Lonneke Bos 6979416
Master Student Medical Imaging
Minor Research Project
Supervisor: Georgios Lappas
Examiners: Prof. dr. Frederik Barkhof, dr. Hugo Kuijf

Enhancing lesion segmentation on contrast MR images in Multiple Sclerosis using deep learning

Multiple Sclerosis (MS), a neuroinflammatory disease of the central nervous system, is characterized by accumulation of lesions in the brain. During active inflammation phase, Gadolinium-based contract agents (GBCA's) are used to express active lesions in T1w MR images as hyperintensities. Artificial intelligence (AI) methods have been utilized among others, in MS white matter (enhancing) lesion segmentations; however, the existing methods have limitations making them difficult to be applicable across different institutions with varying protocols. Likewise, the goal of this research is to segment Gd+ lesions on contrast MR images in MS using deep learning. Similarly, a V-net, a 3D CNN architecture, and 2D acquired T1w data on a cross-sectional and longitudinal basis are used. Two annotators manually delineated enhancing lesions in T1w data and fed to the model. According to the inter-observer variability analysis, consensus increases quality of delineation and consequently model performance. Additionally, to compare the optimal duration for training of the model on the given dataset, two models have been evaluated e.g., 50 epochs (shallow training) and using 150 epochs (extensive training). The accuracy of the lesion prediction is increased when the model is shallow trained, compared to extensive training. Model evaluation on external dataset showed that the model is not robust enough to predict lesions deviating from specific protocols and thus, non-generalizable at this stage. Furthermore, based on the findings of this work, there are preliminary findings that the vendor variability can affect the model performance while the pipeline can be improved by using a larger dataset and hyperparameter optimization. Summarizing, this deep learning model stands a proof of concept for enhancing lesion segmentation using limited single institute data.
Keywords: multiple sclerosis, Gadolinium-enhancing lesion, lesion segmentation, deep learning, V-net, MRI, T1w data

1. INTRODUCTION

Multiple sclerosis (MS) is an inflammatory disease of the central nervous system and is characterized by the accumulation of lesions in the brain (Tullman, 2013). Magnetic Resonance Imaging (MRI) plays a crucial role in the detection of the disease activity of MS in the decision to initiate or escalate disease modifying therapy. The inflammatory disease activity is detected on the longer term by demonstrating new lesions on subsequent MRI scans (Confavreux and Vukusic, 2006). Specifically, the enhancing lesions depict inflammation after intravenous administration of a Gadolinium-based contrast agent (GBCA) (Kappos et al., (1999).

The BBB prevents blood-derived products, pathogens, and cells from entering the brain (Zlokovic, 2011). During active inflammation, as during an MS relapse, the blood-brain barrier (BBB) is disrupted, allowing gadolinium to pass though (Wattjes et al., 2015). In T1-weighted and fluid-attenuated inversion recovery (FLAIR) MRI, this is expressed as hyper-intense lesions on T1w images (as represented in Figure 1), also known as enhancing lesions or active lesions. The enhancing lesion count and load provide a measure of the focal inflammatory activity. This is routinely used to monitor breakthrough disease and to evaluate the efficacy and effectiveness of anti-inflammatory agents in MS clinical trials and practices (Simon et al., 2014, Barkhof et al., 2005, Cotton et al., 2006).

Recently, artificial intelligence (AI) methods have been used in medical image segmentations (Inglese et al., 2005, Norman et al., 2018) and MS white matter (enhancing) lesion segmentations (Brosch et al., 2016, Valverde et al., 2017, Valverde et al. 2018). Convolutional neural networks (CNNs), a subclass of networks that can derive a set of image-based features extracting information that are specifically optimized for the task (Salem et al., 2020). Deep learning algorithms are particularly

suited for image segmentation tasks and have been tailored for the segmentation of enhanced MS white matter lesions (Carass et al., 2017).

Up to date, there are available three DL models for Gadolinium-enhancing (Gd+) lesion segmentation presenting limitations in model agility and required sequences to achieve a good performance. The most recent work exploits a 3D CNN model for segmentation of Gd+ lesions, using multispectral MR data trained extensively over a large dataset of 1006 patients. However, the limited capabilities to identify and segment lesions larger than 70 mm^3 and the pre-request of having available five different MR sequences complicates it to apply this on a larger scale (Coronado et al., 2021).

Moreover, Brugnara et al. (2020) using the nnUnet, a promising CNN-based model with automatic hyper-parameter optimization, investigated the capabilities of such models to segment Gd+ lesions. Brugnara et al. trained the model utilizing a single-institutional dataset of 334 MS patients and tested with a longitudinal dataset with 82 patients at three timepoints (266 MRI exams). As a downside, the well-controlled parameters of the scanned protocols used in combination with limited magnetic field variability and the lower threshold of segmenting Gd+ lesions larger than 7 mm^3 restricted its applicability on real-world datasets.

Additionally, Valverde et al. (2018) used an 11-layer fully convolutional neural network (FCN) patch-based adaptive model called NicMSLesion. A patch-based adaptive model uses known label maps. For each patch in the testing image, similar patches are retrieved from the dataset. To produce an initial segmentation map, the corresponding labels for these patches are combined (Mechrez et al., 2016). The model was trained on a single-center clinical data (n=30) and tested on the public ISBI2015 challenge database consisting of 21 patients' longitudinal data. The small amount of data, the poorer performance architecture, and the lack of hyperparameter optimization of this model led to an ambiguous adoption of such model from the clinical and research community for Gd+ lesion segmentation.

Similarly, the goal of this research is to segment Gd+ lesions on contrast MR images in MS using deep learning mitigating the abovementioned drawbacks. Therefore, the authors used a V-net, 3D CNN architecture, and 2D acquired T1w data on a cross-sectional and longitudinal basis over a period of 10 years.

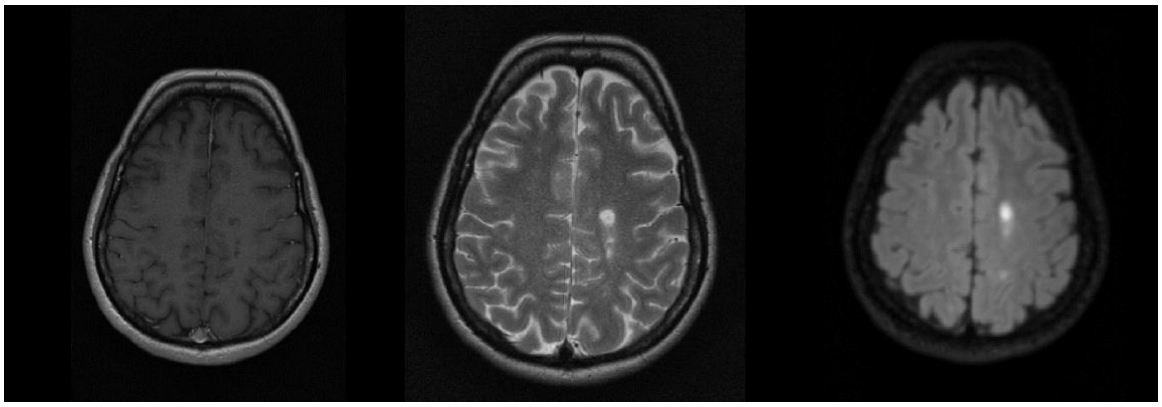


Figure 1. Example of enhancing lesions in respectively T1w, T2w and FLAIR images from same patient.

