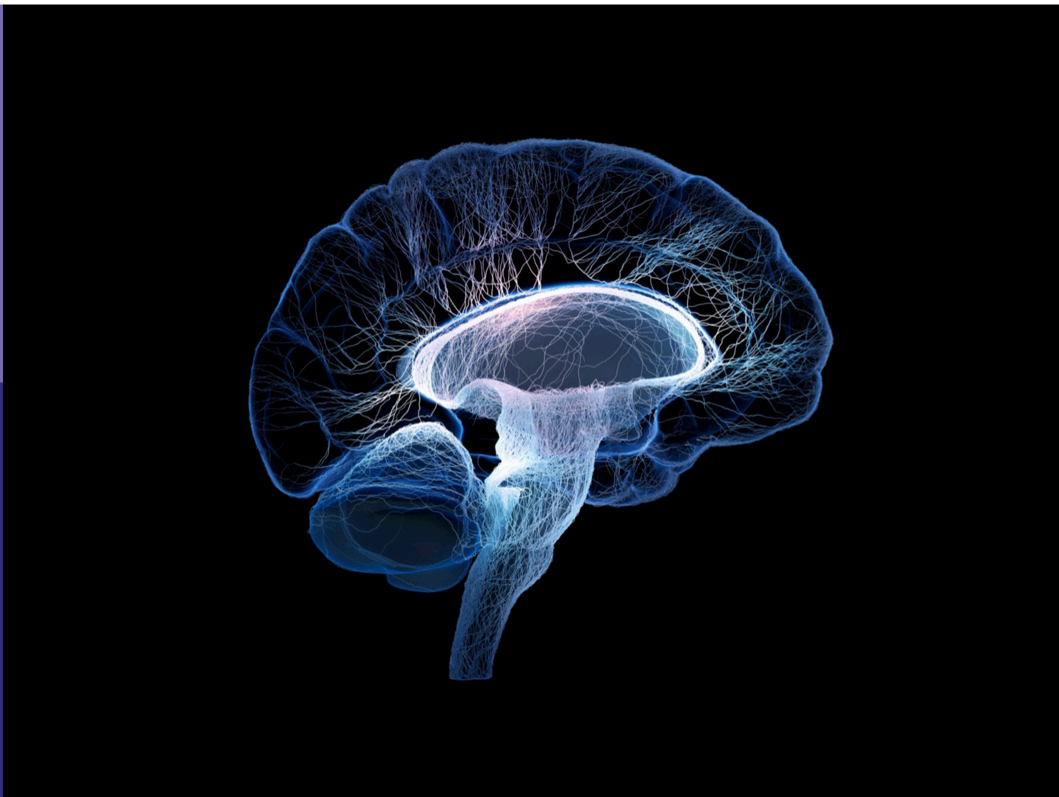
Research Topics

Automatic methods for multiple sclerosis new lesions detection and segmentation

Edited by
Olivier Commowick, Benoit Combès, Frederic Cervenansky
and Michel Dojat

Published in
Frontiers in Neuroscience
Frontiers in Neuroimaging



ISSN 1664-8714
ISBN 978-2-83252-037-6
DOI 10.3389/978-2-83252-037-6

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Automatic methods for multiple sclerosis new lesions detection and segmentation

Topic editors

Olivier Commowick — Inria Rennes - Bretagne Atlantique Research Centre, France
Benoit Combès — Inria Rennes - Bretagne Atlantique Research Centre, France
Frederic Cervenansky — Université Claude Bernard Lyon 1, France
Michel Dojat — Institut National de la Santé et de la Recherche Médicale (INSERM), France

Citation

Commowick, O., Combès, B., Cervenansky, F., Dojat, M., eds. (2023). *Automatic methods for multiple sclerosis new lesions detection and segmentation*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83252-037-6

Table of contents

04 **Editorial: Automatic methods for multiple sclerosis new lesions detection and segmentation**
Olivier Commowick, Benoît Combès, Frédéric Cervenansky and Michel Dojat

08 **New MS lesion segmentation with deep residual attention gate U-Net utilizing 2D slices of 3D MR images**
Beytullah Sarica and Dursun Zafer Seker

22 **Triplanar U-Net with lesion-wise voting for the segmentation of new lesions on longitudinal MRI studies**
Sebastian Hitziger, Wen Xin Ling, Thomas Fritz, Tiziano D'Albis, Andreas Lemke and Joana Grilo

35 **Longitudinal detection of new MS lesions using deep learning**
Reda Abdellah Kamraoui, Boris Mansencal, José V. Manjon and Pierrick Coupé

49 **Image registration and appearance adaptation in non-correspondent image regions for new MS lesions detection**
Julia Andresen, Hristina Uzunova, Jan Ehrhardt, Timo Kepp and Heinz Handels

59 **Evaluating the use of synthetic T1-w images in new T2 lesion detection in multiple sclerosis**
Liliana Valencia, Albert Clèrigues, Sergi Valverde, Mostafa Salem, Arnau Oliver, Àlex Rovira and Xavier Lladó

70 **New lesion segmentation for multiple sclerosis brain images with imaging and lesion-aware augmentation**
Berke Doga Basaran, Paul M. Matthews and Wenjia Bai

84 **New multiple sclerosis lesion segmentation and detection using pre-activation U-Net**
Pooya Ashtari, Berardino Barile, Sabine Van Huffel and Dominique Sappey-Marinier

93 **Online hard example mining vs. fixed oversampling strategy for segmentation of new multiple sclerosis lesions from longitudinal FLAIR MRI**
Marius Schmidt-Mengin, Théodore Soulier, Mariem Hamzaoui, Arya Yazdan-Panah, Benedetta Bodini, Nicholas Ayache, Bruno Stankoff and Olivier Colliot

103 **Improving the detection of new lesions in multiple sclerosis with a cascaded 3D fully convolutional neural network approach**
Mostafa Salem, Marwa Ahmed Ryan, Arnau Oliver, Khaled Fathy Hussain and Xavier Lladó

116 **A unified framework for focal intensity change detection and deformable image registration. Application to the monitoring of multiple sclerosis lesions in longitudinal 3D brain MRI**
Eléonore Dufresne, Denis Fortun, Stéphane Kremer and Vincent Noblet

frontiers | Frontiers in Neuroscience

TYPE Editorial
PUBLISHED 14 March 2023
DOI 10.3389/fnins.2023.1176625

Editorial: Automatic methods for multiple sclerosis new lesions detection and segmentation

Olivier Commowick[1†], Benoît Combès[1], Frédéric Cervenansky[2] and Michel Dojat[3]*

[1]Empenn INSERM U1228, CNRS UMR6074, Inria, University of Rennes I, Rennes, France, [2]Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1206, Lyon, France, [3]Univ Grenoble Alpes, Inserm, U1216, CHU Grenoble Alpes, Grenoble Institut Neurosciences, GIN, Grenoble, France

KEYWORDS

neurodegenerative diseases, MRI, image processing, deep learning, neurology - clinical

Editorial on the Research Topic
Automatic methods for multiple sclerosis new lesions detection and segmentation

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Vince D. Calhoun,
Georgia State University, United States

*CORRESPONDENCE
Michel Dojat
✉ michel.dojat@inserm.fr

†Deceased

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 28 February 2023
ACCEPTED 01 March 2023
PUBLISHED 14 March 2023

CITATION
Commowick O, Combès B, Cervenansky F and Dojat M (2023) Editorial: Automatic methods for multiple sclerosis new lesions detection and segmentation. *Front. Neurosci.* 17:1176625. doi: 10.3389/fnins.2023.1176625

Multiple Sclerosis (MS) is a chronic inflammatory disease of the central nervous system (CNS) affecting more than half a million persons in Europe, with a prevalence rate of 83 per 100,000 with higher rates in northern countries and a female/male ratio around 2.0 (Pugliatti et al., 2006). Today, conventional MR imaging (MRI) is widely used for the patient follow-up, the monitoring of the therapy effects, and more generally in a perspective of personalized medicine, for the understanding of the individual MS progression (Thompson et al., 2018). One of the major challenges in using MRI for MS is the segmentation of lesions whose number, location and appearance at a given time point, are crucial indicators for diagnostic and to tailor treatment to the specific individual disease's evolution.

To cope with inter- and intra-observer variability and reduce the burden and complexity of lesions identification for clinicians, a large number of techniques have been proposed in the literature for the automatic segmentation of MS lesions (see Garcia-Lorenzo et al., 2013; Valverde et al., 2017; Danelakis et al., 2018 for reviews). Several challenges have been proposed to evaluate the performances of these methods (e.g., Carass et al., 2017; Commowick et al., 2021 to cite the most recent ones). Moreover, recently Bonacchi et al. (2022) proposed an overview of Artificial Intelligence applications for MS clinical practice.

A growing literature focuses on the delineation of new MS lesions on T2/FLAIR occurring between two consecutive exams. Detecting the apparition of new MS lesions is of central interest in clinical practice. Indeed, while the palette of Disease Modifying Drugs (DMDs) approved for MS has presently an unknown impact on the compartmentalized neurodegenerative process within the CNS, they aim to substantially reduce, or even stop, the accumulation of new lesions. Consequently, the assessment of such an accumulation allows the clinician to monitor the efficiency of a given DMD on each patient it follows, and therefore to consider a change of treatment in case of insufficient efficiency. Moreover, there is a direct link between accumulation of new lesions and increasing handicap (Sormani et al., 2013). Automating the detection of these new lesions or helping clinicians to identify them would therefore be a major advance for evaluating the patient disease progression and response to treatment.

In 2021, we launched a MICCAI challenge, MSSEG-II (see https://www.ofsep.org/fr/etudes/msseg-ii-challenge-miccai-2021), to compare automated solutions for this specific task i.e., the detection of new lesions appearing at the second time point of two T2/FLAIR images of the patient. For that purpose, we used a large database: 100 patients, each with two time points, the time between the two time points varying between 1 and 3 years. Data were extracted from the national OFSEP cohort (Vukusic et al., 2020), the national French MS registry (https://clinicaltrials.gov/ct2/show/NCT03603457), with 3D FLAIR images from different centers and scanners (15 different scanners in total) using the OFSEP specific protocol (Cotton et al., 2015; Brisset et al., 2020). Only 3D FLAIR images—that is the mostly used clinical sequence for MS brain—were considered. As in our previous challenge (Commowick et al., 2021), the evaluation of solutions was performed on the dedicated FLI-IAM infrastructure (https://www.francelifeimaging.fr/en/about/noeuds/iam/), which comprises Shanoir, a web-oriented solution for imaging data storage and sharing for preclinical and clinical research studies (Barillot et al., 2016; Kain et al., 2020); and the VIP platform (Glatard et al., 2013) for the execution of the corresponding docker of each image processing algorithm/pipeline on EGI infrastructures (https://www.egi.eu/). The use of FLI-IAM allows to automate the competition's process through a sustainable framework and remove the potential biases (e.g., challengers manually optimizing their parameters for each provided case). The ground truth was defined based on the manual delineation, using ITK Snap, of the 100 cases by four neuroradiologists with an MS expertise. Then, a consensus was formed in two steps: a senior expert neuroradiologist examined and confirmed (or declined) disputed lesions among the experts; then a fusion using the STAPLE (Warfield et al., 2004) algorithm was performed. This consensus was then the reference for the evaluation procedure. Forty cases were provided to challengers (e.g., for algorithm training) and 60 cases for algorithm testing. The manual segmentations were provided with the former and unknown to the challengers for the latter.

The present RT gathers 10 papers about solutions for the automatized detection of new lesions in MS subsequent images. All but one (Dufresne et al.) competed during MSSEG-II challenge and were executed on FLI-IAM infrastructure. They are based on a deep learning approach, the U-net architecture (Ronneberger et al., 2015) with its 2D or 3D versions. We may distinguish two classes of approaches, ones that use exclusively the examples provided by the Miccai challenge organizers and those which introduce additional real (Hitziger et al.) or synthetic (Andresen et al.; Kamraoui et al.; Valencia et al.) datasets. Finally, joint modeling, mixing both a registration and a segmentation task, have been investigated (Andresen et al.; Dufresne et al.; Salem et al.).

Then, Hitziger et al. train a 2D U-net with residual units with axial, coronal and sagittal slices. The corresponding slices from the two time-point volumes are paired and introduced to the system as a two-channel input. The predictions from each orientation are then merged with different strategies. The best performances are obtained for the unanimous voting strategy where lesions are confirmed in each orientation. The gain in performance by introducing additional datasets (25 supplementary patients to the initial 40 patients training set) seems weak.

In the same line, Sarica and Seker propose a 2D U-net solution where the standard plain blocks are replaced by residual units and attention gates are introduced to, respectively, enhance the model performances and focalize on new MS lesions on each 2D slice. A majority voting generates the final 3D binary output.

Similarly, Ashtari et al. introduce residual units, this time in a 3D U-net version and data augmentation methods to improve robustness and generalizability of the obtained model.

Basaran et al. consider the recent 3D U-Net version ("No-NewU-Net") combined with several image preprocessing step brain extraction, bias correction, registration and multiple data augmentation methods.

To overcome the difficulty of a supervised training based on scarce new lesion annotated examples, Kamraoui et al. interestingly propose to first pretrain a 3D U-Net on a large one time-point MS dataset (transfer learning), second to pretrain the model used for time-points by introducing realistic synthetic data, and finally to fine-tune the obtained network with the real two time-points data as provided by MSSEG-II.

To tackle class imbalance between voxels belonging to new lesions or not, Schmidt-Mengin et al. introduce a two-stage training strategy to iteratively define a fixed number of patches (30%) containing lesions. This "online hard example mining" strategy is implemented with two 3D U-Nets applied patch-wise in cascade. Such a strategy, applied for the first time on 3D brain scans, seems to emphasize false positive rate.

Instead of using a unique intensity-based approach, Andresen et al., Salem et al., and Dufresne et al. propose to consider a deformation-based approach. Maps of non-corresponding regions between subsequent images are generated during the registration process. In Andresen et al. such maps are then used by a fully convolutional network to segment new lesions that occur across time. Offset maps with baseline allow exploring morphology appearance of new lesions. New lesions are rare and similarly to the previous paper (Kamraoui et al.) the authors insert synthetic lesions during the network training. In Salem et al. the authors introduce a cascade of two 3D U-net patch-wise fully convolutional neural networks. The first registration network learns the deformation field to register the individual sequence of FLAIR images, while the second performs new lesions segmentation. The latter is fed by registered FLAIR images and the deformation maps. Indeed, the first network allows to filter the majority of non-lesion voxels and reveals the possible new lesion candidates, while the second refine the detection in reducing misclassified voxels. The simultaneous training of registration and segmentation modules improves the performances compared to a sequential learning. Valencia et al. propose to improve the previous results in adding synthetic images. The hypothesis is that the introduction of T1-weighted images (T1w), artificially generated, in addition to the FLAIR images improves new MS lesions detection. They use a generative adversarial network (GAN) with an additional MS FLAIR dataset (136 cases) in order to generate T1w corresponding images. The trained GAN is then used to generate the T1w corresponding to the provided MSSEG-II FLAIR images. They show an improvement of the sensitivity performance compared to the only use of FLAIR images.

TABLE 1 Averaged (patient-wise) score for the four experts.

	New lesion cases ($n = 32$)								No new lesion ($n = 28$)	
	DSC	F1	Sensitivity	Specificity	PPV	ηTP	ηFP	ηFN	ηFP	ηFP
Expert 1	0.629	0.709	0.650	1.000	0.707	6.063	1.281	1.094	0.036	1.453
Expert 2	0.597	0.601	0.526	1.000	0.813	4.500	0.844	2.375	0.000	0.000
Expert 3	0.535	0.637	0.580	1.000	0.760	4.313	1.094	2.500	0.107	3.981
Expert 4	0.459	0.519	0.407	1.000	0.801	4.469	0.594	2.375	0.036	0.623

DSC, dice score; PPV, positive predictive value; ηTP, mean number of true positives; ηFP, mean number of false positive; ηFN, mean number of false negative. The provided sensitivity, precision, and PPV are computed at the voxel scale.

Finally, in Dufresne et al., a different deformation-based approach is proposed where deformable registration and local intensity change detection are jointly estimated as a unified optimization problem solving. The joint method is evaluated on synthetic and real MS datasets and compared to the sequential version, where registration and change detection are performed successively, to demonstrate the performance improvement obtained by the former. Such an optimization approach cannot discriminate between new lesions from evolving lesions. It is interesting to note that this is the only non-Deep Learning-based method presented in this RT.

In Table 1, we provide several indexes for the readers in order to have a flavor of the current performances reached by the different solutions described in this RT compared to human experts.

To conclude, MS new lesions detection and segmentation remain very difficult tasks. Presently, automatic methods can be more sensitive for detecting new lesions, but produce more false positive compare to manual delineation by experts. Thus, in spite of slight persistent differences, performances between automatic solutions and human experts are closer than in the previous challenge (see Commowick et al., 2021). However, in order to be used in clinical routine, several steps need to be completed, such as the integration of computerized solutions in the hospital information flow and the quantification of the uncertainty associated to the automatic lesion detection, in place of the standard binary output, to leverage the clinician's work for obvious lesion and requiring his/her expertise only for difficult cases (Lambert et al., 2022). This will lead to the design of a new family of computerized medical assistants for care improvement.

Data from the MSSEG challenges are available here https://shanoir.irisa.fr/shanoir-ng/welcome and can be used to evaluate new solutions.

Author contributions

All authors listed have made a substantial, direct, intellectual contribution to the work. BC, FC, and MD approved it for publication.

Funding

This work was partly funded by France Life Imaging (grant ANR-11-INBS-0006 from the French Investissements d'Avenir program). It has also been partly supported by a grant (OFSEP) provided by the French State and handled by the Agence Nationale de la Recherche, within the framework of the Investissements d'Avenir program, under the reference ANR-10-COHO-002.

Dedication

This editorial is dedicated to OC, our young and talented colleague who prematurely passed away in December 2022.

Acknowledgments

We thank the French national cohort OFSEP (a French Investissements d'Avenir program) and particularly the imaging group inside this cohort consortium for their constant support, fruitful discussions on the challenge, and providing the MR images.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Barillot, C., Bannier, E., Commowick, O., Corouge, I., Baire, A., Fakhfakh, I., et al. (2016). Shanoir: applying the software as a service distribution model to manage brain imaging research repositories. *Front. ICT* 3, 25. doi: 10.3389/fict.2016.00025

Bonacchi, R., Filippi, M., and Rocca, M. A. (2022). Role of artificial intelligence in MS clinical practice. *Neuroimage Clin.* 35, 103065. doi: 10.1016/j.nicl.2022.103065

Brisset, J. C., Kremer, S., Hannoun, S., Bonneville, F., Durand-Dubief, F., Tourdias, T., et al. (2020). New OFSEP recommendations for MRI assessment of multiple sclerosis patients: special consideration for gadolinium deposition and frequent acquisitions. *J. Neuroradiol.* 47, 250–258. doi: 10.1016/j.neurad.2020.01.083

Carass, A., Roy, S., Jog, A., Cuzzocreo, J. L., Magrath, E., Gherman, A., et al. (2017). Longitudinal multiple sclerosis lesion segmentation data resource. *Data Brief* 12, 346–350. doi: 10.1016/j.dib.2017.04.004

Commowick, O., Kain, M., Casey, R., Ameli, R., Ferré, J. C., Kerbrat, A., et al. (2021). Multiple sclerosis lesions segmentation from multiple experts: the MICCAI 2016 challenge dataset. *Neuroimage* 244, 118589. doi: 10.1016/j.neuroimage.2021.118589

Cotton, F., Kremer, S., Hannoun, S., Vukusic, S., and Dousset, V. (2015). OFSEP, a nationwide cohort of people with multiple sclerosis: consensus minimal MRI protocol. *J. Neuroradiol.* 42, 133–140. doi: 10.1016/j.neurad.2014.12.001

Danelakis, A., Theoharis, T., and Verganelakis, D. A. (2018). Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging. *Comput. Med. Imaging Graph.* 70, 83–100. doi: 10.1016/j.compmedimag.2018.10.002

Garcia-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D. L., and Collins, D. L. (2013). Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Med. Image Anal.* 17, 1–18. doi: 10.1016/j.media.2012.09.004

Glatard, T., Lartizien, C., Gibaud, B., da Silva, R. F., Forestier, G., Cervenansky, F., et al. (2013). A virtual imaging platform for multi-modality medical image simulation. *IEEE Trans. Med. Imaging* 32, 110–118. doi: 10.1109/TMI.2012.2220154

Kain, M., Bodin, M., Loury, S., Chi, Y., Louis, J., Simon, M., et al. (2020). Small Animal Shanoir (SAS): a cloud-based solution for managing preclinical MR brain imaging studies. *Front. Neuroinformat.* 14, 20. doi: 10.3389/fninf.2020.00020

Lambert, B., Forbes, F., Tucholka, A., Doyle, S., Dehaene, H., and Dojat, M. (2022). *Trustworthy Clinical AI Solutions: A Unified Review of Uncertainty Quantification in Deep-Learning Models for Medical Image Analysis*. Available online at: https://arxiv.org/abs/2210.03736 (accessed March 2023).

Pugliatti, M., Rosati, G., Carton, H., Riise, T., Drulovic, J., Vécsei, L., et al. (2006). The epidemiology of multiple sclerosis in Europe. *Eur. J. Neurol.* 13, 700–722. doi: 10.1111/j.1468-1331.2006.01342.x

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015*, eds N. Navab, J. Hornegger, W. Wells, and A. Frangi (Cham: Springer). doi: 10.1007/978-3-319-24574-4_28

Sormani, M. P., Rio, J., Tintorè, M., Signori, A., Li, D., Cornelisse, P., et al. (2013). Scoring treatment response in patients with relapsing multiple sclerosis. *Mult. Scler.* 19, 605–612. doi: 10.1177/1352458512460605

Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., et al. (2018). Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol.* 17, 162–173. doi: 10.1016/S1474-4422(17)30470-2

Valverde, S., Cabezas, M., Roura, E., Gonzalez-Villa, S., Pareto, D., Vilanova, J. C., et al. (2017). Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *Neuroimage* 155, 159–168. doi: 10.1016/j.neuroimage.2017.04.034

Vukusic, S., Casey, R., Rollot, F., Brochet, B., Pelletier, J., Laplaud, D. A., et al. (2020). Observatoire Français de la Sclérose en Plaques (OFSEP): a unique multimodal nationwide MS registry in France. *Mult. Scler.* 26, 118–122. doi: 10.1177/1352458518815602

Warfield, S., Zou, K., and Wells, W. (2004). Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans. Med. Imaging* 23, 903–921. doi: 10.1109/TMI.2004.828354

TYPE Original Research
PUBLISHED 22 July 2022
DOI 10.3389/fnins.2022.912000

New MS lesion segmentation with deep residual attention gate U-Net utilizing 2D slices of 3D MR images

Beytullah Sarica[1]* and Dursun Zafer Seker[2]

[1]Department of Applied Informatics, Graduate School, Istanbul Technical University, Istanbul, Turkey, [2]Department of Geomatics Engineering, Faculty of Civil Engineering, Istanbul Technical University, Istanbul, Turkey

Multiple sclerosis (MS) is an autoimmune disease that causes lesions in the central nervous system of humans due to demyelinating axons. Magnetic resonance imaging (MRI) is widely used for monitoring and measuring MS lesions. Automated methods for MS lesion segmentation have usually been performed on individual MRI scans. Recently, tracking lesion activity for quantifying and monitoring MS disease progression, especially detecting new lesions, has become an important biomarker. In this study, a unique pipeline with a deep neural network that combines U-Net, attention gate, and residual learning is proposed to perform better new MS lesion segmentation using baseline and follow-up 3D FLAIR MR images. The proposed network has a similar architecture to U-Net and is formed from residual units which facilitate the training of deep networks. Networks with fewer parameters are designed with better performance through the skip connections of U-Net and residual units, which facilitate information propagation without degradation. Attention gates also learn to focus on salient features of the target structures of various sizes and shapes. The MSSEG-2 dataset was used for training and testing the proposed pipeline, and the results were compared with those of other proposed pipelines of the challenge and experts who participated in the same challenge. According to the results over the testing set, the lesion-wise F1 and dice scores were obtained as a mean of 48 and 44.30%. For the no-lesion cases, the number of tested and volume of tested lesions were obtained as a mean of 0.148 and 1.488, respectively. The proposed pipeline outperformed 22 proposed pipelines and ranked 8th in the challenge.

KEYWORDS

deep residual learning, U-Net, attention gate, convolutional neural networks, multiple sclerosis (MS), MS lesion activity segmentation, lesion activity, MS new lesions segmentation

OPEN ACCESS

EDITED BY
Olivier Commowick,
Inria Rennes-Bretagne Atlantique Research Centre, France

REVIEWED BY
Francesca Galassi,
University of Rennes 1, France
Refaat E. Gabr,
Biogen, United States

*CORRESPONDENCE
Beytullah Sarica
saricab@itu.edu.tr

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 03 April 2022
ACCEPTED 27 June 2022
PUBLISHED 22 July 2022

CITATION
Sarica B and Seker DZ (2022) New MS lesion segmentation with deep residual attention gate U-Net utilizing 2D slices of 3D MR images.
Front. Neurosci. 16:912000.
doi: 10.3389/fnins.2022.912000

1. Introduction

Multiple sclerosis (MS) is an autoimmune disease characterized by demyelinating axons in the central nervous system, resulting in white matter (WM) lesions (Steinman, 1996; Calabresi, 2004). Magnetic resonance imaging (MRI) is widely utilized for various purposes, such as disease diagnosis, patient follow-up, and therapy monitoring. In clinical practice, MRI data can be used to diagnose and assess MS lesions, which helps physicians better understand the natural history of MS (Lladó et al., 2012; Combès et al., 2021). Fluid Attenuated Inversion Recovery (FLAIR) is an MRI technique that provides images in which WM lesions emerge as high-intensity areas, allowing for tracking of the disease progression (Rovira et al., 2015). In particular, this technique facilitates lesion segmentation to acquire quantitative features such as the number and volume of lesions (Roy et al., 2018). Since manual segmentation of such lesions is prone to high interobserver variability and time-consuming processes (Egger et al., 2017; Commowick et al., 2018), accurate automated segmentation methods are required to perform this process (Ma et al., 2022).

