MULTIPLE SCLEROSIS JOURNAL | MSJ

Original Research Paper

Generalizable spinal cord multiple sclerosis lesion segmentation across MRI contrasts, protocols, and centers

Multiple Sclerosis Journal

1–16

DOI: 10.1177/13524585261427333

Pierre-Louis Benveniste, Laurent Létourneau-Guillon, David Araujo, Lydia Chougar, Dumitru Fetco, Masaaki Hori, Kouhei Kamiya, Steven Messina, Charidimos Tsagkas, Bertrand Audoin, Rohit Bakshi, Elise Bannier, Daniel Blezek, Jean-Christophe Brisset, Virginie Callot, Erik Charlson, Michelle Chen, Olga Ciccarelli, Sarah Demortière, Gilles Edan, Massimo Filippi, Tobias Granberg, Cristina Granziera, Christopher C. Hemond, B. Mark Keegan, Anne Kerbrat, Jan Kirschke, Shannon Kolind, Pierre Labauge, Lisa Eunyoung Lee, Yaou Liu, Caterina Mainero, Julian McGinnis, Nilser Laines Medina, Mark Mühlau, Govind Nair, Kristin P. O'Grady, Jiwon Oh, Russell Ouellette, Alexandre Prat, Daniel S. Reich, Maria A. Rocca, Timothy M. Shepherd, Seth A. Smith, Leszek Stawiarz, Jason Talbott, Roger Tam, Shahamat Tauhid, Anthony Traboulsee, Constantina Andrada Treaba, Paola Valsasina, Zachary Vavasour, Marios Yiannakas, Hervé Lombaert and Julien Cohen-Adad

Abstract

Background/Objectives: Characterizing spinal cord multiple sclerosis (MS) lesions in MRI is critical for diagnosis, monitoring, and treatment evaluation. However, current automated approaches for lesion detection and segmentation are typically designed for specific MRI contrasts or acquisition sites, limiting their generalizability in real-world clinical settings where imaging protocols vary widely. This work proposes a robust multi-site, multi-contrast segmentation framework for spinal cord lesions.

Methods: The segmentation model was trained and evaluated on a large-scale dataset comprising 4428 annotated images from 1849 persons with MS across 23 imaging centers, encompassing six MRI contrasts (T1w, T2w, T2*w, PSIR, STIR, and UNIT1) acquired at 1.5 tesla (T), 3 T, and 7 T.

Results: Likert-type assessment performed by neuroradiologist ratings demonstrated superior generalization of the model compared to existing contrast-specific pipelines ($p < 0.01$). Additional experiments evaluated robustness across spinal levels, acquisition resolutions, binarization thresholds, and quantitative evaluation on external labeled datasets.

Conclusions: The proposed model can achieve accurate and reliable spinal cord MS lesion segmentation across heterogeneous MRI data, addressing a key barrier to clinical translation. The model is available in the Spinal Cord Toolbox v7.2 and higher.

Code repository: https://github.com/ivadomed/seg-sc-ms-lesion-multicontrast

Keywords: Spinal cord, multiple sclerosis, lesion, segmentation, magnetic resonance imaging, MRI, deep learning

Date received: 31 October 2025; revised: 22 January 2026; accepted: 7 February 2026

Correspondence to:
P-L Benveniste
NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, 2500 Chem. de Polytechnique, Montréal, QC H3T 0A3, Canada.
pierre-louis-2.benveniste@polymtl.ca

Pierre-Louis Benveniste
NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, Canada; Mila—Quebec AI Institute, Montreal, QC, Canada

Laurent Létourneau-Guillon
Centre de Recherche du Centre Hospitalier de l'Université de Montréal (CHUM), Montréal, QC, Canada

David Araujo
McConnell Brain Imaging Center, McGill, Montréal, QC, Canada

Lydia Chougar
The Neuro—Montreal Neurological Institute and Hospital, McGill University, Montreal, QC, Canada; Department of Neuroradiology, Sorbonne Université, Institut du Cerveau—Paris Brain Institute—ICM, AP-HP, CNRS, Inserm, Hôpital de la Pitié Salpêtrière, Paris, France

Dumitru Fetco
McConnell Brain Imaging

Introduction

Context

Multiple sclerosis (MS) is the leading cause of non-traumatic neurological disability in young adults, with increasing global prevalence.[1] While research has historically focused on brain lesions, spinal cord (SC) lesions, particularly in the cervical region, disrupt motor and sensory pathways and strongly correlate with disability progression.[2]

Center, McGill University, Montréal, QC, Canada; NeuroRx, A Clario Company, Montréal, QC, Canada; 7T MRI MS Working Group, North American Imaging in MS, MNI, Montréal, QC, Canada

Masaaki Hori
Kouhei Kamiya
Department of Radiology, Toho University Omori Medical Center, Tokyo, Japan

Steven Messina
Department of Radiology, Mayo Clinic College of Medicine and Science, Rochester, MN, USA

Charidimos Tsagkas
Department of Neurology, University Hospital Basel, University of Basel, Basel, Switzerland; National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA

Bertrand Audoin
Aix-Marseille Univ, CNRS, CRMBM, Marseille, France; AP-HM, Hôpital Universitaire Timone, CEMEREM, Marseille, France; Neurology Department, AP-HM, Hôpital Universitaire Timone, Marseille, France

Rohit Bakshi
Shahamat Tauhid
Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

Elise Bannier
Gilles Edan
Anne Kerbrat
Univ Rennes, Inria, CNRS, Inserm, IRISA UMR 6074, Visages U1128, Rennes, France; Radiology Department, CHU Rennes, Rennes, France

Daniel Blezek
Departments of Radiology and Biomedical Engineering, Mayo Clinic College of Medicine and Science, Rochester, MN, USA

Jean-Christophe Brisset
Brisset JC Ph.D.—Medical Imaging Consulting, Sophia Antipolis, Valbonne, France

Virginie Callot
Aix-Marseille Univ, CNRS, CRMBM, Marseille, France; AP-HM, Hôpital Universitaire Timone, CEMEREM, Marseille, France

Erik Charlson
Timothy M. Shepherd
Departments of Neurology & Radiology, NYU Langone Medical Center, New York, NY, USA

Michelle Chen
NeuroPoly Lab, Institute of Biomedical Engineering,

Magnetic resonance imaging (MRI) is central to MS diagnosis and monitoring, underpinning the McDonald criteria and their revisions.[3–5] Beyond lesion count, lesion segmentation provides precise volumetric and spatial information that maps lesions to specific anatomical structures (like the corticospinal tract), enabling better prediction of motor outcomes through structure-function analysis.[6–8]

A wide range of MRI sequences, including T1w, T2w, T2*w, STIR, PSIR, and MP2RAGE, performed at various magnetic field strengths from different manufacturers, are used to visualize MS lesions.[9] However, SC imaging remains technically challenging due to its small size, deformability, and susceptibility to magnetic field inhomogeneities.[10] Despite recent harmonization efforts and international guidelines,[11–14] adoption of those guidelines remains very uneven for SC imaging in MS, and lesion appearance continues to vary across contrasts.

These challenges highlight the need for a robust, generalizable model that can automatically segment SC MS lesions across diverse imaging conditions.

Related works

While automated segmentation of MS lesions in the brain has been extensively studied for over two decades,[15–18] SC lesion segmentation remains comparatively underexplored.

The advent of convolutional neural networks (CNNs) marked a paradigm shift in automated brain MS lesion segmentation, with U-Net and its variants establishing state-of-the-art performance[19–23] and becoming the de facto standard.[16,24] Despite the growing interest in transformer-based architectures, CNNs remain particularly well suited for medical imaging tasks due to their strong spatial inductive biases, computational efficiency, and robustness in data-limited regimes.

In contrast, only a limited number of SC-specific methods have been proposed, many of which are not publicly available[25,26] or require advanced technical expertise,[27] limiting their clinical adoption. Moreover, most existing approaches are tailored to specific MRI contrasts and do not generalize well across acquisition protocols. For instance, *sct_deepseg_lesion*[28] has been developed and validated for T2w and T2*w images, while other variants have been proposed for PSIR and STIR,[29] MP2RAGE,[30] or axial T2w scans.[31]

Existing methods are predominantly based on CNNs, often derived from U-Net or its optimized implementation, such as the nnU-Net framework. Gros et al.[28] combined three U-Nets for centerline detection, cord segmentation, and lesion delineation on T2w and T2*w scans. More recently, U-Net pipelines optimized with nnU-Net were proposed for MP2RAGE, PSIR, and STIR sequences,[29,30] while Karthik et al.[31] developed a region-based nnU-Net model, explicitly restricting predictions to the SC. Another recent study[25] featured segmentation on dual-contrast inputs; however, the lack of code or model availability has limited reproducibility and widespread adoption.

A major challenge in SC lesion segmentation models lies in the variability of manual annotations, where inter- and intra-rater variability is substantial for small or ambiguous lesions. This variability is further amplified in the SC compared to the brain due to lower lesion conspicuity and image artifacts.[25,26,28] Such inconsistencies produce noisy ground truth labels that impair model training as deep learning models may learn rater-specific biases rather than lesion-specific features.[32] Evaluation itself is also hindered by noisy annotations, as standard metrics computed against imperfect ground truth labels may not fully reflect the true performance of segmentation models. To mitigate this limitation, complementary evaluation strategies such as blind expert evaluation of predicted segmentations—as we do here with the Likert-type assessment—can provide a more reliable assessment of clinical plausibility.

In this study, we introduce the first segmentation model explicitly designed to operate reliably across a broad spectrum of contrasts. Model performances are assessed using both quantitative segmentation metrics and expert neuroradiologist reviews, ensuring clinical relevance. Rather than aiming for uniform performance across all MRI contrasts, this work focuses on developing a single segmentation model that generalizes robustly across heterogeneous acquisition protocols and contrasts, reflecting real-world clinical variability.

