

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

2

3 Tun Wiltgen^{1,2}, Julian McGinnis^{1,2,3}, Sarah Schlaeger⁴, Florian Kofler^{3,4,5,6}, CuiCi Voon^{1,2}, Achim
4 Berthele¹, Daria Bischl⁴, Lioba Grundl⁴, Nikolaus Will⁴, Marie Metz⁴, David Schinz^{4,7}, Dominik
5 Sepp⁴, Philipp Prucker⁴, Benita Schmitz-Koep⁴, Claus Zimmer⁴, Bjoern Menze⁸, Daniel
6 Rueckert^{3,9}, Bernhard Hemmer^{1,10}, Jan Kirschke⁴, Mark Mühlau^{1,2*}, Benedikt Wiestler^{4,5*}

7

8 1 Department of Neurology, School of Medicine, Klinikum rechts der Isar, Technical University
9 of Munich, Munich, Germany

10 2 TUM-Neuroimaging Center, School of Medicine, Technical University of Munich, Munich,
11 Germany

12 3 Department of Computer Science, Institute for AI in Medicine, Technical University of Munich,
13 Munich, Germany

14 4 Department of Diagnostic and Interventional Neuroradiology, School of Medicine, Klinikum
15 rechts der Isar, Technical University of Munich, Munich, Germany

16 5 TranslaTUM, Center for Translational Cancer Research, Munich, Germany

17 6 Helmholtz AI, Helmholtz Munich, Neuherberg, Germany

18 7 Institute of Radiology, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-
19 Nürnberg, Erlangen, Germany

20 8 Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland

21 9 Department of Computing, Imperial College London, London, United Kingdom

22 10 Munich Cluster for Systems Neurology (SyNergy), Munich, Germany

23 * indicates equal contribution

24

25

26

27

28

29

30

31

32

33 Corresponding Author:

34 Mark Mühlau

35 mark.muehlau@tum.de

36 Department of Neurology, School of Medicine, Technical University of Munich

37 Ismaninger Str. 22, 81675 Munich, Germany

38 **Abstract**

39

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

46

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

57

58 Our empirical analysis shows that LST-AI achieves superior performance compared to existing
59 methods. Its Dice and F1 scores exceeded 0.62, outperforming LST, SAMSEG (Sequence
60 Adaptive Multimodal SEGmentation), and the popular nnUNet framework, which all scored
61 below 0.56. Notably, LST-AI demonstrated exceptional performance on the MSSEG-1 challenge
62 dataset, an international WM lesion segmentation challenge, with a Dice score of 0.65 and an
63 F1 score of 0.63—surpassing all other competing models at the time of the challenge. With
64 increasing lesion volume, the lesion detection rate rapidly increased with a detection rate of
65 >75% for lesions with a volume between 10mm³ and 100mm³.

66

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

71

72

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

75 1. Introduction

76 Multiple sclerosis (MS) is a complex chronic inflammatory disease of the central nervous
77 system. Clinically, MS typically manifests through neurological deficits which are mainly driven
78 by inflammatory demyelinating lesions occurring in brain white matter and in the spinal cord and
79 by neurodegeneration (axonal and neuronal loss). To date, inflammatory white matter lesions
80 are a hallmark of MS and their identification on magnetic resonance imaging (MRI) plays a
81 crucial role in the diagnosis and follow-up of MS (Filippi et al., 2018; Thompson, Banwell, et al.,
82 2018; Thompson, Baranzini, et al., 2018). In addition, the location of lesions within the brain
83 plays a role in diagnosing MS, as lesions in periventricular, juxtacortical, and infratentorial
84 regions are part of the MS diagnostic criteria by indicating dissemination in space. In contrast,
85 lesions in the subcortical region are only used to determine longitudinal dissemination and to
86 monitor disease progression (Thompson, Banwell, et al., 2018).

87
88 In clinical routine and research, the gold standard of lesion identification and segmentation is
89 manual segmentation by trained neuroradiological experts. However, this constitutes a time-
90 consuming task with both relevant inter- and intra-rater variability, thereby hampering studies
91 with large datasets aiming to improve our understanding of MS.

92
93 In past years, many algorithms and tools have been developed and published to accurately
94 automate lesion segmentation. As one of the early contributions to this field, we published the
95 Lesion Segmentation Toolbox (LST), which has since been applied in numerous scholarly
96 publications (Schmidt et al., 2012). While early segmentation algorithms have been designed
97 primarily using statistical and early machine learning models such as Support Vector Machines,
98 Gaussian Mixture Models or engineered by using manually selected features (Schmidt et al.,
99 2012), more recent approaches incorporate learning-based features via encoder/decoder model
100 stages (Cerri et al., 2021) or learn these end to end in fully convolutional models in (semi-)
101 supervised settings (Commowick et al., 2018). With the advent of artificial intelligence (AI),
102 automated lesion segmentation tools based on convolutional neural networks (CNN) have
103 become increasingly popular and indeed provide similar or higher segmentation accuracy than
104 earlier, machine learning-based methods (Diaz-Hurtado et al., 2022; H. Li et al., 2018; Ma et al.,
105 2022; Zeng et al., 2020). This is also reflected in the rankings of published MS lesion
106 segmentation challenges, e.g., MICCAI 2016 (Commowick et al., 2018) and ISBI 2015 (Carass
107 et al., 2017). While CNN-based models often outperform earlier models in challenges, they only
108 excel with a sufficient number of training data, as they are designed to learn priors and features
109 automatically and do not incorporate manual feature selection. Consequently, they are
110 especially prone to overfitting to the training data. Moreover, and in contrast to earlier machine
111 learning models, CNNs are comparatively harder to regularize, as they have higher model and
112 learning capacity, larger number of model parameters and thus more complex loss landscapes.
113 Therefore, a large performance gap between training set and test set is often noticeable and
114 highlights the need to evaluate the performance of CNN-based models on heterogeneous
115 external test data. Overcoming this gap and generalizing segmentation models in order to be
116 applicable to data from multiple protocols and centers is one of the main on-going challenges for
117 AI-based approaches. In this context, some AI-based approaches that have previously been

118 published are optimized towards transferability: Valverde et al. have provided *nicMSLesions*, a
119 CNN-based lesion segmentation method that is able to adjust to a new image domain by
120 retraining their model on a single image (Valverde et al., 2019). An important CNN-based
121 architecture is the UNet, which has been applied in many previous lesion segmentation studies
122 (Ashtari et al., 2022; Hashemi et al., 2022; Krishnan et al., 2023; La Rosa et al., 2020;
123 Ronneberger et al., 2015). Furthermore, recent studies successfully train their models on one
124 dataset and test it on another, external dataset, for which the MICCAI 2016 (Commowick et al.,
125 2021) and ISBI 2015 (Carass et al., 2017) datasets are often selected (Cerri et al., 2021; Gentile
126 et al., 2023; Kamraoui et al., 2022; Krishnan et al., 2023; X. Li et al., 2022; McKinley et al.,
127 2021). Hence, the research field is moving towards more generalized segmentation tools, which
128 is an important step towards clinical applicability of these methods.
129

130 In this study, we introduce a deep learning-based extension of LST. The main contributions can
131 be outlined in three aspects: 1) We provide an open-source lesion segmentation tool (with
132 network weights) that is easy to use and maintained; 2) The tool has been validated on external
133 datasets; 3) Lesion segmentation performance is comparable to or better than state of the art.
134 We carefully explain our selection of model architecture and describe the training and test set
135 used, and show how our composite loss function allows us to optimize our model for
136 generalizability on MRIs of unseen test centers. We also compare the performance of our model
137 against existing MS lesion segmentation algorithms. To facilitate studies and applications in MS
138 research, we provide this enhanced toolkit as open source to the imaging community
139 (<https://github.com/Complmg/LST-AI>).

140 2. Methods

141 2.1. Datasets

142 In the following section, we characterize and define training and test set, including details on
143 image acquisition. With regard to in-house data, we respected the Code of Ethics of the World
144 Medical Association (Declaration of Helsinki) for experiments involving humans; the study was
145 approved by the local ethics committee.
146

147 For the training set, we used an in-house dataset consisting of 491 paired 3D FLAIR and 3D
148 T1w images acquired on a 3.0T Achieva scanner (Philips Medical Systems, Best, The
149 Netherlands) to train both our proposed LST-AI segmentation model and the nnUNet baseline.
150 Testing and evaluation of segmentation performance of all methods was conducted on multiple
151 datasets. The test set includes four publicly available datasets: (i) msisbi: ISBI 2015 training
152 data (Carass et al., 2017) (<https://smart-stats-tools.org/lesion-challenge-2015>); (ii) msljub:
153 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
154 (Commowick et al., 2021) (<https://shanoir.irisa.fr/shanoir-ng/welcome>) and (iv) mssegtrain:
155 MICCAI 2016 challenge training dataset (Commowick et al., 2021)
156 (<https://shanoir.irisa.fr/shanoir-ng/welcome>). One case (msseg-test-center07-08) was removed
157 from the mssegtest dataset because it included incorrect ground truth data. In total, the test set
158

159 consists of 103 images from 87 subjects (note that the publicly available ISBI dataset is a
160 longitudinal dataset). Further characteristics of the datasets, including data on lesion load, are
161 provided in Table 1. Details on image acquisition are provided in Table 2.

162
163

dataset	#subjects	#scans	age (years) mean +/- sd	female / male	Diagnosis (number of images)	number of lesions i) mean +/- sd ii) median (IQR)	total lesion volume (mm ³) i) mean +/- sd ii) median (IQR)	publication	link
in-house training	491	491	34.3 +/- 9.5	330/161	RRMS (261) CIS (227) ON (3)	i) 25.54 +/- 30.59 ii) 15.0 (6.0-33.0)	i) 3492.96 +/- 7300.31 ii) 1244.0 (419.5-3767.5)	N/A	N/A
msisbi	5	21	43.5 +/- 10.3	4/1	RRMS (4) PPMS (1)	i) 45.95 +/- 20.92 ii) 41.0 (34.0-47.0)	i) 12889.76 +/- 11095.38 ii) 7354.0 (3678.0-18425.0)	(Carass et al., 2017)	(1)
msljob	30	30	39.0 +/- 25-64	23/7	RRMS (24) SPMS (2) PRMS (1) CIS (2) Unspecified (1)	i) 111.23 +/- 106.68 ii) 92.0 (31.25-125.0)	i) 17336.87 +/- 16115.41 ii) 14046.5 (1758.0-28430.25)	(Lesjak et al., 2018)	(2)
mssegtest	37	37	46.8 +/- 10.3	29/8	N/A	i) 44.89 +/- 42.11 ii) 29.0 (13.0-64.0)	i) 12672.73 +/- 15099.75 ii) 7348.0 (1453.0-17271.0)	(Commowick et al., 2021)	(3)
mssegtrain	15	15	41.6 +/- 9.8	8/7	N/A	i) 41.67 +/- 30.21 ii) 39.0 (18.0-56.5)	i) 20729.87 +/- 20606.48 ii) 12366.0 (3783.0-33198.5)	(Commowick et al., 2021)	(3)

164 Table 1

165 Characteristics of the datasets. One in-house (training) dataset was used, as well as the public datasets
166 msisbi from the ISBI 2015 challenge (Carass et al., 2017), msljob published by the Laboratory of Imaging
167 Technologies (Lesjak et al., 2018), and mssegtest and mssegtrain which are the testing and training
168 datasets from the MICCAI 2016 challenge, respectively (Commowick et al., 2021).

