LST-AI: a Deep Learning Ensemble for Accurate MS Lesion Segmentation

Tun Wiltgen[1,2], Julian McGinnis[1,2,3], Sarah Schlaeger[4], CuiCi Voon[1,2], Achim Berthele[1], Daria Bischl[4], Lioba Grundl[4], Nikolaus Will[4], Marie Metz[4], David Schinz[4,5], Dominik Sepp[4], Philipp Prucker[4], Benita Schmitz-Koep[4], Claus Zimmer[4], Bjoern Menze[6], Daniel Rueckert[3,7], Bernhard Hemmer[1,8], Jan Kirschke[4], Mark Mühlau[1,2]*, Benedikt Wiestler[4,9]*

1 Department of Neurology, School of Medicine, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany
2 TUM-Neuroimaging Center, School of Medicine, Technical University of Munich, Munich, Germany
3 Department of Computer Science, Institute for AI in Medicine, Technical University of Munich, Munich, Germany
4 Department of Diagnostic and Interventional Neuroradiology, School of Medicine, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany
5 Institute of Radiology, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
6 Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland
7 Department of Computing, Imperial College London, London, United Kingdom
8 Munich Cluster for Systems Neurology (SyNergy), Munich, Germany
9 TranslaTUM, Center for Translational Cancer Research, Munich, Germany
* indicates equal contribution

Abstract

Automated segmentation of brain white matter lesions is crucial for both clinical assessment and scientific research in multiple sclerosis (MS). Over a decade ago, we introduced a lesion segmentation tool, LST, engineered with a lesion growth algorithm (LST-LGA). While recent lesion segmentation approaches have leveraged artificial intelligence (AI), they often remain proprietary and difficult to adopt. Here, we present LST-AI, an advanced deep learning-based extension of LST that consists of an ensemble of three 3D-UNets.

LST-AI specifically addresses the imbalance between white matter (WM) lesions and non-lesioned WM. It employs a composite loss function incorporating binary cross-entropy and Tversky loss to improve segmentation of the highly heterogeneous MS lesions. We train the network ensemble on 491 MS pairs of T1w and FLAIR images, collected in-house from a 3T MRI scanner, and expert neuroradiologists manually segmented the utilized lesion maps for training. LST-AI additionally includes a lesion location annotation tool, labeling lesion location according to the 2017 McDonald criteria (periventricular, infratentorial, juxtacortical, subcortical). We conduct evaluations on 270 test cases —comprising both in-house (n=167) and publicly available data (n=103)—using the Anima segmentation validation tools and compare LST-AI with several publicly available lesion segmentation models.

Our empirical analysis shows that LST-AI achieves superior performance compared to existing methods. Its Dice and F1 scores exceeded 0.5, outperforming LST-LGA, LST-LPA, SAMSEG, and the popular nnUNet framework, which all scored below 0.45. Notably, LST-AI demonstrated exceptional performance on the MSSEG-1 challenge dataset, an international WM lesion segmentation challenge, with a Dice score of 0.65 and an F1 score of 0.63—surpassing all other competing models at the time of the challenge. With increasing lesion volume, the lesion detection rate rapidly increased with a detection rate of >75% for lesions larger than 60mm^3.

Given its higher segmentation performance, we recommend that research groups currently using LST-LGA transition to LST-AI. To facilitate broad adoption, we are releasing LST-AI as an open-source model, available as a command-line tool, dockerized container, or Python script, enabling diverse applications across multiple platforms.

Keywords: Multiple Sclerosis, Artificial Intelligence, Lesion Segmentation, Magnetic Resonance Imaging, White Matter Lesions, Deep Learning

1. Introduction

Multiple sclerosis (MS) is a complex chronic inflammatory disease of the central nervous system. Clinically, MS typically manifests through neurological deficits which are mainly driven by inflammatory demyelinating lesions occurring in brain white matter and in the spinal cord and by neurodegeneration (axonal and neuronal loss). To date, inflammatory white matter lesions are a hallmark of MS and their identification on magnetic resonance imaging (MRI) plays a crucial role in the diagnosis and follow-up of MS (Filippi et al., 2018; Thompson, Banwell, et al., 2018; Thompson, Baranzini, et al., 2018). In addition, the location of lesions within the brain plays a role in diagnosing MS, as lesions in periventricular, juxtacortical, and infratentorial regions are part of the MS diagnostic criteria by indicating dissemination in space. In contrast, lesions in the subcortical region are solely considered to monitor disease progression (Thompson, Banwell, et al., 2018).

In clinical routine and research, the gold standard of lesion identification and segmentation is manual segmentation by trained neuroradiological experts. However, this constitutes a time-consuming task with both relevant inter- and intra-rater variability, thereby hampering studies with large datasets aiming to improve our understanding of MS.

In past years, many algorithms and tools have been developed and published with the goal of accurate automated lesion segmentation. As one of the early contributions to this field, we published the Lesion Segmentation Toolbox (LST), which has since been applied in numerous scholarly publications (Schmidt et al., 2012). While early segmentation algorithms have been designed primarily using statistical and early machine learning models such as Support Vector Machines, Gaussian Mixture Models or engineered by using manually selected features (Schmidt et al., 2012), more recent approaches incorporate learning-based features via encoder/decoder model stages (Cerri et al., 2021) or learn these end to end in fully convolutional models in (semi-) supervised settings (Commowick et al., 2018). With the advent of artificial intelligence (AI), automated lesion segmentation tools based on convolutional neural networks (CNN) have become increasingly popular and indeed provide similar or higher segmentation accuracy than earlier, machine learning-based methods (Diaz-Hurtado et al., 2022; H. Li et al., 2018; Ma et al., 2022; Zeng et al., 2020). This is also reflected in the rankings of published MS lesion segmentation challenges, e.g., MICCAI 2016 (Commowick et al., 2018) and ISBI 2015 (Carass et al., 2017). While CNN-based models often outperform earlier models in challenges, they only excel with a sufficient number of training data, as they are designed to learn priors and features automatically and do not incorporate manual feature selection. Consequently, they are especially prone to overfitting to the training data. Moreover, and in contrast to earlier machine learning models, CNNs are comparatively harder to regularize, as they have higher model and learning capacity, larger number of model parameters and thus more complex loss landscapes. Therefore, a large performance gap between training set and test set is often noticeable and highlights the need to evaluate the performance of CNN-based models on heterogeneous, external test data. Overcoming this gap and generalizing segmentation models in order to be applicable to data from multiple protocols and centers is one of the main on-going challenges for AI-based approaches. In this context, some AI-based

approaches that have previously been published are optimized towards transferability: Valverde et al. have provided *nicMSlesions*, a CNN-based lesion segmentation method that is able to adjust to a new image domain by retraining their model on a single image (Valverde et al., 2019). Furthermore, recent studies successfully train their models on one dataset and test it on another, external dataset, for which the MICCAI 2016 (Commowick et al., 2021) and ISBI 2015 (Carass et al., 2017) datasets are often selected (Cerri et al., 2021; Gentile et al., 2023; Kamraoui et al., 2022; Krishnan et al., 2023; X. Li et al., 2022; McKinley et al., 2021). Hence, the research field is moving towards more generalized segmentation tools, which is an important step towards clinical applicability of these methods.

In this study, we introduce a deep learning-based extension of LST. We carefully explain our selection of model architecture and describe the training and test set used, and show how our composite loss function allows us to optimize our model for generalizability on MRIs of unseen test centers. To support our claim of generalizability, we also compare the performance of our model against existing MS lesion segmentation algorithms. To facilitate studies and applications in MS research, we provide this enhanced toolkit as open source to the imaging community (https://github.com/CompImg/LST-AI).

2. Methods

2.1. Datasets

In the following section, we characterize and define training and test set, including details on image acquisition. With regard to in-house datasets, we respected the Code of Ethics of the World Medical Association (Declaration of Helsinki) for experiments involving humans; the study was approved by the local ethics committee.

For the training set, we used an in-house dataset consisting of 491 paired 3D FLAIR and 3D T1w images acquired on a 3.0T Achieva scanner (Philips Medical Systems, Best, The Netherlands) to train both our proposed LST-AI segmentation model and the nnUNet baseline. Testing and evaluation of segmentation performance of all methods was conducted on a combination of multiple datasets and on each dataset individually. The test set includes two in-house datasets and four publicly available datasets. The two in-house test datasets consist of data acquired on the same 3.0T Achieva scanner used for training data acquisition and on two further 3.0T scanners (Achieva dStream and Ingenia, Philips Medical Systems, Best, The Netherlands), respectively. The four publicly available datasets are: (i) msisbi: ISBI 2015 training data (Carass et al., 2017) (https://smart-stats-tools.org/lesion-challenge-2015); (ii) msljub: dataset published by Laboratory of Imaging Technologies (Lesjak et al., 2018) (https://lit.fe.uni-lj.si/en/research/resources/3D-MR-MS/); (iii) mssegtest: MICCAI 2016 challenge test dataset (Commowick et al., 2021) (https://shanoir.irisa.fr/shanoir-ng/welcome) and (iv) mssegtrain: MICCAI 2016 challenge training dataset (Commowick et al., 2021) (https://shanoir.irisa.fr/shanoir-ng/welcome). One case (msseg-test-center07-08) was removed from the mssegtest dataset because it included incorrect ground truth data. In total, the test set consists of 270 images from 254 subjects (note that the publicly available ISBI dataset is a

longitudinal dataset). Further characteristics of the datasets, including data on lesion load, are provided in Table 1. Details on image acquisition are provided in Table 2.