Methods

Data

Experiments were conducted on a large-scale, heterogeneous multi-site SC MRI dataset (Figure 1). The dataset comprises 4428 annotated images acquired from 1849 unique participants across 23 imaging centers, spanning a wide range of acquisition protocols and scanner configurations. Table 1 lists relevant acquisition parameters for each site. The dataset

Table 1. MRI dataset characteristics across projects/sites, contrasts, acquisitions, orientations, resolutions, participants, and field strength.

Site	#Participants	Field strength	Contrast	Acq.	Orien.	Resolution (RPI)	#Images
Annotated data							
FR AMU	15	3 T	T2*w	2D	ax	0.4 ± 0.1 × 0.4 ± 0.1 × 4.7 ± 1.4	30
		3 T	T2w	2D	sag	2.8 ± 0.0 × 0.7 ± 0.0 × 0.7 ± 0.0	7
		3 T	T2w	3D	sag	1.0 ± 0.0 × 1.0 ± 0.0 × 1.0 ± 0.0	8
CH Basel-2018	23	3 T	T1w	3D	sag	1.0 ± 0.0 × 1.0 ± 0.0 × 1.0 ± 0.0	22
		3 T	T2w	2D	sag	3.0 ± 0.0 × 0.6 ± 0.0 × 0.6 ± 0.0	23
CH Basel-2021	180	3 T	UNIT1	3D	sag	1.0 ± 0.0 × 1.0 ± 0.0 × 1.0 ± 0.0	180
US BWH	80	3 T	T2w	2D	ax	0.6 ± 0.0 × 0.6 ± 0.0 × 3.0 ± 0.0	80
		3 T	T2w	2D	sag	3.0 ± 0.0 × 0.6 ± 0.1 × 0.6 ± 0.1	97
CA CanProCo-Calgary	92	3 T	STIR	2D	sag	3.0 ± 0.0 × 0.7 ± 0.0 × 0.7 ± 0.0	92
CA CanProCo-Edmonton	71	3 T	PSIR	2D	sag	3.0 ± 0.0 × 0.7 ± 0.0 × 0.7 ± 0.0	77
CA CanProCo-Montreal	96	3 T	PSIR	2D	sag	3.0 ± 0.0 × 0.7 ± 0.0 × 0.7 ± 0.0	106
CA CanProCo-Toronto	89	3 T	PSIR	2D	sag	3.0 ± 0.0 × 0.7 ± 0.0 × 0.7 ± 0.0	100
CA CanProCo-Vancouver	80	3 T	PSIR	2D	sag	3.0 ± 0.0 × 0.7 ± 0.0 × 0.7 ± 0.0	80
IT IRCCS	116	3 T	T2*w	2D	ax	0.5 ± 0.0 × 0.5 ± 0.0 × 2.5 ± 0.0	115
		3 T	T2w	2D	sag	2.5 ± 0.0 × 0.5 ± 0.0 × 0.5 ± 0.0	116
SE Karolinska-2019	51	3 T	T2*w	2D	ax	0.4 ± 0.0 × 0.4 ± 0.0 × 4.4 ± 0.0	51
SE Karolinska-2020	28	3 T	T2*w	2D	ax	0.4 ± 0.0 × 0.4 ± 0.0 × 3.3 ± 0.0	22
		3 T	T2w	2D	ax	0.6 ± 0.0 × 0.6 ± 0.0 × 4.4 ± 0.0	27
		3 T	T2w	2D	sag	3.3 ± 0.0 × 0.8 ± 0.0 × 0.8 ± 0.0	21
US MGH	18	7 T	T2*w	2D	ax	0.4 ± 0.0 × 0.4 ± 0.0 × 3.6 ± 0.0	36
US NIH-2017	34	3 T	T2*w	2D	ax	0.5 ± 0.1 × 0.5 ± 0.1 × 4.8 ± 0.3	38
		3 T	T2w	2D	sag	1.7 ± 0.4 × 0.7 ± 0.2 × 0.7 ± 0.2	34
US NIH-2023	163	3 T	UNIT1	3D	sag	1.0 ± 0.0 × 1.0 ± 0.0 × 1.0 ± 0.0	163
US NYU	153	3 T	T2w	2D	ax	0.6 ± 0.0 × 0.6 ± 0.0 × 4.4 ± 1.3	209
		3 T	T2w	2D	sag	3.0 ± 0.1 × 0.7 ± 0.1 × 0.7 ± 0.1	153
FR OFSEP-Lyon	60	1.5 T/3 T	T2*w	2D	ax	0.6 ± 0.0 × 0.6 ± 0.0 × 4.1 ± 0.5	60
		1.5 T/3 T	T2w	2D	sag	4.3 ± 0.4 × 0.7 ± 0.0 × 0.7 ± 0.0	60
FR OFSEP-Montpellier	14	1.5 T/3 T	T2*w	2D	ax	0.7 ± 0.0 × 0.7 ± 0.0 × 3.3 ± 0.0	28
		1.5 T/3 T	T2w	2D	sag	2.7 ± 0.0 × 0.7 ± 0.0 × 0.7 ± 0.0	28
FR OFSEP-Rennes	55	3 T	T2*w	2D	ax	0.4 ± 0.0 × 0.4 ± 0.0 × 3.3 ± 0.0	107
		3 T	T2w	2D	sag	2.7 ± 0.0 × 0.7 ± 0.0 × 0.7 ± 0.0	104
DE TUM	337	1.5 T	T2w	2D	ax	0.6 ± 0.1 × 0.6 ± 0.1 × 4.8 ± 1.1	22
		3 T	T2w	2D	ax	0.3 ± 0.1 × 0.3 ± 0.1 × 5.0 ± 0.2	1977
GB UCL	39	3 T	T2*w	2D	ax	0.5 ± 0.0 × 0.5 ± 0.0 × 5.0 ± 0.0	39
		3 T	T2w	2D	sag	3.0 ± 0.0 × 1.0 ± 0.0 × 1.0 ± 0.0	39
US UCSF	32	3 T	T2w	2D	ax	0.4 ± 0.1 × 0.4 ± 0.1 × 3.8 ± 0.7	32
US Vanderbilt	23	3 T	T2*w	2D	ax	0.3 ± 0.0 × 0.3 ± 0.0 × 5.0 ± 0.0	22
		3 T	T2w	2D	sag	2.0 ± 0.0 × 0.5 ± 0.0 × 0.5 ± 0.0	23
Un-annotated data							
US Mayo	219	3 T	T2w	2D	ax	0.5 ± 0.2 × 0.5 ± 0.2 × 5.1 ± 2.8	219
US UMass-GE-Excite	22	1.5 T	STIR	2D	sag	3.6 ± 0.4 × 0.4 ± 0.0 × 0.4 ± 0.0	36
		1.5 T	T1w	2D	sag	3.6 ± 0.4 × 0.6 ± 0.2 × 0.6 ± 0.2	36
		1.5 T	T2w	2D	ax	0.6 ± 0.2 × 0.6 ± 0.2 × 3.7 ± 0.3	36
		1.5 T	T2w	2D	sag	3.6 ± 0.4 × 0.4 ± 0.0 × 0.4 ± 0.0	36

(continued)

Polytechnique Montreal, Montreal, QC, Canada

Olga Ciccarelli
Marios Yiannakas
Queen Square MS Centre, UCL Queen Square Institute of Neurology, Faculty of Brain Sciences, University College London, London, UK

Sarah Demortière
Neurology Department, AP-HM, Hôpital Universitaire Timone, Marseille, France

Massimo Filippi
Neuroimaging Research Unit, Division of Neuroscience, IRCCS San Raffaele Scientific Institute, Milan, Italy; Neurology Unit, IRCCS San Raffaele Scientific Institute, Milan, Italy; Neurorehabilitation Unit, IRCCS San Raffaele Scientific Institute, Milan, Italy; Neurophysiology Service, IRCCS San Raffaele Scientific Institute, Milan, Italy; Vita-Salute San Raffaele University, Milan, Italy

Tobias Granberg
Russell Ouellette
Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden; Department of Neuroradiology, Karolinska University Hospital, Stockholm, Sweden

Cristina Granziera
Department of Neurology, University Hospital Basel, University of Basel, Basel, Switzerland

Christopher C. Hemond
Departments of Neurology, University of Massachusetts Memorial Medical Center and University of Massachusetts Chan Medical School, Worcester, MA, USA

B. Mark Keegan
Department of Neurology, Mayo Clinic College of Medicine and Science, Rochester, MN, USA

Jan Kirschke
Department of Diagnostic and Interventional Neuroradiology, Klinikum rechts der Isar, TUM School of Medicine and Health, Munich, Germany

Shannon Kolind
Anthony Traboulsee
Departments of Medicine (Neurology), Physics, Radiology, University of British Columbia, Vancouver, BC, Canada

Pierre Labauge
MS Unit, Department

of Neurology, University Hospital of Montpellier, Montpellier, France

Lisa Eunyoung Lee
Jiwon Oh
Department of Medicine (Neurology), University of Toronto, Toronto, ON, Canada; BARLO Multiple Sclerosis Centre and Keenan Research Centre, St. Michael's Hospital, Toronto, ON, Canada

Yaou Liu
Department of Radiology, Xuanwu Hospital, Capital Medical University, Beijing, China; Department of Radiology, Beijing Tiantan Hospital, Capital Medical University, Beijing, China

Caterina Mainero
Constantina Andrada Treaba
Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, MA, USA

Julian McGinnis
Department of Diagnostic and Interventional Neuroradiology, Klinikum rechts der Isar, TUM School of Medicine and Health, Munich, Germany; Institute for AI in Medicine, Technical University of Munich, Munich, Germany