169 Abbreviations: CIS: clinically isolated syndrome, IQR: interquartile range, N/A: not applicable/available,
170 ON: optic neuritis, PPMS: primary progressive multiple sclerosis, RRMS: relapsing-remitting multiple
171 sclerosis, sd: standard deviation, SPMS: secondary progressive multiple sclerosis

172 (1) <https://smart-stats-tools.org/lesion-challenge-2015>

173 (2) <https://lit.fe.uni-lj.si/en/research/resources/3D-MR-MS/>

174 (3) <https://shanoir.irisa.fr/shanoir-ng/welcome>

175

176

dataset	scanner	field strength	sequence	voxel size	#scans
in-house training	Achieva, Philips Medical Systems	3.0T	T1w: TR=9ms, TE=4ms, FA=8	1x1x1mm ³	491
			FLAIR: TR=10000ms, TE=140ms, TI=2750ms	0.9x0.9x1.5mm ³	
msisbi	Philips Medical Systems	3.0T	T1w: TR=10.3ms, TE=6ms, FA=8	0.82x0.82x1.17mm ³	21
			FLAIR: TE=68ms, TI=835ms	0.82x0.82x2.2mm ³	
msljob	Siemens Magnetom Trio	3.0T	T1w: TR=2000ms, TE=20ms, TI=800ms, FA=120	0.42x0.42x3.3mm ³	30
			FLAIR: TR=5000ms, TE=392ms, TI=1800ms, FA=120	0.47x0.47x0.8mm ³	

mssegtest	Siemens Verio	3.0T	T1w: TR=1900ms, TE=2.26ms, FA=9	1x1x1mm ³	10
			FLAIR: TR=5000ms, TE=400ms, TI=1800ms, FA=120	0.5x0.5x1.1mm ³	
	General Electrics Discovery	3.0T	T1w: TR=[7.5,8]ms, TE=3.2ms, FA=10	0.47x0.47x0.6mm ³	8
			FLAIR: TR=9000ms, TE=[140,145]ms, TI=[2355, 2362]ms, FA=90	0.47x0.47x0.9mm ³	
	Siemens Aera	1.5T	T1w: TR=1860ms, TE=3.37ms, FA=15	1.08x1.08x0.9mm ³	9
			FLAIR:TR=5000ms, TE=336ms, TI=1800ms, FA=120	1.03x1.03x1.25mm ³	
Ingenia, Philips Medical Systems	3.0T	T1w: TR=9.4ms, TE=4.3ms, FA=8	0.74x0.74x0.85mm ³	10	
		FLAIR:TR=5400ms, TE=360ms, TI=1800ms, FA=90	0.74x0.74x0.7mm ³		
mssegtrain	Siemens Verio	3.0T	T1w: TR=1900ms, TE=2.26ms, FA=9	1x1x1mm ³	5
			FLAIR: TR=5000ms, TE=400ms, TI=1800ms, FA=120	0.5x0.5x1.1mm ³	
	Siemens Aera	1.5T	T1w: TR=1860ms, TE=3.37ms, FA=15	1.08x1.08x0.9mm ³	5
			FLAIR:TR=5000ms, TE=336ms, TI=1800ms, FA=120	1.03x1.03x1.25mm ³	
	Ingenia, Philips Medical Systems	3.0T	T1w: TR=9.4ms, TE=4.3ms, FA=8	0.74x0.74x0.85mm ³	5
			FLAIR:TR=5400ms, TE=360ms, TI=1800ms, FA=90	0.74x0.74x0.7mm ³	

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

181 2.2. Preprocessing

182 To guarantee fair comparisons across all baselines, we standardize preprocessing across all
 183 datasets and methods. Firstly, we register (rigid registration) all images to the MNI ICBM152
 184 nonlinear atlas version 2009 template ([https://www.mcgill.ca/bic/neuroinformatics/brain-atlases-](https://www.mcgill.ca/bic/neuroinformatics/brain-atlases-human)
 185 [human](https://www.mcgill.ca/bic/neuroinformatics/brain-atlases-human)) using the Greedy command line tool (P. Yushkevich, 2016/2023; P. A. Yushkevich et
 186 al., 2016). This atlas registration both ensures a consistent voxel resolution (1x1x1mm³) and
 187 image orientation, preprocessing steps well established for deep learning segmentation models
 188 (Kofler et al., 2020; Pati et al., 2022). Subsequently, we use the deep learning-based HD-BET
 189 brain extraction tool to generate skull-stripped images (Isensee et al., 2019). Next, the shape of
 190 the skull-stripped images is cropped to the size that is required for the 3D UNets and intensities
 191 are normalized to [0;1]. To benchmark methods in its intended environment, we opt for non-

192 skull-stripped images for SAMSEG, as well as the legacy algorithms of LST, the Lesion
193 Prediction Algorithm (LST-LPA) and the Lesion Growth Algorithm (LST-LGA), which perform
194 optimally with whole-brain data. Consequently, we omit the HD-BET skull-stripping, cropping,
195 and intensity normalization preprocessing steps for these specific baselines while retaining them
196 for others.

197
198 This standardized preprocessing (including skull-stripping) is also integrated into our LST-AI
199 toolbox, providing users with a streamlined approach.

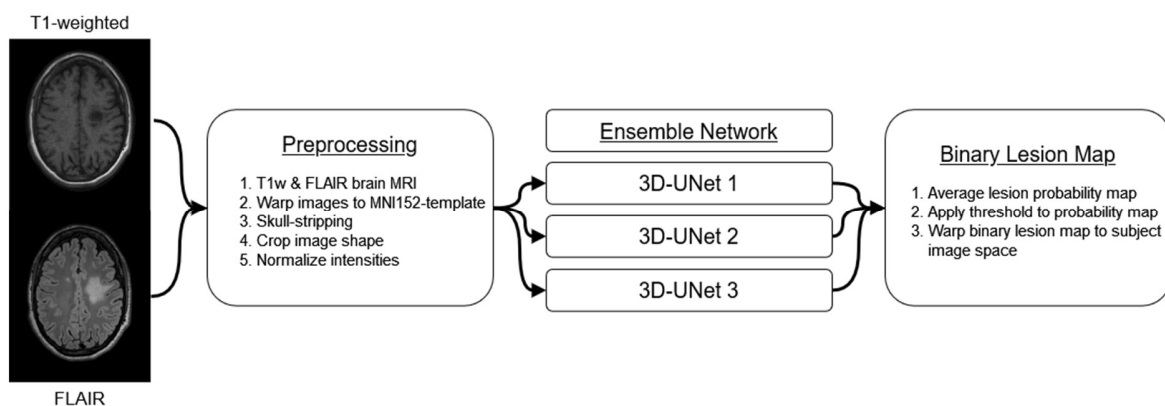
200 2.3. Lesion segmentation

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

205 2.3.1. LST-AI ensemble network

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

208



209

210 Figure 1

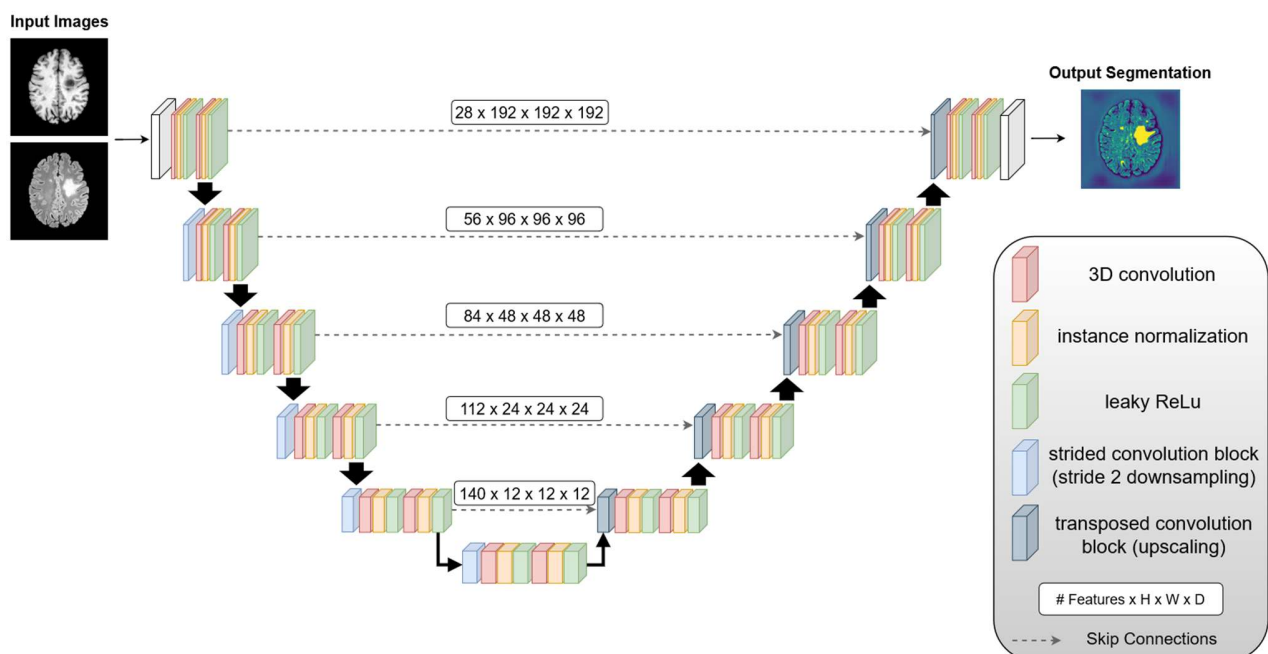
211 The different processing steps of the holistic LST-AI tool are presented. First, a pair of T1w and FLAIR
212 images is warped to MNI space, then skull-stripped, cropped, and intensity-normalized during
213 preprocessing. The resulting images are used as input for the three 3D-UNets of the ensemble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

217

218 The preprocessing functionality included in LST-AI is outlined in section 2.2. Specifically, the
219 T1w and FLAIR images are warped to the MNI ICBM152 template, then skull-stripped, center
220 cropped to shape (192, 192, 192), and, finally, intensities were normalized to [0;1].