dataset	#subjects	#scans	age (years)	female / male	diagnosis	number of lesions	total lesion volume (mm^3)	publication	link
in-house training	491	491	mean(sd) = 34.3 (9.5)	330/161	RRMS (261) CIS (227) ON (3)	mean(sd) = 25.54 (30.59) median(IQR) = 15.0 (6.0-33.0)	mean(sd) = 3492.96 (7300.31) median(IQR) = 1244.0 (419.5-3767.5)	N/A	N/A
in-house test1	83	83	mean(sd) = 35.0 (9.1)	57/26	RRMS (8) CIS (74) ON (1)	mean(sd) = 25.25 (17.97) median(IQR) = 22.0 (13.0-33.5)	mean(sd) = 2828.60 (4108.44) median(IQR) = 1465.0 (639.5-3432.5)	N/A	N/A
in-house test2	84	84	mean(sd) = 33.2 (8.1)	50/34	RRMS (78) PPMS (2) CIS (3) Myelitis (1)	mean(sd) = 29.40 (27.81) median(IQR) = 21.0 (8.0-44.0)	mean(sd) = 13764.67 (18991.13) median(IQR) = 6024.5 (1927.0-18547.75)	N/A	N/A
msisbi	5	21	mean(sd) = 43.5 (10.3)	4/1	RRMS (4) PPMS (1)	mean(sd) = 45.95 (20.92) median(IQR) = 41.0 (34.0-47.0)	mean(sd) = 12889.76 (11095.38) median(IQR) = 7354.0 (3678.0-18425.0)	(Carass et al., 2017)	(1)
msljub	30	30	median(range) = 39 (25-64)	23/7	RRMS (24) SPMS (2) PRMS (1) CIS (2) Unspecified (1)	mean(sd) = 111.23 (106.68) median(IQR) = 92.0 (31.25-125.0)	mean(sd) = 17336.87 (16115.41) median(IQR) = 14046.5 (1758.0-28430.25)	(Lesjak et al., 2018)	(2)
mssegtest	37	37	mean(sd) = 46.8 (10.3)	29/8	N/A	mean(sd) = 44.89 (42.11) median(IQR) = 29.0 (13.0-64.0)	mean(sd) = 12672.73 (15099.75) median(IQR) = 7348.0 (1453.0-17271.0)	(Commowick et al., 2021)	(3)
mssegtrain	15	15	mean(sd) = 41.6 (9.8)	8/7	N/A	mean(sd) = 41.67 (30.21) median(IQR) = 39.0 (18.0-56.5)	mean(sd) = 20729.87 (20606.48) median(IQR) = 12366.0 (3783.0-33198.5)	(Commowick et al., 2021)	(3)

Table 1
Characteristics of the datasets. Three in-house datasets (training, test1, and test2) were used, as well as the public datasets msisbi from the ISBI 2015 challenge (Carass et al., 2017), msljub published by the Laboratory of Imaging Technologies (Lesjak et al., 2018), and mssegtest and mssegtrain which are the testing and training datasets from the MICCAI 2016 challenge, respectively (Commowick et al., 2021).
Abbreviations: CIS: clinically isolated syndrome, IQR: interquartile range, N/A: not applicable/available, ON: optic neuritis, PPMS: primary progressive multiple sclerosis, RRMS: relapsing-remitting multiple sclerosis, sd: standard deviation, SPMS: secondary progressive multiple sclerosis
(1) https://smart-stats-tools.org/lesion-challenge-2015
(2) https://lit.fe.uni-lj.si/en/research/resources/3D-MR-MS/
(3) https://shanoir.irisa.fr/shanoir-ng/welcome

dataset	scanner	field strength	sequence	voxel size	#scans
in-house training	Achieva, Philips Medical Systems	3.0T	T1w: TR=9ms, TE=4ms, FA=8	1x1x1mm3	491

			FLAIR: TR=10000ms, TE=140ms, TI=2750ms	0.9x0.9x1.5mm3	
in-house test1	Achieva, Philips Medical Systems	3.0T	T1w: TR=9ms, TE=4ms, FA=8	1x1x1mm3	83
			FLAIR: TR=10000ms, TE=140ms, TI=2750ms	0.9x0.9x1.5mm3	
in-house test2	Achieva dStream, Philips Medical Systems	3.0T	T1w: TR=9ms, TE=4ms, FA=8	0.75x0.75x0.75mm3	17
			FLAIR: TR=4800ms, TE=270ms, TI=1650ms	0.75x0.75x0.75mm3	
	Ingenia, Philips Medical Systems	3.0T	T1w: TR=9ms, TE=4ms, FA=8	0.75x0.75x0.75mm3	67
			FLAIR: TR=4800ms, TE=320ms, TI=1650ms	0.75x0.75x0.75mm3	
msisbi	Philips Medical Systems	3.0T	T1w: TR=10.3ms, TE=6ms, FA=8	0.82x0.82x1.17mm3	21
			FLAIR: TE=68ms, TI=835ms	0.82x0.82x2.2mm3	
msljub	Siemens Magnetom Trio	3.0T	T1w: TR=2000ms, TE=20ms, TI=800ms, FA=120	0.42x0.42x3.3mm3	30
			FLAIR: TR=5000ms, TE=392ms, TI=1800ms, FA=120	0.47x0.47x0.8mm3	
mssegtest	Siemens Verio	3.0T	T1w: TR=1900ms, TE=2.26ms, FA=9	1x1x1mm3	10
			FLAIR: TR=5000ms, TE=400ms, TI=1800ms, FA=120	0.5x0.5x1.1mm3	
	General Electrics Discovery	3.0T	T1w: TR=[7.5,8]ms, TE=3.2ms, FA=10	0.47x0.47x0.6mm3	8
			FLAIR: TR=9000ms, TE=[140,145]ms, TI=[2355, 2362]ms, FA=90	0.47x0.47x0.9mm3	
	Siemens Aera	1.5T	T1w: TR=1860ms, TE=3.37ms, FA=15	1.08x1.08x0.9mm3	9
			FLAIR:TR=5000ms, TE=336ms, TI=1800ms, FA=120	1.03x1.03x1.25mm3	
	Ingenia, Philips Medical Systems	3.0T	T1w: TR=9.4ms, TE=4.3ms, FA=8	0.74x0.74x0.85mm3	10
			FLAIR:TR=5400ms, TE=360ms, TI=1800ms, FA=90	0.74x0.74x0.7mm3	
mssegtrain	Siemens Verio	3.0T	T1w: TR=1900ms, TE=2.26ms, FA=9	1x1x1mm3	5
			FLAIR: TR=5000ms, TE=400ms, TI=1800ms, FA=120	0.5x0.5x1.1mm3	
	Siemens Aera	1.5T	T1w: TR=1860ms, TE=3.37ms, FA=15	1.08x1.08x0.9mm3	5
			FLAIR:TR=5000ms, TE=336ms, TI=1800ms, FA=120	1.03x1.03x1.25mm3	

	Ingenia, Philips Medical Systems	3.0T	T1w: TR=9.4ms, TE=4.3ms, FA=8	0.74x0.74x0.85mm3	5
			FLAIR:TR=5400ms, TE=360ms, TI=1800ms, FA=90	0.74x0.74x0.7mm3	

Table 2
Acquisition settings of the datasets.
Abbreviations: FA: flip angle, FLAIR: fluid-attenuated inversion recovery, TE: echo time, TI: inversion time, TR: repetition time, T1w: T1-weighted

2.2. Preprocessing

To guarantee fair comparisons across all baselines, we standardize preprocessing across all datasets and methods. Firstly, we register (affine registration) all images to the ICBM 152 nonlinear atlas version 2009 template (https://www.mcgill.ca/bic/neuroinformatics/brain-atlases-human) using the Greedy command line tool (P. Yushkevich, 2016/2023; P. A. Yushkevich et al., 2016). Subsequently, we use the deep learning-based HD-BET brain extraction tool to generate skull-stripped images (Isensee et al., 2019). Next, the shape of the skull-stripped images is cropped to the size that is required for the 3D UNets and intensities are normalized to [0;1]. To benchmark methods in its intended environment, we opt for non-skull-stripped images for SAMSEG, as well as the legacy algorithms of LST, the Lesion Prediction Algorithm (LST-LPA) and the Lesion Growth Algorithm (LST-LGA), which perform optimally with whole-brain data. Consequently, we omit the HD-BET skull-stripping, the cropping, and the intensity normalization preprocessing steps for these specific baselines, while retaining it for others.

This standardized preprocessing (including skull-stripping) is also integrated into our LST-AI toolbox, providing users with a streamlined approach.

2.3. Lesion segmentation

In this section, we first describe the proposed lesion segmentation tool followed by benchmark methods that have been applied in many studies and to which the proposed tool is compared. Finally, we outline the manual lesion segmentation workflows employed across the different datasets.

2.3.1. LST-AI ensemble network

The LST-AI tool encompasses preprocessing, lesion segmentation and, optionally, lesion location annotation. An overview of the workflow is shown in Figure 1.

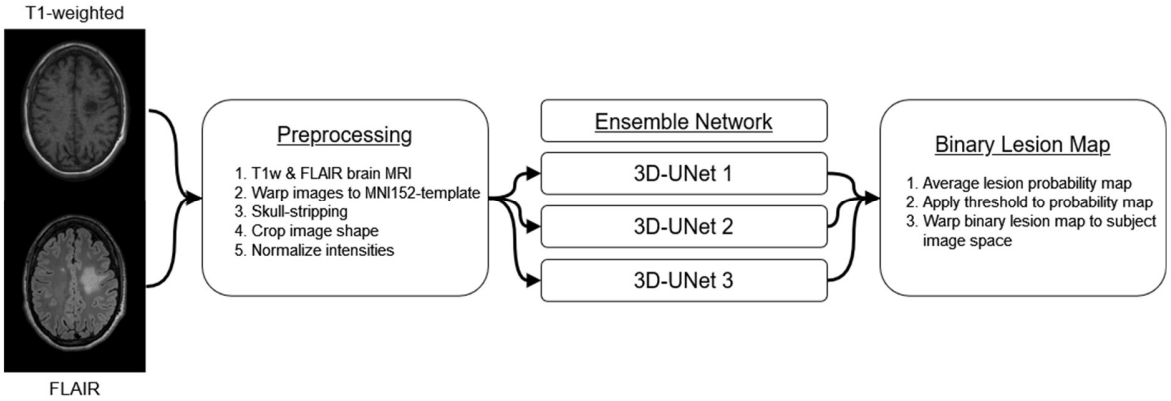

Figure 1
The different processing steps of the holistic LST-AI tool are presented. First, a pair of T1w and FLAIR images is warped to MNI space, then skull-stripped, cropped and intensity-normalized during preprocessing. The resulting images are used as input for the three 3D-UNets of the ensemble network. Each one of the UNets provides a lesion probability map. To generate the binary lesion map, the three lesion probability maps are averaged and a threshold is subsequently applied. Finally, the binary lesion map is warped back to the subject image space (original space of the FLAIR image).

The preprocessing functionality included in LST-AI is outlined in section 2.2. Specifically, the T1w and FLAIR images are warped to the MNI152-template, then skull-stripped, center cropped to shape (192, 192, 192), and, finally, intensities were normalized to [0;1].