2. METHODS AND MATERIALS

2.1 Participants characteristics

All the data used for this research were anonymized and were collected after a data request approved by the VUmc review committee to perform analysis on clinical data. The dataset used in this work is part of a retrospective analysis of MRI data acquired in patients diagnosed with relapse-remitting MS (RRMS) in the active disease phase ensuring that all images include at least one Gd+ lesion under a single institution. The patients used in this research were recruited between 2011 and 2021 following an age group of 38±11.7 years with a male/female ratio of 40/60 and an Expanded Disability Status Scale (EDSS) of 4.35±2.42. The number of lesions varied per case and were categorized in three groups: patients with one lesion, patients with 2-5 lesions and patients with more than 5 lesions.

2.2 Scanner details

This work consists of MRI scans presenting variability in clinical protocol, field strength and scanner variability stimulating the challenging clinical heterogeneity and not only a research well-controlled setup. A summary can be found in Table 1. The advantage of incorporating such heterogeneous parameters enables us to study the scanners' impact on the model performance by presenting examples which capture inter- and intra-scanner variabilities (Takao et al., 2011).

Vendor	*#cases*	*TE (msec)*	*TR (msec)*	*B0 (T)*
SIEMENS Avanto	4	7.8 – 9.1	450 – 480	1.5
SIEMENS Sonata	3	14	580 - 630	1.5
Philips Ingenuity	2	12	600	3.0
GE Healthcare Signa HDxt	12	9 – 20	440 - 640	1.5
GE Healthcare Discovery MR750	7	9	440	3.0
Toshiba Titan3T	5	8	480	3.0

Table 1. Summary of the various vendors and MR scanners used in this research, with the corresponding number of cases, protocol (TE and TR) and magnetic field strength B0.

2.3 Manual delineation instructions

The MRI protocols included among other acquisition of 2D acquired T1-weighted images after intravenous administration of a GBCA which used for this work. To create the masks two annotators followed instructions and used information of neuroradiologists to identify and manually segment lesions in all T1w data. The advantage of multiple annotators is to validate the quality of the labels. All delineations were performed using FSL/FSLeyes, using the edit mode. The instruction for delineation is as follows:

1. If the data is presented as a DICOM file, the files should be converted to a nifti file.
2. Open the FSL/FSLeyes tool, in this study version 5.0.6.1 is used
3. Open the T1w scan (as a .nii.gz file) (File > Add from file > Select file from disk > Open). When using multiple sequences to delineate the lesions, make sure the scans are co-registered and all in the axial view with identical slice thickness and field of view values before beginning the process.
4. Adjust the brightness of the T1w scan to personal preference using the slide bars at the top of FSLeyes.
5. Select the Edit Mode in FSLeyes (Tools > Edit mode).
6. Create a mask (Edit (Ortho View 1) > Create mask).
7. Divide into four parts and go through all the image slices, zooming into a region.
8. To fill the mask, make sure the mask is selected in the overlay list. The toolbar of the edit mode provides various options, such as a pencil, an eraser, and a bucket to fill. The selection size can be altered and should be set to 1. The fill value should also be 1.
9. Try to be most precise with the selection of the voxels that are classified as a lesion. The more precise the delineation is performed, the better the model performs.
10. Inclusion features of lesions in the mask:
    a. Hyperintensity,
    b. Void shaped,
    c. Sharp borders,
    d. Isointense ring, moon-shaped,
    e. "Finger"-like shaped, especially in periventricular regions,
    f. Mostly focal located, affects entire brain and spinal cord, in white and gray matter.
11. Save the delineations and the mask regularly, make sure you make a hard copy of the masks.

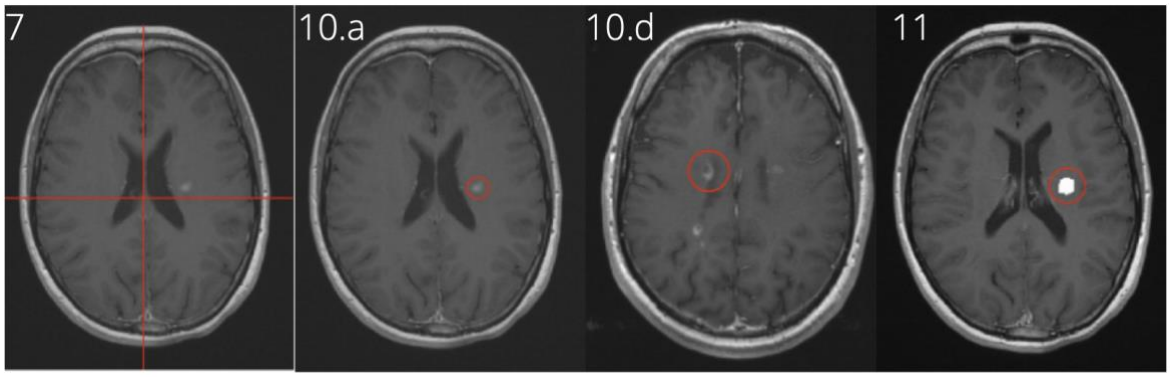


Figure 2. Visual representation of the manual delineation instructions followed in this work.

Subsequently, the intersection of the delineation of the two annotators was taken, causing the model to implicitly trains itself to average the segmentations. After delineation of all T1w-images, two experienced neuroradiologists provided feedback on the quality of the intersection of the manual segmentations of the two annotators. (M.M.S.J., with 10 years of experience in MRI of MS and other neurologic disorders, and B.M., with 10 years of experience in MRI of MS) Comments to improve the quality of the delineations were provided and used to update the delineation. To reduce bias in the delineations, an inter-observer variability study was performed on 10 cases, this IOV is defined as the difference in the delineations between annotators.

2.4 Image Standardization and processing

All images were standardized and preprocessed with an in-house developed automatic pipeline utilizing previously developed software aiming at minimizing the bias on the input-level. To increase the accuracy of downstream analyses, the following pre-processing steps were followed.

First, the skull in all images was stripped using the HD-Brain Extraction Tool (BET) (Isensee et al., 2019). The HD-BET algorithm relies on artificial neural networks. The HD-BET algorithm shows robust performance in the presence of pathology or treatment-induced tissue alterations, is applicable to a broad range of MRI sequence types and is not influenced by variations in MRI hardware and acquisition parameters encountered. Since the data used in this research is also obtained using multiple vendors and in the presence of pathology or treatment-induced alterations, the HD-BET algorithm is a suitable choice for skull-stripping.

Furthermore, as a step to standardize the data, all images were registered to MNI brain space with 1 mm^3 resolution using FSL's/FMRIB's Linear Registration Tool (FLIRT). Since all data is acquired with different vendors, at different timepoints, the dimensions vary for each image (Carass et al., 2017). Lesion segmentation in MR images suffers from the class imbalance problem. To address this problem a loss weighting strategy is used. This strategy uses the binary cross-entropy and Dice loss function to take care of the imbalanced classes.