221

222 With respect to the model architecture, LST-AI is based on an ensemble of three 3D-UNets.
223 Each UNet is built upon the 3D-UNet (Çiçek et al., 2016) architecture and inspired by nnUNet
224 (Isensee et al., 2021). It is composed of 5 encoder and 5 decoder blocks. Each of these blocks
225 is built from two convolution blocks (3D convolution, instance normalization, leaky ReLU
226 activation) and skip connections between respective encoder and decoder blocks (see Figure
227 2). In encoder blocks, downsampling is implemented via strided convolutions with stride 2, while
228 transposed convolutions are used for upscaling in decoder blocks. Following the architectural
229 choices in nnUNet (Isensee et al., 2021), we employ deep supervision layers in the training with
230 the intuition of allowing gradients to flow deeper into the networks' layers (Wang et al., 2015).
231 The number of deep supervision layers differed for the three UNets: one UNet included one
232 deep supervision layer and the two other UNets included two deep supervision layers to allow
233 for some variability in the ensemble predictions. For the loss function, we used a combination of
234 Tversky loss (Salehi et al., 2017) (with higher penalization of false-negative lesion omissions)
235 and binary cross-entropy in the deep supervision layers and a combined dice loss and binary
236 cross-entropy in the full-resolution output. During training, we randomly chained intensity
237 (random Gaussian noise, random Gaussian smoothing, random gamma adjustment) and
238 geometry augmentations (random flips and crops). Each model was trained for a total of 1000
239 epochs, using the stochastic gradient descent optimizer (with Nesterov momentum) and a
240 polynomial learning rate decay, starting at $1e-2$. This training scheme has been adapted from
241 nnUNet and was shown to generalize well in the medical segmentation decathlon (Antonelli et
242 al., 2022). In total, three training runs were started from scratch to create an ensemble of three
243 models, a technique previously reported (H. Li et al., 2018).
244
245



246
247 Figure 2
248 Architecture of the 3D-UNets which constitute the ensemble network of LST-AI. They comprise two

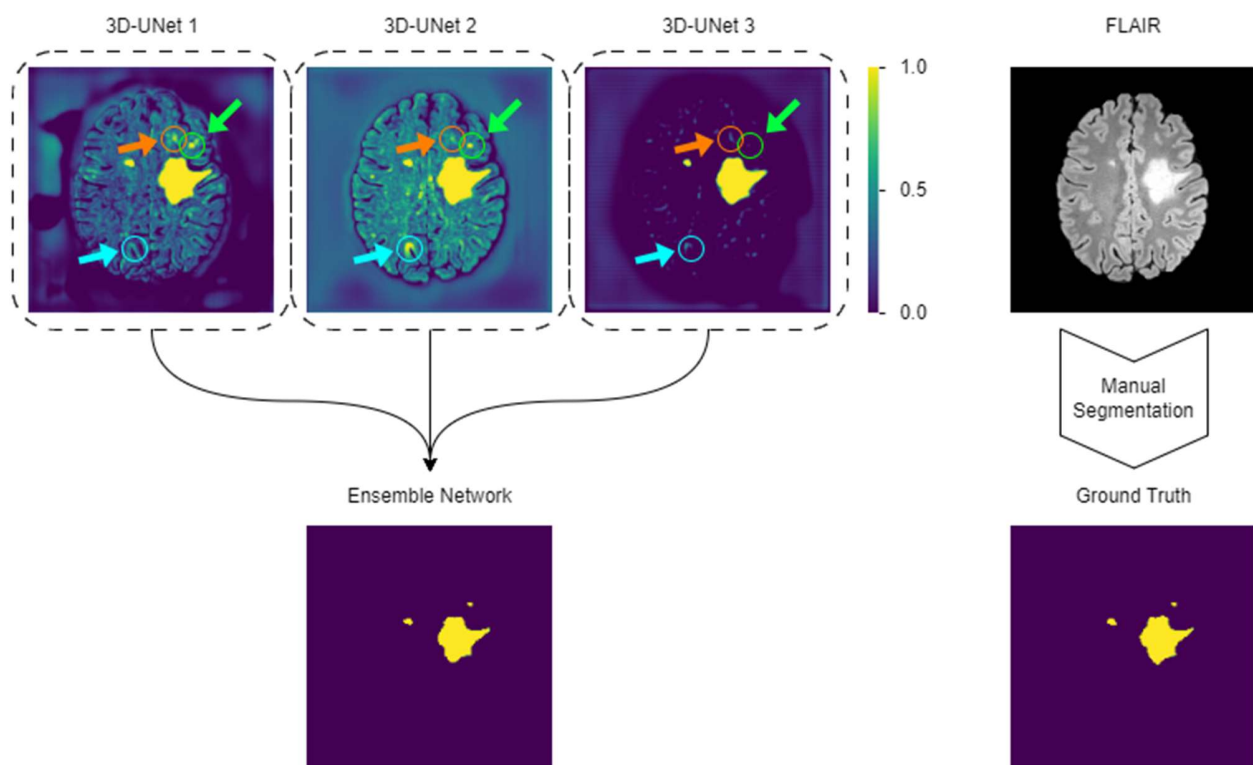
249 channels (one for T1w images and one for FLAIR images) and consist of 5 encoder and 5 decoder
250 blocks. Strided convolutions (stride 2) are used for downsampling and transposed convolutions are used
251 for upscaling. Encoder and decoder blocks are connected via skip connections.

252

253 For the final segmentation output, the preprocessed T1w and FLAIR images are used as input
254 for each one of the 3D UNets which generate three lesion probability maps. The final binary
255 lesion map is obtained by averaging the three lesion probability maps and subsequent
256 thresholding (default threshold of 0.5). This workflow, including the ground truth lesion
257 segmentation mask, is illustrated in Figure 3, using an example of the msljub dataset (subject
258 05).

259

260



261

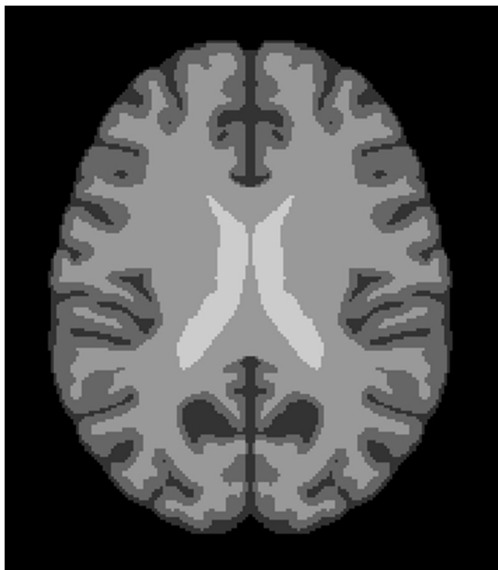
262 Figure 3

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

273

274 As an additional feature, the tool can optionally label lesions according to their location, i.e.,

275 periventricular (PV), juxtacortical (JC), subcortical (SC), or infratentorial (IT). To this end, the
276 same MNI ICBM152 nonlinear T1 atlas used above is first registered deformably (using Greedy)
277 to the skull-stripped T1w image in MNI space. The resulting transformation is applied to a
278 manually labeled anatomical mask indicating different brain regions (inter alia: ventricles for PV
279 labeling, infratentorial region for IT labeling, cortex for JC labeling, and subcortical region for SC
280 labeling), which is thereby registered to the skull-stripped T1w image in MNI space. The
281 anatomical mask is shown in Figure 4. Next, each individual lesion from the binary lesion
282 segmentation map is dilated using a cube as footprint ($3 \times 3 \times 3 \text{mm}^3$), and assigned to the region
283 with which it overlaps by at least one voxel (e.g., if a dilated lesion overlaps with the ventricles of
284 the anatomical mask it is labeled as PV). During this step, lesions are checked to overlap with
285 the four brain regions sequentially so that each lesion can be attributed to only one category.
286 The order of checks is PV, IT, JC, and, finally, SC. By this choice, large lesions overlapping with
287 the inner ventricles and the cortical ribbon are classified as PV (as PV lesions are commonly the
288 largest). In the resulting lesion map, the lesions are labeled according to their location (PV:
289 label=1, JC: label=2, SC: label=3, IT: label=4). Finally, the labeled lesion map is transformed to
290 the original space of the FLAIR image with the inverse of the affine transformation, which was
291 computed earlier, resulting in location-annotated lesion maps in the original subject space as
292 well as in the MNI space.
293



294
295 Figure 4
296 MS-specific anatomical mask indicating four different brain regions: ventricles outlined in light gray (used
297 to label lesions as periventricular), cortex outlined in dark gray (used to label lesions as juxtacortical),
298 subcortical region outlined in gray (used to label lesions as subcortical), or infratentorial region (not visible
299 in the image). Note that lesions are dilated using a $3 \times 3 \times 3 \text{mm}^3$ cube before overlaying with the anatomical
300 mask, which is how lesions can overlap with ventricle or cortex regions, resulting in lesions labeled as
301 periventricular or juxtacortical, respectively.
302

303 We intend to target a diverse user base and provide LST-AI as a set of standalone command
304 line tools and as a dockerized application, including all model checkpoints and required
305 preprocessing tools (Greedy and HD-BET). As LST-AI can be used in similar ways as

306 Freesurfer/FSL command line tools or `nicMSLesions` (docker), we give the opportunity to
307 conveniently integrate our tool into existing workflows.

308
309 For accelerated performance, we recommend using our tool in a GPU-enabled environment but
310 we also provide a fallback method for CPU-only usage. Depending on the exact hardware
311 setup, typical execution time varies between tens of seconds (GPU) and 1-2 minutes on a CPU-
312 only system. We provide LST-AI's functionality for three different workflows: segmentation-only,
313 lesion location annotation-only, or both. Moreover, labels can be exported in the original subject
314 space or in the MNI ICBM152 template space.

315
316 Moreover, we make our source code available, allowing the community to adapt and tailor our
317 tools for different application scenarios, by modifying preprocessing tools or using the
318 checkpoints for pre-training of custom models. We intend to continuously maintain and update
319 our tool in the github repository. In conclusion, while we have high confidence in the
320 generalization capabilities of LST-AI, we want to emphasize that it is explicitly designed for
321 research and non-clinical purposes. It has not undergone the necessary certification or licensing
322 for clinical applications.

323 2.3.2. Benchmark methods

324 Evaluation of the performance of the proposed tool is realized through comparison to other
325 publicly available lesion segmentation methods. This includes the widely used LST version 3.0.0
326 (<https://www.applied-statistics.de/lst.html>) with its lesion growth algorithm (LGA) (Schmidt et al.,
327 2012) and lesion prediction algorithm (LPA) (Vanderbecq et al., 2020), to which our proposed
328 tool presents a complementary, AI-based lesion segmentation method. Additionally, a trained
329 nnUNet and the recently published SAMSEG lesion segmentation tool implemented in
330 Freesurfer version 7.3.2 (Cerri et al., 2021) are used for comparison.