The emergence of new lesions or the expansion of existing lesions is referred to as lesion activity (McFarland et al., 1992). The most important biomarker for monitoring inflammatory changes and disease progression in MS is to track lesion activity between two longitudinal MR images (Patti et al., 2015; Combès et al., 2021). Recently, the delineation of new MS lesions on T2/FLAIR by comparing two time-points MRI data has gained attraction. Determination of new lesions has become even more important than identifying the total number and volume of lesions as it allows clinicians to determine whether a given anti-inflammatory disease modifying drug (DMD) is effective for the patient (Moraal et al., 2010). However, detection and delineation of new lesions appearing at the second-time point are particularly challenging and intra- and inter-rater variability are unavoidable due to small and subtle new lesions (McKinley et al., 2020). Therefore, automating the detection of these new lesions will be a significant improvement in assessing the disease activity of a patient.

Recently, deep learning methods, especially those relying on convolutional neural networks (CNNs) (LeCun et al., 2015), have improved the performance of brain lesion segmentation tasks (Akkus et al., 2017); such as brain tumor segmentation (Havaei et al., 2017), brain extraction (Kleesiek et al., 2016), and MS lesion segmentation (Roy et al., 2018; Aslani et al., 2019; Zhang et al., 2019). Most of these methods rely on encoder-decoder networks, taking MRI data as an input and generating a segmentation output for each pixel (Danelakis et al., 2018). Many CNN-based methods and their variations have also been proposed with different input strategies, such as multi-scale (Brosch et al., 2016), multi-branch (Aslani et al., 2019), and cascaded (Valverde et al., 2017) approaches. However, these together with most of the classical methods perform lesion segmentation on a single MRI data. For determining MS lesion activity, classical image processing approaches have been usually preferred such as image differences, intensity-based approaches, and deformation fields (Ganiler et al., 2014; Lesjak et al., 2016; Salem et al., 2018; Köhler et al., 2019). However, some of these approaches have high variability and inconsistency as they use two different segmentation outputs obtained from the baseline and follow-up images to produce the lesion activity (Krüger et al., 2020). To perform better lesion activity segmentation, deep learning approaches relying on CNNs are essential which take these two images as input; however, these methods have been so far limited for the MS lesion activity segmentation. Salem et al. (2020) who used a classical approach in their previous study proposed the first CNN-based longitudinal approach for detecting new T2-w lesions in brain MRI. In their study, intensity- and deformation- based features from two time-points data were incorporated into the proposed network and trained within an end-to-end procedure. Gessert et al. (2020b) have proposed a CNN-based method using two FLAIR images acquired at two different times to detect lesion activity. They used two-path architectures with attention-guided interactions to process two time-points of MRI data. Furthermore, they extended their work to full 4D deep learning using a history of MRI volumes and proposed a 3D ResNet-based multi-encoder-decoder network in which temporal aggregation was performed by convolutional gated recurrent units (convGRUs) for lesion activity segmentation (Gessert et al., 2020a). However, the dataset of these studies consists of MR images from the same scanner, which decreases the generalizability of these methods toward the intensity and texture characteristics variations, which can be inherited if the data is obtained from different scanners. Thus, there is a need for new deep learning approaches to cope with variations problems that may arise through the use of data from multiple scanners as well.

The patch-based and image-based approaches are generally used in CNN-based medical image segmentation (Aslani et al., 2019). Image-based segmentation approaches exploit the global structure information when processing the entire image; however, the patch-based approaches ignore this information due to the small patch sizes. In image-based segmentation, the 3D MRI data is processed either using slice-based or 3D segmentation methods (Brosch et al., 2016; Tseng et al., 2017). In slice-based image segmentation, each 3D MRI is converted into 2D slices along the x, y, and z axes, and then used as an input for deep learning models. After, these processed slices are aggregated to reconstruct a 3D binary output segmentation. In the 3D segmentation, meaningful information from the original 3D images is extracted with 3D kernels in a CNN. However, applying traditional 3D segmentation with a large number of parameters to a small dataset is prone to a high risk of overfitting issues which is a common issue in medical image analysis (Brosch et al., 2016). To address this overfitting issue, several approaches have been proposed such as defining three 2D kernels for each of the three plane orientations around the voxel (Liu et al., 2017; Tetteh et al., 2020); however, these approaches include more parameters for each plane when compared to the slice-based approach (Aslani et al., 2019).

Training deeper neural networks are challenging due to problems such as degradation problem. To solve these issues, He et al. (2016a,b) presented a deeper residual learning framework that uses identity mapping to ease the network training phase. Ronneberger et al. (2015) modified and extended the fully convolutional network (FCN) architecture (Long et al., 2015) to build the U-Net architecture which works with fewer training images and combines feature maps from multiple levels to enhance the segmentation accuracy. U-Net achieves promising results in medical image segmentation by combining low-level features with high-level semantic features. Combinations of U-Net and residual learning were also used for different image segmentation problems, such as road extraction using remote sensing data (Zhang et al., 2018). In addition, the attention gate (AG) model is proposed for automatically learning to focus on more features related to the target structures of various sizes and shapes (Oktay et al., 2018). AG uses high-level features from skip connections and low-level features from an upsampling operation to emphasize important features. This allows the network to focus on the small and subtle lesions appearing in the target MR images.

In this study, an automated segmentation pipeline with a fully convolutional neural network was used to detect and segment the new lesions observed in follow-up images. This study uses images from "Multiple sclerosis new lesions segmentation challenge (MSSEG-2)" [1] which consists of 3D FLAIR images acquired from different centers and scanners (1.5T and 3T). Residual units and attention gates are incorporated into the U-Net architecture for the new MS lesion activity task. The slice-based approach was preferred as the input strategy due to the above-mentioned advantages. Slices extracted from these pairs of MR scans were combined by stacking corresponding baseline and follow-up slices into the input channel dimension and then utilized as input values for the proposed model. This study has two major contributions to MRI base lesion activity monitoring. First, it is shown that an encoder-decoder-based architecture, namely U-Net, provided acceptable results in detecting and segmenting the lesion activity. Second, it is demonstrated that using a whole-brain slice approach with the U-Net architecture including residual blocks and modified attention gates significantly improves the segmentation of lesion activity on MRI data acquired from different scanners.

2. Materials and methods

2.1. Data, preprocessing, and preparation

In this study, a total of 100 patients' MRI data that was associated with MS disease provided by the MSSEG-2 challenge [2] was utilized. The voxel size of each MRI data in this dataset varies from $0.5 \times 0.5 \times 0.5$ mm^3 to $1.2 \times 1.2 \times 1.2$ mm^3. The dataset was divided into two groups for training and testing. 40 image pairs were used for the training and the remaining were used for testing. For each patient, raw 3D T2/FLAIR MRI pairs were obtained from 15 different MRI scanners at 1.5T and 3T. A rigid registration was applied to these images to bring them into a middle point in which the ground truth data was calculated by the challenge organizers. Thereafter, a consensus delineated ground truth data for the follow-up images were formed by a majority voting among the four experts and validated by a senior expert neuroradiologist.

Data preprocessing is a crucial step for the segmentation task in medical image processing since the raw MRIs may have irrelevant information like non-brain tissues and skulls. Thus, brain extraction followed by N4 bias field correction (Tustison et al., 2010) was performed on these raw 3D images using the Anima MS longitudinal preprocessing script [3]. Intensity normalization was performed on each 3D MRI scan using the 99th percentile and Kernel Density Estimate (KDE) with the Gaussian kernel similar to one described by Reinhold et al. (2019) and Zhang and Oguz (2020). Then, early fusion was performed on the baseline and follow-up images to produce 2-channel input data allowing the proposed model to obtain temporal features from MRI sequences.

The resulting 3D MRI data consists of orthogonal plane orientations which yield three views. From this data, the axial, sagittal, and coronal views along the x, y, and z axes were obtained as 2D slices. Since each generated 2D slice has a different size that depends on the orientation, zero padding was applied to obtain a 512 x 512 slice size for all orientations by centering the brain without affecting the original voxel size. As discussed in detail by Hashemi (2019), zero padding does not deform the patterns in the image and does not affect the network weights during the backpropagation. To restrict excessively unbalanced data and ignore non-informative samples, the slices which have at least one pixel delineated as a new lesion on the follow-up MR images were chosen to create a training subset. As a result, a total of 2,637 2D slices for each time point were derived to be used for training and validation sets. Afterward, the baseline and follow-up images were stacked to generate a 2-channel feature map for each plane orientation. Finally, all 2D stacked slices extracted from all three planes were aggregated to generate a single training input, which allowed to increase training samples and use the contextual information in all directions. Figure 1 shows the raw and preprocessed input data for the two time points dataset with the delineated ground truth data.

2.2. Model architecture

2.2.1. U-Net

U-Net, an encoder-decoder network with skip connections, has shown competitive results in the medical field (Ronneberger

[1] Challenge website: https://portal.fli-iam.irisa.fr/msseg-2/

[2] Challenge Data: https://portal.fli-iam.irisa.fr/msseg-2/data/

[3] Anima scripts: RRID SCR_017072 https://anima.irisa.fr/

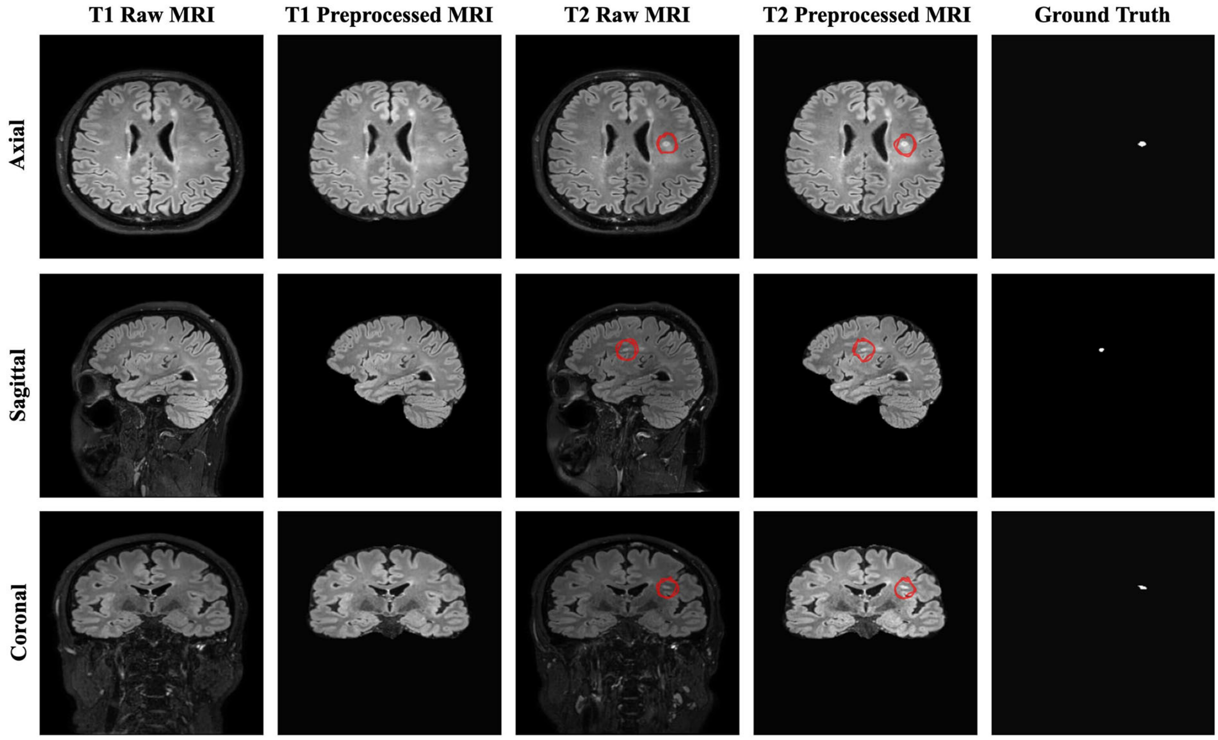


FIGURE 1
The raw, preprocessed, and delineated mask slices including two-time points for the new MS lesions segmentation task.

et al., 2015). This network concatenates features from different levels to enhance segmentation performance. It consists of encoding, bridge, and decoding paths. In the encoding path, the feature map from each layer is downsampled by halving the size to encode the input image into the feature representations. As for the decoding path, the corresponding encoding path which has high-resolution features (semantically low) is combined with the upsampling of the feature maps produced from the lower dimension to better learn representations with the following convolutions. The bridge connects these paths as a transition block. Each block in each layer has two sets of 3 x 3 convolutional layers with a Rectified Linear Unit (ReLU) activation for both downsampling and upsampling operations. The final layer of the U-Net utilizes a 1 x 1 convolution with a sigmoid activation to predict each pixel value ranging from 0 to 1 (Ronneberger et al., 2015). The standard blocks in the U-Net architecture can be replaced with residual units to enhance the model performance.

2.2.2. Residual learning

Adding more layers to build a deeper neural network could enhance the performance of networks; however, increasing the depth of the network may slow down the training process, perhaps resulting in a degradation problem (He et al., 2016a). Deep residual learning uses several residual blocks together in which an identity mapping is created to handle the performance problem, and also address the degradation problem (He et al., 2016a). The residual unit is comprised of two 3 x 3 convolutional blocks, each with Batch Normalization (BN), a ReLU activation, and a convolutional layer, as well as an identity mapping that combines the input and output of the residual unit. Figure 2 shows the residual unit including identity mapping within the proposed model. Each residual unit is formulated according to He et al. (2016b) as the following:

$$y_l = h(x_l) + F(x_l, W_l) \tag{1}$$

$$x_{l+1} = f(y_l) \tag{2}$$

where x_l and x_{l+1} are the input and output of the l-th unit while F, f, and h indicate the residual function, activation function, and identity mapping, respectively. He et al. (2016b) also recommended a full pre-activation as demonstrated in Figure 2. In this study, a full pre-activation residual unit was used to construct and design the deep residual attention gate U-Net.

2.2.3. Attention gate

Attention gates help the models to focus on learning the salient features beneficial for specific tasks while avoiding

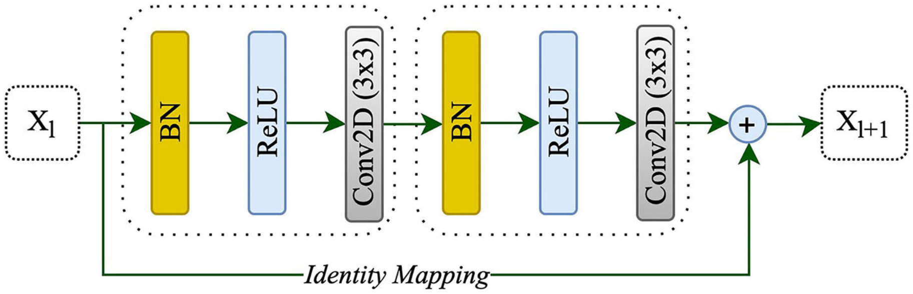


FIGURE 2
A residual unit with identity mapping. x_l and x_{l+1} are the input and output of the l-th unit, respectively.

unnecessary regions in an input image (Oktay et al., 2018). These are used during concatenating skip connection and upsampling to focus more features related to different sizes and shapes on the target structure. Contextual information (gating) obtained at coarser scales is used to achieve feature selectivity in AGs. Figure 3 shows the overview of the attention gate mechanism.

2.2.4. Deep residual attention gate U-Net

In this study, the combination of U-Net, deep residual learning, and attention gate was proposed for the new MS lesion segmentation task. In this combination, the residual unit will facilitate the network training. Information will be able to propagate without degradation thanks to the skip connections within a residual unit and between low and high levels of the network. Thus, deep neural networks are built with fewer parameters while still achieving a competitive segmentation performance. As such, the standard blocks were replaced with residual blocks in the proposed model. AGs, modified by adding BN and a ReLU activation for both input features before convolutional operations, were added between the corresponding encoding part and the upsampling of features maps produced from the lower level. Thus, allowing the model to learn to focus on salient features of various shapes and sizes. Figure 3 demonstrates the details of the designed network with the input data formed by the axial, sagittal, and coronal views extracted from the baseline and follow-up 3D MRI for the new MS lesion segmentation.

2.3. Implementation details

The training set comprised 3D FLAIR images of 40 patients and only 29 had new lesions in their follow-up images. These 29 MR images were divided into the training and validation sets (24 patients for training and 5 patients for validation). To prepare input data, each 3D image was divided into its axial, sagittal, and coronal views. Two-channel input feature data was created using each corresponding 2D slice from both time points as discussed previously. Keras (version=2.4)[4] and TensorFlow (version 2.4)[5] libraries were used for the model development in Python language (version 3.7)[6] (Chollet, 2015; Abadi et al., 2016). The Google Colaboratory, having a Tesla K80 GPU, was used for the training procedure (Bisong, 2019). The proposed model was trained by using the Adam optimizer (Kingma and Ba, 2014), an initial learning rate of 1e-4 (adjusting with patience=10 and factor=0.1 during the training), and a batch size of 8 over 200 epochs, respectively. The validation dice score was also monitored to choose the best model, and model weights were saved based on the best validation dice score during the training. Early stopping (patience=50) was exploited to prevent overfitting as well. Hashemi et al. (2022) used the sum of dice loss with a 1.5 coefficient and binary cross entropy loss as a custom loss function for MS lesion segmentation. Similarly, in this study, a hybrid loss function consisting of binary focal loss and dice loss [dice loss + (1 × binary focal loss)] was employed in order to handle unbalanced labeled data between lesion and background since lesion pixels constitute a minor portion of the whole image. The total loss function is defined as follows:

$$L_t = (1 - \frac{2\,gt\,pr + 1}{gt + pr + 1}) + (1 \times (-gt\alpha(1-pr)^{\gamma}\log(pr) \\ -(1-gt)\alpha pr^{\gamma}\log(1-pr))) \quad (3)$$

where gt denotes the ground truth, and pr indicates prediction. 0.25 and 2.0 default values were used for the parameters of α and γ, respectively.

Keras data generator was used for performing real-time data augmentation such as vertical flipping, horizontal flipping,

[4] https://keras.io/

[5] https://www.tensorflow.org/

[6] https://www.python.org/

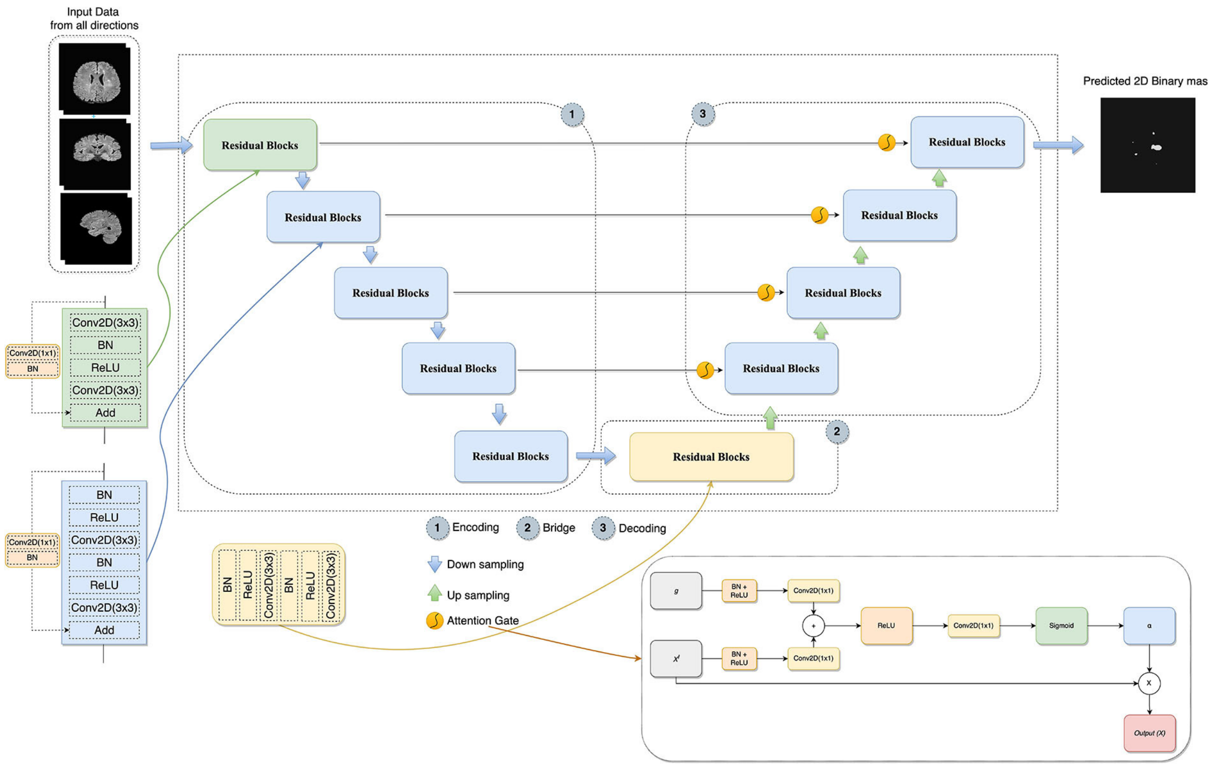


FIGURE 3
The architecture of the proposed model combines U-Net, residual learning, attention gate, and a slice-based approach. In AG adapted from Oktay et al. (2018), x^l is the input features, α is the attention coefficients used to scale the x^l, and g collected from a coarser scale is the gating signal which provides contextual information.

random rotation, and shift range to increase the number of training samples. Figure 4 shows the proposed pipeline for new lesion segmentation of MS activity. First, 3D MRIs were converted into their plane orientations along the x, y, and z axes. Then, 2D slices of two-time points were fused together to create a single input training data for the proposed model. Predicted 2D slices based on the axial, sagittal, and coronal views were converted into the 3D binary segmentation output, and then the final output segmentation mask was generated by using the majority voting among 3D binary outputs obtained from each view.

To compare components of the designed network, a testing subset was created from the MSSEG-2 test dataset provided by the challenge organizers. This subset comprised MRI data of 7 patients by considering the different scanners and new lesion loads. Satisfactory results with the MSSEG-2 dataset could not be obtained by the implementation of the original U-Net. Therefore, this implementation was modified with transpose upsampling instead of a simple upsampling operation, and batch normalization to make the neural network more stable. A hybrid loss function, the summation of binary focal and dice losses, was used for all models.

2.4. Metrics

2.4.1. Dice similarity coefficient

The segmentation of new lesions was considered one of the two most important evaluation criteria for the challenge. This indicates how many new lesions are precisely overlapped in the ground truth which is also known as the Dice score (Commowick et al., 2018). In other words, the Dice Similarity Coefficient (DSC) is used to measure the similarity of the evaluated segmentation and the ground truth. It is formulated as follows:

$$DSC = \frac{2TP}{2TP + FP + FN} \tag{4}$$

where TP, FP, and FN denote the true positive, false positive, and false negative pixels/voxels, respectively.

2.4.2. F1 score

Another important evaluation criterion was the detection of new lesions. This shows the number of new lesions that are correctly detected or not without considering the precision of their contours. Lesion sensitivity, which is the proportion of the detected lesions in the ground truth, and lesion positive

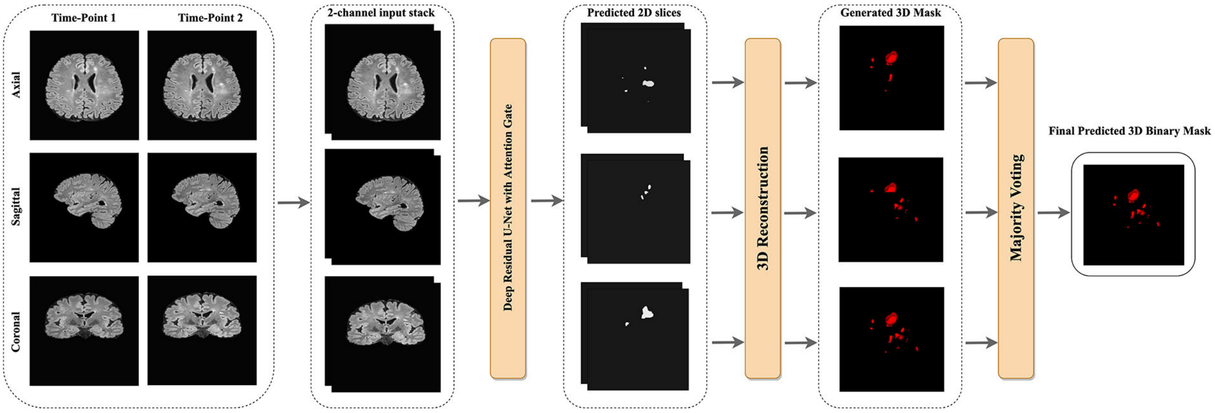


FIGURE 4
The proposed pipeline of new MS lesions segmentation using a slice-based approach including the majority voting for the final 3D segmentation output using the predicted 2D axial, sagittal, and coronal slices.

predictive, which is the proportion of TP lesions in the automatic segmentation, were used to compute the F1 score. Lesion sensitivity (S) and lesion positive predictive (P) can be calculated with the following equations (Commowick et al., 2018):

$$S = \frac{TP_G}{M} \tag{5}$$

$$P = \frac{TP_A}{N} \tag{6}$$

where M and N denote the number of lesions in the ground truth and the automatic segmentation, respectively. TP_G indicates the number of lesions correctly detected by the automatic segmentation among the number of lesions in the ground truth. TP_A denotes the number of lesions correctly detected by the ground truth among the number of lesions in the automatic segmentation. Hereafter, these two metrics can be formulated to calculate the F1 score with the following equation.

$$F_1 = \frac{2SP}{S+P} \tag{7}$$

2.4.3. Metrics for no new lesions

Patients with MS may not have new lesions for their follow-up images. This is usual in clinical cases, and this challenge has also similar cases in both training and test data sets. For example, the testing set is comprised of 28 patients with no new lesions and 32 patients with at least one or more new lesions. The number and volume of new lesions were used as evaluation metrics as well. The volume of new lesions was calculated by multiplying the number of voxels in the segmentation with the voxel volume. A value of zero is the optimal value for these metrics.