Additionally, it is important to correct the image for spatial intensity non-uniformity. This bias field is a slow variant multiplicative noise, that causes inhomogeneity in MR images and results in a decrease of performance of the deep learning model. MONAI is used for intensity normalization and specifically the module including the N4 Advanced Normalization Tools (ANTs) (Tutison et al. 2010). Intensity normalization is the process of mapping intensities of all images into a standard or reference scale (Akkus et al., 2017). Proper intensity normalization is important since improper normalization could lead to false classification of enhancements (Datta et al., 2007). The preprocessed T1w images and corresponding labels were padded to 256 cubic mm^3 volumes as the model used for this work requires a fixed input size and to standardize the data. Besides the data standardization steps applied, further pre-processing steps have been exploited to ensure generalizability that minimizes the likelihood of overfitting to the center data. To achieve that, augmentation on-the-fly was used compromised of affine transformations. All T1w images and corresponding labels were augmented, to improve the performance of the V-net. To prevent a model from memorizing the training data, flipping and rotation of 90 degrees variation was used. Isotropic scaling was applied to have an influence on feature

extraction, because the details of objects appear different when viewed at different scales. An overview of the image pre-processing steps is found in Figure 3.

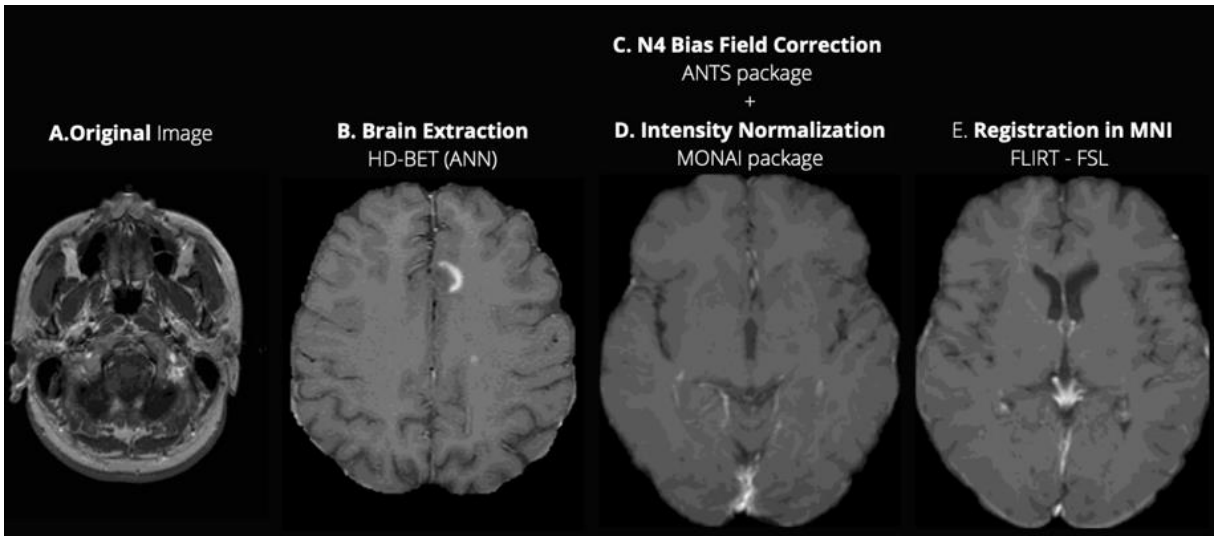


Figure 3. Visual representation of the followed image pre-processing pipeline.

2.5 Network Description

A commonly used deep learning model for segmentation of lesions in MS is the 2D-UNet, a state-of-the-art segmentation model in biomedical image segmentation (Ronneberger et al., 2015). This model contains encoder and decoder networks with skip connections to improve the information flow and preserve low-level spatial features. However, this model is suitable for 2D images. The deep learning model used in this is a 3D-V-net, which is suitable for 3D data and is mainly used for volumetric image segmentations. The network and training strategy relies on the strong use of data augmentation to use the available annotated samples more efficiently. In addition, the network is fast.

The network architecture (as illustrated in Figure 4, a higher resolution image is found in Appendix E) consists of a contracting path and an expansive path. The contracting path follows the same architecture of a convolutional network. There is a repeated application of two 3x3x3 convolutions, each followed by batch normalization, a rectified linear unit (ReLu) and a 2x2x2 max pooling operation with stride 2 for downsampling. At each downsampling step the number of feature channels is doubled.

Every step in the expansive path consists of an upsampling of the feature map followed by a 2x2x2 convolution that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and the two 3x3x3 convolutions, each followed by a ReLu. The cropping is necessary due to the loss of border voxels every convolution. At the final layer a 1x1x1 convolution is used to map each 64-component feature vector to two classes. In total the network has 4 layers consists of 20 convolutional layers. The scripts and other information about this model can be found at *https://gitlab.com/sbig/activemsnet/-/tree/Developing.*

2.6 Hyperparameters

The V-net used comes with default hyperparameters which could be altered for optimal training of the network. The hyperparameters of the model were first acquired using a preliminary dataset consisting of 21 sets of T1w, T2w, FLAIR and PD images provided by the MSLesionChallenge Dataset (Carass et al. 2017). The network was trained for all individual sequences. Using visual interpretation of the outcome the optimal hyperparameters of the model were determined. These are shown in Table 2.

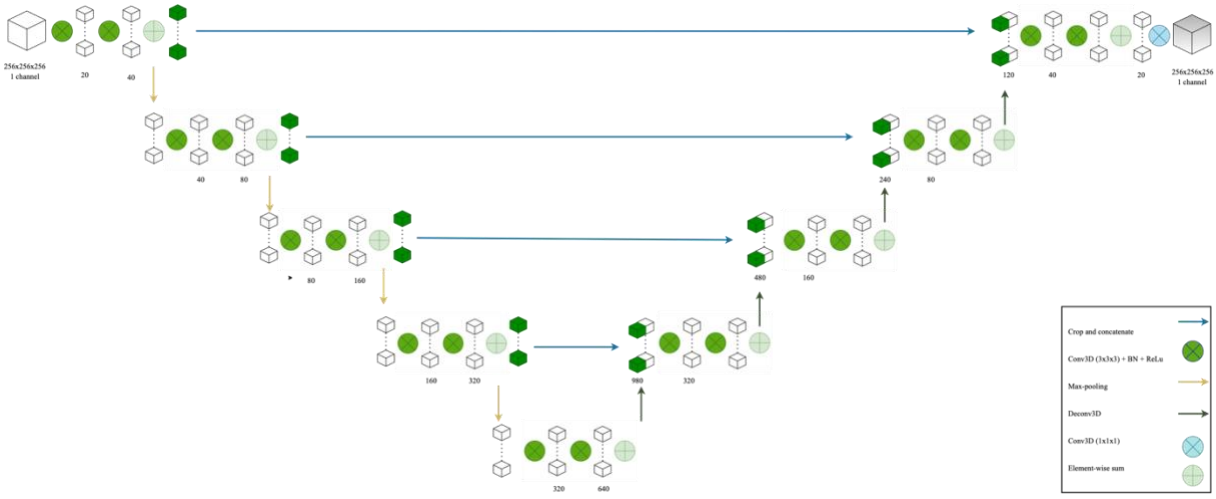


Figure 4. Illustration shows the architecture of V-net used for the prediction of enhancing lesions.

Hyperparameter	
Input shape	(256, 256, 256, 1)
Number of classes	1
Activation function	ReLu
Batch normalization	True
Upsampling mode	Deconvolution
Dropout	0.2
Number of layers	4
Number of filters	20
Output activation function	Sigmoid (threshold = 0.5)

Table 2. The hyperparameters configuration selected to be used during training.