With respect to the model architecture, LST-AI is based on an ensemble of three 3D-UNets. Each UNet is built upon the 3D-UNet (Çiçek et al., 2016) architecture and inspired by nnUNet (Isensee et al., 2021). It is composed of 5 encoder and 5 decoder blocks. Each of these blocks is built from two convolution blocks (3D convolution, instance normalization, leaky ReLU activation), and skip connections between respective encoder and decoder blocks (see Figure 2). In encoder blocks, downsampling is implemented via strided convolutions with stride 2 and transposed convolutions are used for upscaling in decoder blocks. Following the architectural choices in nnUNet (Isensee et al., 2021), we employ deep supervision layers in the training with the intuition of allowing gradients to flow deeper into the networks' layers (Wang et al., 2015). The number of deep supervision layers differed for the three UNets: one UNet included one deep supervision layer and the two other UNets included two deep supervision layers. For the loss function, we used a combination of Tversky loss (Salehi et al., 2017) (with higher penalization of false-negative lesion omissions) and binary cross-entropy in the deep supervision layers and a combined dice loss and binary cross-entropy in the full-resolution output. During training, we randomly chained intensity (random Gaussian noise, random Gaussian smoothing, random gamma adjustment) and geometry augmentations (random flips and crops). Each model was trained for a total of 1000 epochs, using the stochastic gradient descent optimizer (with Nesterov momentum) and a polynomial learning rate decay, starting at 1e-2. In total, three training runs were started from scratch to create an ensemble of three models, a technique previously reported (H. Li et al., 2018).

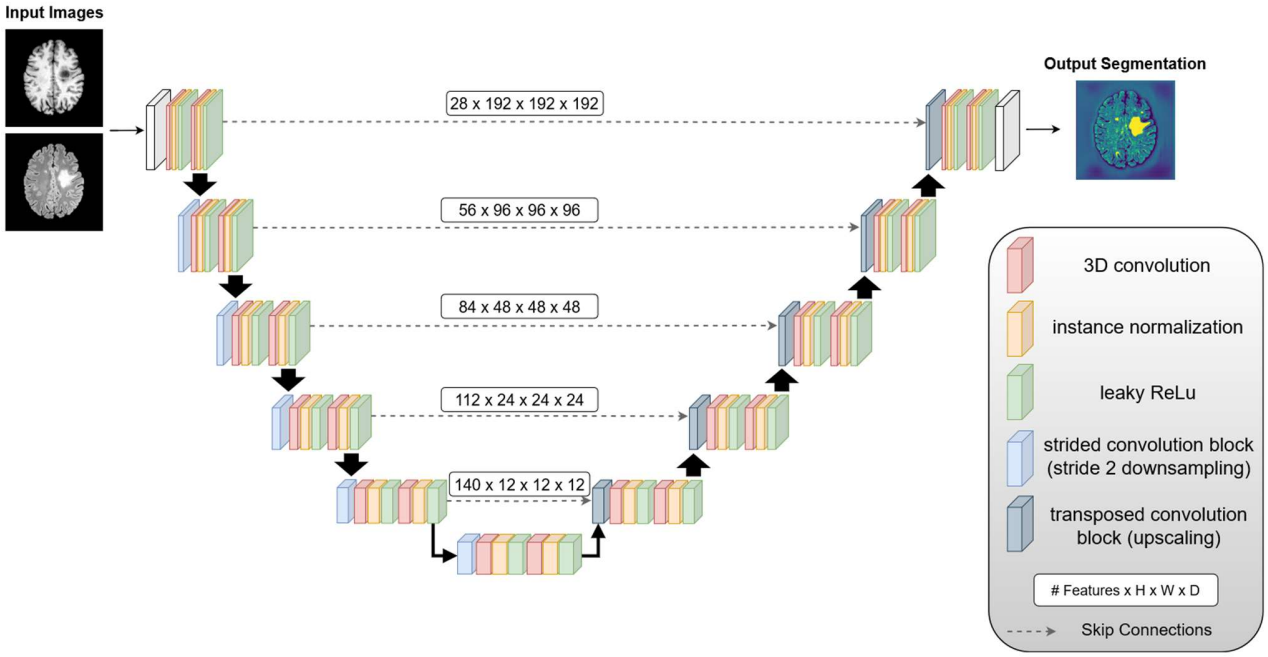

Figure 2
Architecture of the 3D-UNets which constitute the ensemble network of LST-AI. They comprise two channels (one for T1w images and one for FLAIR images) and consist of 5 encoder and 5 decoder blocks. Strided convolutions (stride 2) are used for downsampling and transposed convolutions are used for upscaling. Encoder and decoder blocks are connected via skip connections.

For the final segmentation output, the preprocessed T1w and FLAIR images are used as input for each one of the 3D UNets which generate three lesion probability maps. The final binary lesion map is obtained by averaging the three lesion probability maps and subsequent thresholding (default threshold of 0.5). This workflow, including the ground truth segmentation, is illustrated in Figure 3, using an example of the msljub dataset.

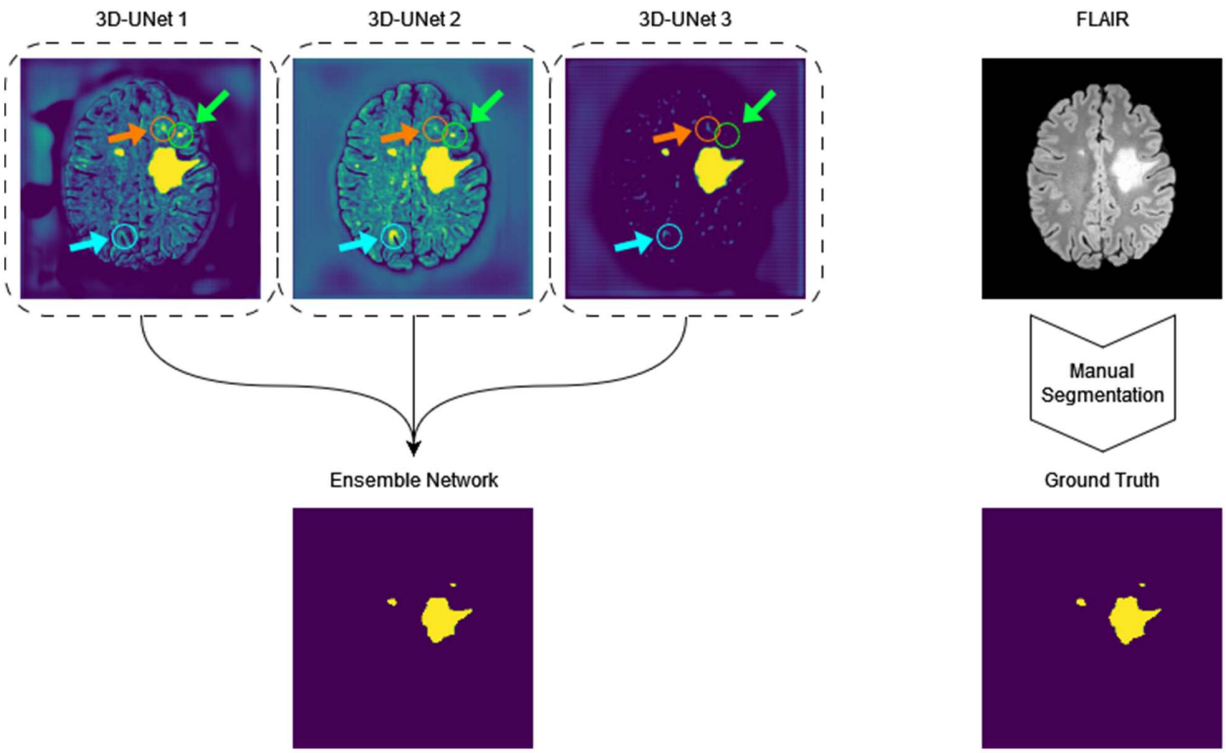

Figure 3
Rationale behind the ensemble network of LST-AI. First, the three 3D-UNets generate a lesion probability map. The mean of the three outputs is calculated and thresholded to generate the final binary lesion map. On the right-hand side, we show a slice of a FLAIR image and the corresponding manual segmentation (i.e., the ground truth). The orange arrow and circle highlight a false positive present in the lesion probability map of 3D-UNet 1, but not in the other lesion probability maps. The light blue arrow and circle highlight a false positive present in the lesion probability map of 3D-UNet 2, but not in the other lesion probability maps. The green arrow and circle highlight a false negative lesion in the lesion probability map of 3D-UNet 3, which is detected by 3D-UNet 1 and 2. Note how the output of the ensemble network is more accurate than the output of the individual networks, as it does not show the false positives and false negatives.

As an additional feature, the tool can optionally label lesions according to their location, i.e., periventricular (PV), juxtacortical (JC), subcortical (SC), or infratentorial (IT). To this end, the same ICBM 152 nonlinear T1 atlas used above is first registered deformably (using Greedy) to the skull-stripped T1w image in MNI space. The resulting transformation is applied to a manually labeled anatomical mask indicating different brain regions (inter alia: PV, IT, JC, and SC), which is thereby registered to the skull-stripped T1w image in MNI space. Next, each individual lesion from the binary lesion segmentation map is automatically labeled using a connected-component analysis, and assigned to the region with which it overlaps (by at least one voxel). During this step, lesions are assigned first to the PV, second to the IT, third to the JC, and, finally, to the SC region. In the resulting lesion map, the lesions are labeled according to their location (PV: label=1, JC: label=2, SC: label=3, IT: label=4). Finally, the labeled lesion map is transformed to the original space of the FLAIR image with the inverse of the affine

transformation, which was computed earlier, resulting in location-annotated lesion maps in the original subject space as well as in the MNI space.

We intend to target a diverse user base and provide LST-AI as a set of standalone command line tools and as a dockerized application, including all model checkpoints and required preprocessing tools (Greedy and HD-BET). As LST-AI can be used in similar ways as Freesurfer/FSL command line tools or nicMSlesions (docker), we give the opportunity to conveniently integrate our tool into existing workflows.

For accelerated performance, we recommend using our tool in a GPU-enabled environment but we also provide a fallback method for CPU-only usage. Depending on the exact hardware setup, typical execution time varies between tens of seconds (GPU) and 1-2 minutes on a CPU-only system. We provide LST-AI's functionality for three different workflows: segmentation-only, lesion location annotation-only, or both. Moreover, labels can be exported in the original subject space or in the MNI-152 template space.

Moreover, we make our source code available, allowing the community to adapt and tailor our tools for different application scenarios, by modifying preprocessing tools or using the checkpoints for pre-training of custom models. We intend to continuously maintain and update our tool in the github repository. In conclusion, while we have high confidence in the generalization capabilities of LST-AI, we want to emphasize that it is explicitly designed for research and non-clinical purposes. It has not undergone the necessary certification or licensing for clinical applications.

2.3.2. Benchmark methods

Evaluation of the performance of the proposed tool is realized through comparison to other publicly available lesion segmentation methods. This includes the widely used LST version 3.0.0 (https://www.applied-statistics.de/lst.html) with its lesion growth algorithm (LGA) (Schmidt et al., 2012) and lesion prediction algorithm (LPA) (Vanderbecq et al., 2020), to which our proposed tool presents a complementary, AI-based lesion segmentation method. Additionally, a trained nnUNet and the recently published SAMSEG lesion segmentation tool implemented in Freesurfer version 7.3.2 (Cerri et al., 2021) are used for comparison.