2.4.4. Other overlap and surface metrics

Overlap metrics consider the voxel-based overlap of the segmentation output (A) and manual annotation mask (G) while surface metric computes the average symmetric surface distance. The surface metric considers contours obtained from the segmentation output and manual annotation mask. As described in Commowick et al. (2018), the MSSEG-2 challenge provides a report on the test data set including some of these measures, such as:

- Positive Predictive Value (PPV):

$$PPV = \frac{A \cap G}{A} \tag{8}$$

- Sensitivity (S_e):

$$S_e = \frac{A \cap G}{G} \tag{9}$$

- Specificity (S_p):

$$S_p = \frac{B - A \cap G}{B - G} \tag{10}$$

where B reveals the entire image.

- Mean Surface Distance (S):

$$S = \frac{\sum_{i \in A_S} d(x_i, G_S) + \sum_{j \in G_S} d(x_j, A_S)}{N_A + N_G} \tag{11}$$

where d indicates the minimal Euclidean distance of a point of one surface to the other surface. N_A and N_G reveal the number of points of each surface, respectively.

2.5. 3D binary image reconstruction

The slices from each view were used to reconstruct the final 3D binary segmentation output. The 3D binary segmentation

TABLE 1 Prediction results of evaluating the challenge test data set published on the challenge website.

Methods	F1 Score	Dice score	Number of tested lesions	Volume of tested lesions (mm^3)
Expert 1	0.712	0.631	0.036	1.453
Expert 3	0.636	0.598	0.000	0.000
Expert 2	0.607	0.536	0.107	3.981
Mediaire-B*	**0.541**	0.437	0.536	29.235
Empenn	*0.532*	0.424	0.286	4.258
Mediaire-A	0.525	0.432	0.429	15.908
Expert 4	0.524	0.461	0.036	0.623
LaBRI-IQDA	0.517	*0.500*	1.143	38.486
SNAC	0.514	0.485	0.321	5.726
MedICL	0.500	**0.507**	0.536	12.713
LaBRI-D&E	0.498	0.472	1.964	177.131
ITU (Ours)	0.480	0.443	0.148	1.488
New Brain	0.477	0.451	0.786	12.371
LYLE	0.441	0.409	**0.036**	**0.470**
SCAN	0.433	0.403	*0.071*	5.373
Neuropoly-2	0.410	0.409	0.107	*0.498*
SCA-withPriors	0.216	0.224	2.464	302.121
IBBM+	0.143	0.155	3.786	123.309

Bold and italic values are the highest and the second-best scores among other proposed methods excluding the experts, respectively. Dice and F1 Score are expected to be a high numerical value while the Number of and Volume of Lesions are expected to be a low numerical value. The Number of and Volume of Lesions metrics are calculated for no new lesion cases. The source data can be accessed at https://doi.org/10.5281/zenodo.5775523. * and + indicate the first and last ranks among the participants, respectively. This table is ordered according to the highest to the lowest based on the F1 score.

was produced by using the 2D predicted slices from each plane orientation. Then, a majority voting was applied to these 3D segmentation outputs to generate the final 3D binary segmentation as shown in Figure 4.

3. Results

The MSSEG-2 challenge aims to segment and detect new MS lesions by comparing the baseline and the follow-up 3D FLAIR images of a patient. Twenty four teams with a total number of 30 pipelines participated in this challenge. Deep learning approaches, most of them relying on the U-Net architecture, were proposed by most of the participants, while only one of the teams used a conventional statistical method and the subtraction between two MR images (Commowick et al., 2021). Table 1 shows the average quantitative metric results of some of the methods presented in the challenge, including the results of the experts[7].

Four metrics were used to evaluate the proposed pipelines for new MS lesion segmentation and detection. The test data set consists of MR images of 60 patients and 32 of them were used for the calculation of the F1 and dice scores due to possessing new lesions at their follow-up images. The remaining patients' data were used for the calculation of the number of tested lesions and volume of tested lesions. According to the challenge results, our proposed pipeline was ranked 8th for F1 and dice scores among the proposed methods. The proposed pipeline produced a mean score of 48% for the F1 score and a mean score of 44.30% for the dice score. For the no-lesion cases, our pipeline was ranked in 5th and 4th places with a mean score of 0.148 and 1.488, respectively for the number of tested and volume of tested lesions. Also, the highest F1 and dice scores including the expert raters were a mean score of 71.20 and 63.10% respectively, which belonged to expert 1. As for the number of tested and volume of tested lesions, the highest score was 0 which belonged to expert 3. On the other hand, the highest F1 and dice scores for the automated methods belonged to teams Mediaire-B and MedICL with a mean score of 54.10 and 50.70%, respectively. The highest score for the number of tested lesions and volume of tested lesions belonged to team LYLE with a mean score of 0.036 and 0.498, respectively. The lowest F1 and dice scores, belonging to the team IBBM, had a mean score of 14.30 and 15.50%, respectively. Figure 5 shows the segmentation performance of the proposed model, consensus, and experts on a slice of an axial view of four patients. As seen in the figure, the proposed model had competitive performance compared to the segmentation output of experts.

[7] Quantitative metrics for all proposed pipeline are available at https://doi.org/10.5281/zenodo.5775523

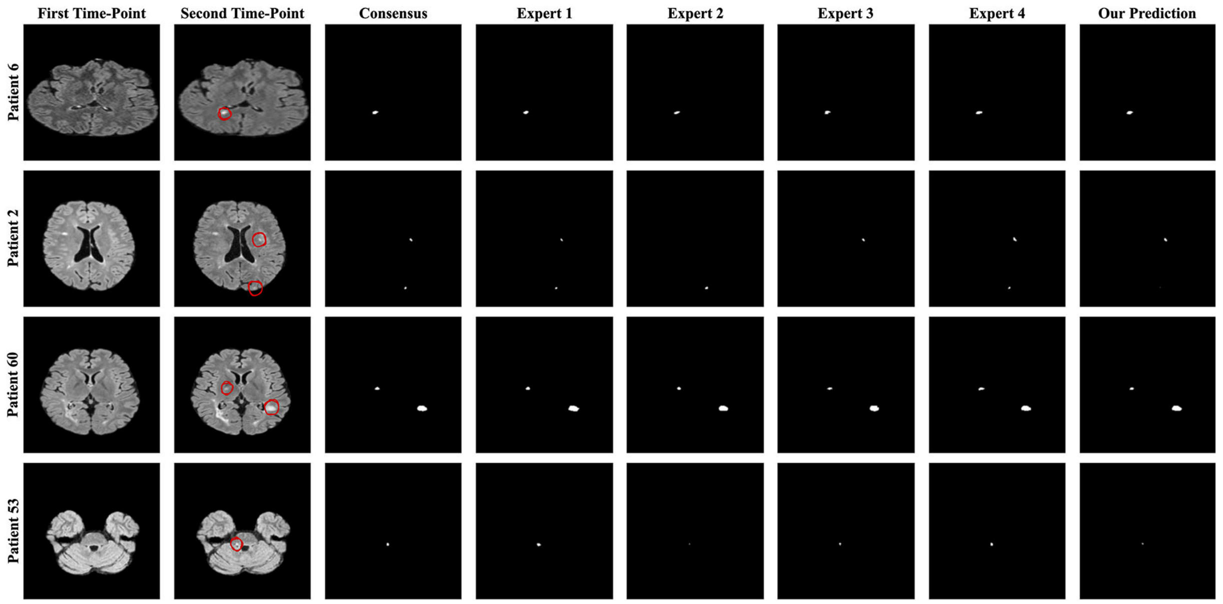


FIGURE 5
The best and worst performances of the proposed model compared to the consensus and each expert segmentation for F1 and dice scores. A slice of axial view from patients 6 and 2 for the F1 score and patients 60 and 53 for the dice score is presented.

The challenge also provides additional metrics discussed in Section 2.4.4 for a complete evaluation although these metrics were not considered for the ranking. The results obtained from some of the proposed methods and experts for additional metrics are given in Table 2. Accordingly, the results of our pipeline with respect to sensitivity, specificity, PPV, and surface distance were a mean score of 0.364, 1.000, 0.675, and 8.548, respectively. Our pipeline had competitive performance compared to experts and other proposed pipelines in some of these metrics. For example, the highest PPV score among experts and proposed methods were a mean of 0.813 and 0.703 for expert 1 and the team LYLE, respectively. Also, the highest score for surface distance belonged to expert 2 and the team LYLE with a mean score of 4.543 and 7.210.

TABLE 2 Prediction results of evaluating the challenge test data set published on the challenge website for other useful metrics.

Methods	Sensitivity	Specificity	PPV	Surface distance
Expert 1	**0.650**	1.000	0.707	*5.907*
Mediaire-B	*0.616*	1.000	0.394	8.803
Expert 3	0.589	1.000	0.760	5.990
Expert 2	0.526	1.000	**0.813**	**4.543**
MedICL	0.514	1.000	0.556	9.194
Expert 4	0.407	1.000	*0.801*	7.885
Proposed model	0.364	1.000	0.675	8.548
LYLE	0.344	1.000	0.703	7.210
SCAN	0.340	1.000	0.678	8.521
IBBM	0.170	1.000	0.242	24.102

Bold and italic values are the highest and the second-best scores among some of the proposed methods and the experts, respectively. Sensitivity, Specificity, and PPV are expected to be a high numerical value while Surface Distance is expected to be a low numerical value. The source data can be accessed at https://doi.org/10.5281/zenodo.5775523. This table is ordered according to the highest to the lowest based on the sensitivity score.

Finally, comparisons between U-Net, U-Net with AGs, U-Net with RUs, U-Net with RUs, and AGs (two types) were realized for the new MS lesion segmentation. The results of U-Net, U-Net + AGs, U-Net + RUs, and U-Net + RUs + AGs are presented in Table 3. As seen in this table, the proposed model achieved the highest dice and F1 scores, a mean score of 58.70 and 61.10%, respectively. U-Net + RUs achieved the highest PPV score, a mean score of 62.40%. Furthermore, this network had fewer training parameters and performed better compared to the U-Net architecture.

TABLE 3 The evaluation results of the proposed method with different components using a subset of the MSSEG-2 test dataset.

Methods	Dice score	F1 Score	PPV	Total parameters
U-Net + RUs + AGs	**0.587**	**0.611**	0.567	4,934,613
U-Net + RUs	0.551	0.441	**0.624**	**4,722,897**
U-Net + AGs	0.505	0.592	0.609	7,947,109
U-Net	0.558	0.490	0.467	7,771,585

Bold values indicate the highest scores in the columns of dice, F1 score, and PPV while the lowest value in the column of total parameters. This table is ordered according to the highest to the lowest based on the dice score.

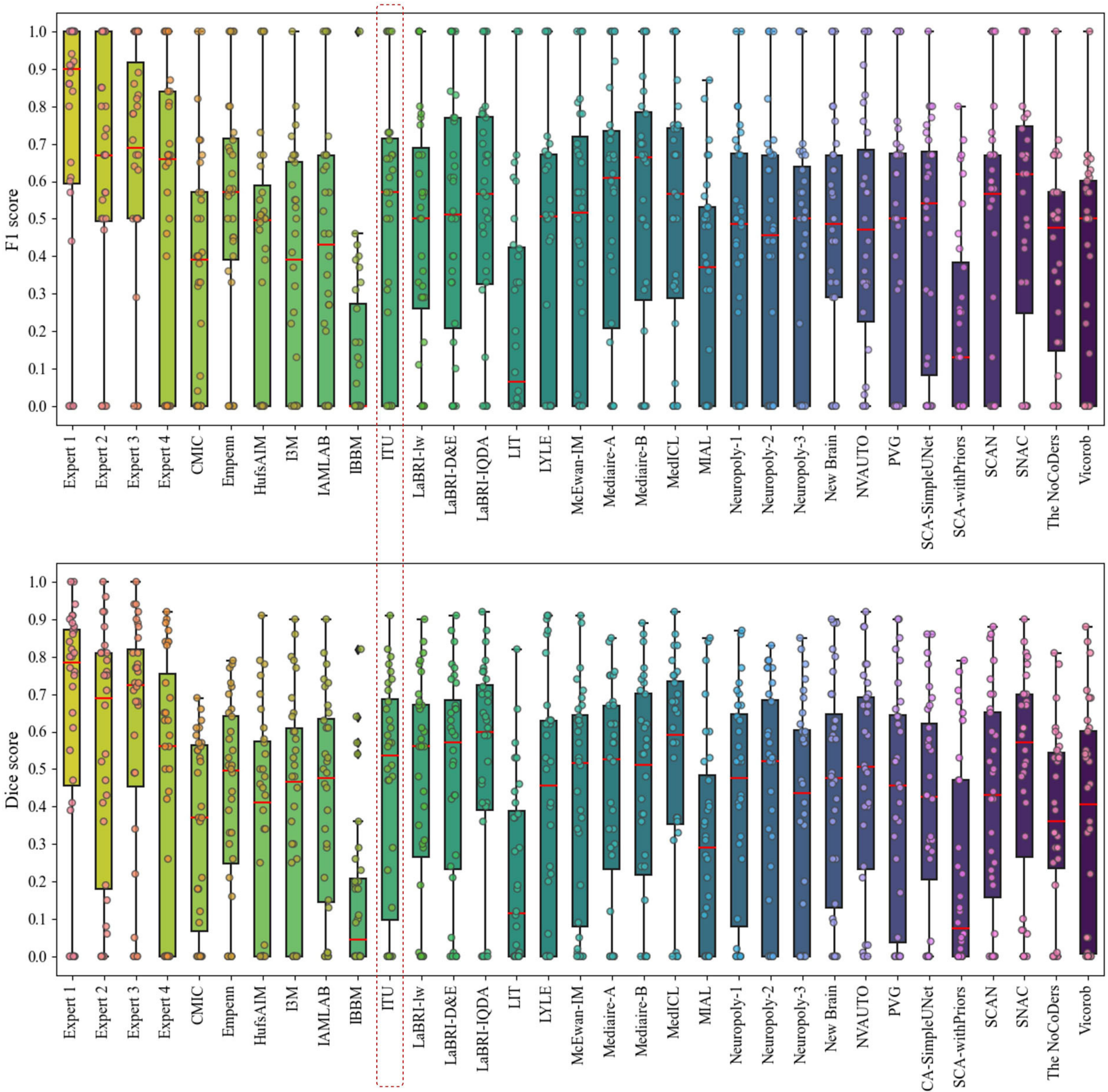


FIGURE 6
Analysis of differences in detection and segmentation by using F1 and dice scores for each expert and each team that participated in the challenge, respectively.

4. Discussions

In this study, a deep learning model was developed to handle the problem of identifying new MS lesions using the baseline and the follow-up 3D FLAIR MR images. Activity segmentation particularly for new lesions is a more challenging task compared to lesion segmentation in a single-time MR scan due to small lesion loads. MS lesion segmentation using traditional and deep learning approaches has usually been studied in a single MRI scan in recent years. However, deep learning approaches for MS lesion activity using the baseline and follow-up MR images still remain limited. In most of these studies, the researchers have been using their own datasets making it difficult to compare and reproduce their results with the proposed pipeline. Thus, in this study, comparisons were performed on the automated methods proposed in the challenge. Moreover, comparisons were performed among components used for building the designed network as well. The proposed

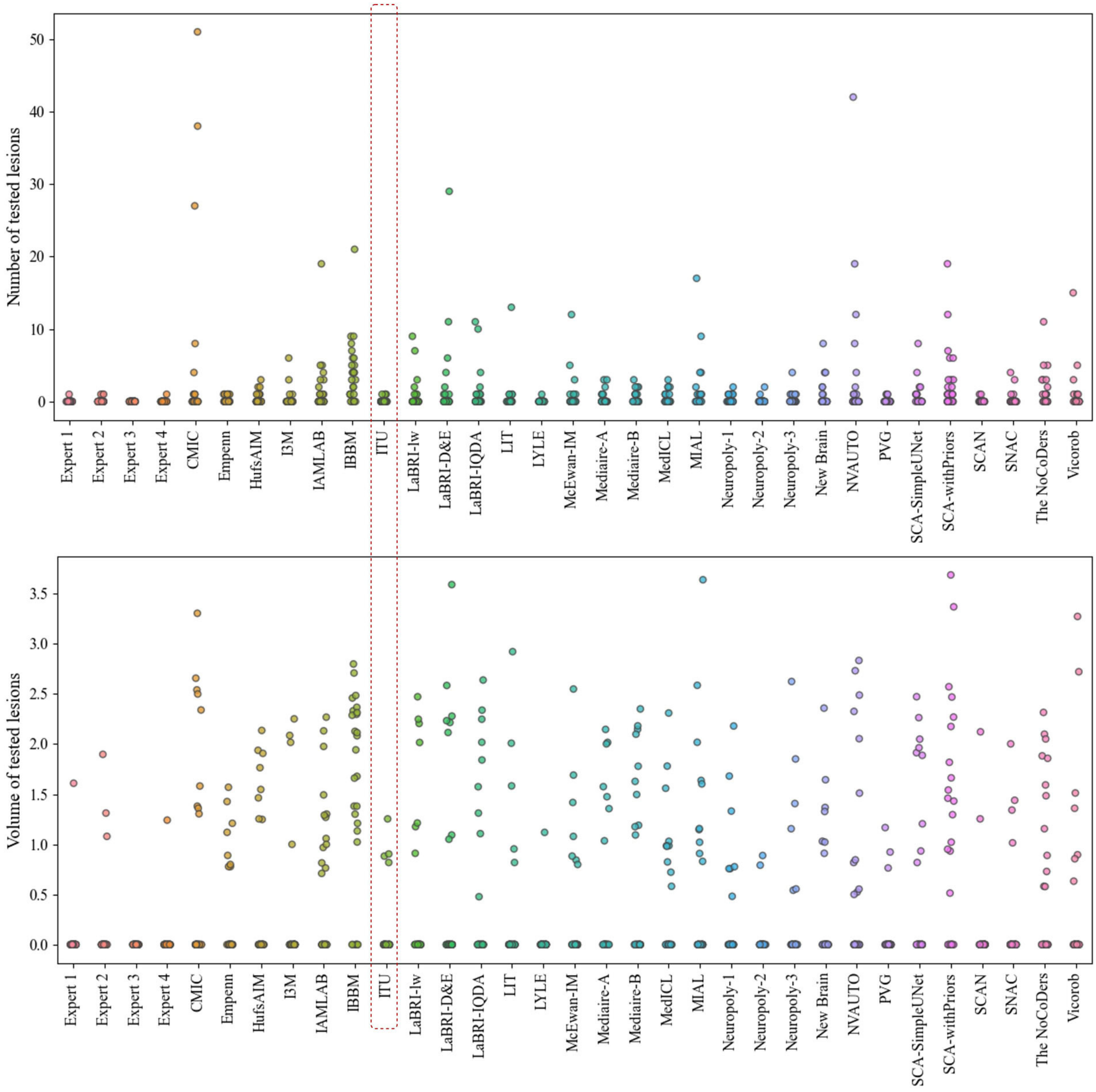


FIGURE 7
Analysis of the number and volume of lesions detection for each expert and each team that participated in the challenge (The data of volume of tested lesions was scaled by $\log_{10}$).

network, which combines the strengths of U-Net, residual units, and attention gates, has outperformed other methods comprising different combinations of components in terms of dice and F1 scores.

A whole-brain slice-based approach was used as patch-based CNNs lack spatial information about MS lesions due to the patch size limitation (Aslani et al., 2019). The results indicated that the proposed pipeline with this approach had a competitive performance for most measures compared to the other pipelines, as given in Table 1. Segmentation performance of new MS lesions improved significantly when baseline and follow-up MRI scans were stacked in the input channel dimension. Thus, baseline and follow-up scans for each patient were stacked as a two-channel input for the proposed pipeline. Furthermore, attention gates modified with BN and ReLU allowed the model to focus on small and subtle new lesions.

Figure 6 presents the analysis of differences in detection and segmentation for F1 and dice scores for each expert and each team that participated in the challenge, respectively. The red box highlighted the team performance of this study for these two metrics. According to F1 and dice scores, proposed methods could not reach the expert level; however, some methods were able to outperform experts which revealed varying scores in different patients [8]. Based on this observation, it was concluded that detection and segmentation of MS new lesions in longitudinal studies is a difficult task even for experts. Therefore, an external reviewer may be needed while analyzing the new lesions with automated methods for the lesion activity.

The evaluation metrics for no new lesions are indicated in Figure 7. The number of connected components in automatic segmentation was used to find the number of lesions detected. Also, the volume of lesions detection (mm^3) was used to evaluate the segmentation performance of both automated and expert delineation outputs. As seen in Figure 7 and Table 1, the proposed pipeline outperformed compared to some of the other proposed methods. The dotted red rectangle highlights the proposed pipeline within this study. Accordingly, some of the proposed methods, including ours, outperformed some experts.

Instead of using a 3D segmentation approach requiring more computational power and learning parameters, the proposed method and the slice-based approach were used together for detecting and segmenting new lesions on the follow-up images. While the appearance of new lesions is of primary interest for the challenge, enlarged or disappearance of MS lesions could be also studied. Different MRI modalities such as T1-and T2-weighted can also be incorporated into the given task to extract more features related to the size or location of new MS lesions even though the FLAIR images reveal lesions as more intense. To achieve a robust automated model for the given task, large datasets from different scanners are needed; however, it is difficult to obtain such datasets.

5. Conclusion

In this study, an automated pipeline for new MS lesion segmentation using the baseline and follow-up 3D FLAIR MRI has been designed with a deep learning-based network that fuses the strengths of U-Net, residual learning, and AG. For more accurate segmentation of new MS lesions, this network architecture was designed as a deep encoder-decoder network to enhance the U-Net by replacing plain blocks with residual blocks and adding attention gates. These residual blocks replaced with the plain blocks facilitate the training. Skip connections within both residual units and U-Net facilitates the propagation of information in both forward and backward phases during the training procedure. AGs integrated into the proposed model emphasize important features propagated over skip connections. A hybrid loss function was introduced as the addition of dice loss and 1 × binary focal loss. The input data for the proposed method was prepared by converting 3D scans into their plane orientations of axial, sagittal, and coronal views which yielded 2D slices. Baseline and follow-up slices were stacked to create a two-channel feature mapping for each plane orientation. Then, all slices extracted from all three planes were grouped into a single input to increase training samples and to use the contextual information in all directions. The predicted 2D slices for each view were aggregated using a majority voting to generate the final 3D binary output. Although new MS lesion segmentation and detection pose a difficult problem due to small lesion sizes, the proposed method has achieved comparable segmentation performance compared to the experts and top-ranked automated methods in the challenge. Finding the appropriate data sets and using the existing ones as publicly available will reduce the gap for the data required in these studies and the lack of which is frequently discussed, and will allow different studies to be carried out. This study provides clues about the recent techniques regarding the MS lesion activity segmentation that can be used as a guide for future studies in this field.

Data availability statement

The dataset used and analyzed in this study can be accessed online at https://portal.fli-iam.irisa.fr/msseg-2/data/.

Author contributions

BS conducted the experiments and organized the main manuscript. DZS participated in the writing and modifying of the English grammar of the manuscript. Both authors analyzed the results and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

Acknowledgments

The authors would like to thank the Empenn team, the French cohort of patients with MS (OFSEP), and France Life Imaging (FLI-IAM) for organizing the MSSEG-2 challenge and providing the dataset. The trained model is available as a Docker image and can be pulled with this command: docker pull beytullahsarica/deep-res-unet-ag-ms-activity-

[8] Evaluation results and analysis slides at https://files.inria.fr/empenn/msseg-2/Challenge_Day_MSSEG2_Results_2021.pdf

segmentation:v1.0.1. Also, our code is available at https://github.com/beytullahsarica/new_ms_lesion_segmentation.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). "Tensorflow: a system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)* (Savannah, GA), 265–283.

Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D. L., and Erickson, B. J. (2017). Deep learning for brain MRI segmentation: state of the art and future directions. *J. Digit. Imaging*.30, 449–459. doi: 10.1007/s10278-017-9983-4

Aslani, S., Dayan, M., Storelli, L., Filippi, M., Murino, V., Rocca, M. A., et al. (2019). Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. *Neuroimage* 196, 1–15. doi: 10.1016/j.neuroimage.2019.03.068

Bisong, E. (2019). *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*. Berkeley, CA: Apress.

Brosch, T., Tang, L. Y., Yoo, Y., Li, D. K., Traboulsee, A., and Tam, R. (2016). Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE Trans. Med. Imaging* 35, 1229–1239. doi: 10.1109/TMI.2016.2528821

Calabresi, P. A. (2004). Diagnosis and management of multiple sclerosis. *Am. Fam. Physician* 70, 1935–1944. Available online at: https://www.aafp.org/pubs/afp/issues/2004/1115/p1935.html

Chollet, F. (2015). *Keras*. Available online at: https://github.com/fchollet/keras.

Combès, B., Kerbrat, A., Pasquier, G., Commowick, O., Le Bon, B., Galassi, F., et al. (2021). A clinically-compatible workflow for computer-aided assessment of brain disease activity in multiple sclerosis patients. *Front. Med.* 8, 740248. doi: 10.3389/fmed.2021.740248

Commowick, O., Cervenansky, F., Cotton, F., and Dojat, M. (2021). "Msseg-2 challenge proceedings: multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure," in *MICCAI 2021-24th International Conference on Medical Image Computing and Computer Assisted Intervention*, 1–118.