Nilser Laines Medina
NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, Canada; Mila—Quebec AI Institute, Montreal, QC, Canada; Aix-Marseille Univ, CNRS, CRMBM, Marseille, France; AP-HM, Hôpital Universitaire Timone, CEMEREM, Marseille, France

Mark Mühlau
Department of Neurology, School of Medicine and Health, Technical University of Munich, Munich, Germany; TUM-Neuroimaging Center, School of Medicine and Health, Technical University of Munich, Munich, Germany

Govind Nair
Daniel S. Reich
National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA

Kristin P. O'Grady
Seth A. Smith
Vanderbilt University Institute of Imaging Science, Nashville, TN, USA

Alexandre Prat
Department of Neuroscience, Université de Montréal, Montreal, QC, Canada; Neuroimmunology Research Laboratory, University of

Table 1. (continued)

Site	#Participants	Field strength	Contrast	Acq.	Orien.	Resolution (RPI)	#Images
US UMass-GE-HDxt	35	1.5 T	T2w-IR	2D	sag	$3.3 \pm 0.0 \times 0.8 \pm 0.1 \times 0.8 \pm 0.1$	45
		1.5 T	T1w	2D	ax	$0.7 \times 0.7 \times 5.0$	1
		1.5 T	T1w	2D	sag	$3.3 \pm 0.0 \times 0.8 \pm 0.1 \times 0.8 \pm 0.1$	45
		1.5 T	T2w	2D	ax	$0.4 \pm 0.0 \times 0.4 \pm 0.0 \times 4.0 \pm 0.1$	45
US UMass-GE-Pioneer	240	3 T	STIR	2D	sag	$3.3 \pm 0.1 \times 0.4 \pm 0.0 \times 0.4 \pm 0.0$	496
		3 T	T1w	2D	sag	$3.3 \pm 0.1 \times 0.4 \pm 0.0 \times 0.4 \pm 0.0$	467
		3 T	T1w	3D	sag	$2.8 \pm 0.6 \times 0.4 \pm 0.0 \times 0.4 \pm 0.0$	32
		3 T	T2w	2D	ax	$0.4 \pm 0.0 \times 0.4 \pm 0.0 \times 3.5 \pm 0.2$	491
US UMass-Siemens	22	1.5 T	STIR	2D	sag	$3.5 \pm 0.1 \times 0.4 \pm 0.0 \times 0.4 \pm 0.0$	24
		1.5 T	T1w	2D	sag	$3.6 \pm 0.7 \times 0.7 \pm 0.1 \times 0.7 \pm 0.1$	24
		1.5 T	T2w	2D	ax	$0.8 \pm 0.0 \times 0.8 \pm 0.0 \times 4.1 \pm 1.5$	24
		1.5 T	T2w	2D	sag	$3.8 \pm 1.4 \times 0.7 \pm 0.0 \times 0.7 \pm 0.0$	24
CN Tiantan	25	3 T	T1w	2D	sag	$5.1 \pm 1.6 \times 0.7 \pm 0.1 \times 0.7 \pm 0.1$	6
		3 T	T1w	3D	sag	$1.1 \pm 0.7 \times 1.0 \pm 0.0 \times 1.0 \pm 0.0$	72
		3 T	T2w	2D	ax	$0.7 \pm 0.0 \times 0.7 \pm 0.0 \times 8.8 \pm 2.6$	52
		3 T	T2w	2D	sag	$3.4 \pm 0.5 \times 0.6 \pm 0.0 \times 0.6 \pm 0.0$	80

AMU: Aix-Marseille Université. Basel: University of Basel. BWH: Brigham and Women's Hospital. CanProCo ($n = 5$ sites): Canadian Prospective Cohort Study for People Living with MS [33]. IRCCS: IRCCS San Raffaele Scientific Institute. Karolinska: Karolinska University Hospital. MGH: Massachusetts General Hospital. NIH: National Institutes of Health. NYU: NYU Langone Medical Center. OFSEP-Lyon: Observatoire Français de la Sclérose en Plaques—Lyon. OFSEP-Montpellier: Observatoire Français de la Sclérose en Plaques—Montpellier. Rennes: Centre hospitalier universitaire de Rennes. TUM: Technical University of Munich. UCL: University College London. UCSF: University of California San Francisco. Vanderbilt: Vanderbilt University Institute of Imaging Science. Tiantan: Beijing Tiantan Hospital, Capital Medical University. Mayo: Mayo Clinic College of Medicine and Science. UMass: University of Massachusetts Memorial Medical Center.

included images acquired on GE, Siemens or Philips MRI systems, at 1.5 T, 3 T or 7 T, using six distinct MRI contrasts: T2w ($n = 3060$), T2*w ($n = 548$), PSIR ($n = 363$), UNIT1 (reconstructed uniform image from MP2RAGE sequence, $n = 343$), STIR ($n = 92$), and T1w ($n = 22$), and spans 2D axial ($n = 2895$), 2D sagittal ($n = 1160$), and 3D ($n = 373$) acquisition planes. Proton-density weighted imaging was excluded because of the poor lesion contrast. The field-of-view coverage varied across sites (brain and upper SC, or SC only). Image resolution exhibited high variability, with an average ($\pm$ standard deviation) of $1.10 \pm 1.13 \times 0.51 \pm 0.24 \times 3.27 \pm 1.95\,\text{mm}^3$ reported in "RPI-" orientation (Right→Left, Posterior→Anterior, Inferior→Superior), and pixel dimensions ranging from 0.19 mm to 11.92 mm, including inter-slice gap.

Lesions were segmented manually at each site and data were organized according to a standard (more details about raters' expertise, segmentation methods, and dataset aggregation can be found in Supplemental Appendix A).

In addition, we also had access to a collection of 2291 unannotated images originating from three independent cohorts (Table 1). These images were used to qualitatively assess the generalization capability of the segmentation model in real-world out-of-distribution clinical data. However, since no lesion segmentations were available for these sites, they could not be included in the quantitative evaluation of model performance.

Model architecture and training

Several deep learning models were benchmarked on our multi-site dataset, and pretrained weights were used when available: Attention U-Net,[33] STUNet,[34] MultiTalent,[35] MedNeXt,[36] and various nnU-Net architectures.[37] Among these architectures, the best-performing model was a 3D Residual Encoder U-Net (ResEnc), trained within the nnUNetv2 framework.[37] The architecture used the nnUNetPlannerResEncL template with an input patch size of $192 \times 192 \times 192$ voxels, with resampled isotropic resolution of $1.0 \times 1.0 \times 1.0\,\text{mm}^3$.

Images were randomly partitioned into training ($n = 3925$), test ($n = 433$), and external validation ($n = 70$) subsets. The dataset was partitioned at the

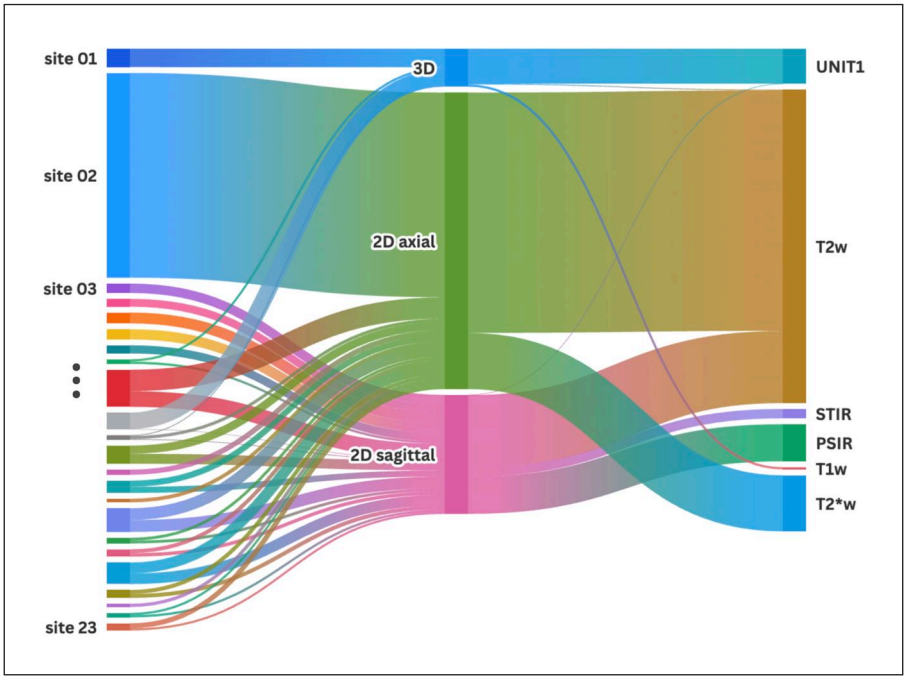

Figure 1. Sankey diagram of annotated MRI scans across clinical sites. Line thickness is associated with the number of scans.
MRI scan distribution is clustered per acquisition type (3D, 2D sagittal, or 2D axial) and per MRI contrast, for each site.

subject level to ensure subject independence between training and test sets, while maintaining the original contrast distribution across both subsets. The external validation set consisted of one separate dataset which was not seen during training or validation. Training was performed using a fivefold cross-validation scheme with an 80%/20% train/validation split in each fold. A small batch size of two was used to maximize generalization and prevent overfitting, consistent with findings from prior work.[38] The loss function was the combination of Dice and Cross-Entropy without label smoothing (DiceCELoss). Standard data augmentation from the nnU-Net framework was used. More details can be found in Supplemental Appendix B.