2.7 Quantitative metrics

The model performance was quantitatively assessed using various (overlap- and distance-based) metrics. To assess the quality of the segmentation the DICE coefficient (DSC), the Mean Surface Distance (MSD) and 95th percentile of Hausdorff distance (HD95) were used. The DSC measures volumetric overlap between segmentation results and annotations and is a widely used metric in volumetric evaluation. The MSD measures a good way of evaluating the accuracy of an image-segmentation when the ground truth is known, and it estimates the error between the outer *S* and *S'* of the segmentations *X* and X'. With the MSD the mean of the vector is taken. When the maximum of the vector is taken, this is the Hausdorff distance. This distance is the maximum distance of a set to the nearest point in the other set. The purpose of the 95 percentile is to eliminate the impact of a very small subset of the outliers. The sensitivity (true positive rate) was used to quantitively examine the quality of segmentation of lesion level.

2.8 Analysis

To find the optimal trained model, the model was trained with varying epochs. An epoch is defined as one complete pass of the training dataset through the algorithm. The model was trained shallow (with 50 epochs) and more extensively (with 150 epochs). The model training level were compared using the quantitative metrics earlier described. Both models were trained with a training dataset of 24 cases and a test data set of 6 cases.

Since inter- and intrascanner variability influences the acquisition of the MR images, the impact of the vendors on the model were also analyzed. The model was rained on 2 vendors, 3 different scanners (*GE*, *Toshiba*, 22 cases) and tested on 2 other vendors, 3 different scanners (*Siemens*, *Philips*, 7 cases).

To qualify the model performance, an external testing dataset was used. This dataset (represented in Table 3) consists of 10 cases, acquired with various vendors and a variability in acquisition protocol.

Vendor	*#cases*	*TE (msec)*	*TR (msec)*	*B0 (T)*
SIEMENS Sonata	1	14	580	1.5
Philips Ingenuity	1	9	440	3.0
GE Healthcare Signa HDxt	6	9	440	1.5
GE Healthcare Discovery MR750	1	12	600	3.0
Toshiba Titan3T	1	8	480	3.0

Table 2. Overview of the external testing dataset acquired by various vendors and MR scanners, with the corresponding number of cases, protocol (TE and TR) and magnetic field strength B0.

3. RESULTS

3.1 Inter-observer variability analysis

To define the differences between the two annotators an inter-observer variability analysis was performed. The two annotators received feedback on the performed delineations by neuroradiologists. Main feedback consists of the incorrect segmentation of hyperintensities that do not represent enhancing lesions. The masks were improved and compared by the Dice coefficient, the 95th percentile of the Hausdorff distance in (mm) and the sensitivity of the delineations of the two annotators before and after consensus. The analysis was made on voxel level.

As shown in Figure 5, the statistical analysis shows an increase in the quality of manual lesion segmentation for all metrics. More specifically, the mean Dice coefficient increased from 0.61 ± 0.30 to 0.86 ± 0.12. The mean Hausdorff Distance decreased from 52.8 mm ± 70.6 mm to 1.15 mm ± 1.01 mm. At last, the sensitivity increased from 0.71 ± 0.33 to 0.89 ± 0.10.

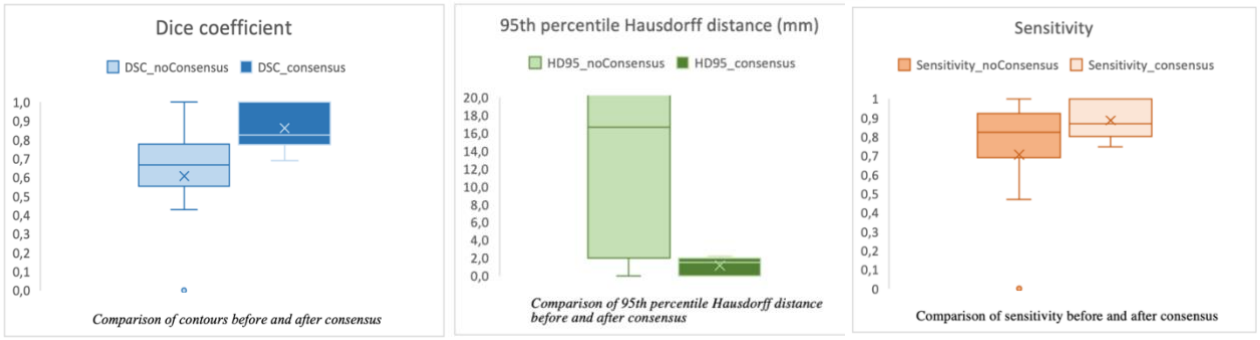


Figure 5. Boxplots of the Dice coefficient, 95th percentile of the Hausdorff distance, and sensitivity for both the manual delineations before and after consensus of the two annotators.

3.2 Model training level comparison

To qualify the model based on the optimal duration of training, the model was trained using 50 epochs and using 150 epochs. Since the output of the V-net is a prediction of the likelihood of the lesion, in voxel level, present in the MR images, this is an output with values between 0 and 1. To exclude the prediction of the lesion segmentation, the model uses a threshold of .95 as a post-processing step. The accuracy of the lesion prediction is measured by the Dice coefficient, the 95th percentile of the Hausdorff distance and the sensitivity. As depicted in Figure 6, the shallow (50 epochs) trained V-net achieves higher accuracy than the extensively (150 epochs) trained model. The overall Dice coefficient for the model trained with 50 epochs is 0.85 ± 0.04 compared to 0.18 ± 0.08 for the model trained with 150 epochs. The mean Hausdorff distance for the model trained with 50 epochs is 1.07 mm ± 0.17 mm compared to 22.3 mm ± 16.0 mm for the model trained with 150 epochs. The mean sensitivity for the model trained with 50 epochs is 0.75 ± 0.06 compared to 0.43 ± 0.28 for the model trained with 150 epochs. To give a visual interpretation of the data Figure 7 represents a bad, average, and good prediction

of the trained models. This figure shows examples (e.g., slices) of the model prediction for enhancing lesions. Shallow trained models are depicted on the left, and extensive trained models are depicted on the right. Each row consists of the model prediction highlighted with blue and the manual delineations highlighted in yellow. Additionally, the input MR image is provided next to them showing the enhancing lesion. First row: poor model performance, second row: average model performance, third row: good model performance.

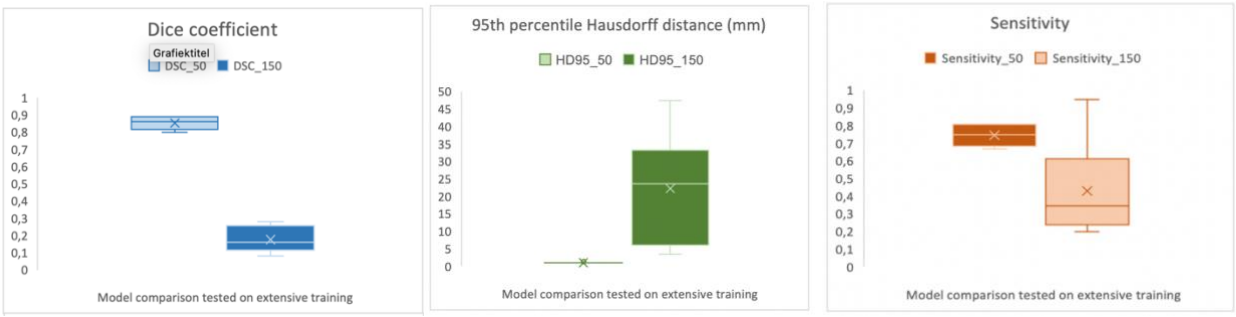


Figure 6. Boxplots of the Dice coefficient, 95th percentile of the Hausdorff distance, and sensitivity for a shallow trained model (50 epochs) compared to a more extensive trained model (150 epochs).