- **LST-LGA** (Schmidt et al., 2012): This method requires T1w and FLAIR images that are not skull-stripped. Before applying the LST-LGA tool, T1w and FLAIR images are preprocessed as described in section 2.2. Additionally, images are denoised using the CAT12 (Gaser et al., 2022) denoising filter implemented in SPM12 (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/). Then, the LST-LGA lesion segmentation algorithm is applied. First, using the methods implemented in SPM12, bias field correction is applied to the FLAIR image, and the T1w image is segmented into white matter, grey matter, and cerebrospinal fluid. Based on the FLAIR intensities, lesion belief maps are generated for each tissue class. The lesion belief map of grey matter is then thresholded (default threshold of 0.3 as suggested in Schmidt et al., 2012), which results in seeds that are used for the lesion growth model. Thereby, lesion seeds are expanded

according to FLAIR hyperintensities, eventually producing a lesion probability map. Finally, a binary lesion map is generated after thresholding the lesion probability map (threshold of 0.5).

- **LST-LPA** (Vanderbecq et al., 2020): This method requires only FLAIR images that are not skull-stripped. Preprocessing is identical to the LST-LGA workflow and includes registration to MNI and denoising. Similarly, bias correction is applied, and a lesion belief map is generated based on FLAIR intensities. The LST-LPA algorithm is a binary regression model that combines the lesion belief map and fixed parameters, which had been learned through logistic regression during the development of the tool in order to calculate the lesion probability map. The binary lesion map is again generated by applying a threshold to the lesion probability map (threshold of 0.5).

- **nnUNet** (Isensee et al., 2021): The UNet's early achievements in deep learning for biomedical segmentation have led to extensive research in refining its architecture for specialized tasks. Building on this, Isensee et al. (2021) have introduced an innovative framework that automates the selection of hyperparameters and data augmentation techniques based on the specific dataset employed. To provide this baseline, we format our training set according to nn-UNet's convention and train the model for 1000 epochs with five-fold cross-validation. We select the stronger 3D-UNet baseline in contrast to a 2D-UNet baseline, and use the full-resolution model as a baseline.

- **SAMSEG** (Cerri et al., 2021): This method requires only one MRI contrast image but it also accepts multiple contrasts. Here, we use T1w and FLAIR image pairs that are not skull-stripped as input. As recommended by the authors (Cerri et al., 2021), preprocessing is minimal, with images only being registered to MNI space using Greedy (P. A. Yushkevich et al., 2016). During the segmentation process, a deformable probabilistic atlas is used as segmentation prior and is iteratively fitted to the input data. Thereby, voxels are assigned to the brain structures with highest probability, including lesions. The binary lesion map is obtained by only selecting the voxels with lesion labels and setting all other voxel values to zero.

For region-specific analyses, all binary lesion maps are annotated with the method implemented in the LST-AI tool. In effect, each lesion is labeled according to its location (i.e., PV, JC, IT, or SC).

2.3.3. Manual segmentation

We make use of multiple datasets. Therefore, the workflows of manual segmentation, i.e., generation of ground truth lesion maps, differ. We describe the manual segmentation of the in-house datasets in detail. For public datasets, we refer to the corresponding publication.

- **in-house:** The training data were first pre-segmented using LST-LGA. Segmented lesions were manually reviewed and corrected by at least two out of four experienced neuroradiologists using ITK-SNAP (P. A. Yushkevich et al., 2006). Regarding test data, lesions were manually segmented and attributed to location labels by an experienced

neuroradiologist.

- **msisbi:** All images were manually delineated by two raters. Here we use the lesion map of rater 2. Since no consensus was available, we arbitrarily selected the lesion maps of one of the two raters as ground truth (rater 2). Protocol details have been described in the original publication (Carass et al., 2017).
- **msljub:** All images were delineated by three raters using a semi-automated approach. A consensus segmentation was obtained through revision of the combined lesion maps by all three raters; a detailed protocol is available in the original publication (Lesjak et al., 2018).
- **mssegtest & mssegtrain:** All images were manually delineated by seven raters, from which a consensus was constructed. Details on the protocol and consensus construction are available in the original publication (Commowick et al., 2021).

2.4. Evaluation

To assess the effectiveness of the LST-AI lesion segmentation tool, we compare its results with manual segmentations and other available tools. The tests exclude the internal training dataset and span the internal test datasets (1 and 2) and multiple external datasets to evaluate the generalizability. These external sets encompass various acquisition protocols, scanners, and originate from different centers. For consistency, we use images and lesion maps in MNI space. Our evaluation covers lesion location annotation, segmentation, and detection methods, applying a minimum lesion volume threshold of 3mm³ corresponding to 3 MNI-space voxels.

2.4.1. Lesion location annotation

The lesion location annotation is evaluated using the manually segmented lesion maps of the in-house test2 cohort. We compute the confusion matrix to analyze the accuracy of the lesion location annotation, considering the manual annotation as ground truth and the automated annotation as prediction. The fraction of lesions that are correctly assigned to the corresponding regions is extracted from the confusion matrix.

2.4.2. Lesion segmentation

Regarding lesion segmentation evaluation, we rely on the animaSegPerfAnalyzer tool from the anima evaluation toolbox (https://anima.irisa.fr/), which was also used in the MICCAI 2016 MS lesion segmentation challenge (Commowick et al., 2018). It requires pairs of ground truth (i.e, manually segmented) and automatically segmented lesion maps. This toolbox computes various metrics to analyze the segmentation performance at both the voxel and lesion level. Regarding voxel-wise analysis, we were interested in the Dice Similarity Coefficient (DSC):

$$DSC = \frac{2TP}{2TP + FP + FN}, \quad (1)$$

the positive predictive value (PPV):

$$PPV = \frac{TP}{TP + FP}, \quad (2)$$

and the sensitivity:

$$sensitivity = \frac{TP}{TP + FN}, \quad (3)$$

where TP denotes the true positives, FP the false positives, FN the false negatives. In addition, we extracted the average surface distance (ASD) with the animaSegPerfAnalyzer tool:

$$ASD = \frac{1}{n+n'}\left[\sum_{x=1}^{n} d(x, S') + \sum_{x'=1}^{n'} d(x', S)\right] \quad (4)$$

$$with \;\; d(x, S') = min||x - x'||_2, \quad (5)$$

where n and n' are the number of points x and x' on the surface S of the manual segmentation and the surface S' of the automated segmentation, respectively, and d() is the minimal Euclidean distance between a point x on surface S and the surface S′.

These metrics are calculated for each image, then averaged within each dataset, and finally averaged across all datasets. Thereby, we provide an overall score across different scanners and centers as well as individual scores for each dataset.

As an additional step, we construct one array by concatenating all images and calculate the DSC across all lesions of all datasets. We will refer to these analyses, neglecting subject-wise information, as first-level analyses (and to those based on subject-wise performance measures as second-level analyses). Thereby, we avoid the per-subject lesion load bias that is introduced when one score is calculated per image. For example, missing a small lesion in an image with only this missed lesion (DSC=0) would have more weight than missing a similar lesion in an image with many other detected lesions (DSC>0).

We further investigate whether the performance of lesion segmentation varies across brain regions. Thereby, we hope to identify the main drivers of the metric values and possible location-dependent variabilities of LST-AI segmentation performance. To this end, we use the location-annotated lesion maps and generate binary lesion maps for each region by only selecting lesion voxels labeled as part of the corresponding region. Using the above evaluation metrics, first-level analysis is conducted for each region and results from different regions and the whole brain are compared to each other.

2.4.3. Lesion detection

In addition to the previous metrics, which quantify the accuracy of lesion segmentation at the voxel level, it is important to evaluate lesion segmentation methods with regard to their ability to detect lesions. In particular, this aspect is crucial in MS, since its diagnosis relies on the detection of lesions (and not on the exact measurement of their volume). To this end, we extract the following scores from the animaSegPerfAnalyzer tool: SensL, the lesion detection sensitivity; PPVL, the positive predictive value for lesions; F1 score, a metric which considers both lesion detection sensitivity and positive predictive value for lesions. SensL and PPVL are calculated according to equations (3) and (2), respectively (on the lesion level rather than on the voxel level). The F1 score is calculated as follows:

$$F1 = 2 * \frac{SensL * PPVL}{SensL + PPVL} = \frac{2TP}{2TP + FP + FN}, \quad (6)$$

which is equal to the equation (1) and can therefore be considered as a lesion-wise DSC.

The anima evaluation toolbox also offers the animaDetectedComponents tool that can be used to investigate the detection of each lesion individually. For each image, the tool generates a list with lesions that are present in the manually segmented lesion map. It indicates, for each lesion, the volume in the manually segmented lesion map and whether it was detected by the automated segmentation method. This enables the assessment of the increase or decrease of lesion detection in relation to lesion volumes.

3. Results

We evaluate LST-AI in multiple aspects; we report both voxel-wise and lesion-wise scores, as both volume and number are established measures of lesion load. We start with lesion location annotation (3.1) as it is also relevant for the description of lesion segmentation (3.2). In 3.2, we report lesion segmentation across the whole brain and across subjects (second-level analyses). We then report the performance across lesions (first-level analyses) both across brain regions (3.3) and in relation to lesion volume (3.4).

3.1. Lesion location annotation

To evaluate the accuracy of the lesion location annotation, lesions of the in-house test2 cohort are manually assigned to four different brain regions and compared with the automatic annotation from LST-AI: PV, JC, SC, and IT. The confusion matrix is provided in Figure 4. In total, 847 lesions are assigned to the PV region during manual annotation, of which 682 (80.5%) are correctly assigned by the automatic method. In the JC region, 812 (86.9%) lesions are correctly classified and 108 (11.6%) are wrongly classified as PV. Less accurate classification is obtained in the SC region as only 285 (33.9%) lesions are identified as such and 220 (26.2%) and 335 (39.8%) are assigned to the PV and JC regions, respectively.

Using LST-AI, we performed lesion location annotation across all test set samples. In all 270 images included in the test set, 11154 lesions are segmented in the manually segmented lesion maps with a total lesion volume of $2.96*10^6 mm^3$. Most of the total lesion volume belongs to PV lesions (lesion volume: $2.33*10^6 mm^3$ (78.8%), lesion number: 3069 (27.5%)), whereas the JC region contains the most lesions (lesion number: 5208 (46.7%), lesion volume: $4.53*10^5 mm^3$ (15.3%)). The other two regions have smaller lesion numbers (IT: 610 (5.5%), SC: 2267 (20.3%)) and lesion volume (IT: $7.42*10^4 mm^3$ (2.5%), SC: $9.93*10^4 mm^3$ (3.4%)).