- 331 • **LST-LGA** (Schmidt et al., 2012): This method requires T1w and FLAIR images that are not
332 skull-stripped. Before applying the LST-LGA tool, T1w and FLAIR images are preprocessed
333 as described in section 2.2. Additionally, images are denoised using the CAT12 (Gaser et
334 al., 2022) denoising filter implemented in SPM12
335 (<https://www.fil.ion.ucl.ac.uk/spm/software/spm12/>). Then, the LST-LGA lesion
336 segmentation algorithm is applied. First, using the methods implemented in SPM12, bias
337 field correction is applied to the FLAIR image, and the T1w image is segmented into white
338 matter, grey matter, and cerebrospinal fluid. Based on the FLAIR intensities, lesion belief
339 maps are generated for each tissue class. The lesion belief map of grey matter is then
340 thresholded (default threshold of 0.3 as suggested in Schmidt et al., 2012), which results in
341 seeds that are used for the lesion growth model. Thereby, lesion seeds are expanded
342 according to FLAIR hyperintensities, eventually producing a lesion probability map. Finally,
343 a binary lesion map is generated after thresholding the lesion probability map (threshold of
344 0.5).
- 345 • **LST-LPA** (Vanderbecq et al., 2020): This method requires only FLAIR images that are not
346 skull-stripped. Preprocessing is identical to the LST-LGA workflow and includes registration

347 to MNI and denoising. Similarly, bias correction is applied, and a lesion belief map is
348 generated based on FLAIR intensities. The LST-LPA algorithm is a binary regression model
349 that combines the lesion belief map and fixed parameters, which had been learned through
350 logistic regression during the development of the tool in order to calculate the lesion
351 probability map. The binary lesion map is again generated by applying a threshold to the
352 lesion probability map (threshold of 0.5).

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

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

370
371 For region-specific analyses, all binary lesion maps are annotated with the method implemented
372 in the LST-AI tool. In effect, each lesion is labeled according to its location (i.e., PV, JC, IT, or
373 SC).

374 2.3.3. Manual segmentation

375 We make use of multiple datasets. Therefore, the workflows of manual segmentation, i.e.,
376 generation of ground truth lesion maps, differ. We describe the manual segmentation protocols
377 of the different datasets and refer to the corresponding publications:

378 • **in-house training:** The training data were first pre-segmented using LST-LGA. Segmented
379 lesions were manually reviewed and, based on FLAIR images, corrected by one out of four
380 experienced neuroradiologists using ITK-SNAP (P. A. Yushkevich et al., 2006). All lesion
381 masks were eventually reviewed by one senior neuroradiologist. The manual lesion
382 segmentation protocol is also described in another publication using the same dataset
383 (Hapfelmeier et al., 2023).

384 • **msisbi:** All images were manually delineated by two raters. Since no consensus was
385 available, we arbitrarily selected the lesion maps of one of the two raters as ground truth
386 (rater 2). Protocol details have been described in the original publication (Carass et al.,

387 2017).

- 388 • **msljob**: All images were delineated by three raters using a semi-automated approach. A
389 consensus segmentation was obtained through revision of the combined lesion maps by all
390 three raters; a detailed protocol is available in the original publication (Lesjak et al., 2018).
- 391 • **mssegtest & mssegtrain**: All images were manually delineated by seven raters, from
392 which a consensus was constructed. Details on the protocol and consensus construction
393 are available in the original publication (Commowick et al., 2021).

394 2.4. Evaluation

395 To assess the effectiveness of the LST-AI lesion segmentation tool, we compare its results with
396 manual segmentations and other available tools in multiple external datasets to evaluate the
397 performance and generalizability. These external sets encompass various acquisition protocols,
398 scanners, and originate from different centers. For consistency, we use images and lesion maps
399 in MNI space. Our evaluation covers lesion segmentation and detection methods, applying a
400 minimum lesion volume threshold of 3mm³ corresponding to 3 MNI-space voxels.

401 2.4.1. Lesion segmentation

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

$$408 \quad DSC = \frac{2TP}{2TP + FP + FN}, \quad (1)$$

409 the positive predictive value (PPV):

$$410 \quad PPV = \frac{TP}{TP + FP}, \quad (2)$$

411 and the sensitivity:

$$412 \quad sensitivity = \frac{TP}{TP + FN}, \quad (3)$$

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

$$415 \quad ASD = \frac{1}{n+n'} [\sum_{x=1}^n d(x, S') + \sum_{x'=1}^{n'} d(x', S)] \quad (4)$$

$$416 \quad \text{with } d(x, S') = \min ||x - x' ||_2, \quad (5)$$

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

420
421 These metrics are calculated for each image, then averaged within each dataset, and finally

422 averaged across all datasets. Thereby, we provide an overall score across different scanners
423 and centers as well as individual scores for each dataset.

424
425 As an additional step, we construct one array by concatenating all images and calculate the
426 DSC across all lesions of all datasets. We will refer to these analyses, neglecting subject-wise
427 information, as first-level analyses (and to those based on subject-wise performance measures
428 as second-level analyses). Thereby, we avoid the per-subject lesion load bias that is introduced
429 when one score is calculated per image. For example, missing a small lesion in an image with
430 only this missed lesion (DSC=0) would have more weight than missing a similar lesion in an
431 image with many other detected lesions (DSC>0).

432
433 We further investigate whether the performance of lesion segmentation varies across brain
434 regions to identify the drivers of the metric values and possible location-dependent variabilities
435 of LST-AI segmentation performance. To this end, we use the location-annotated lesion maps
436 and generate binary lesion maps for each region by only selecting lesion voxels labeled as part
437 of the corresponding region. Using the above evaluation metrics, first-level analysis is
438 conducted for each region and results from different regions and the whole brain are compared
439 to each other.

440 2.4.2. Lesion detection

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

$$450 \quad F1 = 2 * \frac{SensL * PPVL}{SensL + PPVL} = \frac{2TP}{2TP + FP + FN}, \quad (6)$$

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

452
453 The anima evaluation toolbox also offers the animaDetectedComponents tool that can be used
454 to investigate the detection of each lesion individually. For each image, the tool generates a list
455 with lesions that are present in the manually segmented lesion map. It indicates, for each lesion,
456 the volume in the manually segmented lesion map and whether it was detected by the
457 automated segmentation method. This enables the assessment of the increase or decrease of
458 lesion detection in relation to lesion volumes. Both tools (animaSegPerfAnalyzer and
459 animaDetectedComponents) consider a lesion in the manual segmentation as detected if it
460 overlaps with at least 10% with the lesion voxels in the automatically generated lesion map.

461 **3. Results**

462 We evaluate LST-AI in multiple aspects; we report both voxel-wise and lesion-wise scores, as
463 both volume and number are established measures of lesion load. We start with lesion
464 segmentation (3.1) across the whole brain and across subjects (second-level analyses). We
465 then report the performance across lesions (first-level analyses) both across brain regions (3.2)
466 and in relation to lesion size (3.3).

467 **3.1. Second-level lesion segmentation across the whole brain**

468 Lesion segmentation evaluation is conducted across all datasets as well as for each dataset
469 individually. An overview of the results of each segmentation method across all datasets is
470 provided in Table 3. A table with all anima metrics and results per case is included in the
471 supplementary material. In Figure 5, we present the lesion maps (of subject 05 of the msljub
472 dataset) of the different segmentation methods applied in this study.
473

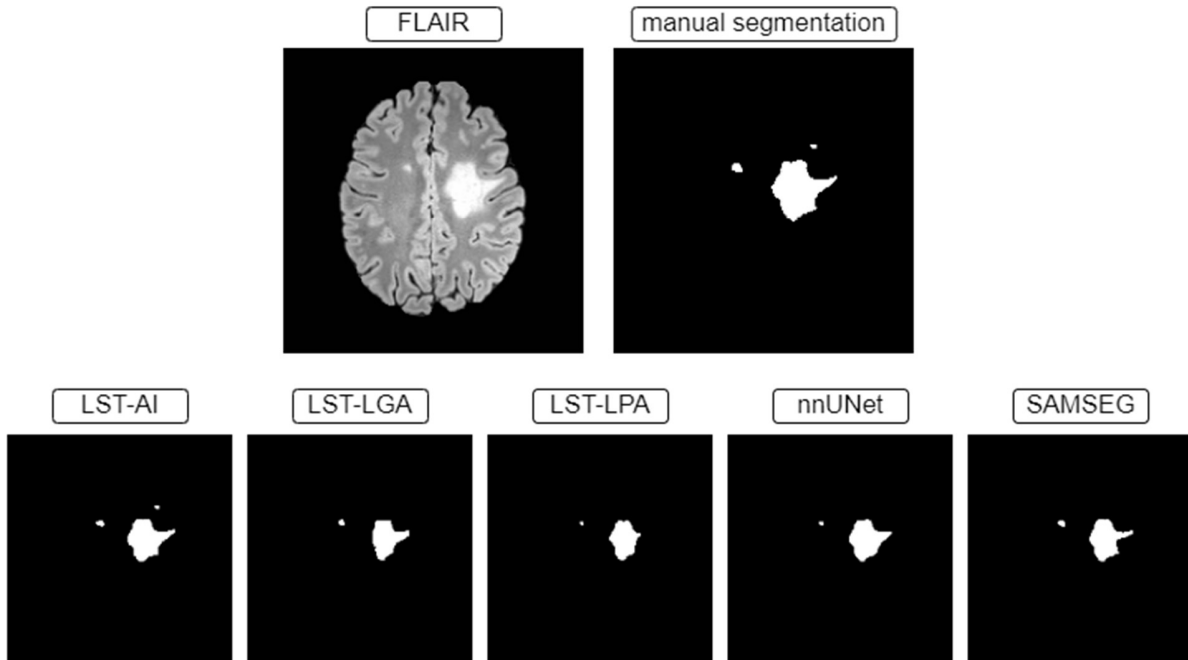
	voxel-wise				lesion-wise		
tool	DSC	PPV	sensitivity	ASD	F1	SensL	PPVL
LST-AI	0.67 +/- 0.14	0.73 +/- 0.15	0.66 +/- 0.17	0.37 +/- 1.12	0.63 +/- 0.15	0.70 +/- 0.19	0.64 +/- 0.20
LST-LGA	0.42 +/- 0.22	0.80 +/- 0.21	0.32 +/- 0.20	1.43 +/- 2.81	0.22 +/- 0.15	0.20 +/- 0.14	0.41 +/- 0.26
LST-LPA	0.44 +/- 0.22	0.79 +/- 0.15	0.34 +/- 0.20	1.35 +/- 2.21	0.23 +/- 0.14	0.25 +/- 0.15	0.34 +/- 0.23
nnUNet	0.51 +/- 0.20	0.90 +/- 0.07	0.38 +/- 0.18	1.36 +/- 4.18	0.46 +/- 0.19	0.40 +/- 0.21	0.64 +/- 0.21
SAMSEG	0.55 +/- 0.20	0.72 +/- 0.21	0.49 +/- 0.19	1.46 +/- 4.57	0.38 +/- 0.18	0.32 +/- 0.18	0.57 +/- 0.21

474
475 Table 3
476 The results of the lesion segmentation evaluation (second-level analysis across all test datasets) of each
477 segmentation tool are presented. The metrics were calculated for each image in the test datasets, and
478 values were subsequently averaged across all images. The averages are reported as mean +/- standard
479 deviation.