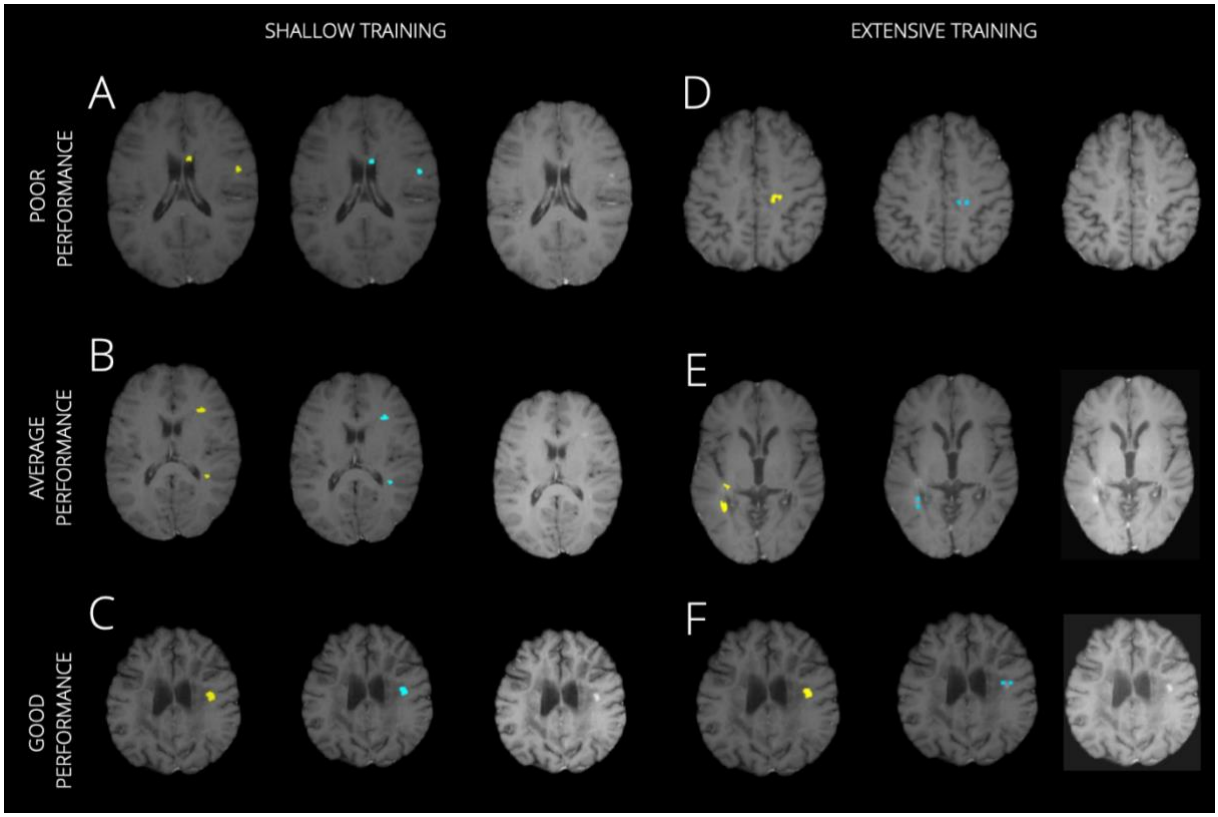


Figure 7. Examples (e.g., slices) of the model prediction for enhancing lesions. Shallow trained models are depicted in A-C, and extensive trained models are depicted in D-F. Each row consists of the model prediction highlighted with blue and the manual delineations highlighted in yellow. Additionally, the input MR image is provided next to them showing the enhancing lesion. First row: poor model performance, second row: average model performance, third row: good model performance.

3.3 Vendor impact on model analysis

As aforementioned, the inter-scanner variability influences the acquisition of the MR images. To analyze this the shallow trained model, given the higher performance, was used to train the data divided in two groups. The training group composed of 2 vendors, 3 different scanners (GE, Toshiba, 22 cases) and the testing group with 2 other vendors, 3 different scanners (Siemens, Philips, 7 cases). The analysis with the Dice coefficient, Hausdorff distance and the sensitivity resulted in a mean of 0.40 ± 0.13, 3.65 mm ± 0.24 mm and 0.25 ± 0.10, respectively. A visual representation of the model, a case with a prediction of the model, the label and the input image are presented in Figure 9. The figure shows differences between the prediction of the model and the label.

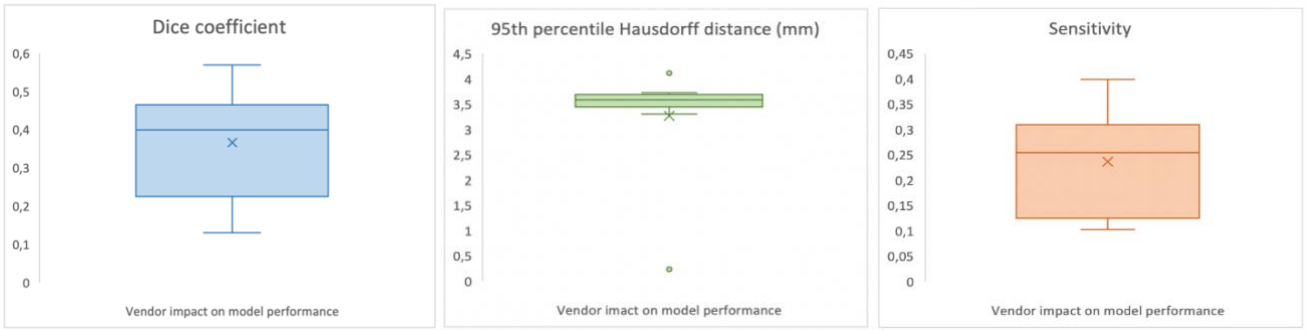


Figure 8. Boxplots of the Dice coefficient, 95th percentile of the Hausdorff distance, and sensitivity for training and testing the highest performance model (50 epochs) with different vendors.

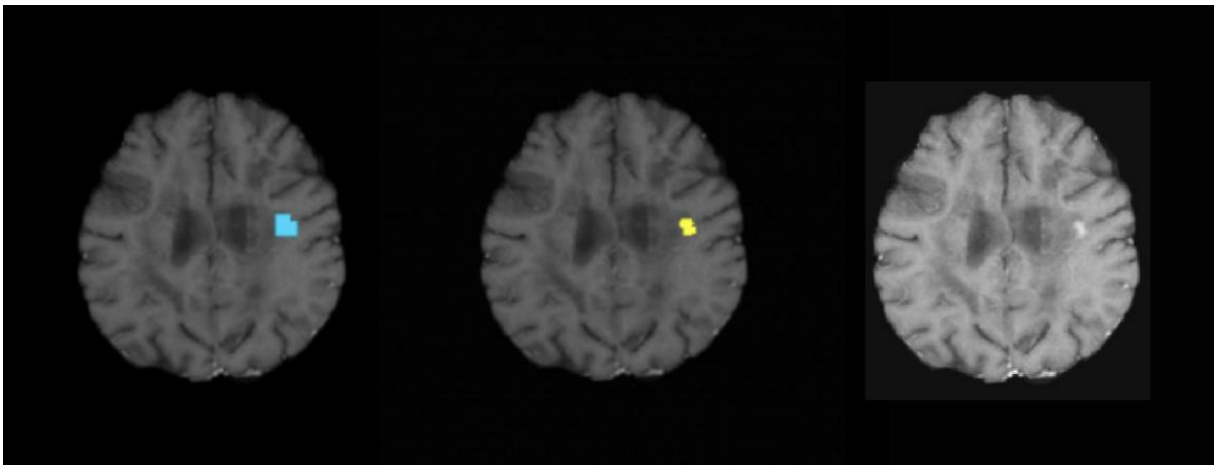


Figure 9. Example of highest performance model (50 epochs) prediction for one case used for the vendor impact analysis. This figure shows the prediction of the model of case 02 on the left in blue, the label of the lesions in yellow in the middle and the input image with the enhancing lesion on the right.