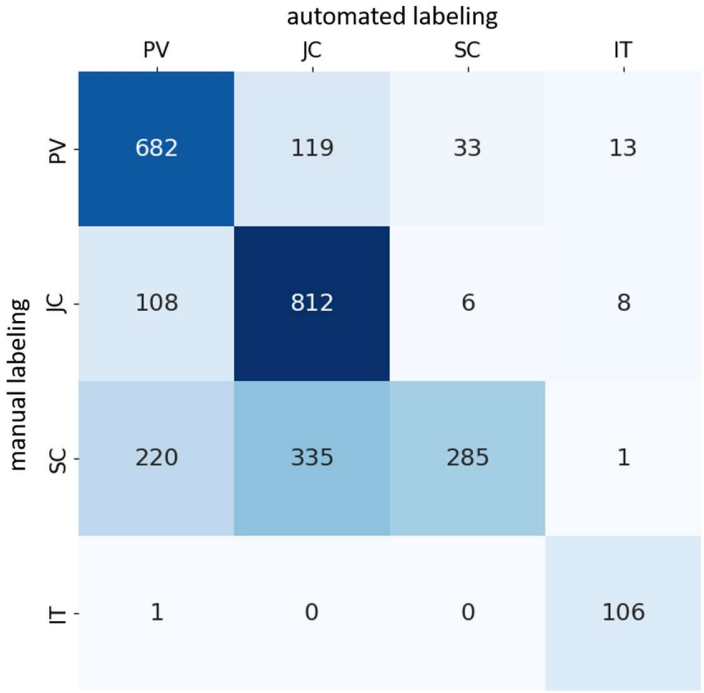

Figure 4
The confusion matrix shows the accuracy of the automated lesion location annotation. The values represent the number of manually labeled lesions that were correctly and incorrectly assigned to the different brain regions through automated labeling. The in-house test2 cohort was used for this evaluation.
Abbreviations: IT: infratentorial, JC: juxtacortical, PV: periventricular, SC: subcortical.

3.2. Second-level lesion segmentation across the whole brain

Lesion segmentation evaluation is conducted across all datasets as well as for each dataset individually. An overview of the results of each segmentation method across all datasets is provided in Table 3. A table with all anima metrics and results per case is included in the supplementary material.

	voxel-wise				lesion-wise		
tool	**DSC**	**PPV**	**sensitivity**	**ASD**	**F1**	**SensL**	**PPVL**
LST-AI	0.55 (0.18)	0.75 (0.19)	0.49 (0.22)	1.15 (3.34)	0.51 (0.21)	0.67 (0.24)	0.49 (0.24)
LST-LGA	0.36 (0.20)	0.68 (0.30)	0.29 (0.19)	2.06 (4.64)	0.20 (0.15)	0.27 (0.19)	0.26 (0.23)
LST-LPA	0.34 (0.20)	0.59 (0.32)	0.29 (0.18)	1.49 (2.26)	0.18 (0.14)	0.30 (0.19)	0.20 (0.20)
nnUNet	0.43 (0.20)	0.84 (0.20)	0.32 (0.19)	2.34 (5.90)	0.36 (0.24)	0.42 (0.22)	0.44 (0.30)
SAMSEG	0.41 (0.21)	0.58 (0.27)	0.37 (0.20)	2.63 (6.87)	0.29 (0.18)	0.31 (0.19)	0.38 (0.26)

Table 3
The results of the lesion segmentation evaluation (second-level analysis across all test datasets) of each segmentation tool are presented. The metrics were calculated for each image in the test datasets, and values were subsequently averaged across all images. The averages are reported as mean (standard deviation).
Abbreviations: ASD: average surface distance, DSC: dice similarity coefficient, PPV: positive predictive value, PPVL: lesion-wise positive predictive value, SensL: lesion-wise sensitivity

The proposed method outperforms the benchmark methods in all categories except for PPV, where only the nnUNet yields higher values (LST-AI: PPV=0.75 (0.19); nnUNet: PPV=0.84 (0.20)). Notably, LST-AI achieves higher DSC and F1 scores (DSC=0.55 (0.18), F1=0.51 (0.21)) compared to the other methods (DSC=0.34-0.43, F1=0.18-0.36), indicating superior segmentation performance both on a voxel-wise and on a lesion-wise level. The lowest ASD is also obtained with LST-AI, indicating more accurate lesion contouring compared to the benchmark methods. Overall the results show that LST-AI is able to identify more true lesions while increasing the fraction of correctly identified lesions among all segmented lesions compared to the benchmark methods.

Evaluating each dataset individually, we can observe some variability for each method across datasets. For LST-AI, this is shown in Table 4. Of note, the performance on the in-house datasets is inferior to that on the public dataset (in-house: DSC=0.47-0.48 and F1=0.41-0.46; public datasets: DSC=0.61-0.74 and F1=0.57-0.70), which, however, is paralleled by lower lesion load in the in-house datasets compared to the other datasets.

	voxel-wise				lesion-wise		
dataset	**DSC**	**PPV**	**sensitivity**	**ASD**	**F1**	**SensL**	**PPVL**
All datasets n=270	0.55 (0.18)	0.75 (0.19)	0.49 (0.22)	1.15 (3.34)	0.51 (0.21)	0.67 (0.24)	0.49 (0.24)
in-house test1 n=83	0.48 (0.20)	0.60 (0.18)	0.45 (0.22)	3.06 (5.48)	0.46 (0.21)	0.46 (0.24)	0.52 (0.22)
in-house test2	0.47 (0.11)	0.91 (0.12)	0.33 (0.09)	0.30 (0.77)	0.41 (0.19)	0.84 (0.14)	0.29 (0.16)

n=84							
msisbi n=21	0.61 (0.13)	0.72 (0.11)	0.54 (0.15)	0.41 (0.66)	0.57 (0.12)	0.55 (0.15)	0.61 (0.13)
msljub n=30	0.74 (0.10)	0.80 (0.07)	0.70 (0.14)	0.21 (0.88)	0.70 (0.10)	0.62 (0.13)	0.83 (0.11)
mssgtest n=37	0.65 (0.16)	0.68 (0.19)	0.68 (0.16)	0.59 (1.60)	0.63 (0.17)	0.83 (0.14)	0.55 (0.22)
mssegtrain n=15	0.67 (0.16)	0.72 (0.16)	0.67 (0.19)	0.12 (0.24)	0.61 (0.15)	0.77 (0.23)	0.53 (0.09)

Table 4
The results of the LST-AI lesion segmentation evaluation (second-level analysis) of each test dataset are presented. The metrics were calculated for each image in the respective test dataset, and values were subsequently averaged across all images. The averages are reported as mean (standard deviation). Abbreviations: ASD: average surface distance, DSC: dice similarity coefficient, PPV: positive predictive value, PPVL: lesion-wise positive predictive value, SensL: lesion-wise sensitivity

3.3. First-level segmentation across brain regions

LST-AI shows the highest first-level DSC scores in the PV region. However, DSC scores differ most in the other three regions with only LST-AI reaching DSC >0.38. Similarly, the highest first-level DSC score within the whole brain is obtained with LST-AI. The results of the different lesion segmentation methods are presented in Table 5 and Figure 5.

tool	Periventricular (PV)	Infratentorial (IT)	Juxtacortical (JC)	Subcortical (SC)	Whole brain
LST-AI	0.70	0.44	0.47	0.39	0.68
LST-LGA	0.54	0.10	0.19	0.15	0.51
LST-LPA	0.56	0.03	0.12	0.18	0.51
nnUNet	0.56	0.29	0.28	0.24	0.54
SAMSEG	0.60	0.20	0.16	0.22	0.56

Table 5
The first-level DSC score (across all test datasets) of each segmentation tool in different brain regions are presented in this table.

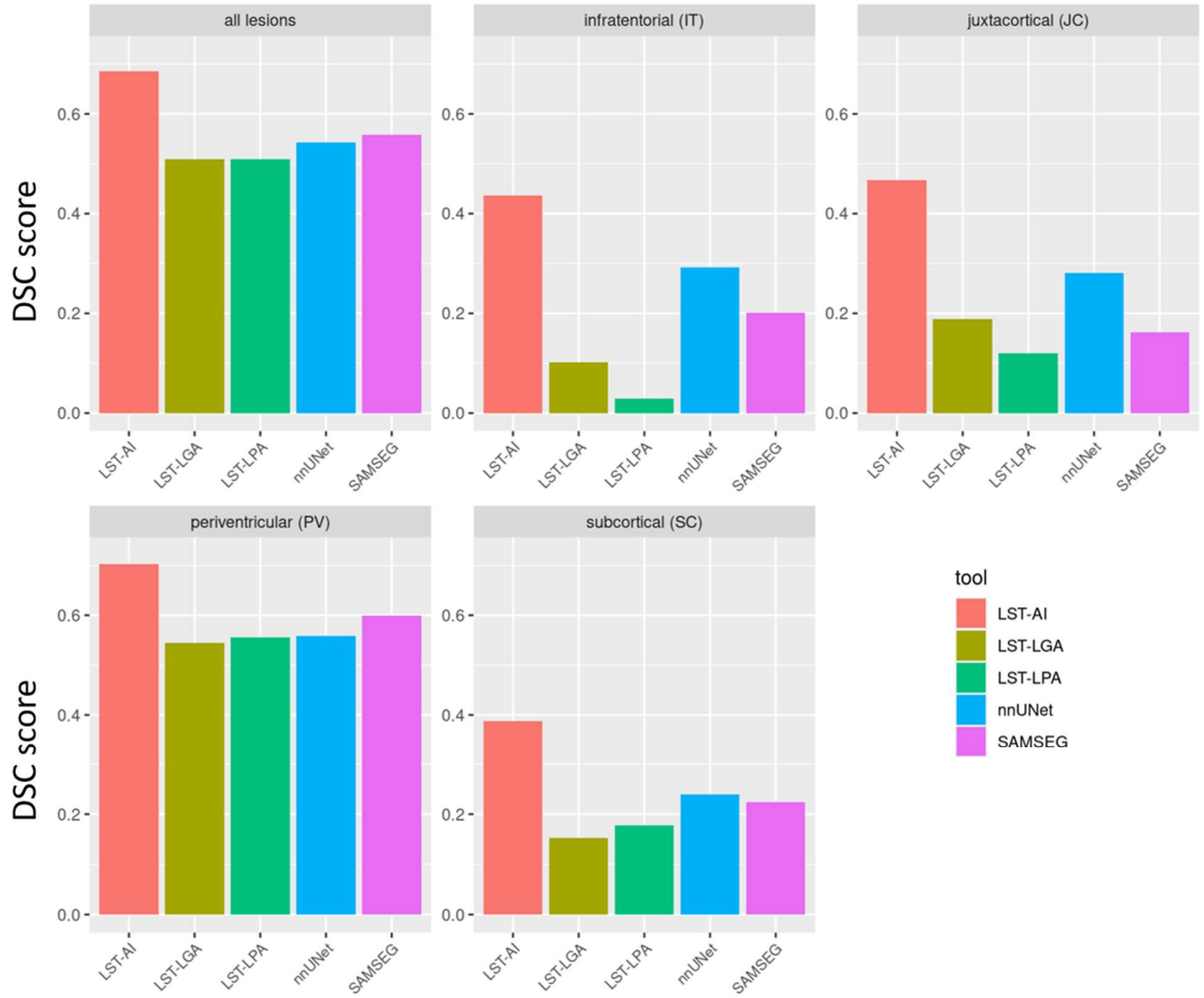

Figure 5
First-level DSC scores (across all test datasets) of each lesion segmentation tool are provided for lesions in different brain regions: all lesions in the whole brain, infratentorial lesions, juxtacortical lesions, periventricular lesions, and subcortical lesions.

