L., 2018. Multiple Sclerosis Lesion Segmentation from Brain MRI via Fully Convolutional Neural Networks URL: http://arxiv.org/abs/1803.09172, arXiv:1803.09172.

Roy, S., Carass, A., Shiee, N., Pham, D.L., Calabresi, P., Reich, D., Prince, J.L., 2013. Longitudinal intensity normalization in the presence of multiple sclerosis lesions. 2013 IEEE 10th International Symposium on Biomedical Imaging , 1384–1387doi:10.1109/ISBI.2013.6556791.

Salehi, S.S.M., Erdogmus, D., Gholipour, A., 2017. Tversky loss function for image segmentation using 3D fully convolutional deep networks, in: International Workshop on Machine Learning in Medical Imaging, pp. 379–387. doi:10.1007/978-3-319-67389-9_44, arXiv:1706.05721.

Salem, M., Valverde, S., Cabezas, M., Pareto, D., Oliver, A., Salvi, J., Rovira, A., Llado, X., 2019. Multiple Sclerosis Lesion Synthesis in MRI Using an Encoder-Decoder U-NET. IEEE Access 7, 25171–25184. doi:10.1109/ACCESS.2019.2900198, arXiv:1901.05733.

Salem, M., Valverde, S., Cabezas, M., Pareto, D., Oliver, A., Salvi, J., Rovira, À., Lladó, X., 2020. A fully convolutional neural network for new T2-w lesion detection in multiple sclerosis. NeuroImage: Clinical 25, 102149. URL: https://doi.org/10.1016/j.nicl.2019.102149, doi:10.1016/j.nicl.2019.102149.

Shinohara, R.T., Sweeney, E.M., Goldsmith, J., Shiee, N., Mateen, F.J., Calabresi, P.A., Jarso, S., Pham, D.L., Reich, D.S., Crainiceanu, C.M., 2014. Statistical normalization techniques for magnetic resonance imaging. NeuroImage: Clinical 6, 9–19. URL: http://dx.doi.org/10.1016/j.nicl.2014.08.008, doi:10.1016/j.nicl.2014.08.008.

Sweeney, E.M., Shinohara, R.T., Reich, D.S., Crainiceanu, C.M., Mri, M.L., 2013. Automatic Lesion Incidence Estimation and Detection in Multiple Sclerosis Using Multisequence Longitudinal MRI. American Journal of Neuroradiology 34, 68–73.

Vaidya, S., Chunduru, A., Muthuganapathy, R., Krishnamurthi, G., 2015. Longitudinal Multiple Sclerosis Lesion Segmentation using 3D Convolutional Neural Networks, in: Proceedings of the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge, pp. 1–2.

Valverde, S., Cabezas, M., Roura, E., González-Villà, S., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Rovira, À., Oliver, A., Lladó, X., 2017. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. NeuroImage 155, 159–168. doi:10.1016/j.neuroimage.2017.04.034, arXiv:1702.04869.

Valverde, S., Salem, M., Cabezas, M., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Rovira, À., Salvi, J., Oliver, A., Lladó, X., 2019. One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. NeuroImage: Clinical 21, 101638. URL: https://doi.org/10.1016/j.nicl.2018.101638, doi:10.1016/j.nicl.2018.101638, arXiv:1805.12415.

Wallin, M.T., Culpepper, W.J., Others, 2019. Global burden of diseases, injuries, and risk factors study (GBD). The Lancet Neurology 18, 269–285. doi:https://doi.org/10.1016/S1474-4422(18)30443-5.

Wang, J., Liu, M., Zhang, C., Xu, H., Zhang, L., Zhao, Y., 2020. An adaptive sparse Bayesian model combined with probabilistic label fusion for multiple sclerosis lesion segmentation in brain MRI. Future Generation Computer Systems 105, 695–704. URL: https://doi.org/10.1016/j.future.2019.12.035, doi:10.1016/j.future.2019.12.035.

Weaver, K.F., Morales, V., Dunn, S.L., Godde, K., Weaver, P.F., 2017. An Introduction to Statistical Analysis in Research. John Wiley Sons, Inc., Hoboken, NJ, USA. URL: http://doi.wiley.com/10.1002/9781119454205, doi:10.1002/9781119454205.

Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c., 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting, in: Advances in neural information processing systems, pp. 802–810.

Yoo, Y., Brosch, T., Traboulsee, A., Li, D.K., Tam, R., 2014. Deep learning of image features from unlabeled data for multiple sclero-

sis lesion segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8679, 117–124. doi:10.1007/978-3-319-10581-9_15.