3.4 Validation

To qualify the robustness of the model for all lesion enhancing T1w data, the trained V-net was used to analyze the quality of automated segmentation using an external dataset. This dataset consists of 10 cases, including various venders, scanning protocols and magnetic field strengths. No quantitative measures were performed due to poorer model performance. However, a visual interpretation of the results states enough about the poorer quality of the automated lesion segmentation. As depicted in Figure 10 the prediction of the model states lesions, whereas in the input images obviously no enhancing lesions, are present.

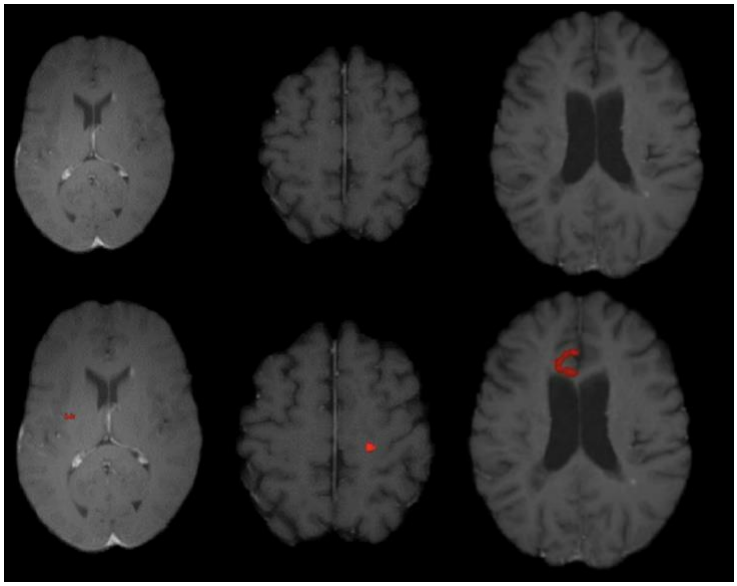


Figure 10. Visual representation of testing the shallow trained V-net with an external dataset. The images on top are the input images, the images on the bottom are the corresponding predictions of the model in red.

4. DISCUSSION

The performance of the model was analyzed using various analysis. First an inter-observer variability study was performed to qualify the accuracy of the manually segmentation of the enhancing lesions. Subsequently, the model training level comparison was performed to qualify the model on the optimal duration of training. To analyze the influence of the scanner variability of MR acquisition, the vendor impact on the model was qualified. Last, to validate the model, the trained V-net was tested with an external dataset to qualify the robustness of the model for all lesion enhancing T1w data.

The inter-observer analysis yielded in a good agreement after consensus for the two annotators used in this work suggesting a small bias on input level. Even though the use of two annotators following instructions of neuroradiologists increases the precision of the delineation of the lesions, a bias arises from caused by manual annotation. To detect the similarities between the two annotators the inter-observer variability study was performed. Quantitative analysis of the quality of the delineations by the two annotators shows that consensus was very important. The main reason for the quality improvement was the removal of wrongly delineated hyperintense areas that were mistaken for lesions. This is mainly visible by the strong decrease of the HD95. An example of a non-lesion hyperintense region is the choroid plexus, a network of blood vessels and cells around the ventricles of the brain.

Different number of epochs were used to optimize the training of the V-net while the best performance was given by the shallow trained model, i.e., 50 epochs compared to an extensive trained model, i.e., 150 epochs. All three measuring methods show an increase in accuracy of the automated segmentation. This is mainly due to overfitting. Overfitting is when the model is trained too well. It is a common pitfall in deep learning algorithms in which a model tries to fit the training data entirely and ends up memorizing the data patterns, noise, and random fluctuations, instead of the lesions. When the dataset changes, the model can't correlate to the previous information anymore. This is clearly shown when the model is validated using an external dataset. The model predicts a lesion when there obviously is no hyperintense area. This limitation can be overcome by increasing the dataset. Each new dataset contains new information what the model can use to create a more accurate prediction of the lesion segmentation.

To analyze the influence of the vendors to the variability in acquisition of the MR images the training and testing group were divided with both 3 different scanners. The quantitative metrics show that the quality of the model is influenced by the different scanners, as the results differ negatively from the quantitative metrics for a shallow trained model. This means that the vendors in this research do have impact on the robustness of the model. This can be verified to increase the training and testing dataset.

As earlier stated, the lesions are represented by hyperintense void shapes in the brain. These lesions are enhancing due to the administration of GBCA's. However, the intensity of the hyperintense region can vary due to differences in dose of GBCA's per patient during acquisition of the T1w images. With a larger dataset the model should pick up on this feature and should not influence the results. The data consists of varying vendors, data acquisition protocol, lesions present in the MR images and of course the gender. The gender ratio present in this study was a male/female ratio of 40/60. This is not representative for the prevalence universally, which is a male/female ratio of 25/75 (Alonso & Hernán, 2008). This could influence the quality of the automated lesion segmentation by the model, thus is a drawback of the method. Another drawback of the method is that only 2D acquired data is used in this research. To increase the variability in the dataset it is needed to also have data acquisitioned in 3D. As well as the multiple sequences the MR scanners can gather, such as FLAIR, T2w and PD as an addition to the T1w data. This data also contains information that is useful for the model to make a more accurate localization and segmentation of the lesion.

To decrease the likelihood of overfitting, augmentation was performed to the data. However, augmentation also has its limitations. By applying augmentation to the dataset, the inherent bias of the original data persists in the augmented data. In addition, identification of the optimal data augmentation strategy is a challenge itself. The augmentation used during training of the model was a combination of affine transformations. The addition of Gaussian noise was also tested to the performance of the model. However, instead of recognizing the lesion patterns, the model focused too much on the noise patterns and proposed an incorrect lesion segmentation as shown in Figure 11. Another way to improve the quality of the V-net is to implement hyperparameter optimization in the research method. A drawback in this method is that hyperparameters in this research are chosen based on visual inspection and compared with earlier research.

Future work that can improve the quality of the automated segmentation method is to include 5-fold cross-validation. This is because cross-validation can be applied to estimate the skill of a deep learning model on unseen data. In addition, transfer learning, which increases efficiency when training, can be used to have pre-training on gliomas and fine-tune it on Gd+ lesions. Next to the work that can be included, hyperparameter optimization can be used to improve the quality of the V-net in the research method.

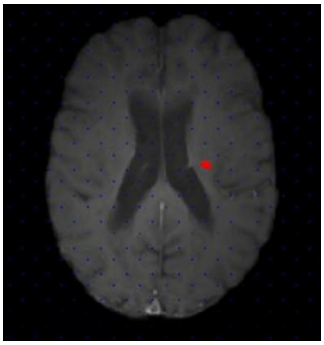


Figure 11. Prediction of the V-net when Gaussian noise is applied as augmentation. In red the label is presented, in blue the prediction of the model, which follows the Gaussian noise pattern.