480 Abbreviations: ASD: average surface distance, DSC: dice similarity coefficient, PPV: positive predictive
481 value, PPVL: lesion-wise positive predictive value, SensL: lesion-wise sensitivity

482
483 The proposed method outperforms the benchmark methods in all categories except for PPV and
484 PPVL. LST-LGA, LST-LPA, and the nnUNet yield higher PPV values (PPV=0.79-0.90) than
485 LST-AI (PPV=0.73), and only the nnUNet yields a PPVL value as high as LST-AI (PPVL=0.64).
486 Notably, LST-AI achieves higher DSC and F1 scores (DSC=0.67 +/- 0.14; F1=0.63 +/- 0.15)
487 compared to the other methods (DSC=0.42-0.55; F1=0.22-0.46), indicating superior
488 segmentation performance both on a voxel-wise and on a lesion-wise level. The lowest ASD is
489 also obtained with LST-AI, indicating more accurate lesion contouring compared to the

490 benchmark methods. Overall the results show that LST-AI is able to identify more true lesions
 491 while increasing the fraction of correctly identified lesions among all segmented lesions
 492 compared to the benchmark methods.
 493



494
 495 **Figure 5**
 496 Binary lesion maps generated by the different lesion segmentation methods applied in this study. As
 497 reference, the first row shows the underlying FLAIR image as well as the manual segmentation (which is
 498 the ground truth). Each method provides slightly different lesion maps, and, in the slice presented here,
 499 only LST-AI detects all lesions present in the ground truth.

500
 501 Evaluating datasets individually (Table 4), we observe the most variability across datasets in
 502 ASD.
 503

dataset	voxel-wise				lesion-wise		
	DSC	PPV	sensitivity	ASD	F1	SensL	PPVL
All datasets n=103	0.67 +/- 0.14	0.73 +/- 0.15	0.66 +/- 0.17	0.37 +/- 1.12	0.63 +/- 0.15	0.70 +/- 0.19	0.64 +/- 0.20
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

504 Table 4

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

510 3.2. First-level segmentation across brain regions

511 In the PV region, LST-AI shows slightly higher first-level DSC scores than the other methods.
512 The difference in terms of first level DSC scores is more pronounced in the other three regions,
513 with only LST-AI reaching DSC >0.47 (other methods: DSC=0.03-0.31). Similarly, the highest
514 first-level DSC score within the whole brain is obtained with LST-AI. The results of the different
515 lesion segmentation methods are presented in Table 5 and Figure 6.

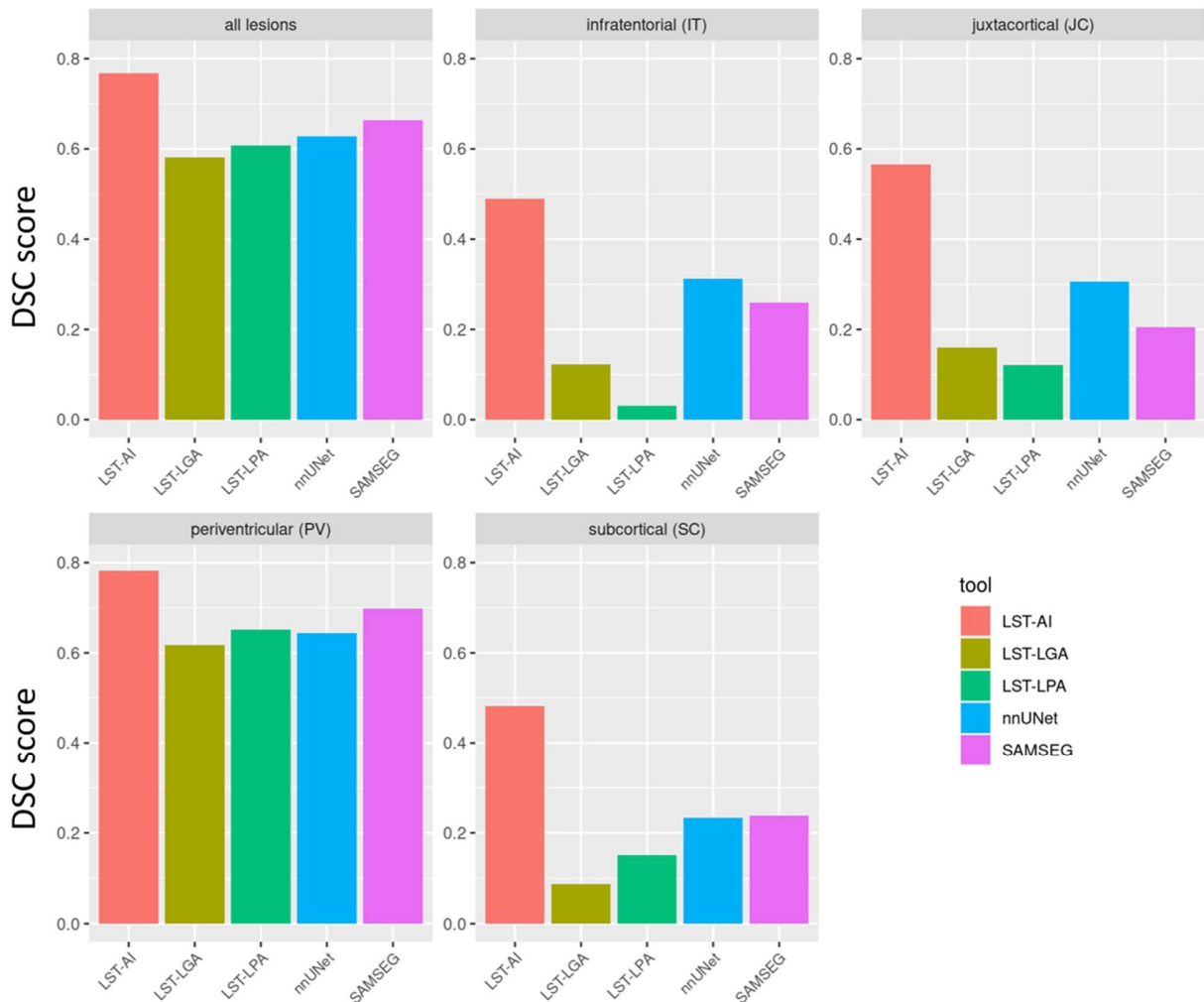
516
517

tool	Periventricular (PV)	Infratentorial (IT)	Juxtacortical (JC)	Subcortical (SC)	Whole brain
LST-AI	0.78	0.49	0.57	0.48	0.77
LST-LGA	0.62	0.12	0.16	0.09	0.58
LST-LPA	0.65	0.03	0.12	0.15	0.61
nnUNet	0.64	0.31	0.31	0.23	0.63
SAMSEG	0.70	0.26	0.21	0.24	0.66

518 Table 5

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

521
522



523
524
525
526
527

Figure 6
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.