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., et al. (2018). Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Sci. Rep.* 8, 1–17. doi: 10.1038/s41598-018-31911-7

Danelakis, A., Theoharis, T., and Verganelakis, D. A. (2018). Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging. *Comput. Med. Imaging Graphics* 70, 83–100. doi: 10.1016/j.compmedimag.2018.10.002

Egger, C., Opfer, R., Wang, C., Kepp, T., Sormani, M. P., Spies, L., et al. (2017). Mri flair lesion segmentation in multiple sclerosis: does automated segmentation hold up with manual annotation? *Neuroimage Clin.* 13:264–270. doi: 10.1016/j.nicl.2016.11.020

Ganiler, O., Oliver, A., Diez, Y., Freixenet, J., Vilanova, J. C., Beltran, B., et al. (2014). A subtraction pipeline for automatic detection of new appearing multiple sclerosis lesions in longitudinal studies. *Neuroradiology* 56, 363–374. doi: 10.1007/s00234-014-1343-1

Gessert, N., Bengs, M., Krüger, J., Opfer, R., Ostwaldt, A.-C., Manogaran, P., et al. (2020a). 4D deep learning for multiple sclerosis lesion activity segmentation. *arXiv preprint arXiv:2004.09216*. doi: 10.48550/arXiv.2004.09216

Gessert, N., Krüger, J., Opfer, R., Ostwaldt, A.-C., Manogaran, P., Kitzler, H. H., et al. (2020b). Multiple sclerosis lesion activity segmentation with attention-guided two-path cnns. *Comput. Med. Imaging Graphics* 84, 101772. doi: 10.1016/j.compmedimag.2020.101772

Hashemi, M. (2019). Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation. *J. Big Data* 6, 1–13. doi: 10.1186/s40537-019-0263-7

Hashemi, M., Akhbari, M., and Jutten, C. (2022). Delve into multiple sclerosis (ms) lesion exploration: a modified attention u-net for ms lesion segmentation in brain mri. *Comput. Biol. Med.* 145, 105402. doi: 10.1016/j.compbiomed.2022.105402

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., et al. (2017). Brain tumor segmentation with deep neural networks. *Med. Image Anal.* 35, 18–31. doi: 10.1016/j.media.2016.05.004

He, K., Zhang, X., Ren, S., and Sun, J. (2016a). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 770–778.

He, K., Zhang, X., Ren, S., and Sun, J. (2016b). "Identity mappings in deep residual networks," in *European Conference on Computer Vision* (Springer), 630–645.

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. doi: 10.48550/arXiv.1412.6980

Kleesiek, J., Urban, G., Hubert, A., Schwarz, D., Maier-Hein, K., Bendszus, M., et al. (2016). Deep mri brain extraction: A 3d convolutional neural network for skull stripping. *Neuroimage* 129, 460–469. doi: 10.1016/j.neuroimage.2016.01.024

Köhle, C., Wahl, H., Ziemssen, T., Linn, J., and Kitzler, H. H. (2019). Exploring individual multiple sclerosis lesion volume change over time: development of an algorithm for the analyses of longitudinal quantitative mri measures. *Neuroimage Clin.* 21, 101623. doi: 10.1016/j.nicl.2018.101623

Krüger, J., Opfer, R., Gessert, N., Ostwaldt, A.-C., Manogaran, P., Kitzler, H. H., et al. (2020). Fully automated longitudinal segmentation of new or enlarged multiple sclerosis lesions using 3d convolutional neural networks. *Neuroimage Clin.* 28, 102445. doi: 10.1016/j.nicl.2020.102445

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lesjak, Ž., Pernuš, F., Likar, B., and Špiclin, Ž. (2016). Validation of white-matter lesion change detection methods on a novel publicly available mri image database. *Neuroinformatics* 14, 403–420. doi: 10.1007/s12021-016-9301-1

Liu, S., Zhang, D., Song, Y., Peng, H., and Cai, W. (2017). "Triple-crossing 2.5 d convolutional neural network for detecting neuronal arbours in 3d microscopic images," in *International Workshop on Machine Learning in Medical Imaging* (Quebec: Springer), 185–193.

Lladó, X., Oliver, A., Cabezas, M., Freixenet, J., Vilanova, J. C., Quiles, A., et al. (2012). Segmentation of multiple sclerosis lesions in brain mri: a review of automated approaches. *Inf. Sci.* 186, 164–185. doi: 10.1016/j.ins.2011.10.011

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA: IEEE), 3431–3440.

Ma, Y., Zhang, C., Cabezas, M., Song, Y., Tang, Z., Liu, D., et al. (2022). Multiple sclerosis lesion analysis in brain magnetic resonance images: techniques and clinical applications. *IEEE J. Biomed. Health Inform.* 26, 2680–2692. doi: 10.1109/JBHI.2022.3151741

McFarland, H. F., Frank, J. A., Albert, P. S., Smith, M. E., Martin, R., Harris, J. O., et al. (1992). Using gadolinium-enhanced magnetic resonance imaging lesions to monitor disease activity in multiple sclerosis. *Ann. Neurol*. 32, 758–766. doi: 10.1002/ana.410320609

McKinley, R., Wepfer, R., Grunder, L., Aschwanden, F., Fischer, T., Friedli, C., et al. (2020). Automatic detection of lesion load change in multiple sclerosis using convolutional neural networks with segmentation confidence. *Neuroimage Clin*. 25, 102104. doi: 10.1016/j.nicl.2019.102104

Moraal, B., van den Elskamp, I. J., Knol, D. L., Uitdehaag, B. M., Geurts, J. J., Vrenken, H., et al. (2010). Long-interval t2-weighted subtraction magnetic resonance imaging: a powerful new outcome measure in multiple sclerosis trials. *Ann. Neurol*. 67, 667–675. doi: 10.1002/ana.21958

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). Attention u-net: learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*. doi: 10.48550/arXiv.1804.03999

Patti, F., De Stefano, M., Lavorgna, L., Messina, S., Chisari, C. G., Ippolito, D., et al. (2015). Lesion load may predict long-term cognitive dysfunction in multiple sclerosis patients. *PLoS ONE* 10, e0120754. doi: 10.1371/journal.pone.0120754

Reinhold, J. C., Dewey, B. E., Carass, A., and Prince, J. L. (2019). Evaluating the impact of intensity normalization on mr image synthesis. *Proc. SPIE Int. Soc. Opt. Eng*. 10949, 890–898. doi: 10.1117/12.2513089

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Munich: Springer), 234–241.

Rovira, A., Wattjes, M., Miller, D., and Study Group, M. (2015). Evidence-basedguidelines: magnimsconsensusguidelinesontheuseof mri in multiple sclerosis-establishing disease prognosis and monitoring patients. *Nat. Rev. Neurol* 11, 597–606. doi: 10.1038/nrneurol.2015.157

Roy, S., Butman, J. A., Reich, D. S., Calabresi, P. A., and Pham, D. L. (2018). Multiple sclerosis lesion segmentation from brain mri via fully convolutional neural networks. *arXiv preprint arXiv:1803.09172.* doi: 10.1109/ISBI.2018.8363545

Salem, M., Cabezas, M., Valverde, S., Pareto, D., Oliver, A., Salvi, J., et al. (2018). A supervised framework with intensity subtraction and deformation field features for the detection of new t2-w lesions in multiple sclerosis. *Neuroimage Clin*. 17, 607–615. doi: 10.1016/j.nicl.2017.11.015

Salem, M., Valverde, S., Cabezas, M., Pareto, D., Oliver, A., Salvi, J., et al. (2020). A fully convolutional neural network for new t2-w lesion detection in multiple sclerosis. *Neuroimage Clin*. 25, 102149. doi: 10.1016/j.nicl.2019.102149

Steinman, L. (1996). Multiple sclerosis: a coordinated immunological attack against myelin in the central nervous system. *Cell* 85, 299–302. doi: 10.1016/S0092-8674(00)81107-1

Tetteh, G., Efremov, V., Forkert, N. D., Schneider, M., Kirschke, J., Weber, B., et al. (2020). Deepvesselnet: vessel segmentation, centerline prediction, and bifurcation detection in 3-d angiographic volumes. *Front. Neurosci*. 1285, 592352. doi: 10.3389/fnins.2020.592352

Tseng, K.-L., Lin, Y.-L., Hsu, W., and Huang, C.-Y. (2017). "Joint sequence learning and cross-modality convolution for 3d biomedical segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 6393–6400.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., et al. (2010). N4itk: improved n3 bias correction. *IEEE Trans. Med. Imaging* 29, 1310–1320. doi: 10.1109/TMI.2010.2046908

Valverde, S., Cabezas, M., Roura, E., González-Villà, S., Pareto, D., Vilanova, J. C., et al. (2017). Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach. *Neuroimage* 155, 159–168. doi: 10.1016/j.neuroimage.2017.04.034

Zhang, H., and Oguz, I. (2020). "Multiple sclerosis lesion segmentation-a survey of supervised cnn-based methods," in *International MICCAI Brainlesion Workshop* (Lima: Springer), 11–29.

Zhang, H., Valcarcel, A. M., Bakshi, R., Chu, R., Bagnato, F., Shinohara, R. T., et al. (2019). "Multiple sclerosis lesion segmentation with tiramisu and 2.5 d stacked slices," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Shenzhen: Springer), 338–346.

Zhang, Z., Liu, Q., and Wang, Y. (2018). Road extraction by deep residual u-net. *IEEE Geosci. Remote Sens. Lett*. 15, 749–753. doi: 10.1109/LGRS.2018.2802944

frontiers | Frontiers in Neuroscience

TYPE Methods
PUBLISHED 12 August 2022
DOI 10.3389/fnins.2022.964250

OPEN ACCESS

EDITED BY
Michel Dojat,
Institut National de la Santé et de la Recherche Médicale (INSERM), France

REVIEWED BY
Ann-Christin Ostwaldt,
University of Helsinki, Finland
Mostafa Salem,
Assiut University, Egypt

*CORRESPONDENCE
Sebastian Hitziger
s.hitziger@mediaire.de

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 08 June 2022
ACCEPTED 22 July 2022
PUBLISHED 12 August 2022

CITATION
Hitziger S, Ling WX, Fritz T, D'Albis T, Lemke A and Grilo J (2022) Triplanar U-Net with lesion-wise voting for the segmentation of new lesions on longitudinal MRI studies. *Front. Neurosci.* 16:964250. doi: 10.3389/fnins.2022.964250

Triplanar U-Net with lesion-wise voting for the segmentation of new lesions on longitudinal MRI studies

Sebastian Hitziger*, Wen Xin Ling, Thomas Fritz,
Tiziano D'Albis, Andreas Lemke and Joana Grilo

Mediaire GmbH, Berlin, Germany

We present a deep learning method for the segmentation of new lesions in longitudinal FLAIR MRI sequences acquired at two different time points. In our approach, the 3D volumes are processed slice-wise across the coronal, axial, and sagittal planes and the predictions from the three orientations are merged using an optimized voting strategy. Our method achieved best F1 score (0.541) among all participating methods in the MICCAI 2021 challenge *Multiple sclerosis new lesions segmentation* (MSSEG-2). Moreover, we show that our method is on par with the challenge's expert neuroradiologists: on an unbiased ground truth, our method achieves results comparable to those of the four experts in terms of detection (F1 score) and segmentation accuracy (Dice score).

KEYWORDS

multiple sclerosis, lesion detection, longitudinal lesion segmentation, biomedical image segmentation, deep learning, MRI

1. Introduction

Multiple Sclerosis (MS) is a chronic, autoimmune disease which causes lesions in the central nervous system (CNS) (Kuhlmann et al., 2017). Magnetic resonance (MR) imagery is routinely used for diagnosis (Thompson et al., 2018) and prognosis (Brownlee et al., 2019) of MS by assessing the dissemination of CNS lesions in space and time. The lesions appear as white matter hyperintensities on T2 or fluid attenuated inversion recovery (FLAIR) weighted MR sequences. Tracking changes in the lesion load over time facilitates monitoring of MS activity and measuring the efficacy of disease modifying therapies (Sormani et al., 2016).

However, manually detecting and delineating lesions on MR images is a time-consuming and error-prone process with high intra- and inter-expert variability (Altay et al., 2013; Egger et al., 2017), especially when the MR acquisitions differ in terms of scanners, sequences, resolution, and quality. For these reasons, a great number of automated methods for lesion detection have been proposed and originally relied on explicit statistical features such as voxel intensities (Van Leemput et al., 2001; Lao et al., 2008; Shiee et al., 2010; Mortazavi et al., 2012; Schmidt et al., 2012; García-Lorenzo et al., 2013). However, most methods target cross-sectional segmentation and although the ISBI 2015 challenge provided longitudinal datasets, methods were not assessed on their ability to segment new or enlarging lesions (Carass et al., 2017). Existing approaches for new lesions segmentation mostly use classical image processing techniques such as image subtraction (Battaglini et al., 2014; Ganiler et al., 2014; Fartaria et al., 2019), deformation fields (Bosc et al., 2003; Salem et al., 2018), or statistical features from the independently segmented time points (Schmidt et al., 2019).

In the recent past, a number of unsupervised (Baur et al., 2021) and supervised deep learning (Zhang and Oguz, 2020; Ma et al., 2022) methods have been suggested for lesion segmentation. Especially convolutional neural networks (CNN) with encoder-decoder architectures and skip connections such as the U-Net (Ronneberger et al., 2015) have shown good performance in the ISBI 2015 and MICCAI 2016 lesion segmentation challenges (Carass et al., 2017; Commowick et al., 2018). Despite the potential of CNNs for lesion segmentation accuracy, their performance has remained below that of human experts (Carass et al., 2017; Commowick et al., 2018). In addition, deep learning based methods have only recently been designed explicitly for the segmentation of new lesions, which only appear in the follow-up but not the baseline scan. The authors of McKinley et al. (2020) independently segment both time point volumes and use the masks and confidence maps to identify new and enlarging lesions. Fully convolutional networks, in contrast, directly take as input the different time points (Krüger et al., 2020). To incorporate correlations between the different time points in the network architecture, Gessert et al. (2020b) use attention-guided interactions and (Gessert et al., 2020a) convolutional gated recurrent units. The authors of Salem et al. (2020) suggest a combined registration and new lesion segmentation network.

To foster the development of methods for assessing temporal lesion activity, the objective of the MICCAI 2021 *Multiple sclerosis new lesions segmentation* (MSSEG-2) challenge was the design of a method for automatic segmentation of new MS lesions on FLAIR MR sequences. Based on two FLAIR time points of a patient, methods had to delineate lesions that had formed on the follow-up but not on the baseline scan. The performance of the submitted algorithms was evaluated in terms of (a) their ability to detect new lesions, measured by the F1 score, and (b) the segmentation accuracy of the new lesions, measured by the Dice score. Pairs of FLAIR volumes from 40 patients were given to the challenge participants for training the algorithms, another 60 patients were held out for validation.

Our approach to this challenge starts with the observation that plain end-to-end CNNs with U-Net like architecture perform exceptionally well in most biomedical image segmentation tasks. This was clearly shown by the authors of the nnU-Net (Isensee et al., 2021), a framework which relies on either 2D or 3D U-Nets (Çiçek et al., 2016) and adjusts its hyperparameters to the given segmentation task. It achieved excellent results in many segmentation challenges, including the ISBI 2015 longitudinal lesion segmentation. While the authors found their 3D version to outperform the 2D counterpart, the performance of 2D models can be enhanced by integrating more 3D information. The triplanar or 2.5D approach processes slices across all three orthogonal directions and then merges the predictions from the different orientations (Roy et al., 2019; Henschel et al., 2020; Sundaresan et al., 2021). A triplanar approach was also used by the winner of the MICCAI 2016 challenge (McKinley et al., 2016).

In this approach, we adapt the triplanar segmentation approach and use a single 2D U-Net (Ronneberger et al., 2015) as base model. This model is trained on slices from the axial, coronal, and sagittal planes. To incorporate information from both time point volumes, corresponding slices from the two volumes are paired and given as a two-channel input. Compared to other triplanar U-Net approaches, our architecture contains two main differences:

- It uses a single U-Net which is trained on sagittal, coronal, and axial slices, allowing to share common features across orientations. This is opposed to the training of three orientation-specific U-Nets in previous approaches (McKinley et al., 2016; Roy et al., 2019; Sundaresan et al., 2021).
- For merging the predictions from different orientations, we observed that single orientation predictions tend to contain many false positive lesions. Hence, we challenge the commonly used softmax averaging and compare it to voting strategies of different sensitivity.

We submitted two segmentation pipelines to the challenge, *mediaire-A* and *mediaire-B*, which use the same model architecture but make use of different data: while the model in mediaire-A is trained only on the official training data, we use additional datasets for training the model in mediaire-B, as described below. Besides this difference, the two pipelines are identical.

Both segmentation pipelines were evaluated by the challenge organizers on the unseen test set, resulting in mediaire-B ranking 1st and mediaire-A 3rd across all submitted models in terms of detection performance (F1 score). In additional validations, where we compare our pipelines to the challenge's annotators, we show that our algorithms are on par with the neuroradiologists in terms of F1 score and segmentation accuracy (Dice score).

2. Materials and equipment

The majority of the 3D FLAIR images used in this study for training and testing the models was provided

TABLE 1 Datasets used for training, validation, and testing, provided by the MSSEG-2 challenge organizers and internal data.

Name	Source	No. of patients	Sequence	Voxel resolutions
TRAIN-MSSEG2	OFSEP HD	40 (29)	3D FLAIR	0.5–1.2 mm anisotropic
TEST-MSSEG2-NL	OFSEP HD	32	3D FLAIR	0.5–1.2 mm anisotropic
TRAIN-B-NL	Internal	25	3D FLAIR	0.5–1.2 mm anisotropic
VAL-A-NL	Internal	20	3D FLAIR	0.5–1.2 mm anisotropic

The suffix -NL denotes that all datasets exhibit new lesions, otherwise the number of patients with new lesions is denoted in parentheses. Each dataset contains 3D FLAIR images of a patient, a baseline and a follow-up scan. Note that the internal datasets in VAL-A-NL are a subset of TRAIN-B-NL. The datasets are described in more detail in Section 2.

by the MSSEG-2 challenge organizers. In addition, 25 internal datasets with pairs of 3D FLAIR images were used for training and validation. All used data, including the corresponding ground truth masks, is described in the following paragraphs. An overview of the datasets is provided in Table 1.

2.1. MSSEG-2 datasets

The data provided by the organizers of the MSSEG-2 challenge consists of 100 pairs of 3D FLAIR weighted MRI sequences from the OFSEP HD cohort[1], each corresponding to two scans of the same patient acquired at different time points (1–3 years apart). The images had been acquired on 15 MRI scanners from different manufacturers (GE, Philips, Siemens) in different locations and exhibited varying resolutions and anisotropic voxel sizes, with resolutions between 0.5 and 1.2 mm. Besides the 3D FLAIR sequences, no other sequences were used for the creation of the ground truth or provided to the participants. For each data pair, a consensus ground truth mask was created from the delineations of four expert neuroradiologists using the protocol described in the following paragraph. Forty of the 100 3D FLAIR image pairs were provided to the challenge participants for training their models, together with the four experts' new lesion segmentations and the consensus ground truth masks. We will refer to these datasets as *TRAIN-MSSEG2*. The remaining 60 pairs were used for evaluating the submitted models. These datasets, including consensus ground truth and the experts' segmentation masks, were provided to the participants after publication of the official challenge results for further analysis. For the calculation of the challenge's main metrics, i.e., the Dice and the F1 score, only the 32 of the 60 dataset pairs that exhibited new lesions were taken into account. We will denote this subset, which is used for the evaluations in Section 4, as *TEST-MSSEG2-NL*. The remainder of the MSSEG-2 test datasets were used by the challenge organizers for further evaluations which are outside the scope of this study and are not used here. The information on data, data access, and annotations is also available on the challenge websites.[2–4]

2.1.1. Consensus reading protocol

For every dataset, manual delineations of new lesions were performed by four expert neuroradiologists, medically trained for MS and at the start of their career (a few years after taking their permanent position). They received instructions to delineate lesions not in contact with other lesions and above 3 mm in size in one of the image planes. The delineation was performed using the software ITK Snap, for which the experts had received a user manual.

Based on the resulting four expert segmentation masks, a consensus ground truth was created with the help of a senior expert neuroradiologist with much longer experience in neuroradiology and MS than the other four experts. The ground truth creation was done in two steps: (i) lesion approval or rejection and (ii) delineation. In step (i), every *majority lesion*, i.e., found by at least three of the four experts, automatically transferred to the ground truth; for any *disputed lesion*, i.e., found by at most two of the experts, the senior expert decided whether to accept or reject it. In step (ii), the delineation of every *accepted lesion* was calculated using the STAPLE (Akhondi-Asl and Warfield, 2013) algorithm based on the concerned experts' lesion segmentations.

As the ground consensus ground truth masks were created by the experts, a direct evaluation of the experts on this same ground truth would be biased. Therefore, we additionally created expert-specific unbiased ground truth masks to compare our pipelines to the experts (see Section 3.6).

[1] https://www.ofsep.org/en/hd-cohort

[2] https://portal.fli-iam.irisa.fr/msseg-2/data/

[3] https://gitlab.inria.fr/amasson/lesion-segmentation-challenge-miccai21/-/blob/master/DATASET.md

[4] https://files.inria.fr/empenn/msseg-2/Challenge_Day_MSSEG2_Introduction.pdf

2.2. Internal datasets

While our model in pipeline *mediaire-A* was trained only on the challenge's official 40 patient volumes, we added internal datasets to train pipeline *mediaire-B*. These consisted of 25 pairs of all 3D FLAIR images from 25 patients, where each pair exhibited new lesions, and will be referred to as *TRAIN-B-NL*. The datasets had been acquired on different scanners by Siemens (Aera 1.5 T, Magnetom Vida 3.0 T, Skyra 1.5 T, Skyra 3.0 T) and Philips (Achieva 1.5 T, Achieva 3.0 T) and had anisotropic voxel resolutions between 0.5 and 1.2 mm. In order to match the challenge data, we also only used the 3D FLAIR sequences for ground truth creation and model training without any additional sequences. Each of the image pairs was annotated by up to four experts (a medical doctor and neuroscientist, a radiologist, and two radiographers with special training in segmenting MS lesions, all of them with more than 2 years of experience with MS-specific MRI interpretation and annotation), and a consensus ground truth had been formed similar to the one used in the challenge, as described in Section 2.1.1. The segmentations were performed using an annotation application integrated into an internal image viewer.

As we trained the models in pipelines mediaire-A and mediaire-B on 5 data folds (see Section 3.1.2) with random 80–20% train-validation splits, we validated the individual fold models on these validation splits. However, for the validation of the orientation merging strategies (cf. Section 3.4), we required the final ensemble model of all folds. For this purpose, we used a subset of 20 patients from TRAIN-B-NL. Since pipeline mediaire-B was trained on these datasets, they could only be used to validate mediaire-A and we denote them as *VAL-A-NL*. We assume that the results of comparing the orientation merging strategies transfer qualitatively from mediaire-A to mediaire-B, as the pipelines are very similar.

2.3. Pre-processing

For each patient in the datasets provided by the MSSEG-2 challenge, the organizers had transformed the two scans onto a common middle point through rigid registration.

We further applied the following preprocessing steps to all 3D FLAIR image pairs in the challenge's and internal training, validation, and test sets: (1) affine registration of each pair of 3D FLAIR images to the MNI template, (2) cropping the FOV to an area around the brain, (3) resampling the volume to 256 × 256 × 256 voxels, and (4) pixel normalization through mean subtraction and division by the standard deviation.

To increase the generalization ability of the model, data augmentation was performed on the preprocessed 3D volumes of the training sets during training, including contrast augmentation, rotations, flipping across the three orthogonal planes, elastic deformations, and bias field augmentation.

3. Methods

The basis for our segmentation pipeline, which we refer to as triplanar U-Net, is a 2D U-Net (Ronneberger et al., 2015). It has two input channels with corresponding slices—either axial, coronal, or sagittal—from the two different time points of each patient. The output of the model is a single-channel 2D binary mask, representing the segmentation of the new lesions found in the corresponding slice. For an illustration and the dimensions of the network (see Figure 1).

Compared to previously suggested triplanar U-Net architectures (Roy et al., 2019; Sundaresan et al., 2021), our approach has two main differences:

1. It uses a single U-Net which is trained on sagittal, coronal, and axial slices, allowing to share common features across orientations. This is opposed to training three orientation-specific U-Nets in the former approaches. Note that this procedure requires the all slices to be of the same dimensions, which is ensured by resampling the volume to a regular cube, as described in Section 2.3.
2. For merging the predictions from different orientations, we test different techniques. In addition to softmax averaging (i.e., averaging the predicted probabilities), we implement and validate three voting strategies of different sensitivities to optimize the method's recall and precision. The best strategy is then implemented.

3.1. Model training

The U-Nets trained for pipelines mediaire-A and mediaire-B use exactly the same training protocol and hyperparameters. However, only the 40 datasets in TRAIN-MSSEG2 were used for training mediaire-A, while mediaire-B was trained on TRAIN-MSSEG2 plus the additional 25 datasets in TRAIN-B-NL (see Section 2).

We trained the triplanar U-Net on batches, each combining a total of 20 axial, coronal, and sagittal slices from different patient volumes for robustness. For the updates of the model weights, we used stochastic gradient descent with momentum and an initial learning rate of 0.0001, which was reduced when the validation loss plateaued. Training was performed with early stopping when the validation loss stopped decreasing, which was usually the case after around 50 epochs.