3.4. First-level lesion detection in relation to lesion size

The lesion volume distribution of the test set is illustrated in Figure 6. The distribution shows a fast and steep decline with the most frequent lesions being small. This is critical as there is no commonly accepted minimum lesion volume (Grahl et al., 2019); moreover, accurate manual lesion segmentation is challenging, cumbersome, and sometimes overwhelming, even for expert readers. In Figure 7, we illustrate the accuracy of lesion detection in relation to lesion volume (bin width of 10mm^3). Small lesions (< 100 mm^3) are detected worse. With increasing lesion volume, the detection rate increases for all methods, with LST-AI showing the steepest incline. Hence, the advantage of LST-AI also applies to small lesions. Notably, the overall performance scores are considerably better for lesions > 100mm^3 than suggested by mere SensL scores.

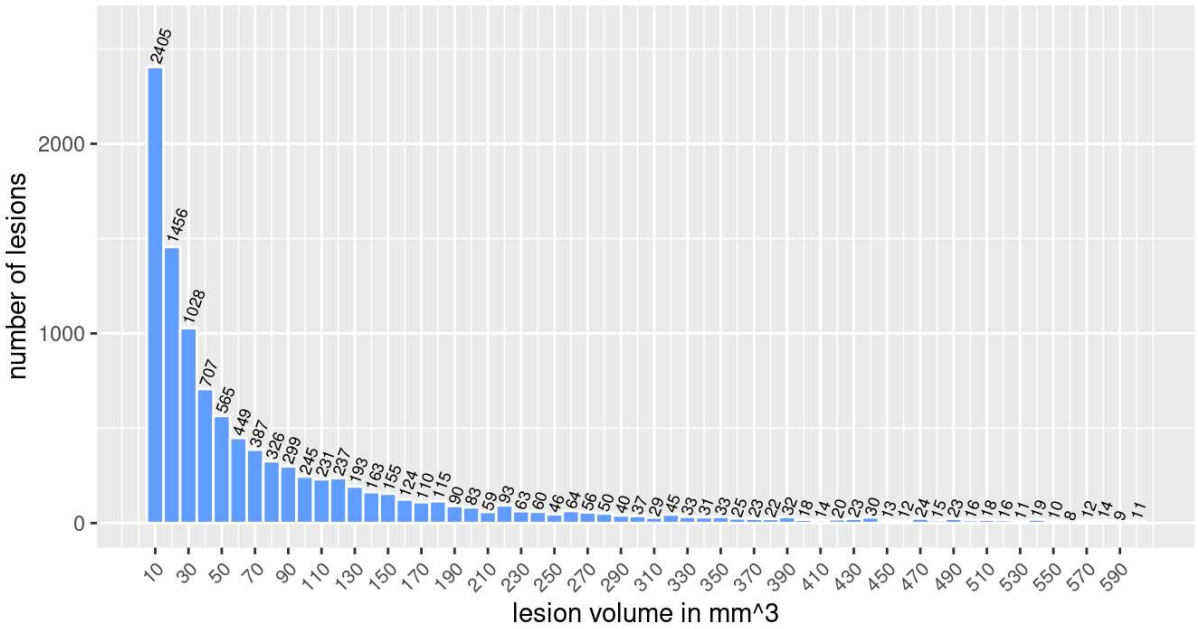

Figure 6

This graph shows the distribution of lesions per volume. The bars and numbers indicate how many lesions are in each volume group. We divided the lesions into groups with a volume range of 10mm^3 and the first bar from the left shows the number of lesions with a volume between 0mm^3 and 10mm^3.

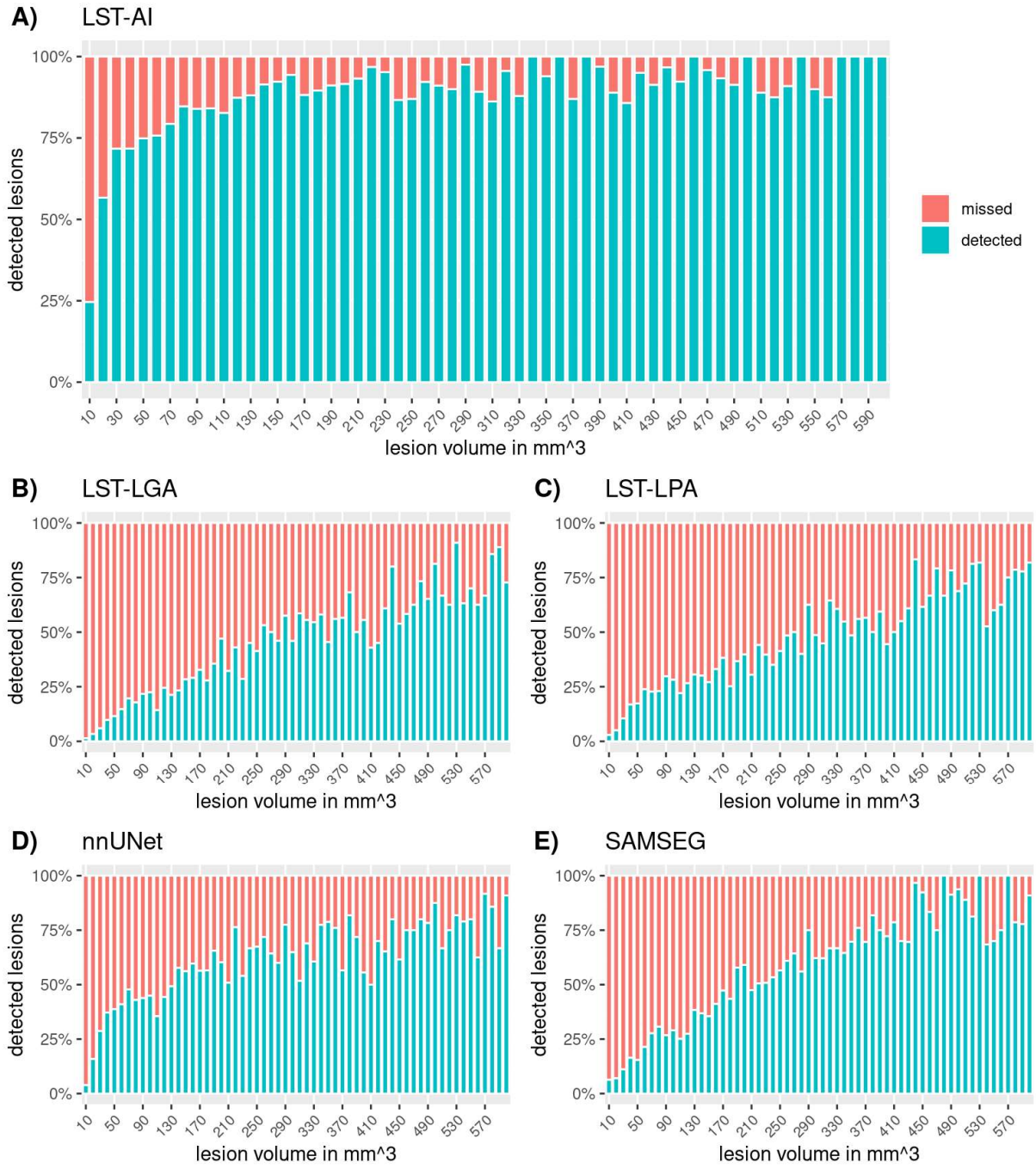

Figure 7

These graphs illustrate the proportion of lesions that are detected in each volume group. We divided the lesions into groups with a volume range of 10mm^3. The first bar from the left indicates the detection rate for lesions with a volume between 0mm^3 and 10mm^3. Note, how the detection rate increases with increasing lesion volume for each segmentation, whereby LST-AI yields the highest detection rates. The detection rate is given in %.

4. Discussion

We propose LST-AI, a new deep learning-based segmentation method for white-matter lesions in MS. It is built from an ensemble of three 3D UNets. Using LST-AI and a pair of T1w and FLAIR MRI images as input, it is possible to accurately segment lesions. We analyze the segmentation performance on multiple datasets, thereby showing that LST-AI generalizes to data from different centers and scanners without retraining. We also compare our method to benchmark methods for validation and find excellent lesion segmentation performance of our method. In addition, LST-AI can label lesions according to their location, thereby providing further possibilities for lesion characterization in multiple sclerosis.