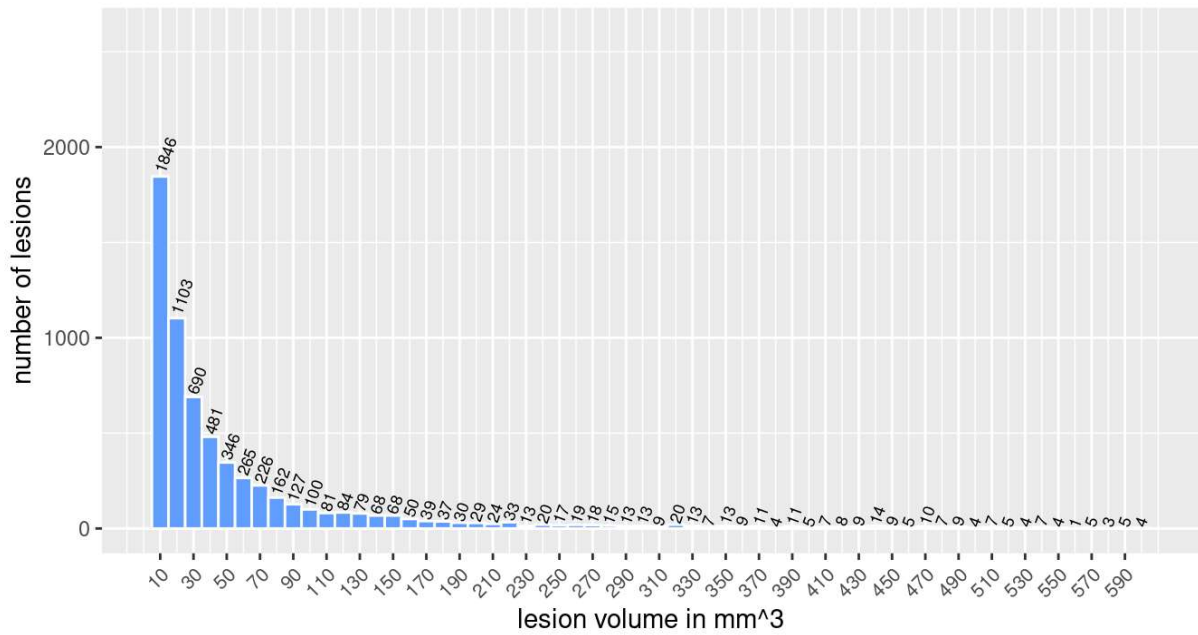
528

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

529
530
531
532
533
534
535
536
537
538
539

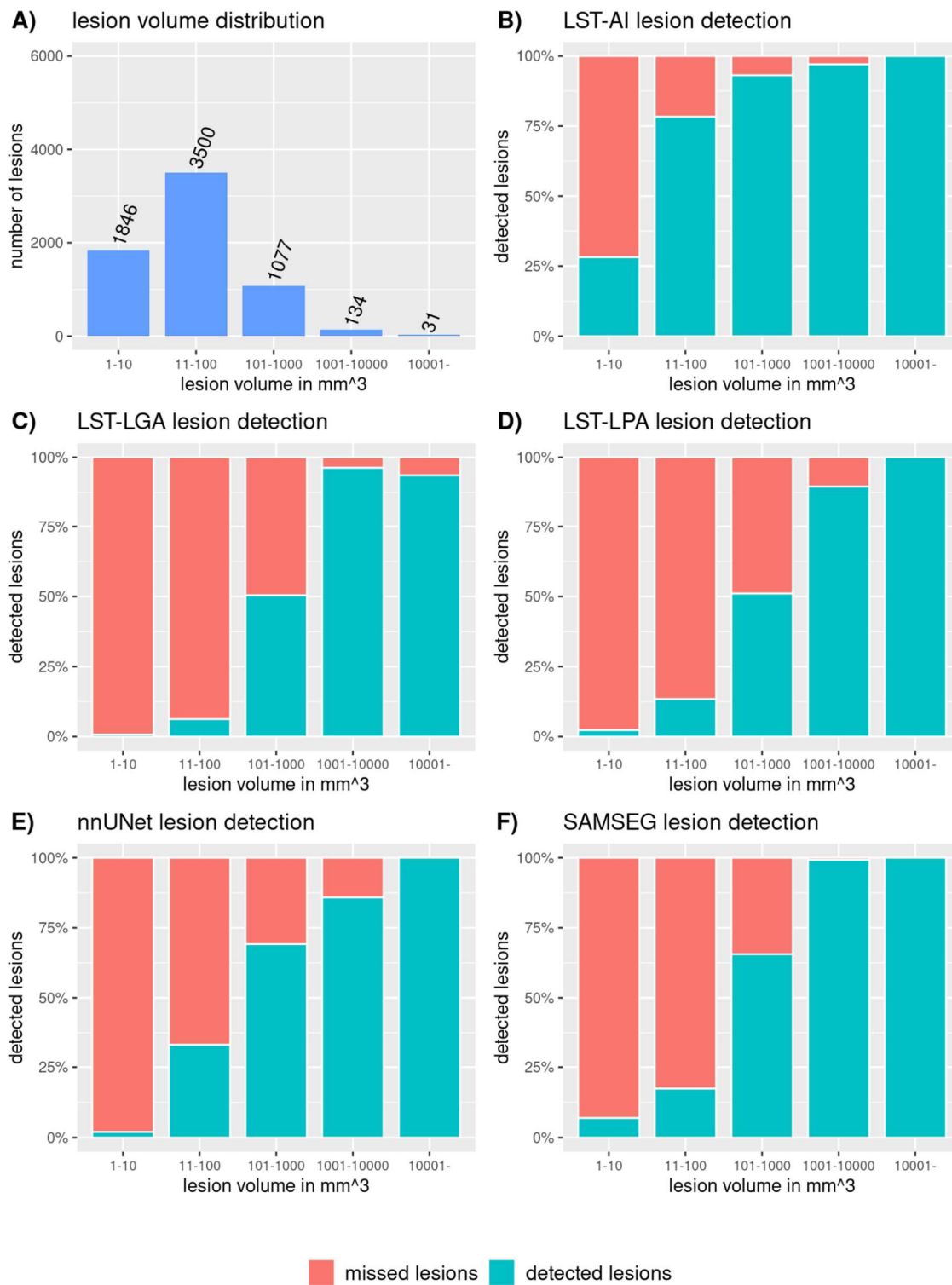
The lesion volume distribution of the test set is illustrated in Figure 7 (bin width of 10mm^3). The distribution shows a fast and steep decline with the most frequent lesions being small. This is critical as there is no commonly accepted minimum lesion volume (Grahl et al., 2019); moreover, accurate manual lesion segmentation is challenging, cumbersome, and sometimes overwhelming, even for expert readers. In Figure 8, we illustrate the accuracy of lesion detection in relation to lesion volume. For this, we divided lesions into groups according to their size: $1\text{-}10\text{ mm}^3$, $11\text{-}100\text{ mm}^3$, $101\text{-}1000\text{ mm}^3$, $1001\text{-}10000\text{ mm}^3$, and larger than 10000 mm^3 . Small lesions ($< 10\text{ mm}^3$) are detected worse. With increasing lesion size, 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 $> 10\text{mm}^3$ than suggested by mere SensL scores.

540
541



542
543
544
545
546
547
548

Figure 7
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 10mm³ and the first bar from the left shows the number of lesions with a volume between 1mm³ and 10mm³.



549
 550 Figure 8
 551 These graphs illustrate the proportion of lesions that are detected in each volume group. We divided the
 552 lesions into groups according to their volume (on the logarithmic scale): 1-10 mm³, 11-100 mm³, 101-
 553 1000 mm³, 1001-10000 mm³, and larger than 10000 mm³. A) shows the number of lesions

554 distribution across the volume groups; B) - F) show the lesion detection ratios of LST-AI, LST-LGA, LST-
555 LPA, nnUNet, and SAMSEG for the different lesion volumes. Note, how the detection rate increases with
556 increasing lesion volume for each segmentation, whereby LST-AI yields the highest detection rates. The
557 detection rate is given in %.

558 4. Discussion

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

567
568 LST-AI is pre-trained on an in-house dataset consisting of 491 images and does not need to be
569 retrained before it is applied to new data. This makes it possible to use the tool even in smaller
570 centers, where data is scarce and only small cohorts are available. Valverde et al., 2019, have
571 previously optimized retraining on small datasets, as their tool only requires a single case to
572 adapt their model to new datasets. They also validated their method on the ISBI 2015 test
573 dataset and achieved a mean DSC of 0.58 (Valverde et al., 2019). In general, high-performing
574 segmentation models in the ISBI 2015 challenge were CNN-based (trained on ISBI 2015
575 training dataset) and reported DSC scores ranging between 0.50 and 0.68 (Ma et al., 2022;
576 Zhang & Oguz, 2021). However, assessing generalizability of segmentation models requires
577 validation on external datasets. This has been done in recent studies, which used different train
578 and test set pairings, including in-house and publicly available data such as ISBI 2015 and
579 MICCAI 2016 data (e.g., train on in-house data and test on MICCAI 2016 data) (Billot et al.,
580 2021; Cerri et al., 2021; Gentile et al., 2023; Kamraoui et al., 2022; X. Li et al., 2022; McKinley
581 et al., 2021; Rakić et al., 2021). Overall, using train and test sets from different image domains
582 led to lower and more variable DSC scores. For example, in the study by Kamraoui et al.
583 (2022), the segmentation performance on the ISBI 2015 test dataset drops when models are
584 trained on in-house data (DSC=0.13-0.48) compared to when they are trained on the ISBI
585 training dataset (DSC=0.64-0.67). On the MICCAI 2016 dataset, however, the models trained
586 on the in-house training dataset showed robust and high DSC scores (0.65-0.72) (Kamraoui et
587 al., 2022). This highlights the impact of differing image domains in train and test sets and the
588 need for validation on multiple test datasets, which can provide a more realistic representation
589 of a model's generalizability. In this study, image domain heterogeneity is simulated by the
590 validation of our method on multiple datasets, which were also part of MS lesion segmentation
591 challenges of the ISBI 2015 conference and the MICCAI 2016 conference (Carass et al., 2017;
592 Commowick et al., 2018, 2021). While our model achieves similar scores (mean DSC of 0.61
593 and 0.65 for ISBI 2015 and MICCAI 2016, respectively) as the top-performing models in both
594 challenges, we want to emphasize that, in contrast to the participating models, our model is not
595 specifically trained on the corresponding training datasets provided in the challenges. These two

596 scores are also close to the inter-rater DSC scores of the expert segmentation used in the
597 challenges (DSC of 0.63 and 0.66-0.76 in ISBI 2015 and MICCAI 2016, respectively) (Carass et
598 al., 2017; Commowick et al., 2021). Other studies investigating the generalizability of their
599 model on external data reported similar DSC scores in the range of 0.48 - 0.72 (Cerri et al.,
600 2021; Kamraoui et al., 2022; McKinley et al., 2021; Rakić et al., 2021). Regarding LST-AI, the
601 DSC scores for the three external datasets (range: 0.61-0.74) underline the good generalization
602 of our model and its reliable application to multicenter data acquired with different scanners and
603 protocols. Overall, results from both second- and first-level analysis show high segmentation
604 performance of LST-AI on unseen data. In contrast, the lower performance of the other
605 methods, e.g., the pre-trained nnUNet, suggests the need for adaptation of these methods
606 through retraining. We believe that using an ensemble approach including multiple pre-trained
607 UNets translates into robustness against performance variability of individual 3D UNets and,
608 therefore, generalizes better across different imaging protocols and centers. Of note, the mean
609 PPV and PPVL values of the benchmark methods are comparable to those of LST-AI. However,
610 this appears to happen at the expense of sensitivity, where LST-AI clearly outperforms the other
611 methods at the voxel and lesion level. Compared to the literature, lesion-wise sensitivity of LST-
612 AI on MICCAI 2016 data (SensL=0.83) and ISBI 2015 data (SensL=0.55) is in the same range
613 as previously reported values (Carass et al., 2017; Commowick et al., 2018; Kamraoui et al.,
614 2022; Krishnan et al., 2023; Ma et al., 2022; Zhang & Oguz, 2021). With regard to clinical
615 applicability of automated lesion segmentation tools, the sensitivity is crucial as diagnosing and
616 monitoring MS relies on the detection of (new) lesions. A newly published method, namely
617 BIANCA-MS (Gentile et al., 2023), has also been validated using the MICCAI 2016 test dataset
618 and yielded results similar to ours in terms of DSC and false positives (in terms of lesion
619 detection). However, the median number of false negatives was equal to 11(IQR: 18) for
620 BIANCA-MS, whereas LST-AI yields a median number of false negatives equal to 4 (IQR: 8),
621 again highlighting the high sensitivity of our proposed method towards lesion detection.

622
623 In MS, lesion location within the brain may play an important role in identifying different disease
624 patterns (Pongratz et al., 2023). In the LST-AI toolbox, a method is included which is able to
625 classify lesions into four categories according to their location (PV, IT, JC, and SC). This makes
626 it possible to seamlessly analyze the lesion load in different brain regions relevant to MS. When
627 looking at the segmentation performance in the four different brain regions, it stands out that,
628 among all methods included in this publication, LST-AI shows the highest (first-level) DSC score
629 in all regions. The increased lesion segmentation performance in the JC region is a particularly
630 relevant finding, since segmentation of lesions close to the cortex based on T1w and FLAIR
631 images has always been a challenge in MS. Also, juxtacortical lesions are thought to be very
632 specific for MS and are strongly associated with clinical disability (Calabrese et al., 2012),
633 making their detection very important.

634
635 We also investigated the lesion detection in relation to lesion volume and we found that LST-AI
636 has a higher lesion detection sensitivity for small lesions than the benchmark methods. Similar
637 to previous reports by Commowick et al. (2018) and Rakić et al. (2021), we also found that it is
638 particularly hard to detect small lesions (<10mm³). Nonetheless, the steep incline of lesion
639 detection with lesion size provides a promising perspective for the integration of automated

640 lesion segmentation tools in clinical settings, since it can help clinicians to detect lesions faster
641 and to diagnose and monitor MS more accurately.

642
643 Our study does not come without limitations. First, our model requires T1w and FLAIR image
644 pairs, which might not always be available. Second, although less pronounced than in the
645 benchmark methods, our model still shows a decrease in lesion detection efficiency with
646 decreasing lesion volumes. Even though the explainability of features learned via CNNs and
647 more specifically U-Nets have been comparatively well studied, they still lack some
648 interpretability in contrast to methods leveraging manually selected features. In addition,
649 preprocessing is included in the LST-AI toolbox and includes registration to MNI space, which
650 ensures identical image dimensions and orientation before segmenting lesions. However,
651 preprocessing steps are known to be crucial in segmentation tasks. Hence, exploring and
652 applying different preprocessing steps could possibly change the performance on some
653 datasets.