3.1.1. Loss function

Recently, it has been observed that combined loss functions tend to be more robust and accurate, especially in segmentation tasks with high class imbalance. For instance, the self-configuring segmentation network nnU-Net (Isensee et al., 2021) uses the combo loss as a default, which is the sum of the

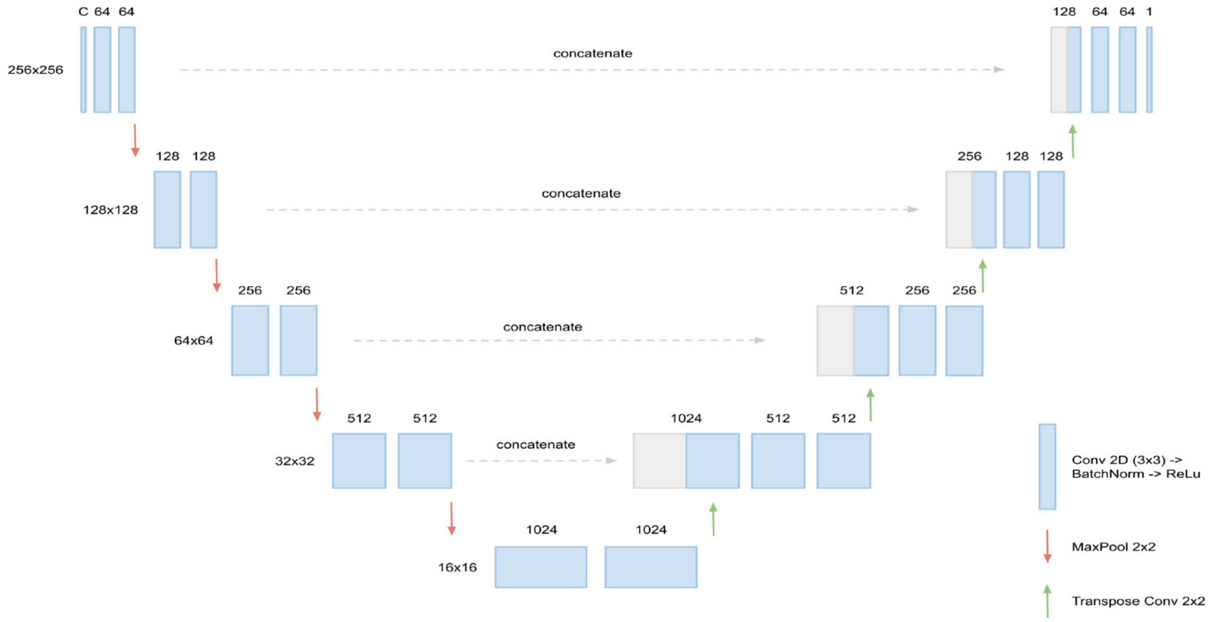


FIGURE 1
Architecture of the 2D U-Net used. Input are $C = 2$ slices (axial, coronal, or sagittal) corresponding of dimension 256 × 256, corresponding to the two FLAIR time points. In every layer of the encoding branch (left), two convolution blocks, consisting of Conv2D (3 × 3), BatchNorm, and ReLU activation, are applied. When passing to a new layer, dimensions are reduced to half by max pooling while the number of channels is doubled in the first convolution block. In the decoding branch (right), the max pooling operations are replaced by transpose convolutions for upsampling and the data from the corresponding layer of the encoding branch is concatenated through skip connections. Output of the U-Net is the 256 × 256 binary mask containing the new lesions segmentation.

Dice loss and the cross entropy loss. For training our models, we use a combination of Dice loss and the TopK loss (Wu et al., 2016), which has shown good performance, for example in the winning and runner up model (Ma, 2021) of the Miccai 2020 ADAM segmentation[5] challenge. The TopK is a hard-mining variant of the cross-entropy loss, focussing only on the k% hardest voxels. We denote with $g_{ic} \in \{0,1\}, p_{ic} \in (0,1)$ the ground truth index and the softmax prediction for voxel i and class c, respectively, and by select_top$_k$ the function that returns the k% largest values. Then the partial loss functions $L_{\text{Dice}}, L_{\text{TopK}}$, and the total loss function L_{total} are defined by

$$L_{\text{Dice}} = 1 - \frac{2\sum_{i,c} g_{ic} \cdot p_{ic}}{\sum_{i,c} g_{ic}^2 + \sum_{i,c} p_{ic}^2}$$

$$L_{\text{TopK}} = -\text{mean}\left(\text{select\_top}_k\left(S_{CE}\right)\right)$$

$$L_{\text{total}} = L_{\text{Dice}} + L_{\text{TopK}}$$

where $S_{CE} = \{g_{ic}\log(p_{ic})\}_{i,c}$ is the set of cross entropy scores for all voxels and classes. Note that L_{TopK} reduces to the cross entropy loss for $k = 100$. In our experiments, we chose $k = 10$.

3.1.2. Cross validation

For each pipeline, mediaire-A and mediaire-B, we train the triplanar U-Net five different data folds, resulting in models M_0, ..., M_4. For each M_i, we hold out 20% of the training data. For inference, the ensemble of all five fold models will be used for segmentation, as explained in Section 3.2.

3.2. Inference

The segmentation process at inference is depicted in Figure 2. From the two 3D FLAIR volumes of each patient, three datasets are created, consisting of pairs of axial, coronal, and sagittal slices, respectively. For each such dataset, inference is performed slice-wise with every fold model M_0, ..., M_4 and the ensemble average of the resulting softmax slices is calculated, resulting in axial, coronal, and sagittal predictions. Then, the three single-orientation predictions are merged to produce the final segmentation mask. This is explained in detail in Section 3.4, where different merging strategies are compared.

3.3. Metrics

To evaluate our experiments, we use the official performance metrics from the MSSEG-2 challenge. These are defined *via* the true positives (TP), false positives (FP), and false negatives (FN) on the lesion level ($\text{TP}_\text{l}, \ldots$) and the voxel level ($\text{TP}_\text{v}, \ldots$). The

[5] https://adam.isi.uu.nl/

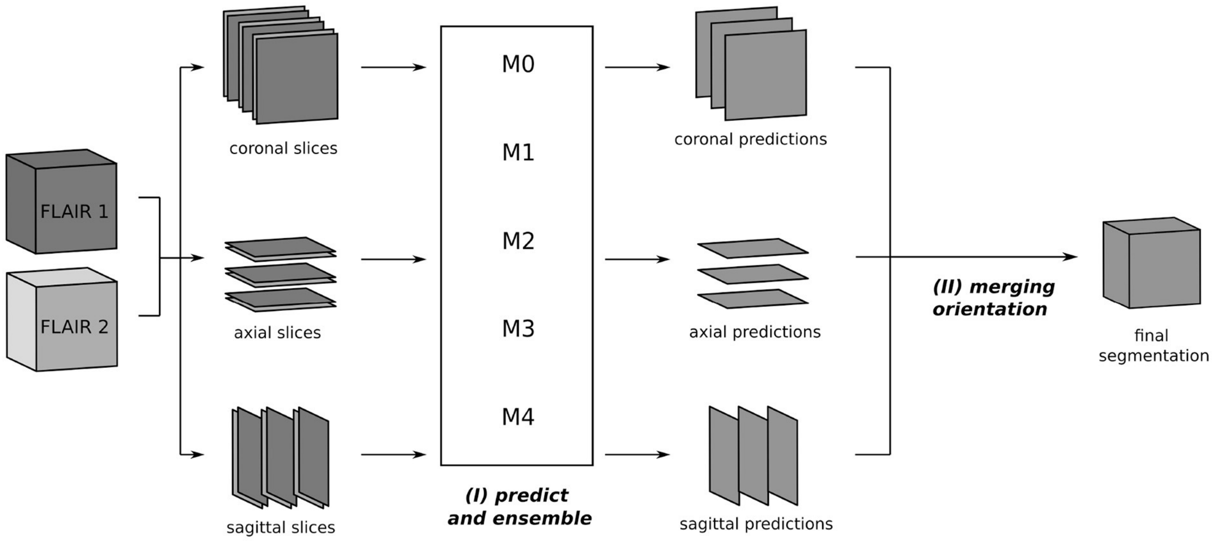


FIGURE 2
Segmentation process using the triplanar U-Nets M_0, ..., M_4 trained on slices from the three orthogonal planes of the different training folds. The 3D FLAIR input volumes are sliced along the coronal, axial, and sagittal planes and grouped together in pairs of corresponding slices. For every orientations, the segmentation is now performed independently: (I) each slice pair is given as a two-channel 2D input to the models M_0, ..., M_4 and predicted softmax scores are averaged. (II) In the final step, the predictions of the individual orientations are merged to yield the final segmentation.

principal new lesion *detection metric* is the F1 score, but we also investigate precision and recall. They are defined as

$$\begin{aligned}\textbf{F1 score} &= \frac{2 * \mathrm{TP_l}}{\mathrm{FP_l} + 2 * \mathrm{TP_l} + \mathrm{FN_l}}\\ \textbf{Recall} &= \frac{\mathrm{TP_l}}{\mathrm{TP_l} + \mathrm{FN_l}}\\ \textbf{Precision} &= \frac{\mathrm{TP_l}}{\mathrm{TP_l} + \mathrm{FP_l}}\end{aligned}$$

The principal *segmentation metric* used is the Dice score, which is the equivalent of the F1 score on a voxel level:

$$\textbf{Dice score} = \frac{2 * \mathrm{TP_v}}{\mathrm{FP_v} + 2 * \mathrm{TP_v} + \mathrm{FN_v}}$$

We note that the quantities $\mathrm{TP_l}$, $\mathrm{FP_l}$, and $\mathrm{FN_l}$ depend on the definition of when lesions in the prediction and the ground truth shall be matched. In the competition evaluation, a match requires certain overlap thresholds to be fulfilled. This is described in detail in the official documentation[6]. All metrics in this paper were calculated using the "animaSegPerfAnalyzer" command from the Anima toolbox[7], which was also used by the challenge organizers to calculate the official results for the leaderboard.

3.4. Validation of orientation merging strategies

As described in Section 3.2, the inference pipeline requires to merge predictions from different orientations. While softmax averaging is commonly used for this step (McKinley et al., 2016; Roy et al., 2019; Sundaresan et al., 2021), we additionally compare three different voting strategies in order to find the optimal balance of recall and precision. This step is depicted in Figure 2(II).

Starting from the predicted probability maps (softmax scores) of each orientation, we first calculated the softmax average as a baseline approach. For the other approaches, which operated on a lesion level, we first thresholded the softmax scores of each of the three orientations to yield hard predictions. Then three different lesion-selection strategies were applied: A lesion was predicted if detected in (a) at least one orientation (union); (b) at least two orientations (majority); (c) all orientations (unanimous voting). The exact segmentation of each selected lesion was defined as the union of the corresponding positive voxels across orientation predictions.

The four approaches were implemented into pipeline mediaire-A and used for segmenting the internal datasets VAL-A-NL (see Section 2.2). The segmentation masks were then rated against the corresponding expert annotations and the results in terms of F1 score, precision, recall, and Dice are shown in Figure 4. The optimal strategy was chosen based on the best F1 score. As mediaire-B was trained on datasets containing

[6] https://portal.fli-iam.irisa.fr/files/2021/06/MS_Challenge_Evaluation_Challengers.pdf

[7] https://anima.irisa.fr/

VAL-A-NL, this validation could not be performed directly for this pipeline. However, due to the similarity of both pipelines, it was assumed that the optimal strategy for mediaire-A would also be the best strategy for mediaire-B.

3.5. Challenge evaluation: Comparison to other participants

Both pipelines, mediaire-A and mediaire-B were submitted to the challenge among a total of 29 rated submissions from 24 teams. For all submitted pipelines, the organizers calculated the predictions on the test dataset with 60 patients. However, for calculating the scores on the detection (F1 score) and the segmentation leaderboard (Dice score), only the 32 patients in TEST-MSSEG2-NL, i.e., those with at least one new lesion, were taken into account (cf. Section 2.1).

The official raw scores for all submissions are publically available[8]. In addition to the official average scores across patients, we visualize the score distributions across patients.

3.6. Challenge evaluation: Comparison on unbiased ground truth

For assessing how our pipelines mediaire-A and mediaire-B perform compared to human neuroradiologists, we evaluate the lesion delineations conducted by the four challenge experts. A naive approach would simply rate the human segmentation masks against the consensus ground truth, as it has been done for the segmentation masks produced by the algorithms. In fact, the resulting scores from this approach are published by the challenge organizers and correspond to those in Figure 5. However, this approach comes with a problem: The ground truth has been created based on the individual segmentation masks of the human experts, which makes it biased toward these experts. Thus, the measured human performance is likely to be higher than it would be on an unbiased ground truth.

We therefore suggest a comparison on an unbiased ground truth of the official challenge test datasets with new lesions, TEST-MSSEG2-NL (cf. Section 2.1), constructed from the corresponding experts' segmentation masks and the consensus ground truth masks, which were provided to the participants after the challenge. For a fair comparison, the segmentation mask s_i of Expert i should be rated against a ground truth u_i whose definition is independent of s_i. We create u_i from the segmentation masks of all other experts $S_{\bar{i}} = \{s_j | j \neq i\}$ using the challenge's consensus reading protocol, as described in Section 2.1.1. By doing so, we exclude the minimal information necessary (segmentation s_i) to unbias the ground truth while preserving the maximal expert knowledge available (segmentations $S_{\bar{i}}$ and senior expert decisions). The protocol involves (i) the acceptance or rejection of lesions found by any expert and (ii) calculating the segmentation of each accepted lesion through majority voting. While (ii) is a simple voxel-wise calculation, the decisions (i) on disputed lesions are taken by a senior expert. We cannot consult the senior expert, however, we can derive the decisions as they are implicitly contained in the consensus ground truth c. The only assumption we make for this derivation is that of constant decisions: if a disputed lesion l was approved (rejected) by the senior expert in the original reading, this same lesion l is also approved (rejected) in a different reading (where the number of total expert masks may be different).

The complete lesion selection process (i) is illustrated in Figure 3 for the example of creating an unbiased ground truth u_4 for Expert 4: First, all lesions in the segmentation masks s_1, s_2, s_3 are grouped into *majority lesions* (found by at least two experts) and *disputed lesions*. The majority lesions are automatically accepted according to the protocol (cf. Section 2.1.1). If a lesion l is disputed, i.e., found by a single expert, it must have been found by at most two experts in the original reading (as this reading had an additional expert). Hence, it was already a disputed lesion in the original reading (cf. Section 2.1.1) and we can derive the senior expert's decision from the consensus ground truth c: if c contains lesion l, it has previously been approved and we include it into the unbiased ground truth. Otherwise, it has previously been rejected and we exclude it.

Having selected all relevant lesions, their exact segmentations are calculated as (ii) the voxel-wise majority vote across the segmentation masks s_1, s_2, s_3, resulting in the unbiased ground truth u_4.

We apply the protocol (i, ii) defined above to generate unbiased ground truth masks $u_1, \ldots, u_4$ for all experts and all patients in the test set. Each expert i is now evaluated by rating the segmentation s_i against u_i in terms of recall, precision, F1 score, and Dice score. As each expert is now rated against a different ground truth, their scores are not directly comparable. Hence, for every expert i, we also rate pipeline mediaire-A and mediaire-B on u_i and compare the resulting scores to this expert. From these four individual assessments, we then calculate the mean scores for experts, mediaire-A, and mediaire-B to compare the average performance model vs. human performance.

3.6.1. Statistical testing

In order to assess whether our pipeline performance is comparable to or better than the expert performance, we tested for statistical significance. To this end, we first defined a margin $d = 0.05$ and regarded performances as *comparable* if their absolute difference was below d. If we wanted to test for comparability only, we could use equivalence tests with margin d. However, since we want to investigate if the models are

[8] https://zenodo.org/record/5775523#.YkVWKjxCRhE

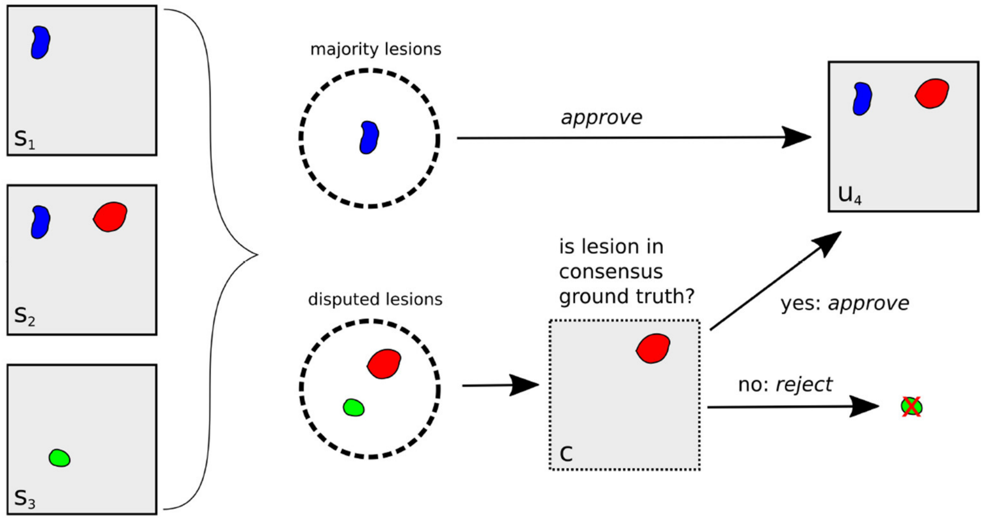


FIGURE 3
Illustration of unbiased ground truth creation u_4 for Expert 4, (right) from the segmentation masks of experts 1, 2, and 3 (left). If a lesion is found by at least two experts (blue lesion), it is automatically selected for u_4. Otherwise, it is a disputed lesion (green, red) and has to be decided on by the senior expert. If it is contained in the consensus ground truth c (red lesion), it has been accepted by the senior expert before, so we approve it. Otherwise, it is rejected.

comparable *or better* than the experts, we choose to conduct *noninferiority tests* (Walker and Nowacki, 2011) with margin d. This leads to the null hypothesis H_0 that the expected difference $E(Y - X)$ between expert performance Y and pipeline performance X is above d. We assess the validity of the null hypothesis H_0 using paired difference Student's t-tests with significance level $\alpha = 0.05$. The tests are conducted for F1 score and Dice and for every combination of pipeline and expert. In addition, we test the pipelines average performance across the different masks u_i against the average performance across experts.

3.7. Implementation

The model is implemented and trained in Python using the PyTorch package.

4. Results

4.1. Validation of orientation merging strategies

As described in Section 3.4, we tested four different strategies for merging the predictions from the three orthogonal orientations: three lesion voting procedures and softmax averaging. Figure 4 shows the performances of the respective methods validated on the datasets in VAL-A-NL (Section 2.2). The most inclusive strategy, union of all lesions, achieves the best recall but a very low precision and thus a bad overall F1 score. While majority voting is significantly better, unanimous voting clearly achieves best F1 score due to a high precision. The baseline method, softmax averaging, shows a performance similar to majority voting. It is only in terms of segmentation accuracy (Dice), that softmax averaging outperforms all voting strategies.

It is interesting that the precision gain when using the very restrictive unanimous voting strategy largely outweighs the slight loss in recall. Apparently, the weakness of a single-orientation model is not its capability to find enough lesions—it rather bears the risk of classifying too many confounding hyperintensities as lesions. The unanimous voting strategy could also be reformulated as: Accept only lesions which have been "seen" in all three orientations.

Since the focus of the challenge is on the detection performance and the most important metric is the F1 score, we implement unanimous voting in both pipelines mediaire-A and mediaire-B.

4.2. Challenge evaluation: Comparison to other participants

The boxplot in Figure 5 shows the official F1 scores of the challenge's main leaderboard for all 29 submissions rated against the consensus ground truth of the datasets in TEST-MSSEG2-NL. It also includes the scores obtained by the experts' segmentation masks when rated against the consensus ground

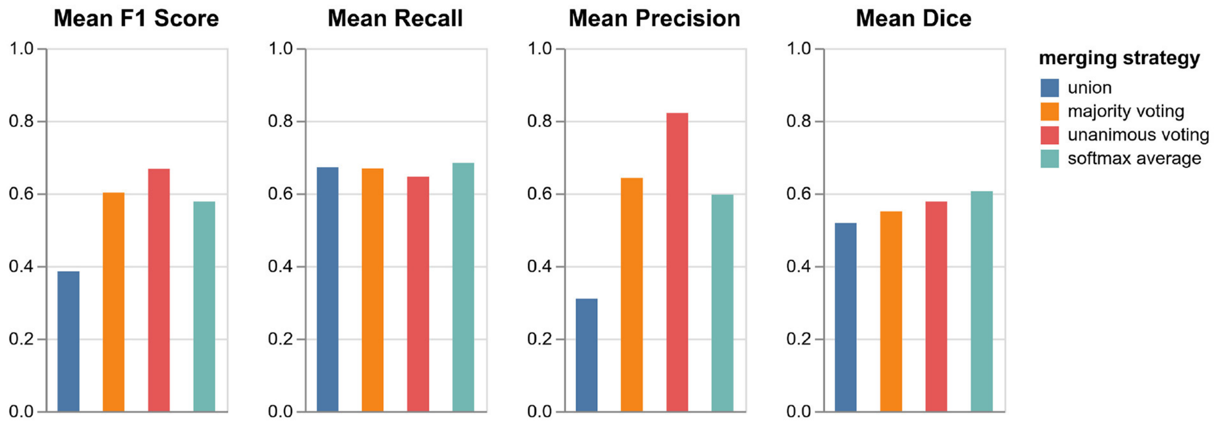


FIGURE 4
Performance of pipeline mediaire-A with different strategies for merging the predictions from axial, coronal, and sagittal orientation. Scores are calculated on the 20 datasets in VAL-A-NL. The approaches union, majority voting, unanimous voting, and softmax averaging are compared in terms of F1 score, precision, recall, and Dice.

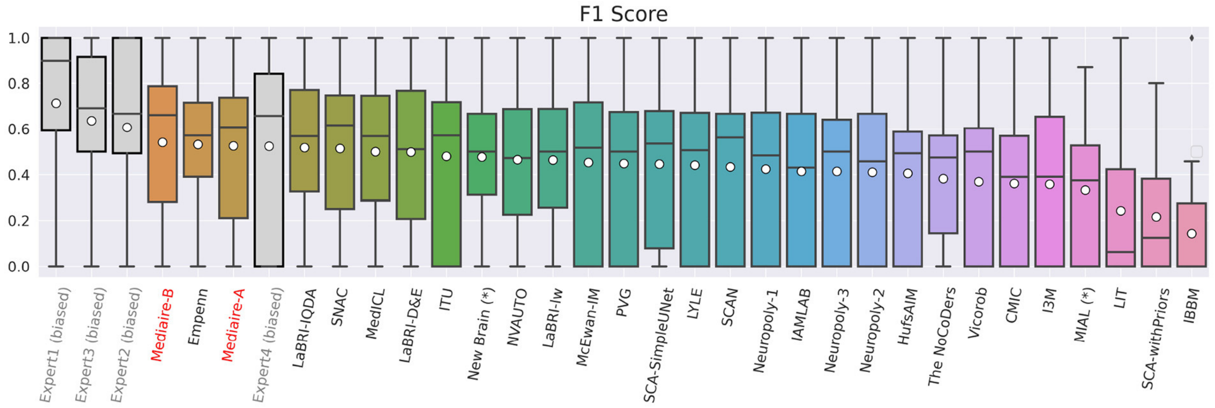


FIGURE 5
F1 scores of experts (gray), our pipelines mediaire-A and mediaire-B (red), and the models submitted by the other MSSEG-2 participants (black), calculated on the 32 datasets in TEST-MSSEG2-NL, all of which exhibited new MS lesions. Horizontal bars indicate the median and white circles the mean values. All experts and models are ordered by their mean F1 score, which also determined the ranking of the main challenge leaderboard. The three best performing methods are our pipeline mediaire-B, Empenn, and pipeline mediaire-A. The scores of the expert segmentations are shown for reference, however, these scores are biased as discussed in Section 3.6.

truth. As discussed in Section 3.6, the latter scores are positively biased, as the rated segmentation masks were the basis for the ground truth creation. An unbiased comparison between our submissions and expert performance is therefore done in Section 4.3.

In terms of detection performance (F1 score), the three best methods are mediaire-B (0.541), Empenn (0.532), and mediaire-A (0.525), respectively. The second best submission Empenn performed segmentation with a 3D nnU-Net (Isensee et al., 2021) trained on official and internal datasets. The great majority of submissions, including all top 10 methods, used deep learning with 3D or 2.5D U-Net-like architectures.

4.3. Challenge evaluation: Comparison on unbiased ground truth

The results of the comparison between algorithms and experts on the unbiased ground truth of the TEST-MSSEG2-NL data (cf. Section 4.3) are shown in Figure 6. Clearly, both pipelines mediaire-A and mediaire-B have higher recall but lower precision than the experts (second and third plot, respectively). In the overall detection performance, the algorithms slightly outperform the experts on average (first plot, last block) and only Expert 1 achieves a slightly higher F1 score. This is in contrast to the evaluation on the (biased) consensus

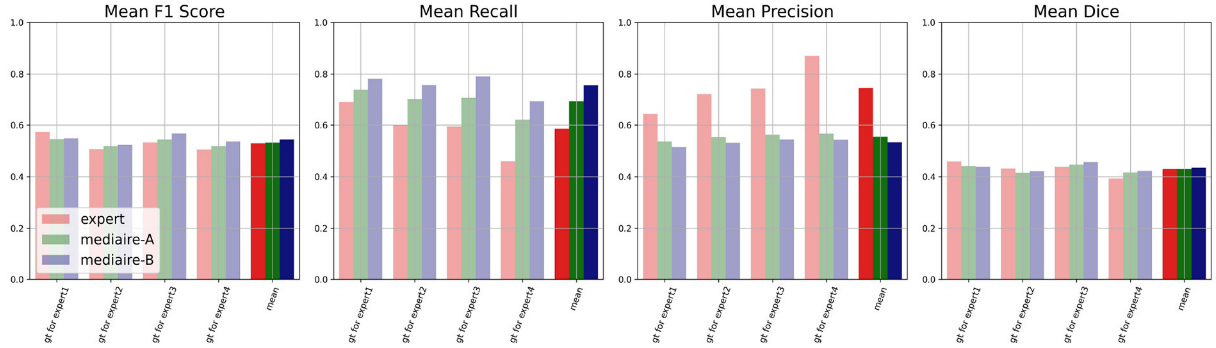


FIGURE 6
Unbiased comparison of each of the experts 1, . . . , 4 to segmentation pipelines mediaire-A and mediaire-B, in terms of F1 score, recall, precision, and Dice. The scores are calculated on the unbiased ground truth masks of the 32 patients in TEST-MSSEG2-NL. For every expert, this unbiased mask is constructed from the other expert masks (cf. Section 3.6) resulting in four individual comparisons with different ground truths. In the last block of each plot, we show the average score across the four experts against the average score of each pipeline across the different ground truth masks. Clearly, the segmentation pipelines have a higher recall while the experts have higher precision. In terms of F1 and Dice scores, the pipelines achieve slightly higher average results.

TABLE 2 *p*-values for testing non-inferiority of the performance of a pipeline (rows) compared to an expert (columns) with a margin of $d = 0.05$.