LST-AI is pre-trained on an in-house dataset consisting of 491 images and does not need to be retrained before it is applied to new data. This makes it possible to use the tool even in smaller centers, where data is scarce and only small cohorts are available. Valverde et al., 2019, have previously optimized retraining on small datasets, as their tool only requires a single case to adapt their model to new datasets. They also validated their method on the ISBI 2015 test dataset and achieved a mean DSC of 0.58 (Valverde et al., 2019). In general, high-performing segmentation models in the ISBI 2015 challenge were CNN-based (trained on ISBI 2015 training dataset) and reported DSC scores ranging between 0.50 and 0.68 (Ma et al., 2022; Zhang & Oguz, 2021). However, assessing generalizability of segmentation models requires validation on external datasets. This has been done in recent studies, which used different train and test set pairings, including in-house and publicly available data such as ISBI 2015 and MICCAI 2016 data (e.g., train on in-house data and test on MICCAI 2016 data) (Billot et al., 2021; Cerri et al., 2021; Gentile et al., 2023; Kamraoui et al., 2022; X. Li et al., 2022; McKinley et al., 2021; Rakić et al., 2021). Overall, using train and test sets from different image domains led to lower and more variable DSC scores. For example, in the study by Kamraoui et al. (2022), the segmentation performance on the ISBI 2015 test dataset drops when models are trained on in-house data (DSC=0.13-0.48) compared to when they are trained on the ISBI training dataset (DSC=0.64-0.67). On the MICCAI 2016 dataset, however, the models trained on the in-house training dataset showed robust and high DSC scores (0.65-0.72) (Kamraoui et al., 2022). This highlights the impact of differing image domains in train and test sets and the need for validation on multiple test datasets, which can provide a more realistic representation of a model's generalizability. In this study, image domain heterogeneity is simulated by the validation of our method on multiple datasets, which were also part of MS lesion segmentation challenges of the ISBI 2015 conference and the MICCAI 2016 conference (Carass et al., 2017; Commowick et al., 2018, 2021). While our model achieves similar scores (mean DSC of 0.61 and 0.65 for ISBI 2015 and MICCAI 2016, respectively) as the top-performing models in both challenges, we want to emphasize that, in contrast to the participating models, our model is not specifically trained on the corresponding training datasets provided in the challenges. These two scores are also close to the inter-rater DSC scores of the expert segmentation used in the challenges (DSC of 0.63 and 0.66-0.76 in ISBI 2015 and MICCAI 2016, respectively) (Carass et al., 2017; Commowick et al., 2021). Other studies investigating the generalizability of their model on external data reported similar DSC scores in the range of [0.48 - 0.72] (Cerri et al., 2021; Kamraoui et al., 2022; McKinley et al., 2021; Rakić et al., 2021). Regarding LST-AI, the

DSC scores for the three external datasets (range: 0.61-0.74) underline the good generalization of our model and its reliable application to multicenter data acquired with different scanners and protocols. Although being trained on data from the same scanner and with the same acquisition protocol, the performance on the in-house test sets is inferior to that on the public dataset. At first sight, this might be unexpected, but one has to consider that the lesion load of subjects in the in-house dataset is lower. In this context, we and others (Commowick et al., 2018) observed a considerable influence of lesion size on lesion segmentation performance, which likely contributed to this counter-intuitive observation. Overall, results from both second- and first-level analysis show high segmentation performance of LST-AI on unseen data. In contrast, the lower performance of the other methods, e.g., the pre-trained nnUNet, suggests the need for adaptation of these methods through retraining. We hypothesize that using an ensemble approach including multiple pre-trained UNets translates into robustness against performance variability of individual 3D UNets and, therefore, generalizes better across different imaging protocols and centers. Of note, the mean PPV and PPVL values of the benchmark methods are comparable to those of LST-AI. However, this appears to happen at the cost of sensitivity, where LST-AI clearly outperforms the other methods at the voxel and lesion level. Compared to the literature, lesion-wise sensitivity of LST-AI on MICCAI 2016 data (SensL=0.83) and ISBI 2015 data (SensL=0.55) is in the same range as previously reported values (Carass et al., 2017; Commowick et al., 2018; Kamraoui et al., 2022; Krishnan et al., 2023; Ma et al., 2022; Zhang & Oguz, 2021). With regard to clinical applicability of automated lesion segmentation tools, the sensitivity is crucial as diagnosing and monitoring MS relies on the detection of (new) lesions. A newly published method, namely BIANCA-MS (Gentile et al., 2023), has also been validated using the MICCAI 2016 test dataset and yielded results similar to ours in terms of DSC and false positives (in terms of lesion detection). However, the median number of false negatives was equal to 11(IQR: 18) for BIANCA-MS, whereas LST-AI yields a median number of false negatives equal to 4 (IQR: 8), again highlighting the high sensitivity of our proposed method towards lesion detection.

In MS, lesion location within the brain may play an important role in identifying different disease patterns (Pongratz et al., 2023). In the LST-AI toolbox, a method is included which is able to classify lesions into four categories according to their location (PV, IT, JC, and SC). This makes it possible to seamlessly analyze the lesion load in different brain regions relevant to MS. In the in-house test2 cohort, the automated lesion location annotation is well in accordance with the manual lesion location annotation, except for SC lesions, where many lesions were classified as either PV (26.2% of all SC lesions) or JC (39.8% of SC all lesions) lesions. However, this accuracy drop should not be overly alarming, since the lesions are sometimes overlapping with multiple regions and are assigned to the SC region in the last iteration of the annotation. In other words, if a lesion is overlapping with multiple regions, it is always assigned to a region other than SC. When looking at the segmentation performance in the four different brain regions, it stands out that, among all methods included in this publication, LST-AI shows the highest DSC score in all regions. The increased lesion segmentation performance in the JC region is a particularly relevant finding, since segmentation of lesions close to the cortex based on T1w and FLAIR images has always been a challenge in MS. Also, juxtacortical lesions are thought to be

very specific for MS and are strongly associated with clinical disability (Calabrese et al., 2012), making their detection very important.

We also investigated the lesion detection in relation to lesion volume and we found that LST-AI has a higher lesion detection sensitivity for small lesions than the benchmark methods. Similar to previous reports by Commowick et al. (2018) and Rakić et al. (2021), we also found that it is particularly hard to detect small lesions (<10mm^3). Nonetheless, the steep incline of lesion detection with lesion size provides a promising perspective for the integration of automated lesion segmentation tools in clinical settings, since it can help clinicians to detect lesions faster and to diagnose and monitor MS more accurately.

Our study does not come without limitations. First, our model requires T1w and FLAIR image pairs, which might not always be available. Second, although less pronounced than in the benchmark methods, our model still shows a decrease in lesion detection efficiency with decreasing lesion volumes. Even though the explainability of features learned via CNNs and more specifically U-Nets have been comparatively well studied, they still lack some interpretability in contrast to methods leveraging manually selected features. In addition, preprocessing is included in the LST-AI toolbox and includes registration to MNI space, which ensures identical image dimensions and orientation before segmenting lesions. However, preprocessing steps are known to be crucial in segmentation tasks. Hence, exploring and applying different preprocessing steps could possibly change the performance on some datasets.

In conclusion, we introduce LST-AI, a new lesion segmentation toolbox and make it publicly available on GitHub (https://github.com/CompImg/LST-AI). It includes a preprocessing pipeline as well as an ensemble of three 3D UNets with binary cross-entropy and Tversky loss, making it a holistic lesion segmentation tool, enabling easy-to-implement, quick, and accurate automated lesion segmentation for MS research without retraining and fine-tuning. We validated its robustness on multiple datasets (in-house and publicly available datasets) and found excellent performance. We believe that, in future studies, LST-AI should replace LST.

References

Billot, B., Cerri, S., Leemput, K. V., Dalca, A. V., & Iglesias, J. E. (2021). Joint Segmentation Of Multiple Sclerosis Lesions And Brain Anatomy In MRI Scans Of Any Contrast And Resolution With CNNs. *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1971–1974. https://doi.org/10.1109/ISBI48211.2021.9434127

Calabrese, M., Poretto, V., Favaretto, A., Alessio, S., Bernardi, V., Romualdi, C., Rinaldi, F., Perini, P., & Gallo, P. (2012). Cortical lesion load associates with progression of

disability in multiple sclerosis. *Brain: A Journal of Neurology*, *135*(Pt 10), 2952–2961. https://doi.org/10.1093/brain/aws246

Carass, A., Roy, S., Jog, A., Cuzzocreo, J. L., Magrath, E., Gherman, A., Button, J., Nguyen, J., Prados, F., Sudre, C. H., Jorge Cardoso, M., Cawley, N., Ciccarelli, O., Wheeler-Kingshott, C. A. M., Ourselin, S., Catanese, L., Deshpande, H., Maurel, P., Commowick, O., … Pham, D. L. (2017). Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. *Neuroimage*, *148*, 77–102. https://doi.org/10.1016/j.neuroimage.2016.12.064

Cerri, S., Puonti, O., Meier, D. S., Wuerfel, J., Mühlau, M., Siebner, H. R., & Van Leemput, K. (2021). A contrast-adaptive method for simultaneous whole-brain and lesion segmentation in multiple sclerosis. *NeuroImage*, *225*, 117471. https://doi.org/10.1016/j.neuroimage.2020.117471

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, & W. Wells (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* (pp. 424–432). Springer International Publishing. https://doi.org/10.1007/978-3-319-46723-8_49

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., Pop, S. C., Girard, P., Ameli, R., Ferre, J. C., Kerbrat, A., Tourdias, T., Cervenansky, F., Glatard, T., Beaumont, J., Doyle, S., Forbes, F., Knight, J., Khademi, A., … Barillot, C. (2018). Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure. *Sci Rep*, *8*(1), 13650. https://doi.org/10.1038/s41598-018-31911-7

Commowick, O., Kain, M., Casey, R., Ameli, R., Ferré, J.-C., Kerbrat, A., Tourdias, T., Cervenansky, F., Camarasu-Pop, S., Glatard, T., Vukusic, S., Edan, G., Barillot, C., Dojat, M., & Cotton, F. (2021). Multiple sclerosis lesions segmentation from multiple

693 experts: The MICCAI 2016 challenge dataset. *NeuroImage*, *244*, 118589.
694 https://doi.org/10.1016/j.neuroimage.2021.118589
695 Diaz-Hurtado, M., Martínez-Heras, E., Solana, E., Casas-Roma, J., Llufriu, S., Kanber, B., &
696 Prados, F. (2022). Recent advances in the longitudinal segmentation of multiple
697 sclerosis lesions on magnetic resonance imaging: A review. *Neuroradiology*, *64*(11),
698 2103–2117. https://doi.org/10.1007/s00234-022-03019-3
699 Filippi, M., Bar-Or, A., Piehl, F., Preziosa, P., Solari, A., Vukusic, S., & Rocca, M. A. (2018).
700 Multiple sclerosis. *Nature Reviews Disease Primers*, *4*(1), Article 1.
701 https://doi.org/10.1038/s41572-018-0041-4
702 Gaser, C., Dahnke, R., Thompson, P. M., Kurth, F., Luders, E., & Initiative, A. D. N. (2022). *CAT*
703 *– A Computational Anatomy Toolbox for the Analysis of Structural MRI Data* (p.
704 2022.06.11.495736). bioRxiv. https://doi.org/10.1101/2022.06.11.495736
705 Gentile, G., Jenkinson, M., Griffanti, L., Luchetti, L., Leoncini, M., Inderyas, M., Mortilla, M.,
706 Cortese, R., De Stefano, N., & Battaglini, M. (2023). BIANCA-MS: An optimized tool for
707 automated multiple sclerosis lesion segmentation. *Human Brain Mapping*.
708 https://doi.org/10.1002/hbm.26424
709 Grahl, S., Pongratz, V., Schmidt, P., Engl, C., Bussas, M., Radetz, A., Gonzalez-Escamilla, G.,
710 Groppa, S., Zipp, F., Lukas, C., Kirschke, J., Zimmer, C., Hoshi, M., Berthele, A.,
711 Hemmer, B., & Mühlau, M. (2019). Evidence for a white matter lesion size threshold to
712 support the diagnosis of relapsing remitting multiple sclerosis. *Multiple Sclerosis and*
713 *Related Disorders*, *29*, 124–129. https://doi.org/10.1016/j.msard.2019.01.042
714 *Https://anima.irisa.fr/*. (n.d.). ANIMA. Retrieved August 8, 2023, from https://anima.irisa.fr/
715 *Https://www.applied-statistics.de/lst.html*. (n.d.). LST – Lesion Segmentation for SPM. Retrieved
716 July 28, 2023, from https://www.applied-statistics.de/lst.html
717 *Https://www.fil.ion.ucl.ac.uk/spm/software/spm12/*. (n.d.). SPM12 Software - Statistical
718 Parametric Mapping. Retrieved July 28, 2023, from

https://www.fil.ion.ucl.ac.uk/spm/software/spm12/

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, *18*(2), Article 2. https://doi.org/10.1038/s41592-020-01008-z

Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H.-P., Heiland, S., Wick, W., Bendszus, M., Maier-Hein, K. H., & Kickingereder, P. (2019). Automated brain extraction of multisequence MRI using artificial neural networks. *Human Brain Mapping*, *40*(17), 4952–4964. https://doi.org/10.1002/hbm.24750

Kamraoui, R. A., Ta, V.-T., Tourdias, T., Mansencal, B., Manjon, J. V., & Coup, P. (2022). DeepLesionBrain: Towards a broader deep-learning generalization for multiple sclerosis lesion segmentation. *Medical Image Analysis*, *76*, 102312. https://doi.org/10.1016/j.media.2021.102312

Krishnan, A. P., Song, Z., Clayton, D., Jia, X., de Crespigny, A., & Carano, R. A. D. (2023). Multi-arm U-Net with dense input and skip connectivity for T2 lesion segmentation in clinical trials of multiple sclerosis. *Scientific Reports*, *13*(1), Article 1. https://doi.org/10.1038/s41598-023-31207-5

Lesjak, Z., Galimzianova, A., Koren, A., Lukin, M., Pernus, F., Likar, B., & Spiclin, Z. (2018). A Novel Public MR Image Dataset of Multiple Sclerosis Patients With Lesion Segmentations Based on Multi-rater Consensus. *Neuroinformatics*, *16*(1), 51–63. https://doi.org/10.1007/s12021-017-9348-7

Li, H., Jiang, G., Zhang, J., Wang, R., Wang, Z., Zheng, W.-S., & Menze, B. (2018). Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. *NeuroImage*, *183*, 650–665. https://doi.org/10.1016/j.neuroimage.2018.07.005

Li, X., Zhao, Y., Jiang, J., Cheng, J., Zhu, W., Wu, Z., Jing, J., Zhang, Z., Wen, W., Sachdev, P. S., Wang, Y., Liu, T., & Li, Z. (2022). White matter hyperintensities segmentation using

745 an ensemble of neural networks. *Human Brain Mapping*, *43*(3), 929–939.
746 https://doi.org/10.1002/hbm.25695
747 Ma, Y., Zhang, C., Cabezas, M., Song, Y., Tang, Z., Liu, D., Cai, W., Barnett, M., & Wang, C.
748 (2022). Multiple Sclerosis Lesion Analysis in Brain Magnetic Resonance Images:
749 Techniques and Clinical Applications. *IEEE Journal of Biomedical and Health*
750 *Informatics*, *26*(6), 2680–2692. https://doi.org/10.1109/JBHI.2022.3151741
751 McKinley, R., Wepfer, R., Aschwanden, F., Grunder, L., Muri, R., Rummel, C., Verma, R.,
752 Weisstanner, C., Reyes, M., Salmen, A., Chan, A., Wagner, F., & Wiest, R. (2021).
753 Simultaneous lesion and brain segmentation in multiple sclerosis using deep neural
754 networks. *Sci Rep*, *11*(1), 1087. https://doi.org/10.1038/s41598-020-79925-4
755 Pongratz, V., Bussas, M., Schmidt, P., Grahl, S., Gasperi, C., El Husseini, M., Harabacz, L.,
756 Pineker, V., Sepp, D., Grundl, L., Wiestler, B., Kirschke, J., Zimmer, C., Berthele, A.,
757 Hemmer, B., & Mühlau, M. (2023). Lesion location across diagnostic regions in multiple
758 sclerosis. *NeuroImage: Clinical*, *37*, 103311. https://doi.org/10.1016/j.nicl.2022.103311
759 Rakić, M., Vercruyssen, S., Van Eyndhoven, S., de la Rosa, E., Jain, S., Van Huffel, S., Maes,
760 F., Smeets, D., & Sima, D. M. (2021). icobrain ms 5.1: Combining unsupervised and
761 supervised approaches for improving the detection of multiple sclerosis lesions.
762 *NeuroImage: Clinical*, *31*, 102707. https://doi.org/10.1016/j.nicl.2021.102707
763 Salehi, S. S. M., Erdogmus, D., & Gholipour, A. (2017). *Tversky loss function for image*
764 *segmentation using 3D fully convolutional deep networks* (arXiv:1706.05721). arXiv.
765 https://doi.org/10.48550/arXiv.1706.05721
766 Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., Hoshi, M., Ilg, R.,
767 Schmid, V. J., Zimmer, C., Hemmer, B., & Mühlau, M. (2012). An automated tool for
768 detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. *NeuroImage*,
769 *59*(4), 3774–3783. https://doi.org/10.1016/j.neuroimage.2011.11.032
770 Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., Correale, J.,

771 Fazekas, F., Filippi, M., Freedman, M. S., Fujihara, K., Galetta, S. L., Hartung, H. P.,
772 Kappos, L., Lublin, F. D., Marrie, R. A., Miller, A. E., Miller, D. H., Montalban, X., …
773 Cohen, J. A. (2018). Diagnosis of multiple sclerosis: 2017 revisions of the McDonald
774 criteria. *The Lancet Neurology*, *17*(2), 162–173. https://doi.org/10.1016/S1474-
775 4422(17)30470-2
776 Thompson, A. J., Baranzini, S. E., Geurts, J., Hemmer, B., & Ciccarelli, O. (2018). Multiple
777 sclerosis. *Lancet (London, England)*, *391*(10130), 1622–1636.
778 https://doi.org/10.1016/S0140-6736(18)30481-1
779 Valverde, S., Salem, M., Cabezas, M., Pareto, D., Vilanova, J. C., Ramió-Torrentà, L., Rovira,
780 À., Salvi, J., Oliver, A., & Lladó, X. (2019). One-shot domain adaptation in multiple
781 sclerosis lesion segmentation using convolutional neural networks. *NeuroImage:*
782 *Clinical*, *21*, 101638. https://doi.org/10.1016/j.nicl.2018.101638
783 Vanderbecq, Q., Xu, E., Stroer, S., Couvy-Duchesne, B., Diaz Melo, M., Dormont, D., Colliot,
784 O., & Alzheimer's Disease Neuroimaging, I. (2020). Comparison and validation of seven
785 white matter hyperintensities segmentation software in elderly patients. *Neuroimage*
786 *Clin*, *27*, 102357. https://doi.org/10.1016/j.nicl.2020.102357
787 Wang, L., Lee, C.-Y., Tu, Z., & Lazebnik, S. (2015). *Training Deeper Convolutional Networks*
788 *with Deep Supervision* (arXiv:1505.02496). arXiv.
789 https://doi.org/10.48550/arXiv.1505.02496
790 Yushkevich, P. (2023). *Greedy* [C++]. https://github.com/pyushkevich/greedy (Original work
791 published 2016)
792 Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., & Gerig, G. (2006).
793 User-guided 3D active contour segmentation of anatomical structures: Significantly
794 improved efficiency and reliability. *NeuroImage*, *31*(3), 1116–1128.
795 https://doi.org/10.1016/j.neuroimage.2006.01.015
796 Yushkevich, P. A., Pluta, J., Wang, H., Wisse, L. E. M., Das, S., & Wolk, D. (2016). Fast

Automatic Segmentation of Hippocampal Subfields and Medial Temporal Lobe Subregions In 3 Tesla and 7 Tesla T2-Weighted MRI. *Alzheimer's & Dementia*, *12*(7S_Part_2), P126–P127. https://doi.org/10.1016/j.jalz.2016.06.205

Zeng, C., Gu, L., Liu, Z., & Zhao, S. (2020). Review of Deep Learning Approaches for the Segmentation of Multiple Sclerosis Lesions on Brain MRI. *Frontiers in Neuroinformatics*, *14*. https://www.frontiersin.org/articles/10.3389/fninf.2020.610967

Zhang, H., & Oguz, I. (2021). Multiple Sclerosis Lesion Segmentation—A Survey of Supervised CNN-Based Methods. In A. Crimi & S. Bakas (Eds.), *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (pp. 11–29). Springer International Publishing. https://doi.org/10.1007/978-3-030-72084-1_2

Acknowledgments:
We thank Naga Karthik Enamundram and Joshua Newton for helpful discussions around the packaging of LST-AI, the evaluation of the different algorithms using the anima toolbox, and for visualization of the UNet architecture.

Funding:
MM received funding by a research grant of the National Institutes of Health (grant 1R01NS112161-01). MM received funding by the Bavarian State Ministry for Science and Art (Collaborative Bilateral Research Program Bavaria – Québec: AI in medicine, grant F.4-V0134.K5.1/86/34). BM, DR, MM and BW received funding from the DFG, SPP Radiomics (project number 428223038).

Data and code availability:
We provide our toolbox as source code, command line tool and dockerized application at https://github.com/CompImg/LST-AI.

Author contributions:
Tun Wiltgen: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization
Julian McGinnis: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing - original draft, Visualization
Sarah Schlaeger: Investigation, Resources, Data curation, Writing - review & editing
Cuici Voon: Investigation, Data curation, Writing - review & editing
Achim Berthele: Resources, Writing - review & editing
Daria Bischl: Resources, Data curation, Writing - review & editing
Lioba Grundl: Resources, Data curation, Writing - review & editing

Nikolaus Will: Resources, Data curation, Writing - review & editing
Marie Metz: Resources, Data curation, Writing - review & editing
David Schinz: Resources, Data curation, Writing - review & editing
Dominik Sepp: Resources, Data curation, Writing - review & editing
Philipp Prucker: Resources, Data curation, Writing - review & editing
Benita Schmitz-Koep: Resources, Data curation, Writing - review & editing
Claus Zimmer: Resources, Writing - review & editing
Bjoern Menze: Resources, Writing - review & editing
Daniel Rückert: Resources, Writing - review & editing
Bernhard Hemmer: Resources, Writing - review & editing
Jan Kirschke: Investigation, Resources, Data curation, Writing - review & editing
Mark Mühlau: Conceptualization, Methodology, Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition
Benedikt Wiestler: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing - original draft, Supervision, Project administration, Funding acquisition

Declaration of Competing Interests:
The authors declare no competing interests.