5. CONCLUSION

To conclude, this deep learning model stands a proof of concept for enhancing lesion segmentation on MR contrast images, trained on a small amount of T1w data with limited performance in terms of generalizability across other centers. Additionally, there are preliminary findings that the vendor variability can affect the model performance. Moreover, based on the findings of this research, shallow model training (i.e., for a few epochs) indicates better results on enhancing lesion segmentation using limited single institute data. Future work could focus on 5-fold cross validation, transfer learning and hyperparameter optimization with an increased dataset size.

REFERENCES

Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D. L., & Erickson, B. J. (2017). Deep learning for. brain MRI segmentation: state of the art and future directions. Journal of digital imaging, 30(4), 449-459.

Alonso, A., & Hernán, M. A. (2008). Temporal trends in the incidence of multiple sclerosis: a systematic review. Neurology, 71(2), 129-135.

Barkhof, F., Held, U., Simon, J. H., Daumer, M., Fazekas, F., Filippi, M., ... & Wolinsky, J. (2005). Predicting gadolinium enhancement status in MS patients eligible for randomized clinical trials. Neurology, 65(9), 1447-1454.

Brosch, T., Tang, L. Y., Yoo, Y., Li, D. K., Traboulsee, A., & Tam, R. (2016). Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. IEEE transactions on medical imaging, 35(5), 1229-1239

Brugnara, G., Isensee, F., Neuberger, U., Bonekamp, D., Petersen, J., Diem, R., ... & 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(4), 2356-2364.

Carass, A. Snehashis Roy, Amod Jog, Jennifer L. Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, et al. 2017. "Longitudinal Multiple Sclerosis Lesion Segmentation Data Resource.". Data in Brief, 12, 346-50.

Carass, A., Roy, S., Jog, A., Cuzzocreo, J. L., Magrath, E., Gherman, A., ... & Pham, D. L. (2017). Longitudinal multiple sclerosis lesion segmentation: resource and challenge. NeuroImage, 148, 77-102.

Confavreux, C., & Vukusic, S. (2006). Natural history of multiple sclerosis: a unifying concept. Brain, 129(3), 606-616.

Coronado, I., Gabr, R. E., & Narayana, P. A. (2021). Deep learning segmentation of gadolinium-enhancing lesions in multiple sclerosis. Multiple Sclerosis Journal, 27(4), 519-527.

Cotton, F., Weiner, H. L., Jolesz, F. A., & Guttmann, C. R. (2003). MRI contrast uptake in new lesions in relapsing-remitting MS followed at weekly intervals. Neurology, 60(4), 640-646.

Datta, S., Sajja, B. R., He, R., Gupta, R. K., Wolinsky, J. S., & Narayana, P. A. (2007). Segmentation of gadolinium-enhanced lesions on MRI in multiple sclerosis. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine, 25(5), 932-937.

Inglese, M., Grossman, R. I., & Filippi, M. (2005). Magnetic resonance imaging monitoring of multiple sclerosis lesion evolution. Journal of Neuroimaging, 15, 22S-29S.

Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., ... & Kickingereder, P. (2019). Automated brain extraction of multisequence MRI using artificial neural networks. Human brain mapping, 40(17), 4952-4964.

Kappos, L., Moeri, D., Radue, E. W., Schoetzau, A., Schweikert, K., Barkhof, F., ... & Filippi, M. (1999). Predictive value of gadolinium-enhanced magnetic resonance imaging for relapse rate and changes in disability or impairment in multiple sclerosis: a meta-analysis. The Lancet, 353(9157), 964-969.

Mechrez, R., Goldberger, J., & Greenspan, H. (2016). Patch-based segmentation with spatial consistency: application to MS lesions in brain MRI. International Journal of Biomedical Imaging, 2016.

Norman, B., Pedoia, V., & Majumdar, S. (2018). Use of 2D U-Net convolutional neural networks for automated cartilage and meniscus segmentation of knee MR imaging data to determine relaxometry and morphometry. Radiology, 288(1), 177.

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). Springer, Cham.

Salem, M., Valverde, S., Cabezas, M., Pareto, D., Oliver, A., Salvi, J., ... & Lladó, X. (2020). A fully convolutional neural network for new T2-w lesion detection in multiple sclerosis. NeuroImage: Clinical, 25, 102149.

Simon, J. H. (2014). MRI outcomes in the diagnosis and disease course of multiple sclerosis. Handbook of clinical neurology, 122, 405-425.
Takao, H., Hayashi, N., & Ohtomo, K. (2011). Effect of scanner in longitudinal studies of brain volume changes. Journal of Magnetic Resonance Imaging, 34(2), 438-444.
Tullman, M. J. (2013). Overview of the epidemiology, diagnosis, and disease progression associated with multiple sclerosis. Am J Manag Care, 19(2 Suppl), S15-20.
Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: improved N3 bias correction. IEEE transactions on medical imaging, 29(6), 1310-1320.
Valverde, S., Cabezas, M., Roura, E., González-Villà, S., Pareto, D., Vilanova, J. C., ... & Lladó, X. (2017). Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. NeuroImage, 155, 159-168.
Valverde, S., Salem, M., Cabezas, M., Pareto, D., Vilanova, J. C., & Ramio-Torrenta, L. (2018). One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. Neuroimage Clin. 2019; 21: 101638.
Wattjes, M. P., Rovira, À., Miller, D., Yousry, T. A., Sormani, M. P., De Stefano, N., ... & Montalban, X. (2015). MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis--establishing disease prognosis and monitoring patients. Nature Reviews Neurology, 11(10), 597-607.
Zlokovic, B. V. (2011). Neurovascular pathways to neurodegeneration in Alzheimer's disease and other disorders. Nature Reviews Neuroscience, 12(12), 723-738.

APPENDIX A

To define the differences between the two annotators an inter-observer variability study was performed. The two annotators received feedback on the performed delineations by neuroradiologists. Main feedback consists of the incorrect segmentation of hyperintensities that don't represent enhancing lesions. This table represents the values of the Dice score, the Hausdorff Distance 95th percentile and the sensitivity for the cases with (rows in blue) and without consensus.

CASE	DICE SCORE	HD95	SENSITIVITY
1	0,690	67,780	0,94
2	0,000	237,520	0
3	0,600	2,300	0,85
4	0,610	136,250	0,47
5	1,000	0,000	1
6	0,660	74,310	1
7	0,730	1,730	0,49
8	0,650	2,230	0,83
9	0,000	109,730	0
10	0,000	79,000	0
11	1,000	0,000	1
12	0,490	98,820	0
13	1,000	0,000	0,77
14	0,680	12,800	1
15	0,720	14,000	0,67
16	0,770	2,000	0,87
17	0,730	27,330	0,83
18	0,820	2,230	0,87
19	0,690	2,000	0,75
20	0,780	2,000	0,8
21	0,600	20,460	0,87
22	0,660	19,430	0,79
23	0,830	1,000	0,81
24	0,560	167,000	0,93
25	0,000	226,150	0,94
26	0,000	203,820	0
27	1,000	0,000	1
28	0,860	1,000	0,85
29	0,550	62,860	0,78
30	0,620	41,270	0,82
31	0,670	12,480	0,91
32	0,430	60,650	0,8

Supplementary Table 1 Quantitative metrics for manual delineations before and after consensus

Appendix B

To qualify the model based on the optimal duration of training, the model was trained using 50 epochs and using 150 epochs. These tables represent the values of the Dice score, the Hausdorff Distance 95th percentile and the sensitivity for a shallow trained (50 epochs) and an extensive trained (150 epochs) model.