		Expert 1	Expert 2	Expert 3	Expert 4	Mean
mediaire-A	F1 score	0.3186	**0.0367**	0.1160	0.1015	0.0551
	dice	0.2037	**0.0447**	0.0873	**0.0261**	**0.0161**
mediaire-B	F1 score	0.2935	**0.0248**	**0.0361**	**0.0326**	**0.0135**
	dice	0.2254	**0.0245**	0.0532	**0.0204**	**0.0101**

Significant *p*-values ($p < 0.05$) are marked in bold. The last column shows the *p*-values for the scores averaged across the different experts. The values are calculated using paired *t*-tests using the pipeline and expert scores on the 32 datasets in TEST-MSSEG2-NL.

ground truth in Section 4.2, where experts 1, 2, and 3 had significantly higher F1 scores than all submitted methods. In terms of segmentation accuracy (last plot), expert and algorithm performances are very similar.

The differences in F1 score and Dice between experts and models are relatively small and statistically not significant. We therefore tested for non-inferiority, i.e., if each pipeline's performance is within a $d = 0.05$ margin or better than each expert's performance using paired t-tests, as described in Section 3.6.1. The resulting *p*-values are shown in Table 2 with the significant values ($p < 0.05$) in bold. In terms of F1 score, the results for mediare-A are significant only when compared to Expert 2, while for mediaire-B they are significant when compared to expert experts 2, 3, 4 and the average across experts. In terms of Dice score, test results for mediaire-A and mediaire-B are significant when compared to experts 2 and 4 and the mean of experts.

In conclusion, we showed that our better pipeline, mediaire-B, is at least comparable to three (two) of the four experts and the expert average in terms of F1 score (Dice score).

Processing of the segmentations took an average of 97 s per dataset (±2 s standard deviation) on a Laptop with graphics processing unit (CPU: Intel Core i7-10750H, 32 GiB RAM; GPU: NVIDIA GeForce RTX 2080 Super, 8 GiB RAM).

5. Discussion

The detection of new MS lesions is clinically important for diagnosis, prognosis, and treatment monitoring. An automatic method with a detection and segmentation accuracy comparable to that of an expert neuroradiologist can be highly beneficial to improve diagnostic quality by providing a "second pair of eyes," to decrease inter-rater variability, and to reduce the manual reading time and effort. For instance, the study in Altay et al. (2013) assumed a maximal time of 10 min for a clinician to count lesions on an MS dataset and showed significant variability in the results of clinicians of different expertise level.

We presented a deep learning based approach using the U-Net to segment new lesions on 3D FLAIR volumes by processing slices from axial, coronal, and sagittal planes. We showed that our U-Net based segmentation pipelines not only outperform all other competing methods in the MSSEG-2 challenge in terms of detection accuracy measured by lesion-wise F1-score. They are also on par with an average expert neuroradiologist, both in detection (F1 score) and segmentation accuracy (Dice score) when compared on an unbiased ground truth. The automatic lesion segmentation was performed in <2 min on a Laptop with GPU, which is significantly less than the expected annotation time needed by a human annotator.

As a major difference to other triplanar or 2.5D U-Nets with softmax averaging of orientations, our algorithm uses unanimous voting which only accepts lesions that have been confirmed in all three orientations. Even though this approach may seem restrictive, it is actually aligned with the diagnostic guideline for MS lesions detection that lesions should be confirmed on multiple planes to avoid false positive results (Filippi et al., 2019). In addition, we saw in Section 4.3 that our algorithm outperformed the human experts in recall but had lower precision. For any less restrictive strategy than unanimous voting, this discrepancy would have been even more severe, which also becomes clear from the validation in Figure 4. We therefore suggest that unanimous voting is a key factor for the good performance of our algorithm.

Another slight performance gain was achieved through the use of additional training data, leading to a higher recall of the model mediaire-B compared to mediaire-A (cf. Figure 6). While the augmentation of the training size does not always lead to improved model performance in our experience, we took particular care to optimize the distribution of the additional data: (i) we added only patients with new lesions, leading to a recall improvement with only slight decrease in precision, and (ii) the corresponding consensus ground truth was created using a protocol similar to the one used by the challenge organizers.

While the presented outcomes are encouraging, there is still room for improvement: our algorithms had a higher recall than the average neuroradiologist, however, the precision was lower. Future works may therefore focus on an improved false positive reduction. Furthermore, we could observe a performance gain by increasing the relatively small training set from 40 (mediaire-A) to 65 datasets (mediaire-B). Training on a larger set could therefore increase performance even further.

Another limiting factor of this study is the use of only 3D FLAIR datasets acquired with high resolution which does not necessarily reflect the clinical reality. While the presented approach can be applied to 2D, low-resolution, or low-quality datasets, we do not know how well the present results translate to such a data regime. In particular, the information in thick slices may not be sufficient to distinguish a lesion from brain tissue. To this end, we suggest a follow-up study with a larger and more diverse training and test set in order to yield a complete assessment covering a broad range of clinical settings.

Code availability statement

The code for training the models in pipelines mediaire-A and mediaire-B and performing the segmentations is proprietary and not publically available. For studies involving the comparison to or the reproduction of the presented results, the trained models may be provided within a research collaboration. In this case, researchers are invited to reach out to the CEO of mediaire, Andreas Lemke (a.lemke@mediaire.de).

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: The data for training pipeline mediaire-A and calculating the results was obtained as part of the MICCAI 2021 MSSEG-2 challenge (https://portal.fli-iam.irisa.fr/msseg-2/data/). Access was restricted to challenge participants. Requests to access these datasets should be directed to challenges-iam@inria.fr. The additional datasets used for validation and training of pipeline mediaire-B are not readily available because they were provided to mediaire GmbH by its customers' patients for the improvement and validation of its algorithms. mediaire GmbH does not hold the rights to distribute the data to third parties. Questions concerning these datasets should be directed to the corresponding author of this article.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This project was funded partly by the European Regional Development Fund and the City Berlin (Grant number: 10631009926).

Acknowledgments

The authors would like to thank the organizers of the MICCAI 2021 MSSEG-2 challenge for providing the data, conducting the evaluation of the submitted algorithms from all 24 teams and clarifying all our questions.

Conflict of interest

All authors are employed at mediaire GmbH (as machine learning engineers or managers).

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Akhondi-Asl, A., and Warfield, S. K. (2013). Simultaneous truth and performance level estimation through fusion of probabilistic segmentations. *IEEE Trans. Med. Imaging* 32, 1840–1852. doi: 10.1109/TMI.2013.2266258

Altay, E. E., Fisher, E., Jones, S. E., Hara-Cleaver, C., Lee, J.-C., and Rudick, R. A. (2013). Reliability of classifying multiple sclerosis disease activity using magnetic resonance imaging in a multiple sclerosis clinic. *JAMA Neurol.* 70, 338–344. doi: 10.1001/2013.jamaneurol.211

Battaglini, M., Rossi, F., Grove, R. A., Stromillo, M. L., Whitcher, B., Matthews, P. M., et al. (2014). Automated identification of brain new lesions in multiple sclerosis using subtraction images. *J. Magn. Reson. Imaging* 39, 1543–1549. doi: 10.1002/jmri.24293

Baur, C., Denner, S., Wiestler, B., Navab, N., and Albarqouni, S. (2021). Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study. *Med. Image Anal.* 69:101952. doi: 10.1016/j.media.2020.101952

Bosc, M., Heitz, F., Armspach, J.-P., Namer, I., Gounot, D., and Rumbach, L. (2003). Automatic change detection in multimodal serial MRI: application to multiple sclerosis lesion evolution. *NeuroImage* 20, 643–656. doi: 10.1016/S1053-8119(03)00406-3

Brownlee, W. J., Altmann, D. R., Prados, F., Miszkiel, K. A., Eshaghi, A., Gandini Wheeler-Kingshott, C. A., et al. (2019). Early imaging predictors of long-term outcomes in relapse-onset multiple sclerosis. *Brain* 142, 2276–2287. doi: 10.1093/brain/awz156

Carass, A., Roy, S., Jog, A., Cuzzocreo, J. L., Magrath, E., Gherman, A., et al. (2017). Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage* 148, 77–102. doi: 10.1016/j.neuroimage.2016.12.064

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Athens: Springer), 424–432.

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., et al. (2018). Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Sci. Rep.* 8, 1–17. doi: 10.1038/s41598-018-31911-7

Egger, C., Opfer, R., Wang, C., Kepp, T., Sormani, M. P., Spies, L., et al. (2017). MRI FLAIR lesion segmentation in multiple sclerosis: does automated segmentation hold up with manual annotation? *NeuroImage Clin.* 13, 264–270. doi: 10.1016/j.nicl.2016.11.020

Fartaria, M. J., Kober, T., Granziera, C., and Cuadra, M. B. (2019). Longitudinal analysis of white matter and cortical lesions in multiple sclerosis. *NeuroImage Clin.* 23:101938. doi: 10.1016/j.nicl.2019.101938

Filippi, M., Preziosa, P., Banwell, B. L., Barkhof, F., Ciccarelli, O., De Stefano, N., et al. (2019). Assessment of lesions on magnetic resonance imaging in multiple sclerosis: practical guidelines. *Brain* 142, 1858–1875. doi: 10.1093/brain/awz144

Ganiler, O., Oliver, A., Diez, Y., Freixenet, J., Vilanova, J. C., Beltran, B., et al. (2014). A subtraction pipeline for automatic detection of new appearing multiple sclerosis lesions in longitudinal studies. *Neuroradiology* 56, 363–374. doi: 10.1007/s00234-014-1343-1

García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D. L., and Collins, D. L. (2013). Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Med. Image Anal.* 17, 1–18. doi: 10.1016/j.media.2012.09.004

Gessert, N., Bengs, M., Krüger, J., Opfer, R., Ostwaldt, A.-C., Manogaran, P., et al. (2020a). 4D deep learning for multiple sclerosis lesion activity segmentation. *arXiv preprint arXiv:2004.09216*. doi: 10.48550/arXiv.2004.09216

Gessert, N., Krüger, J., Opfer, R., Ostwaldt, A.-C., Manogaran, P., Kitzler, H. H., et al. (2020b). Multiple sclerosis lesion activity segmentation with attention-guided two-path CNNs. *Comput. Med. Imaging Graph.* 84:101772. doi: 10.1016/j.compmedimag.2020.101772

Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., and Reuter, M. (2020). FastSurfer-a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage* 219:117012. doi: 10.1016/j.neuroimage.2020.117012

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* 18, 203–211. doi: 10.1038/s41592-020-01008-z

Krüger, J., Opfer, R., Gessert, N., Ostwaldt, A.-C., Manogaran, P., Kitzler, H. H., et al. (2020). Fully automated longitudinal segmentation of new or enlarged multiple sclerosis lesions using 3D convolutional neural networks. *NeuroImage Clin.* 28:102445. doi: 10.1016/j.nicl.2020.102445

Kuhlmann, T., Ludwin, S., Prat, A., Antel, J., Brück, W., and Lassmann, H. (2017). An updated histological classification system for multiple sclerosis lesions. *Acta Neuropathol.* 133, 13–24. doi: 10.1007/s00401-016-1653-y

Lao, Z., Shen, D., Liu, D., Jawad, A. F., Melhem, E. R., Launer, L. J., et al. (2008). Computer-assisted segmentation of white matter lesions in 3D MR images using support vector machine. *Acad. Radiol.* 15, 300–313. doi: 10.1016/j.acra.2007.10.012

Ma, J. (2021). Cutting-edge 3D medical image segmentation methods in 2020: are happy families all alike? *arXiv preprint arXiv:2101.00232*. doi: 10.48550/arXiv.2101.00232

Ma, Y., Zhang, C., Cabezas, M., Song, Y., Tang, Z., Liu, D., et al. (2022). Multiple sclerosis lesion analysis in brain magnetic resonance images: techniques and clinical applications. *IEEE J. Biomed. Health Inform.* doi: 10.1109/JBHI.2022.3151741

McKinley, R., Wepfer, R., Grunder, L., Aschwanden, F., Fischer, T., Friedli, C., et al. (2020). Automatic detection of lesion load change in multiple sclerosis using convolutional neural networks with segmentation confidence. *NeuroImage Clin.* 25:102104. doi: 10.1016/j.nicl.2019.102104

McKinley, R., Wepfer, R., Gundersen, T., Wagner, F., Chan, A., Wiest, R., et al. (2016). "Nabla-net: A deep dag-like convolutional architecture for biomedical image segmentation," in *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (Athens: Springer), 119–128.

Mortazavi, D., Kouzani, A. Z., and Soltanian-Zadeh, H. (2012). Segmentation of multiple sclerosis lesions in MR images: a review. *Neuroradiology* 54, 299–320. doi: 10.1007/s00234-011-0886-7

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Munich: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Roy, A. G., Conjeti, S., Navab, N., Wachinger, C., Initiative, A. D. N., et al. (2019). QuickNAT: a fully convolutional network for quick and accurate segmentation of neuroanatomy. *NeuroImage* 186, 713–727. doi: 10.1016/j.neuroimage.2018.11.042

Salem, M., Cabezas, M., Valverde, S., Pareto, D., Oliver, A., Salvi, J., et al. (2018). A supervised framework with intensity subtraction and deformation field features for the detection of new T2-w lesions in multiple sclerosis. *NeuroImage Clin*. 17, 607–615. doi: 10.1016/j.nicl.2017.11.015

Salem, M., Valverde, S., Cabezas, M., Pareto, D., Oliver, A., Salvi, J., et al. (2020). A fully convolutional neural network for new T2-w lesion detection in multiple sclerosis. *NeuroImage Clin*. 25:102149. doi: 10.1016/j.nicl.2019.102149

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., et al. (2012). An automated tool for detection of flair-hyperintense white-matter lesions in multiple sclerosis. *Neuroimage* 59, 3774–3783. doi: 10.1016/j.neuroimage.2011.11.032

Schmidt, P., Pongratz, V., Küster, P., Meier, D., Wuerfel, J., Lukas, C., et al. (2019). Automated segmentation of changes in flair-hyperintense white matter lesions in multiple sclerosis on serial magnetic resonance imaging. *NeuroImage Clin*. 23:101849. doi: 10.1016/j.nicl.2019.101849

Shiee, N., Bazin, P.-L., Ozturk, A., Reich, D. S., Calabresi, P. A., and Pham, D. L. (2010). A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. *NeuroImage* 49, 1524–1535. doi: 10.1016/j.neuroimage.2009.09.005

Sormani, M. P., Gasperini, C., Romeo, M., Rio, J., Calabrese, M., Cocco, E., et al. (2016). Assessing response to interferon-β in a multicenter dataset of patients with MS. *Neurology* 87, 134–140. doi: 10.1212/WNL.0000000000002830

Sundaresan, V., Zamboni, G., Rothwell, P. M., Jenkinson, M., and Griffanti, L. (2021). Triplanar ensemble U-Net model for white matter hyperintensities segmentation on MR images. *Med. Image Anal*. 73:102184. doi: 10.1016/j.media.2021.102184

Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., et al. (2018). Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol*. 17, 162–173. doi: 10.1016/S1474-4422(17)30470-2

Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., and Suetens, P. (2001). Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE Trans. Med. Imaging* 20, 677–688. doi: 10.1109/42.938237

Walker, E., and Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *J. Gen. Intern. Med*. 26, 192–196. doi: 10.1007/s11606-010-1513-8

Wu, Z., Shen, C., and Hengel, A. v. d. (2016). Bridging category-level and instance-level semantic image segmentation. *arXiv preprint arXiv:1605.06885*. doi: 10.48550/arXiv.1605.06885

Zhang, H., and Oguz, I. (2020). "Multiple sclerosis lesion segmentation-a survey of supervised CNN-based methods," in *International MICCAI Brainlesion Workshop* (Lima: Springer), 11–29.

frontiers | Frontiers in Neuroimaging

TYPE Methods
PUBLISHED 25 August 2022
DOI 10.3389/fnimg.2022.948235

OPEN ACCESS

EDITED BY
Benoit Combès,
Inria Rennes-Bretagne Atlantique
Research Centre, France

REVIEWED BY
Kalavathi Palanisamy,
The Gandhigram Rural Institute, India
Niharika S. D'Souza,
IBM Research Almaden, United States

*CORRESPONDENCE
Reda Abdellah Kamraoui
reda-abdellah.kamraoui@u-bordeaux.fr

SPECIALTY SECTION
This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroimaging

RECEIVED 19 May 2022
ACCEPTED 11 July 2022
PUBLISHED 25 August 2022

CITATION
Kamraoui RA, Mansencal B, Manjon JV and Coupé P (2022) Longitudinal detection of new MS lesions using deep learning.
Front. Neuroimaging 1:948235.
doi: 10.3389/fnimg.2022.948235

Longitudinal detection of new MS lesions using deep learning

Reda Abdellah Kamraoui[1]*, Boris Mansencal[1], José V. Manjon[2] and Pierrick Coupé[1]

[1]PICTURA, Univ. Bordeaux, Bordeaux INP, CNRS, LaBRI, UMR5800, Talence, France, [2]ITACA, Universitat Politècnica de València, Valencia, Spain

The detection of new multiple sclerosis (MS) lesions is an important marker of the evolution of the disease. The applicability of learning-based methods could automate this task efficiently. However, the lack of annotated longitudinal data with new-appearing lesions is a limiting factor for the training of robust and generalizing models. In this study, we describe a deep-learning-based pipeline addressing the challenging task of detecting and segmenting new MS lesions. First, we propose to use transfer-learning from a model trained on a segmentation task using single time-points. Therefore, we exploit knowledge from an easier task and for which more annotated datasets are available. Second, we propose a data synthesis strategy to generate realistic longitudinal time-points with new lesions using single time-point scans. In this way, we pretrain our detection model on large synthetic annotated datasets. Finally, we use a data-augmentation technique designed to simulate data diversity in MRI. By doing that, we increase the size of the available small annotated longitudinal datasets. Our ablation study showed that each contribution lead to an enhancement of the segmentation accuracy. Using the proposed pipeline, we obtained the best score for the segmentation and the detection of new MS lesions in the MSSEG2 MICCAI challenge.

KEYWORDS

new lesion detection, new lesions segmentation, data augmentation, transfer learning, data synthesis

1. Introduction

Multiple sclerosis (MS) is a chronic autoimmune disease of the central nervous system. The pathology is characterized by inflammatory demyelination and axonal injury, which can lead to irreversible neurodegeneration. The disease activity, such as MS lesions, can be observed using magnetic resonance imaging (MRI). The detection of new MS lesions is one of the important biomarkers that allow clinicians to adapt the patient's treatment and assess the evolution of this disease.

Recently, the automation of single time-point MS lesion segmentation has shown encouraging results. Many techniques showed performance comparable to clinicians in controlled evaluation conditions (refer to Commowick et al., 2016; Carass et al., 2017). These methods use a single time-point scan to segment all appearing lesions at the time of the image acquisition. However, these cross-sectional techniques are not adapted to the longitudinal detection of new lesions. Indeed, using these methods requires repeatedly running the segmentation process for each time-point independently to segment MS lesions before detecting new ones. Unlike the human reader, these methods are not designed to jointly exploit the information contained at each time point. Consequently, single-time MS lesion segmentation methods performance is not optimal for the detection of new lesions between two time-points. Moreover, inconsistencies may appear between segmentations of both time-points since they are processed independently.

To specifically address this detection task using both time-points at the same time, some detection methods have been proposed. In one of the earliest studies, Bosc et al. (2003) used a nonlinear intensity normalization method and statistical hypothesis test methods for change detection. Elliott et al. (2013) used a Bayesian tissue classifier on the time-points to estimate lesion candidates followed by a random-forest-based classification to refine the identification of new lesions. Ganiler et al. (2014) used image subtraction and automated thresholding. Cheng et al. (2018) integrated neighborhood texture in a machine learning framework. Salem et al. (2018) trained a logistic regression model with features from the image intensities, the image subtraction values, and the deformation field operators. Schmidt et al. (2019) used lesion maps of different time-points and FLAIR intensities distribution within normal-appearing white matter to estimate lesion changes. Krüger et al. (2020) used a 3D convolutional neural network (CNN) where each time-point is passed through the same encoder. Then, the produced feature maps are concatenated and fed into the decoder.

Training learning-based methods for the task of new lesions detection require a dataset specifically designed for the task. The most obvious form of the training data would be a longitudinal dataset of MS patients (with two or more successive time-points) with new appearing lesions carefully delineated by experts in the field. However, the construction of such a dataset is very difficult. To begin, new lesions may take several months or even years to appear and be visible in a patient's MR image. Moreover, a time-consuming and costly process is necessary for several experts to annotate new lesions from the two time-points and to obtain an accurate consensus segmentation. Although the organizers of the MICCAI Longitudinal Multiple Sclerosis Lesion Segmentation Challenge (MSSEG2-challenge MICCAI, 2021) provided such a dataset, the training set is severely impacted by class imbalance (refer to Section 2.5.3 for more details) due to the difficulty of finding new lesions in the follow-up scan. This under-representation of new lesions in longitudinal datasets is limiting the training of state-of-the-art deep learning algorithms from scratch on this complex task. Besides, achieving generalizing results on unseen domains (refer to Mårtensson et al., 2020; Bron et al., 2021; Omoumi et al., 2021) may require more data diversity.

Several studies tackled the problem of training data scarcity. First, transfer learning is a strategy used to create high-performance learners trained with more widely available data from different domains when the target domain/task data are expensive or difficult to collect (refer to Torrey and Shavlik, 2009; Weiss et al., 2016). Second, synthetic data generation is performed by using a model able to simulate realistic artificial data that can be used during training (refer to Tremblay et al., 2018; Tripathi et al., 2019; Khan et al., 2021). Third, data-augmentation is a set of techniques used to handle the variability in real-world data by enhancing the size and quality of the training dataset (refer to Shorten and Khoshgoftaar, 2019). Recently, Zhang et al. (2020) showed that applying extensive data augmentation during training also enhances the generalization capability of the methods.

In this article, we propose an innovative strategy integrating these three strategies into a single pipeline for new MS lesion segmentation to tackle data rarity for our task. First, we use transfer-learning to exploit the larger and more diverse datasets available for the task of single-point MS lesion segmentation which does not require longitudinal data. Second, we propose a novel data synthesis technique able to generate two realistic time-points with new MS lesions from a single FLAIR scan. Third, we use a data-augmentation technique to simulate a large variety of artifacts that may occur during the MRI acquisitions. This technique aims to enhance both the variability and size of the training data and to improve the generalization of our model.

2. Methods and materials

2.1. Method overview

To deal with data rarity for new MS lesion segmentation, we proposed a three stage pipeline as shown in Figure 1. In Stage One, an encoder-decoder network is trained on the task of single time-point MS lesions segmentation. This step aims to train the encoder part of the network to extract relevant features related to MS lesions that can be used in the next steps. Stage One enables to indirect use of large datasets dedicated to single time-point MS lesion segmentation for the task of new lesions segmentation. This stage is detailed in Section 2.2. In Stage Two, the new lesions segmentation model composed of the previous task encoder is pretrained with synthetic data. To this end, we trained external models able to generate two realistic time-points from a single image also taken from single time-point MS datasets. It combines the effects of lesion inpainting and lesion generating models to simulate the appearance of new lesions. This strategy is detailed in Section 2.3. In Stage Three, the decoder is fine-tuned with real longitudinal data from the new MS lesion training-set of the MSSEG2 MICCAI challenge.

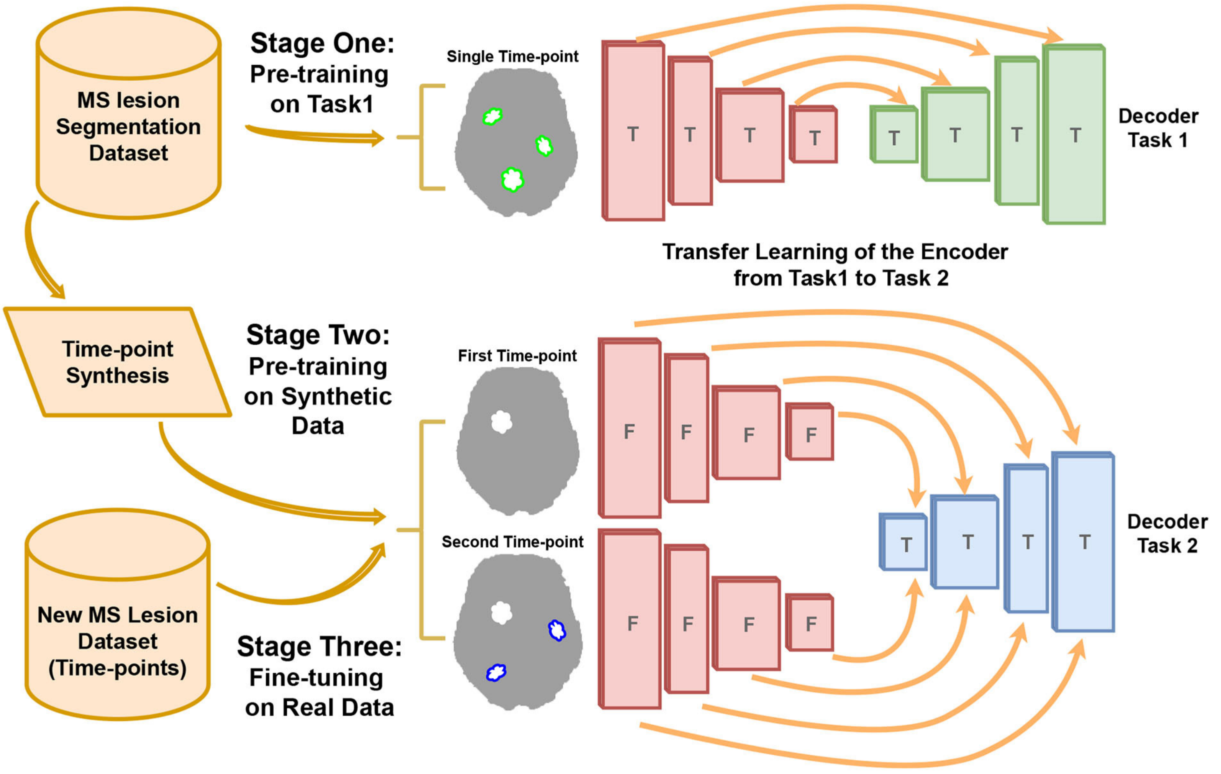

MS lesion Segmentation Dataset
Stage One: Pre-training on Task1
Single Time-point
T
Decoder Task 1
Transfer Learning of the Encoder from Task1 to Task 2
Time-point Synthesis
Stage Two: Pre-training on Synthetic Data
First Time-point
F
Second Time-point
Decoder Task 2
New MS Lesion Dataset (Time-points)
Stage Three: Fine-tuning on Real Data

FIGURE 1
The pipeline of our new MS lesion segmentation method. The three stages include: First, the pre-training on the task of single-time-point MS lesion segmentation (Task 1). Second, pre-training on the task of new MS lesions segmentation (Task 2) with synthetic data. Third, fine-tuning the model with real data. The encoder weights are trained (T) in Stage One and freezed (F) in Stage Two and Stage Three.