654
655 In conclusion, we introduce LST-AI, a new lesion segmentation toolbox and make it publicly
656 available on GitHub (<https://github.com/Complmg/LST-AI>). It includes a preprocessing pipeline
657 as well as an ensemble of three 3D UNets with binary cross-entropy and Tversky loss, making it
658 a holistic lesion segmentation tool, enabling easy-to-implement, quick, and accurate automated
659 lesion segmentation for MS research without retraining and fine-tuning. We validated its
660 robustness on multiple datasets and found excellent performance. We believe that, in future
661 studies, LST-AI should replace LST.

662 References

- 663
664 Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., Litjens,
665 G., Menze, B., Ronneberger, O., Summers, R. M., van Ginneken, B., Bilello, M., Bilic, P.,
666 Christ, P. F., Do, R. K. G., Gollub, M. J., Heckers, S. H., Huisman, H., Jarnagin, W. R.,
667 ... Cardoso, M. J. (2022). The Medical Segmentation Decathlon. *Nature*
668 *Communications*, 13(1), 4128. <https://doi.org/10.1038/s41467-022-30695-9>
669 Ashtari, P., Barile, B., Van Huffel, S., & Sappey-Marini er, D. (2022). New multiple sclerosis
670 lesion segmentation and detection using pre-activation U-Net. *Frontiers in Neuroscience*,
671 16. <https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2022.975862>
672 Billot, B., Cerri, S., Leemput, K. V., Dalca, A. V., & Iglesias, J. E. (2021). Joint Segmentation Of
673 Multiple Sclerosis Lesions And Brain Anatomy In MRI Scans Of Any Contrast And

- 674 Resolution With CNNs. *2021 IEEE 18th International Symposium on Biomedical Imaging*
675 *(ISBI)*, 1971–1974. <https://doi.org/10.1109/ISBI48211.2021.9434127>
- 676 Calabrese, M., Poretto, V., Favaretto, A., Alessio, S., Bernardi, V., Romualdi, C., Rinaldi, F.,
677 Perini, P., & Gallo, P. (2012). Cortical lesion load associates with progression of
678 disability in multiple sclerosis. *Brain: A Journal of Neurology*, *135*(Pt 10), 2952–2961.
679 <https://doi.org/10.1093/brain/aws246>
- 680 Carass, A., Roy, S., Jog, A., Cuzzocreo, J. L., Magrath, E., Gherman, A., Button, J., Nguyen, J.,
681 Prados, F., Sudre, C. H., Jorge Cardoso, M., Cawley, N., Ciccarelli, O., Wheeler-
682 Kingshott, C. A. M., Ourselin, S., Catanese, L., Deshpande, H., Maurel, P., Commowick,
683 O., ... Pham, D. L. (2017). Longitudinal multiple sclerosis lesion segmentation: Resource
684 and challenge. *Neuroimage*, *148*, 77–102.
685 <https://doi.org/10.1016/j.neuroimage.2016.12.064>
- 686 Cerri, S., Puonti, O., Meier, D. S., Wuerfel, J., Mühlau, M., Siebner, H. R., & Van Leemput, K.
687 (2021). A contrast-adaptive method for simultaneous whole-brain and lesion
688 segmentation in multiple sclerosis. *NeuroImage*, *225*, 117471.
689 <https://doi.org/10.1016/j.neuroimage.2020.117471>
- 690 Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net:
691 Learning Dense Volumetric Segmentation from Sparse Annotation. In S. Ourselin, L.
692 Joskowicz, M. R. Sabuncu, G. Unal, & W. Wells (Eds.), *Medical Image Computing and*
693 *Computer-Assisted Intervention – MICCAI 2016* (pp. 424–432). Springer International
694 Publishing. https://doi.org/10.1007/978-3-319-46723-8_49
- 695 Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., Pop, S. C., Girard, P.,
696 Ameli, R., Ferre, J. C., Kerbrat, A., Tourdias, T., Cervenansky, F., Glatard, T.,
697 Beaumont, J., Doyle, S., Forbes, F., Knight, J., Khademi, A., ... Barillot, C. (2018).
698 Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data
699 Management and Processing Infrastructure. *Sci Rep*, *8*(1), 13650.

- 700 <https://doi.org/10.1038/s41598-018-31911-7>
- 701 Commowick, O., Kain, M., Casey, R., Ameli, R., Ferré, J.-C., Kerbrat, A., Tourdias, T.,
702 Cervenansky, F., Camarasu-Pop, S., Glatard, T., Vukusic, S., Edan, G., Barillot, C.,
703 Dojat, M., & Cotton, F. (2021). Multiple sclerosis lesions segmentation from multiple
704 experts: The MICCAI 2016 challenge dataset. *NeuroImage*, *244*, 118589.
705 <https://doi.org/10.1016/j.neuroimage.2021.118589>
- 706 Diaz-Hurtado, M., Martínez-Heras, E., Solana, E., Casas-Roma, J., Llufrú, S., Kanber, B., &
707 Prados, F. (2022). Recent advances in the longitudinal segmentation of multiple
708 sclerosis lesions on magnetic resonance imaging: A review. *Neuroradiology*, *64*(11),
709 2103–2117. <https://doi.org/10.1007/s00234-022-03019-3>
- 710 Filippi, M., Bar-Or, A., Piehl, F., Preziosa, P., Solari, A., Vukusic, S., & Rocca, M. A. (2018).
711 Multiple sclerosis. *Nature Reviews Disease Primers*, *4*(1), Article 1.
712 <https://doi.org/10.1038/s41572-018-0041-4>
- 713 Gaser, C., Dahnke, R., Thompson, P. M., Kurth, F., Luders, E., & Initiative, A. D. N. (2022). *CAT*
714 – *A Computational Anatomy Toolbox for the Analysis of Structural MRI Data* (p.
715 2022.06.11.495736). bioRxiv. <https://doi.org/10.1101/2022.06.11.495736>
- 716 Gentile, G., Jenkinson, M., Griffanti, L., Luchetti, L., Leoncini, M., Inderyas, M., Mortilla, M.,
717 Cortese, R., De Stefano, N., & Battaglini, M. (2023). BIANCA-MS: An optimized tool for
718 automated multiple sclerosis lesion segmentation. *Human Brain Mapping*.
719 <https://doi.org/10.1002/hbm.26424>
- 720 Grahl, S., Pongratz, V., Schmidt, P., Engl, C., Bussas, M., Radetz, A., Gonzalez-Escamilla, G.,
721 Groppa, S., Zipp, F., Lukas, C., Kirschke, J., Zimmer, C., Hoshi, M., Berthele, A.,
722 Hemmer, B., & Mühlau, M. (2019). Evidence for a white matter lesion size threshold to
723 support the diagnosis of relapsing remitting multiple sclerosis. *Multiple Sclerosis and*
724 *Related Disorders*, *29*, 124–129. <https://doi.org/10.1016/j.msard.2019.01.042>
- 725 Hapfelmeier, A., On, B. I., Mühlau, M., Kirschke, J. S., Berthele, A., Gasperi, C., Mansmann, U.,

- 726 Wuschek, A., Bussas, M., Boeker, M., Bayas, A., Senel, M., Havla, J., Kowarik, M. C.,
727 Kuhn, K., Gatz, I., Spengler, H., Wiestler, B., Grundl, L., ... Hemmer, B. (2023).
728 Retrospective cohort study to devise a treatment decision score predicting adverse 24-
729 month radiological activity in early multiple sclerosis. *Therapeutic Advances in*
730 *Neurological Disorders*, 16, 17562864231161892.
731 <https://doi.org/10.1177/17562864231161892>
- 732 Hashemi, M., Akhbari, M., & Jutten, C. (2022). Delve into Multiple Sclerosis (MS) lesion
733 exploration: A modified attention U-Net for MS lesion segmentation in Brain MRI.
734 *Computers in Biology and Medicine*, 145, 105402.
735 <https://doi.org/10.1016/j.combiomed.2022.105402>
- 736 <https://anima.irisa.fr/>. (n.d.). ANIMA. Retrieved August 8, 2023, from <https://anima.irisa.fr/>
737 <https://www.applied-statistics.de/lst.html>. (n.d.). LST – Lesion Segmentation for SPM. Retrieved
738 July 28, 2023, from <https://www.applied-statistics.de/lst.html>
- 739 <https://www.fil.ion.ucl.ac.uk/spm/software/spm12/>. (n.d.). SPM12 Software - Statistical
740 Parametric Mapping. Retrieved July 28, 2023, from
741 <https://www.fil.ion.ucl.ac.uk/spm/software/spm12/>
- 742 Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A
743 self-configuring method for deep learning-based biomedical image segmentation. *Nature*
744 *Methods*, 18(2), Article 2. <https://doi.org/10.1038/s41592-020-01008-z>
- 745 Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A.,
746 Schlemmer, H.-P., Heiland, S., Wick, W., Bendszus, M., Maier-Hein, K. H., &
747 Kickingreder, P. (2019). Automated brain extraction of multisequence MRI using
748 artificial neural networks. *Human Brain Mapping*, 40(17), 4952–4964.
749 <https://doi.org/10.1002/hbm.24750>
- 750 Kamraoui, R. A., Ta, V.-T., Tourdias, T., Mansencal, B., Manjon, J. V., & Coup, P. (2022).
751 DeepLesionBrain: Towards a broader deep-learning generalization for multiple sclerosis

- 752 lesion segmentation. *Medical Image Analysis*, 76, 102312.
753 <https://doi.org/10.1016/j.media.2021.102312>
- 754 Kofler, F., Berger, C., Waldmannstetter, D., Lipkova, J., Ezhov, I., Tetteh, G., Kirschke, J.,
755 Zimmer, C., Wiestler, B., & Menze, B. H. (2020). BraTS Toolkit: Translating BraTS Brain
756 Tumor Segmentation Algorithms Into Clinical and Scientific Practice. *Frontiers in*
757 *Neuroscience*, 14, 125. <https://doi.org/10.3389/fnins.2020.00125>
- 758 Krishnan, A. P., Song, Z., Clayton, D., Jia, X., de Crespigny, A., & Carano, R. A. D. (2023).
759 Multi-arm U-Net with dense input and skip connectivity for T2 lesion segmentation in
760 clinical trials of multiple sclerosis. *Scientific Reports*, 13(1), Article 1.
761 <https://doi.org/10.1038/s41598-023-31207-5>
- 762 La Rosa, F., Abdulkadir, A., Fartaria, M. J., Rahmanzadeh, R., Lu, P.-J., Galbusera, R.,
763 Barakovic, M., Thiran, J.-P., Granziera, C., & Cuadra, M. B. (2020). Multiple sclerosis
764 cortical and WM lesion segmentation at 3T MRI: A deep learning method based on
765 FLAIR and MP2RAGE. *NeuroImage: Clinical*, 27, 102335.
766 <https://doi.org/10.1016/j.nicl.2020.102335>
- 767 Lesjak, Z., Galimzianova, A., Koren, A., Lukin, M., Pernus, F., Likar, B., & Spiclin, Z. (2018). A
768 Novel Public MR Image Dataset of Multiple Sclerosis Patients With Lesion
769 Segmentations Based on Multi-rater Consensus. *Neuroinformatics*, 16(1), 51–63.
770 <https://doi.org/10.1007/s12021-017-9348-7>
- 771 Li, H., Jiang, G., Zhang, J., Wang, R., Wang, Z., Zheng, W.-S., & Menze, B. (2018). Fully
772 convolutional network ensembles for white matter hyperintensities segmentation in MR
773 images. *NeuroImage*, 183, 650–665. <https://doi.org/10.1016/j.neuroimage.2018.07.005>
- 774 Li, X., Zhao, Y., Jiang, J., Cheng, J., Zhu, W., Wu, Z., Jing, J., Zhang, Z., Wen, W., Sachdev, P.
775 S., Wang, Y., Liu, T., & Li, Z. (2022). White matter hyperintensities segmentation using
776 an ensemble of neural networks. *Human Brain Mapping*, 43(3), 929–939.
777 <https://doi.org/10.1002/hbm.25695>