Evaluation

Model performance was assessed on two distinct subsets: (i) an internal test set comprising 10% of each dataset ($n=433$), and (ii) an external test set composed of an entirely independent dataset not seen during training ($n=70$). The internal test set included six MRI contrasts: PSIR ($n=36$), STIR ($n=7$), T1w ($n=3$), T2*w ($n=57$), T2w ($n=295$), and UNIT1 ($n=35$), while the external set included T2*w ($n=22$) and T2w ($n=48$) acquisitions. For each input, the final segmentation was generated by averaging the binary predictions obtained from the five cross-validation folds, followed by binarization using a fixed threshold of 0.5, which corresponds to the cut-off for partial volume effect.

Quantitative evaluation of segmentation quality employed both voxel-wise and lesion-wise metrics.

The proposed model was compared to established segmentation pipelines tailored to specific MRI contrasts: (i) *sct_deepseg_lesion*,[28] which supports sagittal T2w (flag "-c t2"), axial T2w (flag "-c t2_ax") and axial T2*w (flag "-c t2s") contrasts, (ii) *sct_deepseg lesion_ms_mp2rage*, adapted to MP2RAGE UNIT1 acquisitions,[30] and (iii) *sct_deepseg lesion_ms_axial_t2*, recently proposed for axial T2w images.[31]

A complementary qualitative evaluation was conducted by selecting a panel of 8 neuroradiologists who reviewed a subset of 40 randomly selected images from the internal test set (~9%), scoring the quality of the segmentation masks produced by the model as well as the manual segmentation, using a 5-point Likert-type scale (1: very poor, 5: excellent). Manual segmentation and model prediction were anonymized to limit evaluation bias. Inter-rater agreement was also quantified (see details in Supplemental Appendix C).

Model performance along the SC was computed per intervertebral disks in Supplemental Appendix D

Montreal Hospital Research Centre (CRCHUM), Montreal, QC, Canada

Maria A. Rocca
Neuroimaging Research Unit, Division of Neuroscience, IRCCS San Raffaele Scientific Institute, Milan, Italy; Neurology Unit, IRCCS San Raffaele Scientific Institute, Milan, Italy; Vita-Salute San Raffaele University, Milan, Italy

Leszek Stawiarz
Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden

Jason Talbott
Department of Radiology and Biomedical Imaging, Zuckerberg San Francisco General Hospital, University of California, San Francisco, CA, USA

Roger Tam
Zachary Vavasour
School of Biomedical Engineering, University of British Columbia, Vancouver, BC, Canada

Paola Valsasina
Neuroimaging Research Unit, Division of Neuroscience, IRCCS San Raffaele Scientific Institute, Milan, Italy

Hervé Lombaert
Mila—Quebec AI Institute, Montreal, QC, Canada; Polytechnique Montréal, Montreal, QC, Canada

Julien Cohen-Adad
NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, Canada; Mila—Quebec AI Institute, Montreal, QC, Canada; Functional Neuroimaging Unit, CRIUGM, Université de Montréal, Montreal, QC, Canada; Centre de Recherche du CHU Sainte-Justine, Université de Montréal, Montreal, QC, Canada

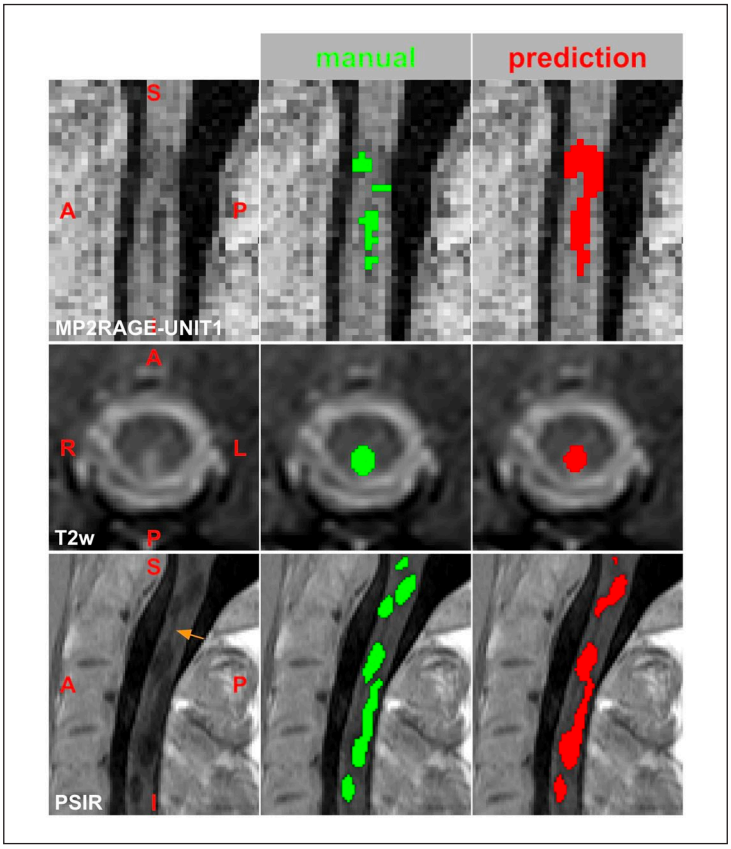

Figure 2. Qualitative examples of SC lesion segmentation across different MRI contrasts and orientations.
From left to right: original MRI, manual segmentation (green), and model prediction (red). Examples are shown for MP2RAGE-UNIT1 (sagittal) from the University of Basel, T2w (axial) from the Technical University of Munich, and PSIR (sagittal) acquisitions from the CanProCo dataset. The model correctly identifies focal and elongated lesions across contrasts, though some discrepancies in lesion extent are observed. The orange arrow points to the central canal.

alongside the prevalence of each disk level and the average lesion total volume. The robustness of the model across spatial resolution and post-processing experiments were also investigated in Supplemental Appendix E and F, respectively.

To assess the generalizability of the model, unlabeled scans from independent sites were passed through the model ("Un-annotated data" in Table 1). The resulting predictions were inspected to evaluate segmentation plausibility and overall robustness on acquisition protocols unseen during training.

Results

Qualitative examples of predicted lesion segmentations

Figure 2 presents examples of predicted lesion segmentations. In some instances, manual annotations appear to underestimate lesion boundaries, whereas the model predictions provide a more complete delineation (e.g. MP2RAGE-UNIT1). The PSIR example illustrates the intrinsic ambiguity in lesion interpretation, where the same hyperintense region could be segmented either as several smaller lesions or as a single larger confluent lesion.

Figure 3 shows examples of predicted segmentations on unannotated MRI scans, illustrating generalization to unseen acquisition protocols. Across diverse sites, field strengths, and contrasts, the model produced reliable segmentations.

Model performance

Quantitative evaluation of the model demonstrated robust segmentation performance across datasets (see Table 2). On the internal test set, the model achieved a mean Dice score of 0.63, while lesion-wise evaluation yielded an L-F1-score of 0.71. On the external test set, the model achieved an L-F1-score of 0.80, confirming its ability to generalize across independent data. Importantly, L-Recall remained consistently high

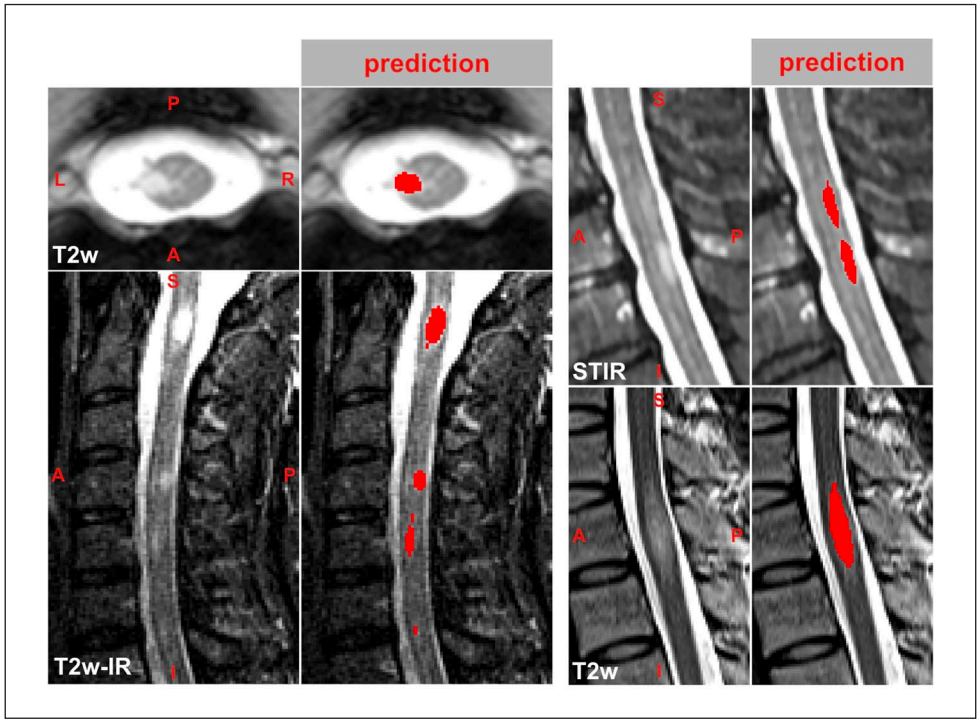

Figure 3. Representative examples of SC lesion segmentations across MRI scans from unannotated dataset.
Predicted segmentations are displayed next to the original MRI. Examples displayed are: axial T2w from Mayo Clinic College of Medicine and Science; sagittal STIR from University of Massachusetts Memorial Medical Center; sagittal T2w-IR from University of Massachusetts Memorial Medical Center; sagittal T2w from Beijing Capital Medical University.

Table 2. Voxel-wise and lesion-wise performance of the proposed segmentation model. Performance is reported on the train set ($n=3925$), test set ($n=433$), and external test set ($n=70$).