2.2. Transfer-learning from single time-point MS lesion segmentation task

The encoder used for new MS lesion segmentation is first trained on single time-point lesion segmentation (refer to Figure 2, from Stage One to Stage Two). This choice is motivated by two reasons. First, we consider that datasets for MS lesion segmentation with lesion mask segmentation by experts are more diverse and larger than available datasets for new lesion segmentation (which requires a longitudinal study). Second, the task of MS lesion segmentation is tightly close to the one of new MS lesion segmentation. By learning to segment lesions, the model implicitly learns the concept of a lesion, either the lesion is considered new or was already existing in the first time-point. To conclude, since there is a proximity between the two tasks, there is likely a gain from exploiting a large amount of training data for the first task to improve the second task's performance.

2.2.1. Model architecture design

Our method is based on the transfer learning from the task of "Single time-point MS lesion segmentation" to the task of "new lesions segmentation from two time-points." Thus, two different architectures are used but with the same building blocks for each task. For the first task, a 3D U-Net shape architecture is used, as shown in Figure 3A. This kind of architecture has been very effective and robust for MS lesion segmentation (Isensee et al., 2021; Kamraoui et al., 2022). It is composed of an encoder and a decoder linked with one another by skip connections.

For the second task, a siamese-encoder followed by a single decoder is used, as shown in Figure 3B. The shared-weights encoders are chosen to extract the same set of features from both time points. Then, these features resulting from the different levels of both encoder paths are aggregated (refer to Figure 3B). The aggregation module is composed of concatenation and a convolution operation. Feature maps are first concatenated by channels (i.e., the result channel size is two times the original size), then the convolution operation aggregates the information back to the original channel size. Finally, the aggregated features are passed through the decoder.

2.3. Time-points synthesis

The data synthesis method is based on the simulation of new MS lesions between two time-points using single time-point

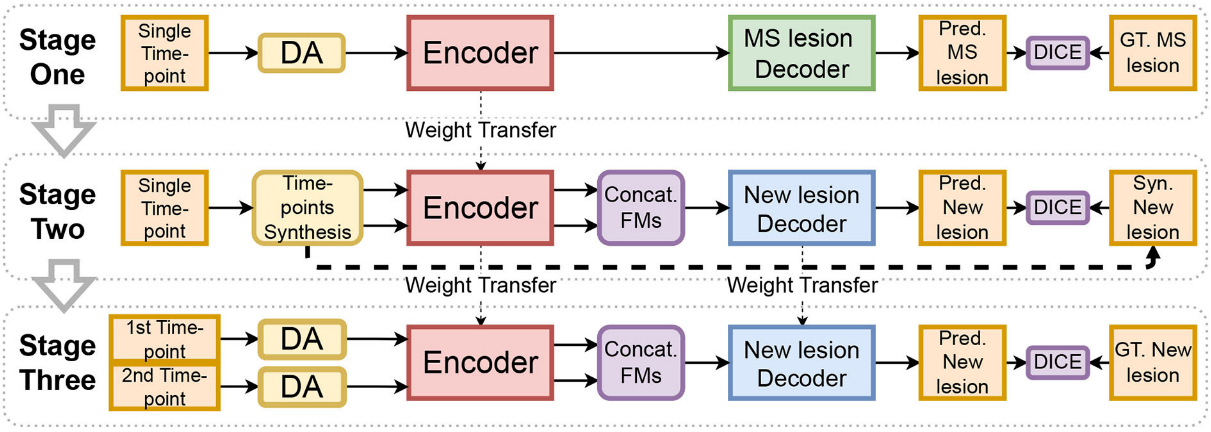


FIGURE 2
The diagram represents our training method. Input images are augmented with the proposed method (DA). The encoder trained in Stage One is used in Stage Two and Stage Three to extract feature maps (FMs) of the two-time points. The aggregation block (Concat. FMs) is used to combine features.

FLAIR images. As shown in Figure 4, our pipeline generates "on the fly" synthetic 3D patches that represent longitudinal scans of the same patient with evolution in their lesion mask. The synthetic data is generated in three steps. In the first step, a 3D FLAIR patch and its MS lesion segmentation mask are randomly sampled from different MS lesion segmentation datasets (refer to Section 2.5.1). Then, the patch and lesion mask are randomly augmented with flipping and rotations. A copy of the FLAIR patch is performed to represent the two time-points. Then, both identical patches are altered with the described data augmentation (refer to Section 2.4) to differentiate the two patches. At this point, the lesion masks of the two synthetic time-points are still identical. Thus, there are no new lesions. In the second step, a connected component operation is used to separate each independent lesion from the lesion mask. Each lesion is either inpainted (i.e., removed) from one of the two time-points or both of them, or it can be kept in both of the time-points. The lesion inpainting model is used to inpaint the lesion region with hallucinated healthy tissue (refer to Section 2.3.1). Next, the new lesion mask is constructed from lesion regions that have been kept in the second time-point but not the first one. In the third step, the lesion generator model is used to simulate new synthetic lesions at realistic locations (using white/gray matter segmentation and a probabilistic distribution of MS lesions on the brain in the MNI space). Synthetic lesions are generated for one of the time-points or both of them (refer to Section 2.3.2). Similar to the previous step, the new lesion mask is updated to include only the generated lesions on the second time-point.

2.3.1. Lesion inpainting model

The lesion inpainting model is trained, independently and priorly to our proposed pipeline, with randomly selected 3D FLAIR patches which do not contain MS lesions or white matter hyperintensities. Similar to Manjón et al. (2020), A 3D U-Net network is optimized to reconstruct altered input images. Specifically, the input patch is corrupted with Gaussian noise (i.e., with a mean and a standard deviation of the image intensities) in lesion-like areas at random locations. When the model is trained, it can be used to synthesize healthy regions in lesion locations that are replaced with random gaussian (refer to Manjón et al., 2020 for details).

2.3.2. Lesion generator model

The lesion generator is trained before our proposed pipeline to simulate realistic lesions. The generator is a 3D U-Net network with two input channels and one output channel. The first input channel receives an augmented version of 3D FLAIR patches containing MS lesions where lesions are replaced with random noise. The second input channel receives the MS lesion mask of the original 3D FLAIR patch. The output channels predict the original 3D FLAIR patch with lesions. Thus, the trained model can simulate synthetic MS lesions from a 3D patch of FLAIR and its corresponding lesion mask.

2.4. Data augmentation

The quality of the MRI greatly varies between datasets. The quality of the images depends on several factors such as signal-to-noise ratio, contrast-to-noise ratio, resolution, or slice thickness. Since our training set is limited, it does not reflect the diversity of real-world images. To make our training stages robust to the large variety of artifacts that may occur during the MRI acquisitions, an extensive Data

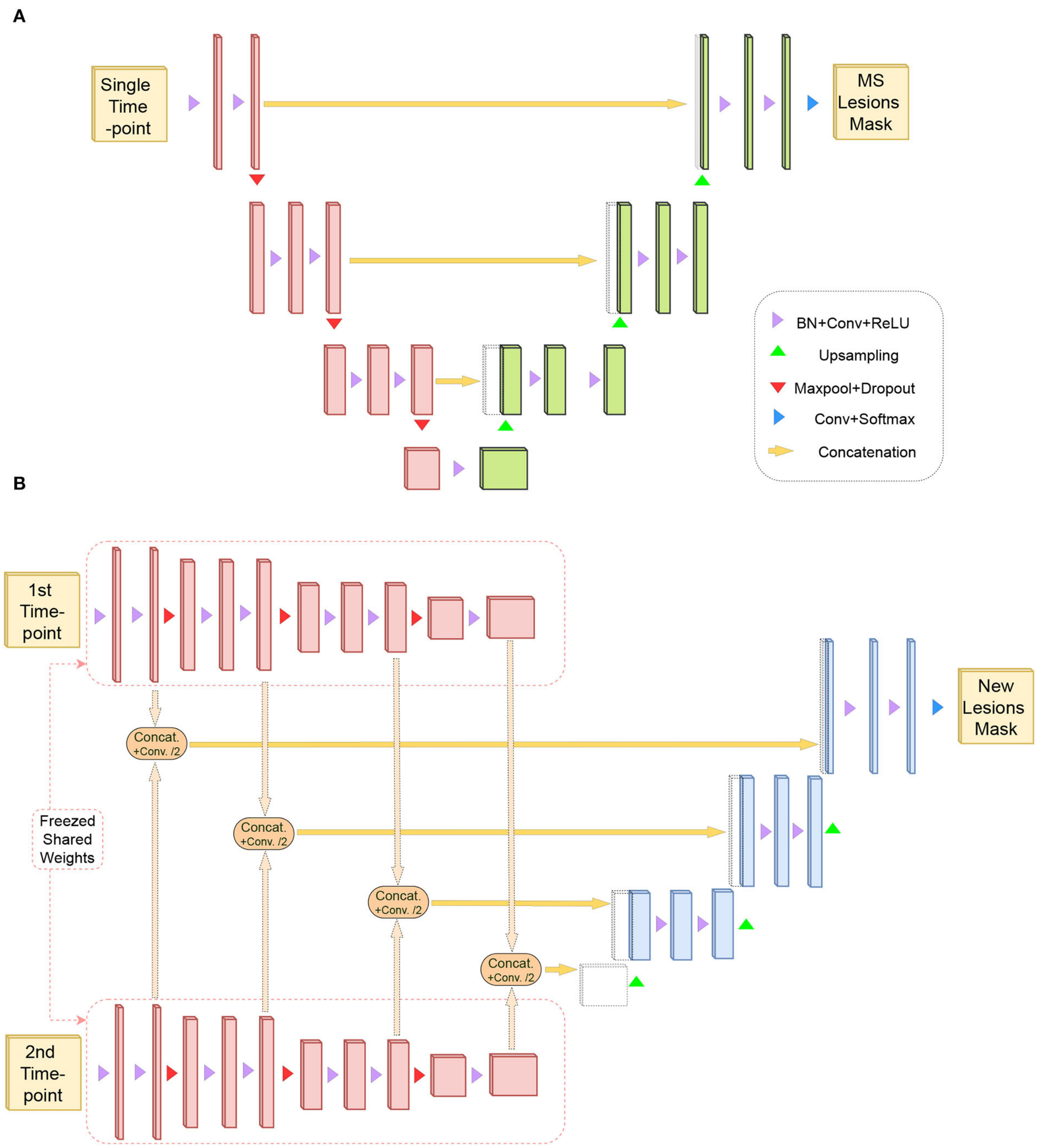


FIGURE 3
(A) Represents U-Net like architecture composed of an encoder (in red) and a decoder for the task of MS lesion segmentation (in green). This task requires a single time-point as input and produces the MS lesion mask. **(B)** Shows a siamese-encoder (in red) to extract the same sets of features from the two time-points. Same-level features are aggregated with a combination module and are forwarded to a decoder for the task of new lesions segmentation (in blue).

Augmentation (DA) is used (refer to "DA" in Figure 2 and "Data Augmentation" in Figure 4). Such DA technique also helps to better oversample the scarce samples with new lesions (refer to Section 2.5.3).

We use an improved version of the data augmentation strategy proposed in Kamraoui et al. (2022), which simulates MRI quality disparity. During training, we simulate "on the fly" altered versions of 3D patches. We randomly introduce a set of

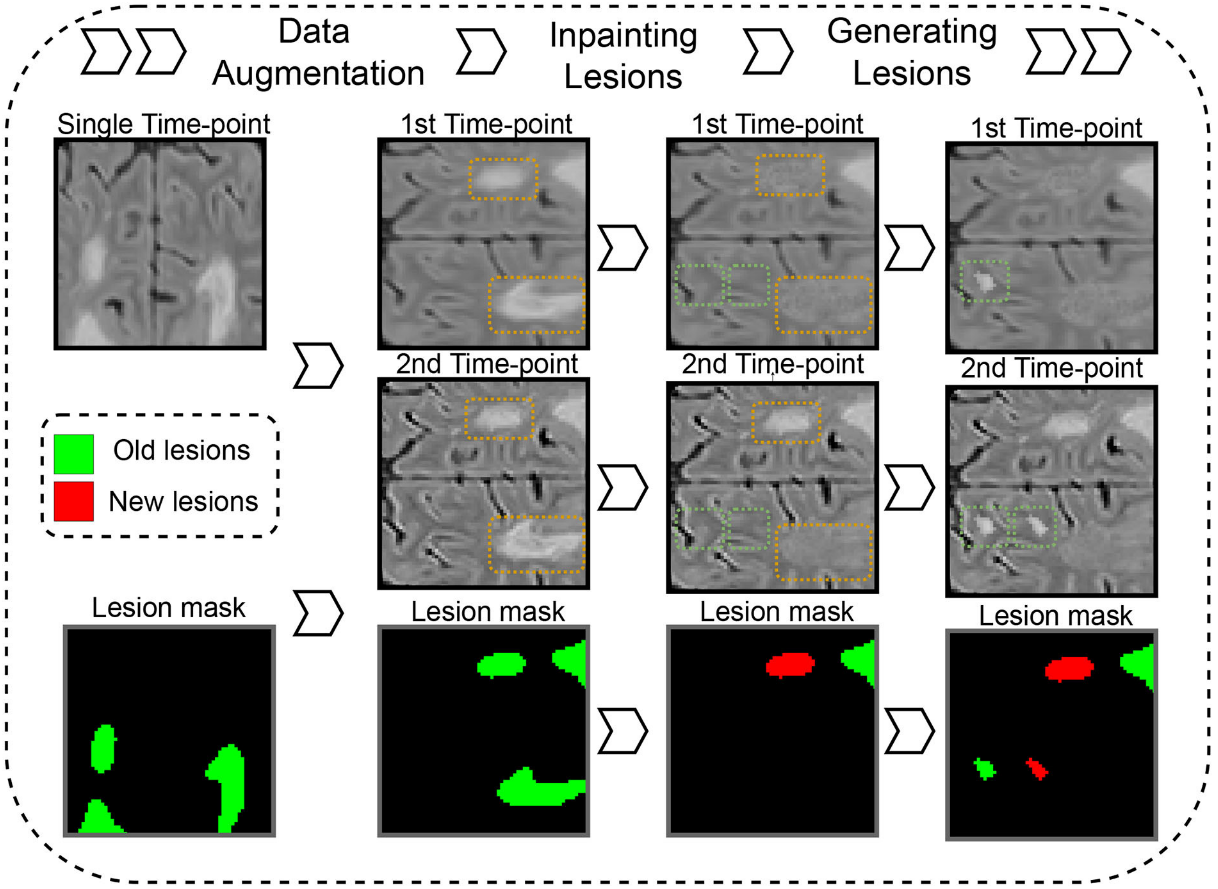


FIGURE 4
Synthetic time points with new MS lesion generation pipeline. Dashed orange and green rectangles on images represent areas where lesions are inpainted or generated.

alterations in the spatial and frequency space (k-space): Blur, edge enhancement, axial subsampling distortion, anisotropic downsampling, noise, bias-field variation, motion effect, MRI spike artifacts, and ghosting effect. Figure 5 shows augmentation samples.

For the blur, a gaussian kernel is used with a randomly selected standard deviation (SD) ranging between $[0.5, 1.75]$. For edge enhancement, we use unsharp masking with the inverse of the blur filter. For axial subsampling distortion, we simulate acquisition artifacts that can result from the varying slice thickness. We use a uniform filter (a.k.a mean filter) along the axial direction with a size of $[1 \times 1 \times sz]$ where $sz \in 2, 3, 4$. For anisotropic downsampling, the image is downsampled through an axis with a random factor ranging between $[1.5, 4]$ and upsampled back again with a B-spline interpolation. For noise, we add to the image patch a Gaussian noise with 0 mean and an SD ranging between $[0.02, 0.1]$. Bias-field variation is generated using the study of Sudre et al. (2017) which considers the bias field as a linear combination of polynomial basis functions. Motion effect has been generated based on the study of Shaw et al. (2018). The movements are simulated by combining in the k-space a sequence of affine transforms with random rotation and translation in the ranges $[-5, 5]$ degrees and $[-4, 4]$ mm, respectively. Both MRI spike artifacts and the ghosting effect have been generated with the implementation of Pérez-García et al. (2021).

2.5. Data

Different datasets are used for the training and validation of the two tasks (refer to Table 1).

2.5.1. Single time-point datasets

For time-points synthesis (refer to 2.3) and encoder pretraining (refer to 2.2), we jointly used three datasets containing single time-points FLAIR and lesion masks. First, the ISBI (Carass et al., 2017) training-set contains 21 FLAIR images with expert annotation done by two raters. Although the dataset

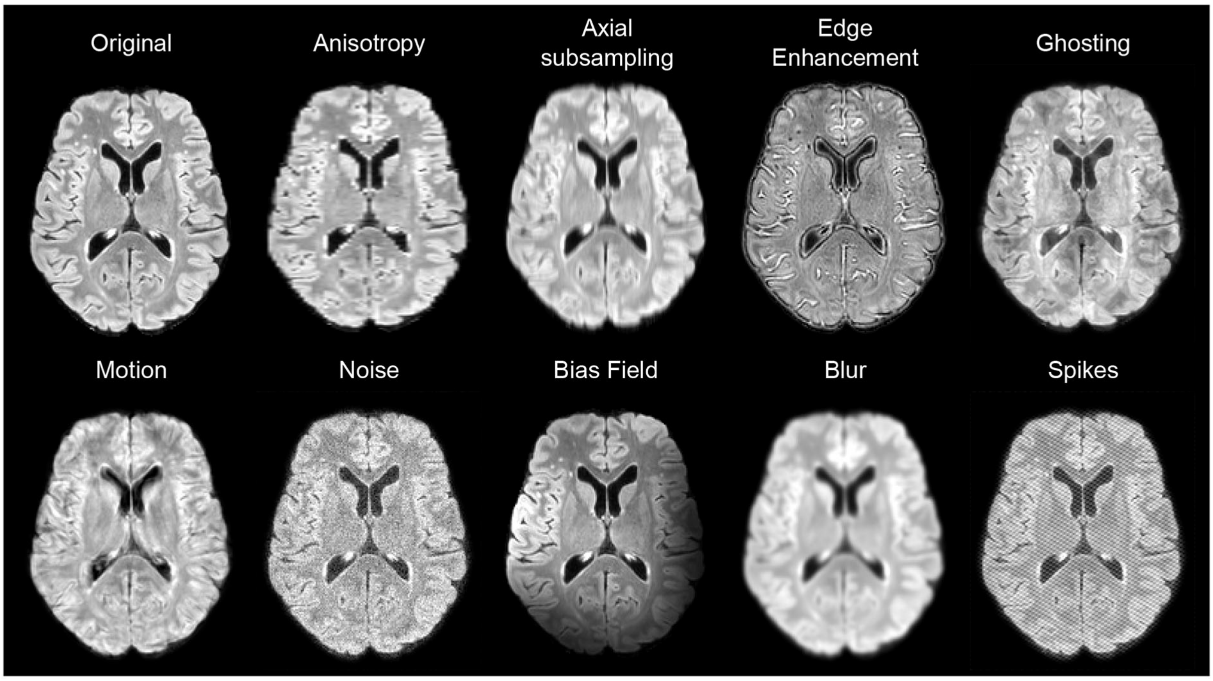


FIGURE 5
Examples of data augmentation applied on FLAIR images.

TABLE 1 Summary of the used datasets. For each dataset, the object count (Obj. Count) and the total volume (Tot. Vol. cm^3) represent, respectively, the total number and the total volume in cm^3 of lesions or new lesions (depending on the task).

Task	Dataset	Patients	Time-point	Raters	Obj. count	Tot. vol. (cm^3)	Clinical site/Scanners
MS lesion segmentation	ISBI	5	4-5	2	514	243	Single-site
	MSSEG' 16	15	1	7	512	367	Multi-site: three sites
	In-house	43	1	2	2,391	1,313	Multi-site
New MS lesion segmentation	MSSEG2 Training-set	40	2	4	123	23	Multi-site: 15 MRI scanners
	MSSEG2 Test-set	60	2	4	174	60	(GE scanners only in Test-set)

is composed of longitudinal time-points from 5 patients, the provided expert annotations focus on the lesion mask of each time-point independently from the others and do not provide new lesion masks. Thus, we use the 21 images independently. Second, the MSSEG'16 training-set (Commowick et al., 2016) contains 15 patients from three different clinical sites. Each FLAIR image is along with a consensus segmentation for MS lesions from seven human experts. Third, our in-house (Coupé et al., 2018) dataset is composed of 43 subjects diagnosed with MS. The images were acquired with different scanners and multiple resolutions and their lesion masks have been obtained by two human experts.

All images were pre-processed using the lesionBrain pipeline from the volBrain platform (Manjón and Coupé, 2016). First, it includes image denoising (Manjón et al., 2010). Second, an affine registration to MNI space is performed using the T1w modality, then the FLAIR is registered to the transformed T1w. Skull stripping and bias correction have been performed on the modalities, followed by the second denoising. Finally, the intensities have been normalized with kernel density estimation.

2.5.2. Two time-points datasets

The dataset provided by the MSSEG2-challenge (MICCAI, 2021) is used to train our method. The challenge dataset features a total of 100 patients with MS. For each patient, two 3D FLAIR sequence time-points have been acquired spaced apart by a 1–3 years period. The dataset has been split into 40 patients for training and 60 patients for testing. A total of 15 different MRI scanners were used for the acquisition of the entire dataset. However, all images from GE scanners have been reserved only for the testing set to see the generalization capability of the algorithms. Reference segmentation on these data was defined by a consensus of four expert neuroradiologists.

For preprocessing, the challenge organizers proposed a docker[1] built with the Anima scripts. It includes bias correction, denoising, and skull stripping. In addition, we added a registration step to the MNI space using a FLAIR template, (i.e., the training and inference are performed in the MNI space, then the segmentation masks are transformed-back to the native space for evaluation).

Before challenge day, the testing set (the 60 patients) was not publicly available. Thus, to test our methods (refer to Section 3.1.1), we defined an internal validation subset from the 40 challenge training data. Of the 40 patients, six cases containing confirmed new lesions were kept out from the training-set and were used as an internal test-set. For the challenge evaluation (refer to Section 3.2), the model submitted to the challenge organizers was trained on the entire MSSEG2 training-set.

2.5.3. Dataset class imbalance

Anomaly detection/segmentation tasks, such as MS lesion segmentation, suffer from class imbalance where the positive class is scarce (refer to Johnson and Khoshgoftaar, 2019). Herein, the MSSEG2-challenge (MICCAI, 2021) dataset is composed of 100 patients (40 for training and 60 for test) and all the MS Lesions Segmentation datasets combined account for 64 patients and 79 images. Therefore, the number of image is similar. However, the class imbalance is highly different when evaluating the class imbalance using the number of objects to detect/segment (which represent MS lesions for the first task and new lesions for the second one) and their total volume for each dataset (refer to Table 1). Indeed, we see that the MSSEG2-challenge datasets (especially training-set) suffer from more severe under-representation of the positive class. Consequently, it will be more difficult to train a model for New MS lesion segmentation than for the task of single time-point MS lesion segmentation. Furthermore, it shows that MS lesion segmentation datasets could significantly enrich the training of New MS lesion segmentation models.

2.6. Implementation details

First, all models are trained on 3D image patches of size [64 × 64 × 64]. For the two time-points new lesion model, an ensemble of five networks (different training/validation data-split) is used. During inference, the consensus (prediction average) of the ensemble segmentation is taken. For each voxel, the two classes, output probabilities of the five networks are averaged, and the class with the highest probability is picked (new lesion voxel or not).

Second, the Dice-loss (soft DICE with probabilities as continuous values) is used as a loss function for the training of the single time-point MS lesion segmentation and the two time-points new lesion models. The mean-squared error is used as a loss function to train time-point synthesis models (inpainting and lesion generator models).

Finally, the experiments have been performed using PyTorch framework version 1.10.0 on Python version 3.7 of Linux environment with NVIDIA Titan Xp GPU 12 GB RAM. All models were optimized with Adam (Kingma and Ba, 2014) using a learning rate of 0.0001 and a momentum of 0.9.

2.7. Validation framework

2.7.1. Evaluation metrics

The assessment of a segmentation method is usually measured by a similarity metric between the predicted segmentation and the human expert ground truth.

First, we use several complementary metrics to assess segmentation performance. Namely, we use the Dice similarity coefficient, the Positive Predictive Value (PPV or the precision), and the true positive rate (TPR, known as recall or Sensitivity).

$$Dice = \frac{2 \times TP}{(TP + FN) + (TP + FP)}, \tag{1}$$

$$PPV = \frac{TP}{TP + FP}, \qquad TPR = \frac{TP}{TP + FN}, \tag{2}$$

where TP, FN, and FP represent, respectively, true positives, false negatives, and false positives.