- 778 Ma, Y., Zhang, C., Cabezas, M., Song, Y., Tang, Z., Liu, D., Cai, W., Barnett, M., & Wang, C.
779 (2022). Multiple Sclerosis Lesion Analysis in Brain Magnetic Resonance Images:
780 Techniques and Clinical Applications. *IEEE Journal of Biomedical and Health*
781 *Informatics*, 26(6), 2680–2692. <https://doi.org/10.1109/JBHI.2022.3151741>
- 782 McKinley, R., Wepfer, R., Aschwanden, F., Grunder, L., Muri, R., Rummel, C., Verma, R.,
783 Weisstanner, C., Reyes, M., Salmen, A., Chan, A., Wagner, F., & Wiest, R. (2021).
784 Simultaneous lesion and brain segmentation in multiple sclerosis using deep neural
785 networks. *Sci Rep*, 11(1), 1087. <https://doi.org/10.1038/s41598-020-79925-4>
- 786 Pati, S., Baid, U., Edwards, B., Sheller, M., Wang, S.-H., Reina, G. A., Foley, P., Gruzdev, A.,
787 Karkada, D., Davatzikos, C., Sako, C., Ghodasara, S., Bilello, M., Mohan, S., Vollmuth,
788 P., Brugnara, G., Preetha, C. J., Sahm, F., Maier-Hein, K., ... Bakas, S. (2022).
789 Federated learning enables big data for rare cancer boundary detection. *Nature*
790 *Communications*, 13(1), 7346. <https://doi.org/10.1038/s41467-022-33407-5>
- 791 Pongratz, V., Bussas, M., Schmidt, P., Grahl, S., Gasperi, C., El Hussein, M., Harabacz, L.,
792 Pineker, V., Sepp, D., Grundl, L., Wiestler, B., Kirschke, J., Zimmer, C., Berthele, A.,
793 Hemmer, B., & Mühlau, M. (2023). Lesion location across diagnostic regions in multiple
794 sclerosis. *NeuroImage: Clinical*, 37, 103311. <https://doi.org/10.1016/j.nicl.2022.103311>
- 795 Rakić, M., Vercruyssen, S., Van Eyndhoven, S., de la Rosa, E., Jain, S., Van Huffel, S., Maes,
796 F., Smeets, D., & Sima, D. M. (2021). icobrain ms 5.1: Combining unsupervised and
797 supervised approaches for improving the detection of multiple sclerosis lesions.
798 *NeuroImage: Clinical*, 31, 102707. <https://doi.org/10.1016/j.nicl.2021.102707>
- 799 Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical
800 Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.),
801 *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp.
802 234–241). Springer International Publishing. [https://doi.org/10.1007/978-3-319-24574-](https://doi.org/10.1007/978-3-319-24574-4_28)
803 [4_28](https://doi.org/10.1007/978-3-319-24574-4_28)

- 804 Salehi, S. S. M., Erdogmus, D., & Gholipour, A. (2017). *Tversky loss function for image*
805 *segmentation using 3D fully convolutional deep networks* (arXiv:1706.05721). arXiv.
806 <https://doi.org/10.48550/arXiv.1706.05721>
- 807 Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förchler, A., Berthele, A., Hoshi, M., Ilg, R.,
808 Schmid, V. J., Zimmer, C., Hemmer, B., & Mühlau, M. (2012). An automated tool for
809 detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. *NeuroImage*,
810 *59*(4), 3774–3783. <https://doi.org/10.1016/j.neuroimage.2011.11.032>
- 811 Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., Correale, J.,
812 Fazekas, F., Filippi, M., Freedman, M. S., Fujihara, K., Galetta, S. L., Hartung, H. P.,
813 Kappos, L., Lublin, F. D., Marrie, R. A., Miller, A. E., Miller, D. H., Montalban, X., ...
814 Cohen, J. A. (2018). Diagnosis of multiple sclerosis: 2017 revisions of the McDonald
815 criteria. *The Lancet Neurology*, *17*(2), 162–173. [https://doi.org/10.1016/S1474-](https://doi.org/10.1016/S1474-4422(17)30470-2)
816 [4422\(17\)30470-2](https://doi.org/10.1016/S1474-4422(17)30470-2)
- 817 Thompson, A. J., Baranzini, S. E., Geurts, J., Hemmer, B., & Ciccarelli, O. (2018). Multiple
818 sclerosis. *Lancet (London, England)*, *391*(10130), 1622–1636.
819 [https://doi.org/10.1016/S0140-6736\(18\)30481-1](https://doi.org/10.1016/S0140-6736(18)30481-1)
- 820 Valverde, S., Salem, M., Cabezas, M., Pareto, D., Vilanova, J. C., Ramió-Torrentà, L., Rovira,
821 À., Salvi, J., Oliver, A., & Lladó, X. (2019). One-shot domain adaptation in multiple
822 sclerosis lesion segmentation using convolutional neural networks. *NeuroImage:*
823 *Clinical*, *21*, 101638. <https://doi.org/10.1016/j.nicl.2018.101638>
- 824 Vanderbecq, Q., Xu, E., Stroer, S., Couvy-Duchesne, B., Diaz Melo, M., Dormont, D., Colliot,
825 O., & Alzheimer's Disease Neuroimaging, I. (2020). Comparison and validation of seven
826 white matter hyperintensities segmentation software in elderly patients. *Neuroimage*
827 *Clin*, *27*, 102357. <https://doi.org/10.1016/j.nicl.2020.102357>
- 828 Wang, L., Lee, C.-Y., Tu, Z., & Lazebnik, S. (2015). *Training Deeper Convolutional Networks*
829 *with Deep Supervision* (arXiv:1505.02496). arXiv.

- 830 <https://doi.org/10.48550/arXiv.1505.02496>
- 831 Yushkevich, P. (2023). *Greedy* [C++]. <https://github.com/pyushkevich/greedy> (Original work
832 published 2016)
- 833 Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., & Gerig, G. (2006).
834 User-guided 3D active contour segmentation of anatomical structures: Significantly
835 improved efficiency and reliability. *NeuroImage*, 31(3), 1116–1128.
836 <https://doi.org/10.1016/j.neuroimage.2006.01.015>
- 837 Yushkevich, P. A., Pluta, J., Wang, H., Wisse, L. E. M., Das, S., & Wolk, D. (2016). Fast
838 Automatic Segmentation of Hippocampal Subfields and Medial Temporal Lobe
839 Subregions In 3 Tesla and 7 Tesla T2-Weighted MRI. *Alzheimer's & Dementia*,
840 12(7S_Part_2), P126–P127. <https://doi.org/10.1016/j.jalz.2016.06.205>
- 841 Zeng, C., Gu, L., Liu, Z., & Zhao, S. (2020). Review of Deep Learning Approaches for the
842 Segmentation of Multiple Sclerosis Lesions on Brain MRI. *Frontiers in Neuroinformatics*,
843 14. <https://www.frontiersin.org/articles/10.3389/fninf.2020.610967>
- 844 Zhang, H., & Oguz, I. (2021). Multiple Sclerosis Lesion Segmentation—A Survey of Supervised
845 CNN-Based Methods. In A. Crimi & S. Bakas (Eds.), *Brainlesion: Glioma, Multiple*
846 *Sclerosis, Stroke and Traumatic Brain Injuries* (pp. 11–29). Springer International
847 Publishing. https://doi.org/10.1007/978-3-030-72084-1_2

848

849 **Acknowledgments:**

850 We thank Naga Karthik Enamundram and Joshua Newton for helpful discussions around the
851 packaging of LST-AI, the evaluation of the different algorithms using the anima toolbox, and for
852 visualization of the U-Net architecture.

853

854 **Funding:**

855 TW received funding by a research grant of the National Institutes of Health (grant
856 1R01NS112161-01). JM, JK, and MM received funding by the Bavarian State Ministry for
857 Science and Art (Collaborative Bilateral Research Program Bavaria – Québec: AI in medicine,
858 grant F.4-V0134.K5.1/86/34). BM, DR, MM and BW received funding from the DFG, SPP
859 Radiomics (project number 428223038).

860

861 **Data and code availability:**

862 We provide our toolbox as source code, command line tool and dockerized application at
863 <https://github.com/Complmg/LST-AI>.

864

865 **Author contributions:**

866 Tun Wiltgen: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data
867 curation, Writing - original draft, Visualization

868 Julian McGinnis: Conceptualization, Methodology, Software, Formal analysis, Data curation,
869 Writing - original draft, Visualization

870 Sarah Schlaeger: Investigation, Resources, Data curation, Writing - review & editing

871 Florian Kofler: Resources, Software

872 Cuici Voon: Investigation, Data curation, Writing - review & editing

873 Achim Berthele: Resources, Writing - review & editing

874 Daria Bischl: Resources, Data curation, Writing - review & editing

875 Lioba Grundl: Resources, Data curation, Writing - review & editing

876 Nikolaus Will: Resources, Data curation, Writing - review & editing

877 Marie Metz: Resources, Data curation, Writing - review & editing

878 David Schinz: Resources, Data curation, Writing - review & editing

879 Dominik Sepp: Resources, Data curation, Writing - review & editing

880 Philipp Prucker: Resources, Data curation, Writing - review & editing

881 Benita Schmitz-Koep: Resources, Data curation, Writing - review & editing

882 Claus Zimmer: Resources, Writing - review & editing

883 Bjoern Menze: Resources, Writing - review & editing

884 Daniel Rückert: Resources, Writing - review & editing

885 Bernhard Hemmer: Resources, Writing - review & editing

886 Jan Kirschke: Investigation, Resources, Data curation, Writing - review & editing

887 Mark Mühlau: Conceptualization, Methodology, Resources, Writing - review & editing,

888 Supervision, Project administration, Funding acquisition

889 Benedikt Wiestler: Conceptualization, Methodology, Software, Formal analysis, Data curation,

890 Writing - original draft, Supervision, Project administration, Funding acquisition

891

892 **Declaration of Competing Interests:**

893 The authors declare no competing interests.