	Train set ($n=3925$)	Test set ($n=433$)	Ext. test set ($n=70$)
Dice ↑	0.72 ± 0.28	0.63 ± 0.34	0.66 ± 0.33
L-Recall ↑	0.85 ± 0.27	0.81 ± 0.31	0.87 ± 0.27
L-PPV ↑	0.86 ± 0.30	0.76 ± 0.37	0.82 ± 0.34
L-F1-score ↑	0.81 ± 0.30	0.71 ± 0.36	0.80 ± 0.33

Evaluation is done on both voxel-wise metrics (Dice score) and lesion-wise metrics (recall, positive predictive value (PPV), and F1-score).

(>0.80 across all sets), indicating that the model rarely missed lesions. In contrast, precision values were lower, especially on the test set (L-PPV = 0.76), reflecting a tendency toward over-detection. The relatively large standard deviations across metrics underscore the challenges posed by SC imaging (heterogeneity of acquisition parameters, motion artifacts, etc.) and inter-rater variability in manual annotations.

Table 3 shows the robustness of the model across MRI contrasts. Among the most represented contrasts, T2w and UNIT1 images achieved the highest Dice scores, with 0.75 on the training set and 0.64 on the test set for T2w, and 0.77 and 0.67 for UNIT1. Conversely, underrepresented contrasts such as T1w ($n=19$ in training) exhibited markedly lower performance, with Dice dropping to 0.42 on the test set. Intermediate performances were obtained for PSIR and T2*w acquisitions, although with higher variability. Notably, despite the limited number of cases, STIR images yielded a relatively high Dice score on the test set, which can be explained by the similarity of T2w and STIR contrasts.

Performance across MRI contrasts should be interpreted in light of the strong imbalance in contrast

Table 3. Dice score per MRI contrast.

Contrast	Train set	Test set	Ext. test set
PSIR	0.59 ± 0.32 (*n* = 327)	0.51 ± 0.33 (*n* = 36)	–
STIR	0.56 ± 0.32 (*n* = 85)	0.73 ± 0.35 (*n* = 7)	–
T1w	0.57 ± 0.31 (*n* = 19)	0.42 ± 0.52 (*n* = 3)	–
T2*w	0.65 ± 0.24 (*n* = 469)	0.61 ± 0.29 (*n* = 57)	0.69 ± 0.22 (*n* = 22)
T2w	0.75 ± 0.28 (*n* = 2717)	0.64 ± 0.35 (*n* = 295)	0.65 ± 0.37 (*n* = 48)
UNIT1	0.77 ± 0.20 (*n* = 308)	0.67 ± 0.26 (*n* = 35)	–

Performance is reported on the train, test, and external test sets for each available sequence.

Table 4. Quantitative comparison with existing segmentation methods on voxel-wise metrics (Dice) and lesion-wise metrics (L-Recall, L-PPV, and L-F1-score) reported for both training (A) and testing sets (B).

A	Our model	sct_deepseg_lesion (T2w and T2*w)	sct_deepseg (axial T2w)	sct_deepseg (MP2RAGE-UNIT1)
Dice ↑	**0.72 ± 0.28**†	0.39 ± 0.38	0.51 ± 0.40	0.23 ± 0.38
L-Recall ↑	**0.85 ± 0.27**†	0.64 ± 0.43	0.72 ± 0.41	0.43 ± 0.48
L-PPV ↑	**0.86 ± 0.30**†	0.49 ± 0.45	0.55 ± 0.44	0.22 ± 0.39
L-F1-score ↑	**0.81 ± 0.30**†	0.45 ± 0.42	0.55 ± 0.43	0.23 ± 0.39
B	Our model	sct_deepseg_lesion (T2w and T2*w)	sct_deepseg (axial T2w)	sct_deepseg (MP2RAGE-UNIT1)
Dice ↑	**0.63 ± 0.34**†	0.36 ± 0.37	0.48 ± 0.40	0.23 ± 0.38
L-Recall ↑	**0.81 ± 0.31**†	0.63 ± 0.43	0.71 ± 0.41	0.44 ± 0.48
L-PPV ↑	**0.76 ± 0.37**†	0.44 ± 0.45	0.52 ± 0.45	0.22 ± 0.39
L-F1-score ↑	**0.71 ± 0.36**†	0.42 ± 0.42	0.51 ± 0.43	0.23 ± 0.39

The top line indicates the method names and the contrasts for which they were designed, listed in parentheses. The proposed model consistently outperforms contrast-specific methods in both train and test data. † indicates significant differences ($p < 0.01$). ↑ Higher score is better.

representation within the dataset. While T2w and T2*w acquisitions dominate the training and evaluation sets, several underrepresented contrasts, including UNIT1 and STIR, still achieved competitive performance, suggesting that the model captures contrast-invariant lesion characteristics rather than relying on contrast-specific cues.

Comparison to baseline methods

The proposed model consistently outperformed existing methods across all evaluated metrics (see Table 4). On the test set, our model achieved a mean Dice score of 0.63 ± 0.34, compared to 0.36 ± 0.37 for *sct_deepseg T2w_ax*, 0.48 ± 0.40 for *sct_deepseg MP2RAGE*, and 0.23 ± 0.38 for *sct_deepseg_lesion*. Comparable trends were observed for L-Recall, L-PPV, and L-F1-score. Specifically, L-Recall remained high (0.81 ± 0.31) while maintaining balanced L-PPV (0.76 ± 0.37), resulting in a superior L-F1-score (0.71 ± 0.36) relative to all baseline approaches. These findings indicate that the model provides substantial improvements in both voxel-wise and lesion-wise detection compared with existing contrast-specific segmentation strategies.

Table 5 compares model performance across MRI contrasts. The proposed model performed better than existing methods on all contrasts, even on contrasts for which existing models had been specifically trained on.

Likert-type grading by expert neuroradiologists

Figure 4 shows the comparative evaluation of Likert-type scores between manual and predicted segmentations. Global comparison showed a significant difference ($p = 0.01$, Wilcoxon signed-rank test) between manual (3.38 ± 1.28) and predicted

Table 5. Dice score per MRI contrast for our model and existing segmentation methods, reported on training and testing sets.

	Our model		sct_deepseg_lesion (T2w and T2*w)		sct_deepseg (axial T2w)		sct_deepseg (MP2RAGE-UNIT1)	
	Train	Test	Train	Test	Train	Test	Train	Test
PSIR	**0.59 ± 0.32**	**0.51 ± 0.33**	0.04 ± 0.16	0.12 ± 0.32	0.19 ± 0.37	0.13 ± 0.28	0.16 ± 0.19	0.15 ± 0.16
STIR	**0.56 ± 0.32**	**0.73 ± 0.35**	0.31 ± 0.27	0.32 ± 0.32	0.12 ± 0.13	0.09 ± 0.10	0.21 ± 0.40	0.43 ± 0.53
T1w	**0.57 ± 0.31**	**0.42 ± 0.52**	0.07 ± 0.22	0.06 ± 0.09	0.02 ± 0.04	0.01 ± 0.00	0.06 ± 0.22	0.09 ± 0.13
T2*	**0.65 ± 0.24**	**0.61 ± 0.29**	0.46 ± 0.27	0.43 ± 0.28	0.46 ± 0.29	0.49 ± 0.30	0.14 ± 0.34	0.20 ± 0.40
T2w	**0.75 ± 0.28**	**0.64 ± 0.35**	0.46 ± 0.39	0.41 ± 0.38	0.63 ± 0.37	0.59 ± 0.38	0.24 ± 0.42	0.22 ± 0.40
UNIT1	**0.77 ± 0.20**	**0.67 ± 0.26**	0.15 ± 0.33	0.16 ± 0.36	0.05 ± 0.16	0.08 ± 0.24	0.34 ± 0.26	0.35 ± 0.24

The top line indicates the method names and the contrasts for which they were designed, listed in parentheses. The proposed model consistently achieves higher Dice across contrasts, even outperforming baseline models on their specific contrast.