CASE	DSC	HD95	SENSITIVITY
29	0,89	1	0,81
30	0,85	1	0,73
31	0,8	1,41	0,67
32	0,89	1	0,81
33	0,82	1	0,69
34	0,87	1	0,77

Supplementary Table 2 Quantitative metrics for shallow trained model (50 epochs)

CASE	DSC	HD95	SENSITIVITY
29	0,28	3,46	0,42
30	0,125	28,47	0,2
31	0,157	26,12	0,25
32	0,164	21,33	0,5
33	0,08	7	0,27
34	0,25	47,46	0,95

Supplementary Table 3 Quantitative metrics for extensive trained model (150 epochs)

APPENDIX C

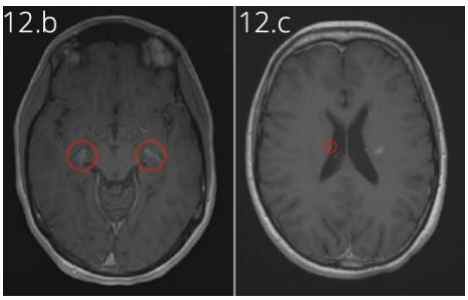
To analyze the influence on the acquisition of the MR images due to inter-scanner variability the shallow trained model, given the higher performance, was used to train the data divided in two groups. This table represents the quantitative metrics of the results of the tested cases.

CASE	DSC	HD95	SENSITIVITY
2	0,2	3,6	0,11
10	0,25	4,12	0,14
20	0,57	3,6	0,4
29	0,42	3,74	0,26
32	0,50	3,60	0,34
33	0,43	3,31	0,28
34	0,40	3,60	0,25

Supplementary Table 4 Quantitative metrics for vendor impact on model analysis

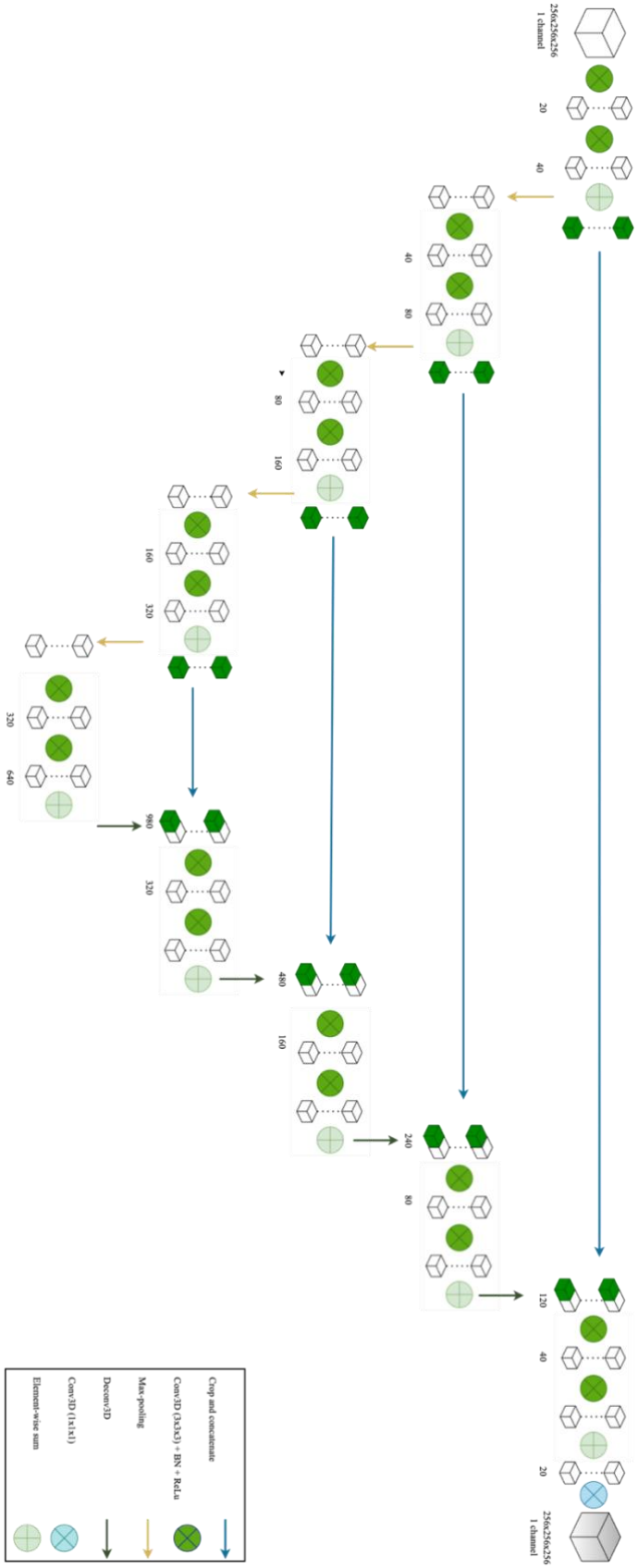
APPENDIX D

12. Exclusion criteria of instruction from neuroradiologists.
    1. Hypo-intensity, this is usually oedema or atrophy,
    2. In case of mirroring and similarity around the ventricular area, this is most likely plexus,
    3. In case of a small hyperintensity around the periventricular area, this is most likely a vessel.



APPENDIX E

A higher quality image of the visual representation of the model.



Layman's Summary

Multiple Sclerosis (MS), a neuroinflammatory disease of the central nervous system, is characterized by lesions in the brain. Lesions are inflammations in the brain and are visible in Magnetic Resonance images as void shaped hyperintensities after administration of a Gadolinium-based contrast agent (GBCA's). Artificial intelligence (AI) methods have been used among other researchers, in MS lesion segmentations; however, the existing methods have limitations making them difficult to be applicable to different MR scanners. Likewise, the goal of this research is to segment lesions on MR images using deep learning. Deep learning is an artificial intelligence method where data processing is are used to extract features from data. Features of lesions are for example their shape, intensity, and location in the brain. The deep learning method used in this research is called a V-net, which is a 3D structure that is suitable for volumetric image segmentations. A deep learning model is first trained with a dataset, here it learns all the features needed for lesion segmentation. Subsequently, the trained model is tested with a similar dataset to test how well it performs. To acquire the dataset two annotators manually segmented lesions in MR data and fed it to the model. The goal of the deep learning model is to predict an accurate lesion segmentation. Analysis of the segmentation quality before and after discussion between the two annotators concluded that this increased the quality of the delineation and therefore the model performance. A model can be trained with different number of epochs, this is defined as the number of times that the algorithm will work through the dataset. The model is analyzed with 50 epochs (shallow training) and 150 epochs (extensive training). The accuracy of the lesion segmentation is increased when the model is shallow trained, compared to extensive training. The quality of the model when used on a completely new, unseen dataset was tested and it showed that the model is not good enough to predict lesions when data with other MR protocols are used. Additionally, based on the findings of this work, there are preliminary findings that using different MR scanners can influence the model performance. The method can be improved by using a larger dataset. Summarizing, this deep learning model stands a proof of concept for lesion segmentation using limited data.