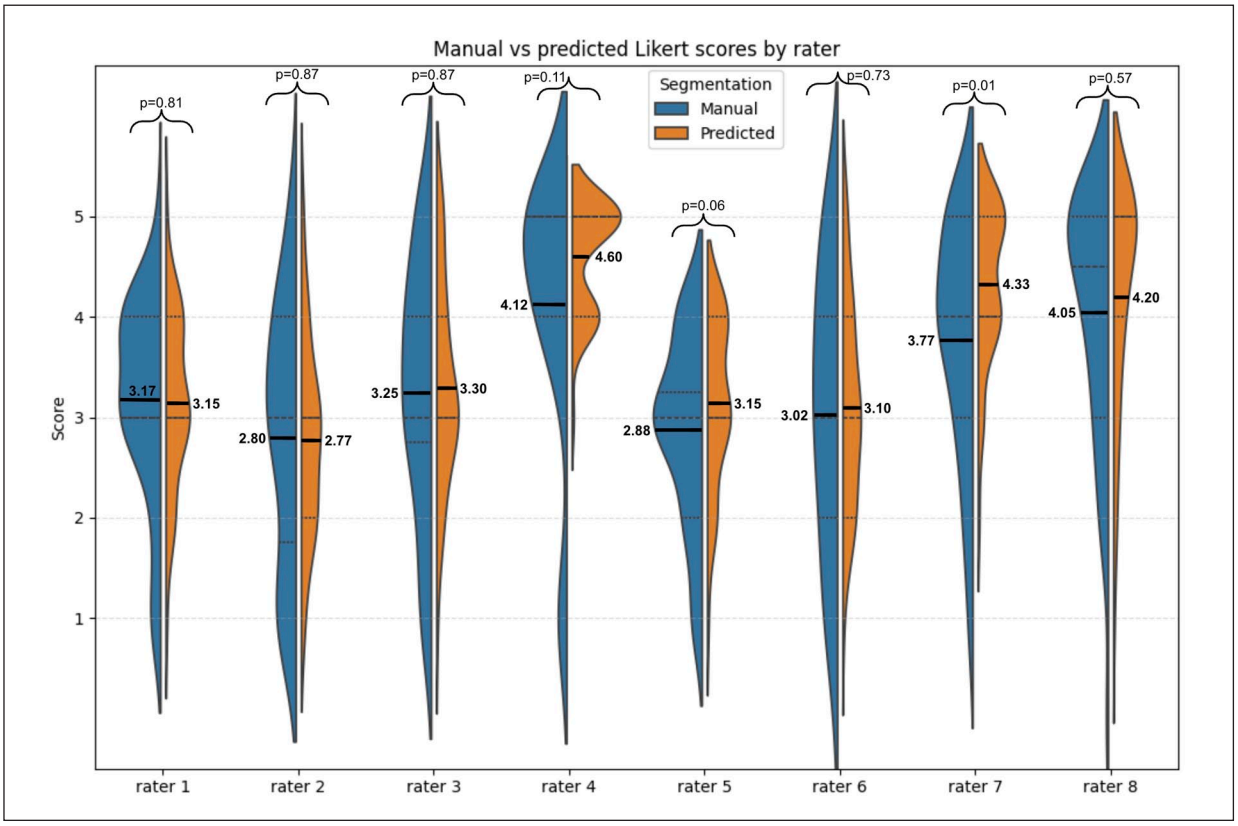

Figure 4. Violin plot of Likert-type scores comparing manual and predicted segmentations across six raters.
Blind grading was performed on a 5-point scale (1 = poor, 5 = excellent). Predicted segmentations received similar or higher scores compared to manual annotations, supporting their clinical acceptability. *P*-values corresponding to Wilcoxon signed-rank tests are reported at the top of each rater. Thick horizontal lines represent the mean, large dashes the median and small dashes the quartiles.

segmentation (3.58 ± 1.09). For most raters, no significant difference was observed between scores assigned to predicted segmentations and manual annotations. Exceptions were noted for rater 7, who assigned significantly higher scores to the predicted vs the manual segmentations (4.33 vs 3.77, $p = 0.01$). In addition to an overall higher mean score, the variance of Likert-type ratings was lower for predicted segmentations, suggesting greater consistency across raters in their assessment of model outputs.

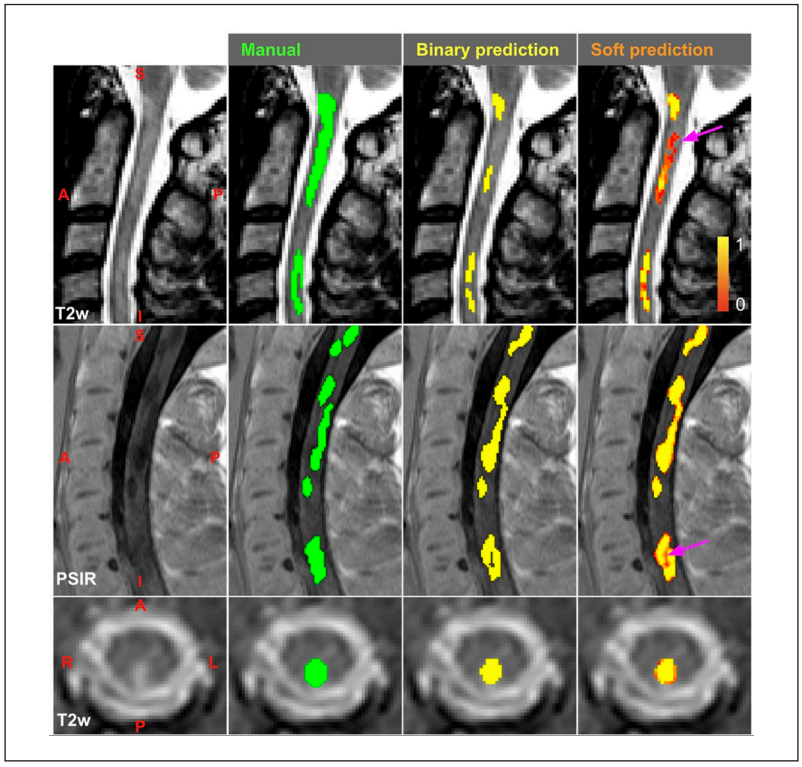

Figure 5. Visual comparison of manual segmentation, predicted soft segmentation and predicted binary segmentation on sagittal T2w, sagittal PSIR and axial T2w (from top to bottom).
From left to right: original image, manual reference segmentation (green), binary prediction (yellow), and soft segmentation (color-coded from yellow = high probability to red = lower probability). In some regions where the binary model predicted no lesion, the soft segmentation predicted values within a ~0.2 to 0.6 range (purple arrows), suggesting the presence of lesions, albeit with lower certainty, which is clinically relevant information as it helps assess the reliability of the segmentation algorithm.

Rater Kappa agreement was also investigated in Supplemental Appendix C.

Soft segmentation

Soft predictions provide voxel-wise probability estimates of model uncertainty and partial volume, enabling clinicians to select either more exhaustive or more conservative segmentations depending on the clinical context.[39] As illustrated in Figure 5, soft segmentations preserve finer boundary details and highlight subtle lesion regions that are not captured by binary predictions. In clinical settings, soft segmentations could also be used to compute more precise volumes, particularly at the boundaries of the lesions where binary segmentation does not account for partial volume effects.

Discussion

This study introduces a multi-contrast deep learning model for SC MS lesion segmentation. Unlike prior approaches tailored to specific acquisition protocols, the proposed framework was trained and validated on a uniquely diverse multi-site dataset ($n = 23$) encompassing six MRI contrasts and acquired on various manufacturers. This design enabled the model to demonstrate strong generalizability and to achieve state-of-the-art performance for SC MS lesion segmentation.

Dataset considerations

The multi-site nature of the dataset presents both opportunities and limitations. A major challenge arises from the strong imbalance in MRI contrast distributions: while T2w dominates clinical practice, newer sequences such as PSIR, STIR, and MP2RAGE remain underrepresented, influencing model learning and performance. Regarding the external test set, further experimentation is necessary to evaluate the model's performance on a broader range of contrasts, particularly those currently underrepresented within the training dataset.

Beyond considerations of model performance, the dataset used in the current study can shed additional light on the best MR contrasts to use for lesion detection. For example, computing metrics such as lesion-to-background contrast and inter-rater variability could inform current efforts in standardizing spinal cord imaging protocols.[12,14] Specifically, small confluent lesions tend to lead to high rater variability as they can be interpreted as many small lesions or a larger confluent lesion. Also, highly anisotropic (such as STIR or PSIR in our study) made interpretation more complex compared to isotropic acquisitions (UNIT1 in our study), leading to more variable interpretations. The latter could explain the relatively high performance of the model compared to the relatively small amount of UNIT1 scans.

Furthermore, the absence of a unified segmentation protocol across sites posed additional challenges. While the variability in ground truth labels negatively impacted model performance,[40] it also represented the "real world" variability in ground truth labels and allowed the model to learn an average of site-specific annotation biases.

Proton-density weighted imaging was a priori excluded from this work. Lesion boundaries were extremely difficult to delineate due to the subtle contrast between abnormal and healthy tissue. Including such data would have compromised the performance of the model.

Although the vast majority of the dataset consists of 3 T acquisitions, the model design and evaluation confer robustness to variations in spatial resolution and image quality. This suggests that the proposed model should perform well on other field strengths (1.5 T and 7 T), although it remains to be further tested extensively.

Model training strategies and performances

The results of this study indicate that a relatively simple architecture, when combined with a well-engineered training pipeline such as nnU-Net, can achieve superior performance compared to more recent state-of-the-art models.[41]

To mitigate the class imbalance between lesion and non-lesion voxels, we investigated an alternative strategy involving training exclusively on volumes containing lesions. Although this approach increased L-Recall, it came at the cost of reduced L-PPV, ultimately lowering overall performance.

An additional consideration is whether developing models across all imaging modalities is necessary, given that some contrasts provide superior lesion sensitivity.[42,43] Although these comparisons are to some extent dependent on factors other than the pulse sequence, our findings indicate that, by leveraging complementary contrasts and larger, more heterogeneous datasets, a multimodal framework enhances generalization and reduces modality-specific biases.

To enhance model robustness and generalizability, various training configurations were tested. Aggressive data augmentation, such as done by,[44] facilitated faster performance improvements during early training epochs, but showed performance plateauing slightly below that of standard data augmentation, while increasing training time by a factor of three to four. Batch sampling strategies, such as those done in,[45] designed to up-weight underrepresented modalities, did not enhance segmentation accuracy and may have inadvertently promoted overfitting by imposing unrealistic class distributions.

Qualitative inspection of segmentation results revealed persisting challenges in lesion delineation. In particular, raters and models alike struggled with whether to delineate boundaries as one large lesion or several smaller ones. Variable visualization of the central canal in the SC further complicated interpretation, as it could easily be mistaken for a lesion (see the orange arrow in the PSIR image of Figure 2). This ambiguity becomes particularly problematic in longitudinal analyses, where lesion "fusion" may confound clinical interpretation of lesion growth or stability.

Performance remained skewed toward contrasts with higher representation in the training data, although strong performance was still achieved for the relatively lower-represented UNIT1 scans. The external test set only contained T2w and T2*w scans. Consequently, broader validation will require subsequent external testing on underrepresented contrasts. High variability in performance was observed, which can be attributed to multiple factors, including variability in manual annotations, image artifacts, and partial volume effects. When comparing our model to baseline tools, it is important to note that prior methods were trained on subsets of both training and test data, potentially inflating their performance.

Importantly, moderate Dice scores are expected due to high inter- and intra-rater variability. Walsh et al.[26] demonstrated that even expert raters achieve median voxel-wise Dice scores below 0.5 when evaluated

against a senior expert-adjudicated ground truth. In this context, very high Dice values would likely indicate overfitting to a specific rater style rather than improved clinical validity. The performance reported here should therefore be interpreted relative to known human variability rather than absolute segmentation accuracy.

Evaluation metrics

The choice of evaluation strategy in SC MS lesion segmentation remains a debated issue as the most used metric, Dice score, is not suited for small object segmentation with high boundary uncertainty.[46] The 10% IoU threshold to match predictions with reference lesions is somewhat arbitrary, but no consensus exists within the community regarding this threshold. While the Free-response Receiver Operating Characteristic (FROC) may provide a more clinically meaningful assessment of detection performance, it relies on lesion-level probability estimates, which are not produced by the current framework.

The Likert-type ratings were used to provide an independent, expert-based qualitative assessment of clinical acceptability of the predicted segmentations relative to manual annotations. This evaluation aimed to assess perceived segmentation quality and consistency, complementing quantitative metrics that are known to be sensitive to annotation variability in spinal cord lesions. Likert-type evaluations by expert neuroradiologists showed that predicted segmentations were perceived as comparable, or in some cases slightly superior, to manual annotations. This reinforces the clinical relevance of the predictions, showing that despite quantitative metrics being relatively modest, the outputs achieve a level of quality acceptable to experts. Furthermore, Likert-type scores showed reduced variability for predicted vs manual segmentations. This reduced variability could indicate that the automated segmentations may exhibit more uniform quality and clearer lesion delineation, contributing to improved inter-rater agreement compared to manual annotations. Nevertheless, potential bias cannot be excluded, as raters might occasionally recognize the source of the segmentation, which could have influenced their evaluations.

Overall, this tool could contribute to improved detection and quantification of lesions in the spinal cord. As demonstrated in other studies,[25] expert evaluations are improved when incorporating model prediction. Furthermore, this tool could mitigate variability in scan interpretation, as it is not subject to inter-rater variability and exhibits robustness across contrasts.

Declaration of Conflicting Interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Laurent Létourneau-Guillon is supported by a Fonds de Recherche Quebec Sante (FRQ-S)/Fondation de L'Association des Radiologistes du Quebec (FARQ) Junior 1 salary award (https://doi.org/10.69777/311203). Shannon Kolind has received grant support or consulting fees from AbbVie, Biogen, Roche, and Sanofi-Genzyme. B. Mark Keegan: consulting from Moderna, EMD Serono, Tr1X Inc, and book royalties from Oxford University Press. Daniel S. Reich—research funding from Abata and Sanofi. Massimo Filippi is Editor-in-Chief of the *Journal of Neurology*, Associate Editor of *Human Brain Mapping*, *Neurological Sciences*, and *Radiology*; received compensation for consulting services from Alexion, Almirall, Biogen, Merck, Novartis, Roche, Sanofi; speaking activities from Bayer, Biogen, Celgene, Chiesi Italia SpA, Eli Lilly, Genzyme, Janssen, Merck-Serono, Neopharmed Gentili, Novartis, Novo Nordisk, Roche, Sanofi, Takeda, and TEVA; participation in Advisory Boards for Alexion, Biogen, Bristol-Myers Squibb, Merck, Novartis, Roche, Sanofi, Sanofi-Aventis, Sanofi-Genzyme, Takeda; scientific direction of educational events for Biogen, Merck, Roche, Celgene, Bristol-Myers Squibb, Lilly, Novartis, Sanofi-Genzyme; he receives research support from Biogen Idec, Merck-Serono, Novartis, Roche, the Italian Ministry of Health, the Italian Ministry of University and Research, and Fondazione Italiana Sclerosi Multipla. Maria A. Rocca received consulting fees from Biogen, Bristol-Myers Squibb, Roche, and speaker honoraria from Alexion, Biogen, Bristol-Myers Squibb, Celgene, Horizon Therapeutics Italy, Merck-Serono SpA, Mitsubishi-Tanabe Pharma, Neuraxpharm, Novartis, Roche, Sandoz, and Sanofi. She receives research support from the MS Society of Canada, the Italian Ministry of Health, the Italian Ministry of University and Research, and Fondazione Italiana Sclerosi Multipla. She is an Associate Editor for *Multiple Sclerosis and Related Disorders*, and Associate Co-Editor for Europe and Africa for Multiple Sclerosis Journal. O. Ciccarelli is an NIHR Research Professor (RP-2017-08-ST2-004); she has been a member of an independent DSMB for Novartis; she acted as a consultant for Merck, Biogen, and Lundbeck; she is Deputy Editor of Neurology®, for which she receives an honorarium; and has received research grant support from the MS Society of Great Britain and Northern Ireland, the NIHR UCLH Biomedical Research Centre, the Rosetree Trust, the National MS Society, and the NIHR-HTA. All other authors report no relevant disclosures. Tobias Granberg—Awardee of the Grant for Multiple Sclerosis Innovation (GMSI) funded by Merck. Dr.

Bakshi has received speaking honoraria from EMD Serono, advisory board consulting fees from Sanofi, and research support from Novartis. Constantina Andrada Treaba has received research support from Genentech. Kristin O'Grady—Dr. O'Grady's research is supported in part by the National Multiple Sclerosis Society under award number JF-2306-41540. The authors not mentioned in this section declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by the Canada Research Chair in Quantitative Magnetic Resonance Imaging [CRC-2020-00179], the Canadian Institute of Health Research [PJT-190258, PJT-203803], the Canada Foundation for Innovation [32454, 34824], the Fonds de Recherche du Québec—Santé [322736, 324636], the Natural Sciences and Engineering Research Council of Canada [RGPIN-2019-07244], the Canada First Research Excellence Fund (IVADO and TransMedTech), the Courtois NeuroMod project, the Quebec BioImaging Network [5886, 35450], INSPIRED (Spinal Research, UK; Wings for Life, Austria; Craig H. Neilsen Foundation, USA), Mila—Tech Transfer Funding Program. This research is supported in part by the FRQNT Strategic Clusters Program (Center UNIQUE—Centre de recherche Neuro-IA du Québec) and Canada Research Chair in Shape Analysis in Medical Imaging. These works were supported by a grant from the Fonds de recherche du Québec (https://doi.org/10.10.69777/370582). This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH). The contributions of the NIH authors are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services. CanProCo funders: MS Canada, Biogen Canada, Brain Canada Foundation, Hoffmann-La Roche Limited, and Government of Alberta.

Ethical Considerations

Data acquisition and storage at each site were authorized by the local IRB. Data were then aggregated at the managing site, under Polytechnique Montréal's IRB (CER-2324-26-D).

Consent to Participate

Research participants in their respective imaging sites signed a consent form as per the local IRB regulations.

Consent for Publication

Not applicable.

ORCID iDs

Pierre-Louis Benveniste https://orcid.org/0009-0003-3122-1957
David Araujo https://orcid.org/0000-0002-1600-6138
Dumitru Fetco https://orcid.org/0000-0003-1335-8274
Masaaki Hori https://orcid.org/0000-0002-1791-8032
Bertrand Audoin https://orcid.org/0000-0002-9860-7657
Rohit Bakshi https://orcid.org/0000-0001-8601-5534
Elise Bannier https://orcid.org/0000-0002-8942-7486
Daniel Blezek https://orcid.org/0000-0002-6498-6273
Jean-Christophe Brisset https://orcid.org/0000-0002-7947-3622
Virginie Callot https://orcid.org/0000-0003-0850-1742
Michelle Chen https://orcid.org/0009-0005-7711-5492
Olga Ciccarelli https://orcid.org/0000-0001-7485-1367
Gilles Edan https://orcid.org/0000-0002-9641-6734
Massimo Filippi https://orcid.org/0000-0002-5485-0479
Tobias Granberg https://orcid.org/0000-0001-6700-1022
Cristina Granziera https://orcid.org/0000-0002-4917-8761
Christopher C. Hemond https://orcid.org/0000-0002-2408-4638
B. Mark Keegan https://orcid.org/0000-0002-2880-935X
Anne Kerbrat https://orcid.org/0000-0002-4530-3553
Jan Kirschke https://orcid.org/0000-0002-7557-0003
Shannon Kolind https://orcid.org/0000-0003-1362-1968
Lisa Eunyoung Lee https://orcid.org/0000-0002-8334-3740
Julian McGinnis https://orcid.org/0009-0000-2224-7600
Nilser Laines Medina https://orcid.org/0000-0003-0677-8991
Mark Mühlau https://orcid.org/0000-0002-9545-2709
Govind Nair https://orcid.org/0000-0003-3725-615X

Kristin P. O'Grady https://orcid.org/0000-0002-4550-2026
Jiwon Oh https://orcid.org/0000-0001-5519-6088
Russell Ouellette https://orcid.org/0000-0001-9217-1445
Daniel S. Reich https://orcid.org/0000-0002-2628-4334
Maria A. Rocca https://orcid.org/0000-0003-2358-4320
Seth A. Smith https://orcid.org/0000-0002-4168-562X
Leszek Stawiarz https://orcid.org/0000-0003-1018-3763
Roger Tam https://orcid.org/0000-0002-4593-2587
Anthony Traboulsee https://orcid.org/0000-0002-0351-9639
Constantina Andrada Treaba https://orcid.org/0000-0001-8260-207X
Paola Valsasina https://orcid.org/0000-0001-5390-2655
Marios Yiannakas https://orcid.org/0000-0003-4986-446X
Julien Cohen-Adad https://orcid.org/0000-0003-3662-9532

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Supplemental Material

Supplemental material for this article is available online.

References

1. Walton C, King R, Rechtman L, et al. Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS, third edition. *Mult Scler* 2020; 26(14): 1816–1821.
2. Waldman AD, Catania C, Pisa M, et al. The prevalence and topography of spinal cord demyelination in multiple sclerosis: A retrospective study. *Acta Neuropathol* 2024; 147: 51.
3. McDonald WI, Compston A, Edan G, et al. Recommended diagnostic criteria for multiple sclerosis: Guidelines from the international panel on the diagnosis of multiple sclerosis. *Ann Neurol* 2001; 50(1): 121–127.
4. Thompson AJ, Banwell BL, Barkhof F, et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol* 2018; 17: 162–173.
5. Montalban X, Lebrun-Frénay C, Oh J, et al. Diagnosis of multiple sclerosis: 2024 revisions of the McDonald criteria. *Lancet Neurol* 2025; 24: 850–865.
6. Kerbrat A, Gros C, Badji A, et al. Multiple sclerosis lesions in motor tracts from brain to cervical cord: Spatial distribution and correlation with disability. *Brain* 2020; 143: 2089–2105.
7. Jackson-Tarlton CS, Flanagan EP, Messina SA, et al. Progressive motor impairment from "critical" demyelinating lesions of the cervicomedullary junction. *Mult Scler* 2023; 29(1): 74–80.
8. Ahmad R, Jackson-Tarlton CS, Flanagan EP, et al. Critical demyelinating lesions in progressive multiple sclerosis: A prospective observational study. *J Neurol* 2025; 272: 677.
9. Demortière S, Lehmann P, Pelletier J, et al. Improved cervical cord lesion detection with 3D-MP2RAGE sequence in patients with Multiple Sclerosis. *AJNR Am J Neuroradiol* 2020; 41(6): 1131–1134.
10. Stroman PW, Wheeler-Kingshott C, Bacon M, et al. The current state-of-the-art of spinal cord imaging: Methods. *Neuroimage* 2014; 84: 1070–1081.
11. Saslow L, Li DKB, Halper J, et al. An international standardized magnetic resonance imaging protocol for diagnosis and follow-up of patients with multiple sclerosis: Advocacy, dissemination, and implementation strategies. *Int J MS Care* 2020; 22(5): 226–232.
12. Cohen-Adad J, Alonso-Ortiz E, Abramovic M, et al. Generic acquisition protocol for quantitative MRI of the spinal cord. *Nat Protoc* 2021; 16(10): 4611–4632.
13. Barkhof F, Reich DS, Oh J, et al. 2024 MAGNIMS-CMSC-NAIMS consensus recommendations on the use of MRI for the diagnosis of multiple sclerosis. *Lancet Neurol* 2025; 24(10): 866–879.
14. Wattjes MP, Ciccarelli O, Reich DS, et al. 2021 MAGNIMS-CMSC-NAIMS consensus recommendations on the use of MRI in patients with multiple sclerosis. *Lancet Neurol* 2021; 20(8): 653–670.
15. Guttmann CR, Kikinis R, Anderson MC, et al. Quantitative follow-up of patients with multiple sclerosis using MRI: Reproducibility. *J Magn Reson Imaging* 1999; 9(4): 509–518.
16. Kaur A, Kaur L and Singh A. State-of-the-art segmentation techniques and future directions for multiple sclerosis brain lesions. *Arch Computat Methods Eng* 2021; 28: 951–977.
17. Aslani S, Dayan M, Storelli L, et al. Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. *Neuroimage* 2019; 196: 1–15.
18. Valverde S, Cabezas M, Roura E, et al. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *Neuroimage* 2017; 155: 159–168.
19. Havaei M, Guizard N, Chapados N, et al. HeMIS: Hetero-modal image segmentation. In: Ourselin S,

Joskowicz L, Sabuncu MR, et al. (eds) *Medical image computing and computer-assisted intervention—MICCAI 2016*. Cham: Springer International Publishing, 2016, pp. 469–477.

20. Essa E, Aldesouky D, Hussein SE, et al. Neuro-fuzzy patch-wise R-CNN for multiple sclerosis segmentation. *Med Biol Eng Comput* 2020; 58(9): 2161–2175.

21. Gessert N, Bengs M, Krüger J, et al. 4D deep learning for multiple sclerosis lesion activity segmentation 2020, http://arxiv.org/abs/2004.09216

22. Kamraoui RA, Ta VT, Tourdias T, et al. DeepLesionBrain: Towards a broader deep-learning generalization for multiple sclerosis lesion segmentation. *Med Image Anal* 2022; 76: 102312.

23. Wiltgen T, McGinnis J, Schlaeger S, et al. LST-AI: A deep learning ensemble for accurate MS lesion segmentation. medRxiv. Epub ahead of print 11 March 2024. DOI: 10.1101/2023.11.23.23298966.

24. Zeng C, Gu L, Liu Z, et al. Review of deep learning approaches for the segmentation of multiple sclerosis lesions on brain MRI. *Front Neuroinform* 2020; 14: 610967.

25. Lodé B, Hussein BR, Meurée C, et al. Evaluation of a deep learning segmentation tool to help detect spinal cord lesions from combined T2 and STIR acquisitions in people with multiple sclerosis. *Eur Radiol* 2025; 35(10): 5954–5964.

26. Walsh R, Meurée C, Kerbrat A, et al. Expert variability and deep learning performance in spinal cord lesion segmentation for multiple sclerosis patients. In: *2023 IEEE 36th international symposium on computer-based medical systems (CBMS)*, L'Aquila, 22–24 June 2023, pp. 463–470. New York: IEEE.

27. Polattímur R, Dandil E, Yildirim MS, et al. FractalSpiNet: Fractal-based U-net for automatic segmentation of cervical spinal cord and MS lesions in MRI. *IEEE Access* 2024; 12: 110955–110976.

28. Gros C, De Leener B, Badji A, et al. Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks. *Neuroimage* 2019; 184: 901–915.

29. Benveniste P-L, Valošek J, Chen M, et al. Automatic segmentation of spinal cord multiple sclerosis lesions across multiple sites, contrasts and vendors. In: *32th annual meeting of ISMRM*, Singapore, 4–9 May 2024.

30. Medina NL, Mchinda S, Testud B, et al. Automatic multiple sclerosis lesion segmentation in the spinal cord on 3T and 7T MP2RAGE images. In: *33th annual meeting of ISMRM*, Honolulu, HI, 10–15 May 2025.

31. Naga Karthik E, McGinnis J, Wurm R, et al. Automatic segmentation of spinal cord lesions in MS: A robust tool for axial T2-weighted MRI scans. *Imaging Neurosci* 2025; 3: IMAG.a.45.

32. Karimi D, Dou H, Warfield SK, et al. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Med Image Anal* 2020; 65: 101759.

33. Oktay O, Schlemper J, Folgoc LL, et al. Attention U-Net: Learning where to look for the pancreas. arXiv. Epub ahead of print 20 May 2018. DOI: 10.48550/arXiv.1804.03999.

34. Huang Z, Wang H, Deng Z, et al. STU-Net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. arXiv. Epub ahead of print 13 April 2023. DOI: 10.48550/arXiv.2304.06716.

35. Ulrich C, Wald T, Isensee F, et al. Large scale supervised pretraining for traumatic brain injury segmentation. arXiv. Epub ahead of print 9 April 2025. DOI: 10.48550/arXiv.2504.06741.

36. Roy S, Koehler G, Ulrich C, et al. MedNeXt: Transformer-driven scaling of ConvNets for medical image segmentation. In: Greenspan H (ed.) *Lecture notes in computer science*. Cham: Springer Nature, 2023, pp. 405–415.

37. Isensee F, Jaeger PF, Kohl SAA, et al. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* 2021; 18(2): 203–211.

38. Keskar NS, Mudigere D, Nocedal J, et al. On large-batch training for deep learning: Generalization gap and sharp minima. arXiv. Epub ahead of print 9 February 2016. DOI: 10.48550/arXiv.1609.04836.

39. Gros C, Lemay A and Cohen-Adad J. SoftSeg: Advantages of soft versus binary training for image segmentation. *Med Image Anal* 2021; 71: 102038.

40. Arpit D, Jastrzebski S, Ballas N, et al. A closer look at memorization in deep networks. *ICML* 2017; 70: 233–242.

41. Isensee F, Wald T, Ulrich C, et al. NnU-net revisited: A call for rigorous validation in 3D medical image segmentation. In: Linguraru MG (ed.) *Lecture notes in computer science*. Cham: Springer Nature, 2024, pp. 488–498.

42. Peters S, Neves FB, Huhndorf M, et al. Detection of spinal cord multiple sclerosis lesions using a 3D-PSIR sequence at 1.5 T. *Clin Neuroradiol* 2024; 34(2): 403–410.

43. Galler S, Stellmann JP, Young KL, et al. Improved lesion detection by using axial T2-weighted MRI with full spinal cord coverage in multiple sclerosis. *AJNR Am J Neuroradiol* 2016; 37(5): 963–969.

44. Warszawer Y, Molinier N, Valošek J, et al. *TotalSpineSeg: Robust segmentation and labeling*

of vertebrae, intervertebral discs, spinal cord, and spinal canal in MRI images using nnU-Net and iterative algorithm. Geneva: Zenodo, 2024.

45. Ulrich C, Isensee F, Wald T, et al. MultiTalent: A multi-dataset approach to medical image segmentation. In: Greenspan H (ed.) *Lecture notes in computer science*. Cham: Springer Nature, 2023, pp. 648–658.

46. Maier-Hein L, Reinke A, Godau P, et al. Metrics reloaded: Recommendations for image analysis validation. *Nat Methods* 2024; 21(2): 195–212.