Second, recent studies (i.e., Commowick et al., 2018) question the relevance of classic metrics (Dice) compared to detection metrics, which are used for MS diagnostic and clinical evaluation of the patient evolution. Thus, in addition to the voxel-wise metrics, we also use lesion-wise metrics that focus on the lesion count. We use the lesion detection F1 ($LesF_1$) score defined as

$$LesF_1 = \frac{2 \times S_L \times P_L}{(S_L + P_L)}, \tag{3}$$

where S_L is lesion sensitivity, i.e., the proportion of detected lesions and P_L is lesion positive predictive value, i.e., the proportion of true positive lesions. For result harmonization with challenge organizers and participants, the same evaluation tool is used, i.e., animaSegPerfAnalyzer (Commowick et al., 2018). All lesions that are smaller in size than $3mm^3$ are removed. For S_L, only ground-truth lesions that overlap at least

[1] https://github.com/Inria-Empenn/lesion-segmentation-challenge-miccai21/

10% with segmented volume are considered positive. For a predicted lesion to be considered positive for P_L, it has to be overlapped by at least 65% and do not go outside by more than 70% of the volume.

Finally, to jointly consider the different metrics (i.e., segmentation and detection performance), it would be convenient to aggregate them into a single score. Thus, we propose the average of DICE and $LesF_1$ (Avg. Score) as an aggregation score for comparing different methods.

2.7.2. Statistical test

To assert the advantage of a technique obtaining the highest average score, we conducted a Wilcoxon test (i.e., paired statistical test) over the lists of metric scores. The significance of the test is established for a p-value below 0.05. In the following tables, * indicates a significantly better average score when compared with the rest of the other approaches.

3. Results

Several experiments were conducted on our methods, including an ablation study and the comparison with state-of-the-art methods in competition during the challenge evaluation.

3.1. Internal validation

3.1.1. Ablation study

To evaluate each contribution of our training pipeline, Table 2 compares our full method with a baseline and other variations of our method on the internal validation dataset. The baseline in this experiment was trained with real time-points only and by using a classic data augmentation composed of orthogonal rotations and mirroring.

First, when using only transfer learning on top of the baseline, we measured an increase in DICE and TPR compared to the baseline but approximately the same $LesF_1$ and PPV. Second, when using only time-point synthesis pretraining on the top of the baseline, we obtained a significantly higher $LesF_1$ compared to the baseline and an increase in DICE. This variation also obtained the highest PPV at the expense of the lowest TPR. Third, when comparing the use of the proposed data augmentation, we see an increase in DICE and PPV but approximately the same $LesF_1$. Finally, when combining the transfer learning, time-point synthesis pre-training, and the proposed data-augmentation, we obtained the highest Avg. Score, DICE, $LesF_1$, and TPR.

3.1.2. The impact of longitudinal dataset size

Figure 6 shows the performance of our method when trained with different longitudinal dataset sizes. From the 34 patients available for the training with two time-points in Internal Validation settings (refer to Section 2.5.2), we tested the performance of our model when training on 34, 36, 17, 8, and 0 patients. In the case of 0 patients, our method performance was obtained using synthetic data only (i.e., Stage Two where only cross-sectional MS segmentation databases were used as described in Table 1). For the rest of the experiments, the reported number of patients with two time-points was used for the fine-tuning step (i.e., Stage Three).

First, for the baseline version (i.e., with neither pre-training nor data augmentation), the graph can be separated into two phases. From 0 to 17 patients, the graph shows an increase in both metrics. From 17 to 34 patients, metrics of baseline versions reach a plateau. Since the baseline is trained from scratch, its performance improves with the increase in dataset size. However, the performance increase is less significant for the second phase since it is more difficult to improve metrics when approaching their optimal value.

Second, for our method, the graph shows two phases. From 0 to 8 patients, the performance decreases slightly. From 8 to 34 patients, the graph shows a slow increase in metrics until plateauing. Since we use transfer learning and pretraining on synthetic data for our method, its performance does not depend only on the number of patients from MSSEG2 Training-set. The drop in performance in the first phase can be explained by the fact that using eight patients for fine-tuning is less effective than using the model trained on synthetic data only.

3.2. Challenge evaluation

To evaluate our method on the challenge dataset, Table 3 compares it to the leader-board state-of-the-art methods. Results of the top performing methods were reported from challenge-day results.

Besides the top-performing methods, Table 3 also includes the expert raters' performance to give an insight into human performance. Their performance is measured compared to each other, contrary to the top methods that are evaluated using consensus segmentation. Raters x vs. y means that we evaluate the performance of rater x when considering rater y segmentations as ground truth. Indeed, we consider that such a strategy can be more meaningful than the consensus segmentation in our case since the expert consensus already encodes the raters' segmentation and, thus, is unfair when compared to other strategies that did not participate in the consensus.

First, from the top five best-performing methods, LaBRI-IQDA (Kamraoui et al., 2021; our team's submission during the challenge-day) obtained the best score for the challenge. This method was similar to the proposed baseline with data augmentation. Second, the proposed method (results obtained after challenge-day) obtained the highest $LesF_1$ and Average

TABLE 2 The internal validation results for the ablation study.

Transfer learning	Time-point synthesis	Data augm.	Avg. Score	DICE	*LesF*$_1$	TPR	PPV
✓	✓	✓	**0.543***	**0.514***	**0.573***	**0.500***	0.546
✓	✗	✗	0.483	0.480	0.486	0.461	0.532
✗	✓	✗	0.501	0.461	0.541	0.384	**0.602***
✗	✗	✓	0.477	0.464	0.488	0.406	0.565
✗	✗	✗	0.469	0.449	0.489	0.413	0.534

✓ and ✗ symbolize using or not each contribution. Bold values indicate the best result for a metric and * indicates that the advantage is statistically significant (Wilcoxon test).

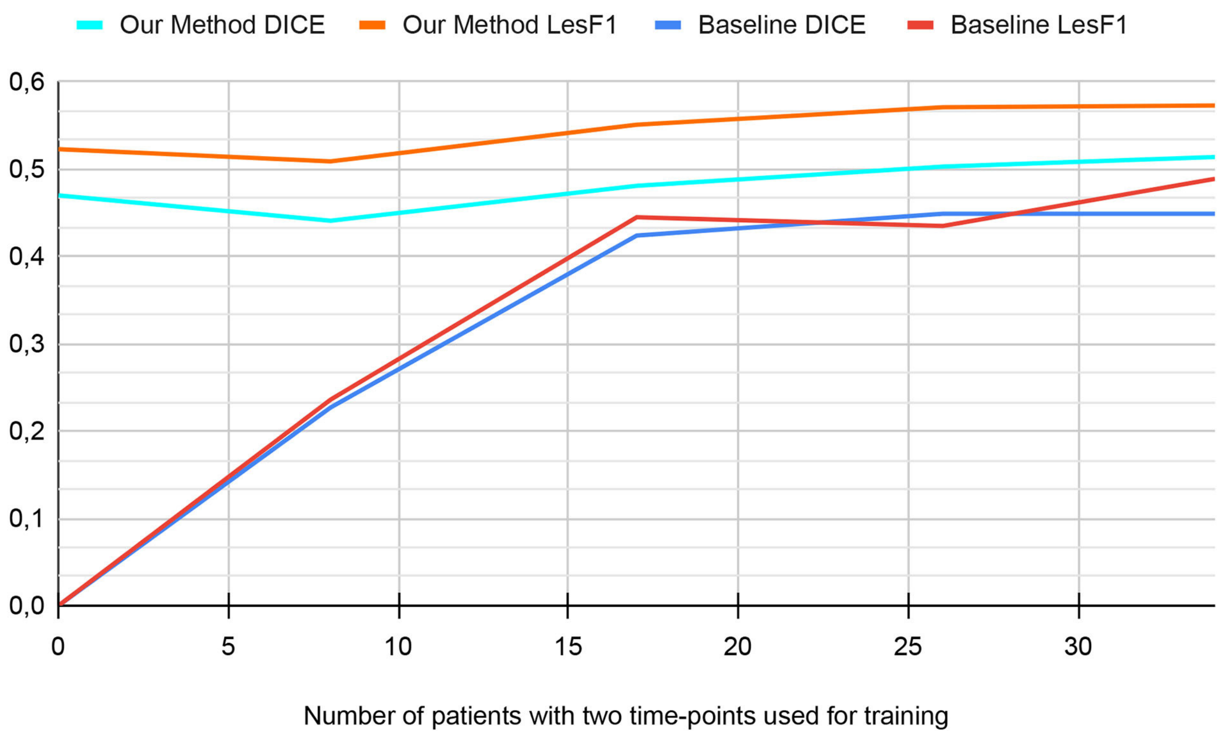


FIGURE 6
The performance in the internal validation of our method and the baseline based on the number of patients used for training (from MSSEG2 Training-set).

scores. Moreover, these both scores are significantly better than all the listed state-of-the-art methods. The DICE score obtained by MedICL was not significantly better than the one obtained by our method. Third, all but one (Empenn) leader-board automatic method obtained better DICE than raters segmentation. Our proposed method, LaBRI-IQDA, and MedICL even surpassed all raters in Average Scores.

Figure 7 shows the segmentation of new lesions by our proposed method. As a ground-truth reference, we compare the segmentation with the consensus segmentation of raters. We also compare each rater segmentation against their consensus. From the five segmentation, we see that our segmentation is the most accurate with the consensus. Each of the human experts Rater 2, Rater 3, and Rater 4 missed one or multiple lesions when segmenting this sample. Although Rater 1 did not miss any lesions, we see that our segmentation is the closest to the consensus.

Overall, our method obtained the best result in the MSSEG2 challenge evaluation (during the challenge and after). Moreover, the result of the experiments showed that our segmentation is objective and can produce more accurate segmentations than human raters.

4. Discussion

The transfer-learning from a single time-point MS lesion segmentation task is an effective method to train the model for the task of two time-points new MS lesion segmentation even with a small dataset. Indeed, it enables us to exploit the large

TABLE 3 Results of MSSEG2-challenge (MICCAI, 2021) evaluation.

	Experiment	Avg. Score	DICE	$LesF_1$
	Raters 1 vs. 2	0.466	0.426	0.507
	Raters 1 vs. 3	0.499	0.434	0.564
	Raters 1 vs. 4	0.434	0.382	0.486
Challenge-day	LaBRI-IQDA (Kamraoui et al., 2021)	0.507	0.498	0.515
	MedICL (Zhang et al., 2021)	0.503	**0.506**	0.5
	SNAC (Cabezas et al., 2021)	0.496	0.484	0.513
	Mediaire-B (Dalbis et al., 2021)	0.489	0.436	0.541
	Empenn (Masson et al., 2021)	0.478	0.423	0.532
	The Proposed Method	**0.523***	0.495	**0.550***

From top to bottom, the table shows the challenge raters' agreement on the segmentation compared to each other, the leader-board results of the challenge-day top methods, and the result of the method described in this article (obtained after challenge-day). For automatic methods, bold values indicate the best result for a metric, and * indicates that the advantage is statistically significant (Wilcoxon test).

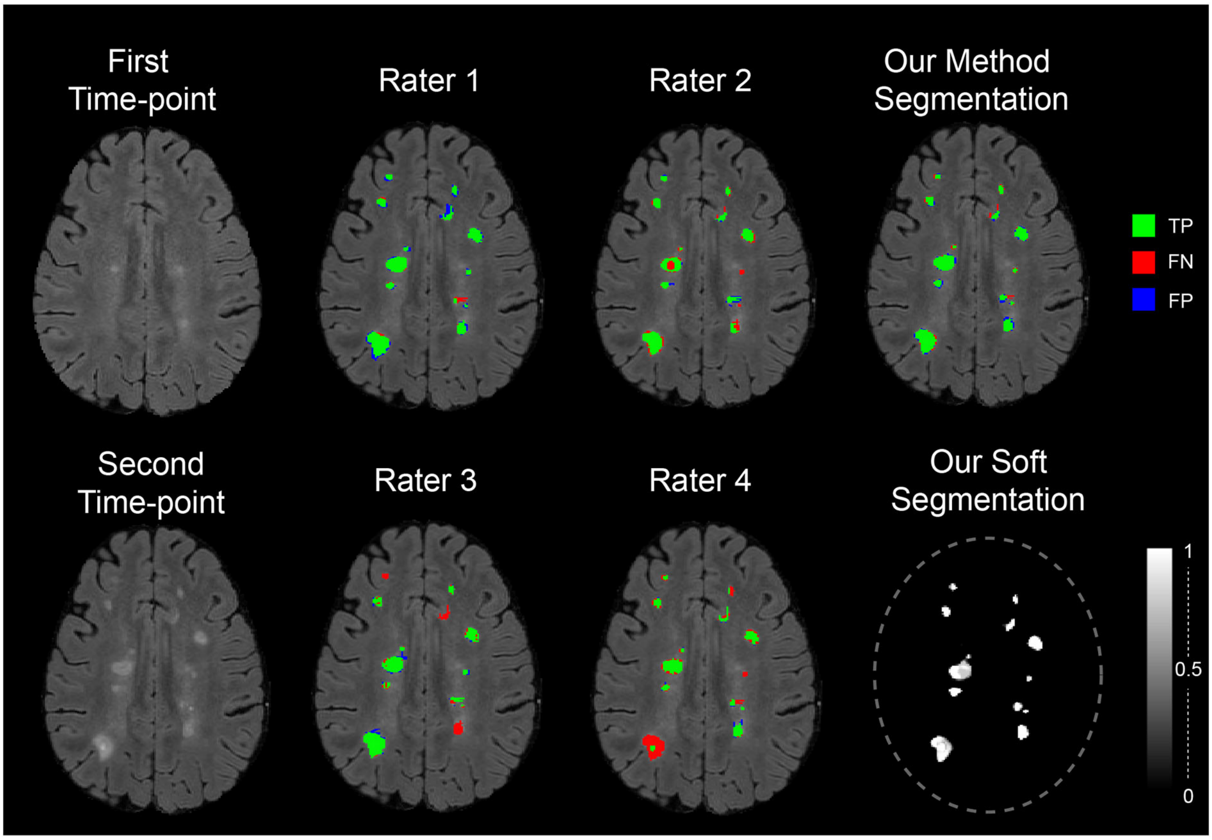


FIGURE 7
The segmentation of the proposed method and the expert rater on a sample image from MICCAI 2021—longitudinal multiple sclerosis lesion segmentation testing dataset. The segmentations are compared against the consensus of the four raters using the colors: green, red, and blue to symbolize TP, FN, and FP regions of new lesions.

available MS cross-sectional datasets compared to longitudinal datasets. In our case, the encoder for the first task was compatible with the siamese-encoder of the second task and thus was used to extract MS-relevant features from the two time-points. Additionally, we used a learnable aggregation module for time-points feature combination. Besides, by freezing the encoder weights after the transfer-learning from the first to the second task, we ensure that the extracted features in the second task are dataset-independent from the second task dataset (smaller dataset). This independence ensures that the high performance of the proposed method is stable and generalizing.

Longitudinal time-points synthesis is an original approach on how to augment data diversity. It can be extended to other change detection tasks where longitudinal data are hard to acquire. According to the results of our experiments, this strategy turns out to be very effective when used as pretraining. Indeed, when the model is first pretrained with time-point synthesis, it is subject to a wider range of diversity, which aims to constrain the model to extract more generalizing features.

The proposed data augmentation method is an effective technique to make our learning process less dependent on MRI quality and acquisition artifacts. It simulates different acquisition conditions to enhance generalization and helps to better over-sample the available new lesions examples. Our data-augmentation comparison (refer to Table 2) showed the proposed augmentation method contributes to segmentation accuracy in both internal validation and challenge evaluation (i.e., MRI from scanners not seen during training).

The ablation study performed using the internal validation process showed that each contribution, taken separately, enhanced the segmentation accuracy. It also showed that when combining all contributions, we achieved the best results. Similarly, the challenge evaluation showed that the proposed method achieved better results than the best-performing methods of the challenge.

Our experiment in Section 3.1.2 has shown interesting behavior of our method when trained on only 8 patients (minor performance decrease compared to using synthetic data only). The fine-tuning and optimization by selecting the best weights combination based on a very limited validation set has foreseeably led to overfitting. Thus, it is advised that the number of samples and their quality (containing enough new MS lesions) are sufficient so the fine-tuning step could enhance the performance. If the labeled dataset is not sufficient, combining both synthetic and real data could also be explored.

Our study explored the possibility of using a similar task such as MS lesion segmentation to better train new MS lesion segmentation models. Transfer learning has led to satisfactory results. However, other methods for instance multi-task learning and consistency regularization should be explored likewise. Other of our experiments (that have not been covered in our paper) investigated such strategies on both single time-point MS and new MS lesion segmentation. Unfortunately, it is difficult to deal with the different class imbalances and complexities of both tasks which makes optimizing jointly over single time-point MS and new MS lesion segmentation harder. We believe that a training-set containing both the segmentation of new lesions and the segmentation of other lesions contained in both time points could lead the community to propose better segmentation/detection models.

Although it is sometimes difficult for experts to agree upon whether a lesion is new or not, their consistency in the segmentation of new lesions is even more difficult. This inconsistency, despite being mitigated by the consensus of several experts, will have repercussions on the quality of the segmentation accuracy. Thus, we believe that if there is interest in the quantification of new lesion volume, the output of models trained only on one modality (FLAIR) and for the task of new lesion segmentation should be taken with precaution. Combining the outputs of this model with another one trained on a single time-point with several modalities (T1w and FLAIR) could lead to better and more accurate segmentation.

Besides the detection of new lesions, another interesting biomarker for MS clinicians is the measurement of disappearing lesions. Our proposed method could potentially be used for this task by inverting the time-point order. However, it has not been validated in our study and requires the appropriate expert annotations.

5. Conclusion

In this article, we propose a training pipeline to deal with the lack of data for new MS lesion segmentation from two time points. The pipeline encompasses transfer learning from single time-point MS lesion segmentation, pretraining with time-point synthesis, and data-augmentation adapted for MR images. Our ablation study showed that each of our contributions enhances the accuracy of the segmentation. Overall, our pipeline was very effective for new MS lesions segmentation (Best score in MSSEG2-challenge; MICCAI, 2021) and can be extended to other tasks that suffer from longitudinal data scarcity.

Data availability statement

The data analyzed in this study was obtained from France Life Imaging (FLI)—Information Analysis and Management (IAM) node, the following licenses/restrictions apply: Users must subscribe to the challenge to get data access via Shanoir-NG (next generation). Requests to access these datasets should be directed to Shanoir-NG, https://shanoir.irisa.fr/shanoir-ng/challenge-request.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the patients/participants or patients/participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

RK: method design and implementation, experiment, coding, writing, and editing. BM: coding, writing, and editing. JM: writing and editing. PC: method design, experiment, writing, and editing. All authors contributed to the article and approved the submitted version.

Funding

This study benefited from the support of the project DeepvolBrain of the French National Research Agency (ANR-18-CE45-0013). This study was achieved within the context of the Laboratory of Excellence TRAIL ANR-10-LABX-57 for the BigDataBrain project. Moreover, we thank the Investments for the future Program IdEx Bordeaux (ANR-10-IDEX-03-02 and RRI IMPACT), the French Ministry of Education and Research, and the CNRS for the DeepMultiBrain project. This study has also been supported by the PID2020-118608RB-I00 grants from the Spanish Ministerio de Economia, Industria Competitividad.

Acknowledgments

The authors gratefully acknowledge the support of NVIDIA Corporation with their donation of the TITAN Xp GPU used in this research.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bosc, M., Heitz, F., Armspach, J.-P., Namer, I., Gounot, D., and Rumbach, L. (2003). Automatic change detection in multimodal serial MRI: application to multiple sclerosis lesion evolution. *NeuroImage* 20, 643–656. doi: 10.1016/S1053-8119(03)00406-3

Bron, E. E., Klein, S., Papma, J. M., Jiskoot, L. C., Venkatraghavan, V., Linders, J., et al. (2021). Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. *NeuroImage* 31:102712. doi: 10.1016/j.nicl.2021.102712

Cabezas, M., Luo, Y., Kyle, K., Ly, L., Wang, C., and Barnett, M. (2021). "Estimating lesion activity through feature similarity: a dual path Unet approach for the MSSEG2 MICCAI challenge," in *MSSEG-2 Challenge Proceedings: Multiple Sclerosis New Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*, 107.

Carass, A., Roy, S., Jog, A., Cuzzocreo, J. L., Magrath, E., Gherman, A., et al. (2017). Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage* 148, 77–102. doi: 10.1016/j.dib.2017.04.004

Cheng, M., Galimzianova, A., Lesjak, Ž., Špiclin, Ž., Lock, C. B., and Rubin, D. L. (2018). "A multi-scale multiple sclerosis lesion change detection in a multi-sequence MRI," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Springer), 353–360. doi: 10.1007/978-3-030-00889-5_40

Commowick, O., Cervenansky, F., and Ameli, R. (2016). "MSSEG challenge proceedings: multiple sclerosis lesions segmentation challenge using a data management and processing infrastructure," in *MICCAI*. Available online at: https://scholar.google.fr/scholar?hl=fr&as_sdt=0%2C5&q=MSSEG+challenge+proceedings&btnG=#d=gs_cit&t=1659870818068&u=%2Fscholar%3Fq%3Dinfo%3AZnRobGdVz6gJ%3Ascholar.google.com%2F%26output%3Dcite%26scirp%3D0%26hl%3Dfr

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., et al. (2018). Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Sci. Rep.* 8, 1–17. doi: 10.1038/s41598-018-31911-7

Coupé, P., Tourdias, T., Linck, P., Romero, J. E., and Manjón, J. V. (2018). "Lesionbrain: an online tool for white matter lesion segmentation," in *International Workshop on Patch-based Techniques in Medical Imaging* (Springer), 95–103. doi: 10.1007/978-3-030-00500-9_11

Dalbis, T., Fritz, T., Grilo, J., Hitziger, S., and Ling, W. X. (2021). "Triplanar U-net with orientation aggregation for new lesions segmentation," in *MSSEG-2 Challenge Proceedings: Multiple Sclerosis New Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*, 57.

Elliott, C., Arnold, D. L., Collins, D. L., and Arbel, T. (2013). Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain MRI. *IEEE Trans. Med. imaging* 32, 1490–1503. doi: 10.1109/TMI.2013.2258403

Ganiler, O., Oliver, A., Diez, Y., Freixenet, J., Vilanova, J. C., Beltran, B., et al. (2014). A subtraction pipeline for automatic detection of new appearing multiple sclerosis lesions in longitudinal studies. *Neuroradiology* 56, 363–374. doi: 10.1007/s00234-014-1343-1

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2021). NNU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* 18, 203–211. doi: 10.1038/s41592-020-01008-z

Johnson, J. M., and Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *J. Big Data* 6, 1–54. doi: 10.1186/s40537-019-0192-5

Kamraoui, R. A., Ta, V.-T., Manjon, J. V., and Coupé, P. (2021). "Image quality data augmentation for new MS lesion segmentation," in *MSSEG-2 Challenge Proceedings: Multiple Sclerosis New Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*, 37.

Kamraoui, R. A., Ta, V.-T., Tourdias, T., Mansencal, B., Manjon, J. V., and Coupé, P. (2022). DeepLesionBrain: towards a broader deep-learning generalization for multiple sclerosis lesion segmentation. *Med. Image Anal.* 76:102312. doi: 10.1016/j.media.2021.102312

Khan, A. R., Khan, S., Harouni, M., Abbasi, R., Iqbal, S., and Mehmood, Z. (2021). Brain tumor segmentation using K-means clustering and deep learning with synthetic data augmentation for classification. *Microsc. Res. Techn.* 84, 1389–1399. doi: 10.1002/jemt.23694

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. Availableo online at: https://arxiv.org/pdf/1412.6980.pdf

Krüger, J., Opfer, R., Gessert, N., Ostwaldt, A.-C., Manogaran, P., Kitzler, H. H., et al. (2020). Fully automated longitudinal segmentation of new or enlarged multiple sclerosis lesions using 3D convolutional neural networks. *NeuroImage* 28:102445. doi: 10.1016/j.nicl.2020.102445

Mårtensson, G., Ferreira, D., Granberg, T., Cavallin, L., Oppedal, K., Padovani, A., et al. (2020). The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. *Med. Image Anal.* 2020:101714. doi: 10.1016/j.media.2020.101714

Manjón, J. V., and Coupé, P. (2016). volBrain: an online MRI brain volumetry system. *Front. Neuroinform.* 10:30. doi: 10.3389/fninf.2016.00030

Manjón, J. V., Coupé, P., Martí-Bonmatí, L., Collins, D. L., and Robles, M. (2010). Adaptive non-local means denoising of MR images with spatially varying noise levels. *J. Magn. Reson. Imaging* 31, 192–203. doi: 10.1002/jmri.22003

Manjón, J. V., Romero, J. E., Vivo-Hernando, R., Rubio, G., Aparici, F., de la Iglesia-Vaya, M., et al. (2020). "Blind MRI brain lesion inpainting using deep learning," in *International Workshop on Simulation and Synthesis in Medical Imaging* (Springer), 41–49. doi: 10.1007/978-3-030-59520-3_5

Masson, A., Le Bon, B., Kerbrat, A., Edan, G., Galassi, F., and Combes, B. (2021). "A NNUnet implementation of new lesions segmentation from serial FLAIR images of MS patients," in *MSSEG-2 Challenge Proceedings: Multiple Sclerosis New Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*, 5.

MICCAI (2021). *Longitudinal Multiple Sclerosis Lesion Segmentation Challenge*. Available online at: https://portal.fli-iam.irisa.fr/msseg-2/data/

Olivas, E. S., Guerrero, J. D. M., Martinez-Sober, M., Magdalena-Benedito, J. R., and Serrano, L. (2009). *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques: Algorithms, Methods, and Techniques*. IGI Global.

Omoumi, P., Ducarouge, A., Tournier, A., Harvey, H., Kahn, C. E., Louvet-de Verchère, F., et al. (2021). To buy or not to buy-evaluating commercial AI solutions in radiology (the ECLAIR guidelines). *Eur. Radiol.* 31, 3786–3796. doi: 10.1007/s00330-020-07684-x

Pérez-García, F., Sparks, R., and Ourselin, S. (2021). Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Comput. Methods Prog. Biomed.* 2021:106236. doi: 10.1016/j.cmpb.2021.106236

Salem, M., Cabezas, M., Valverde, S., Pareto, D., Oliver, A., Salvi, J., et al. (2018). A supervised framework with intensity subtraction and deformation field features for the detection of new T2-w lesions in multiple sclerosis. *NeuroImage* 17, 607–615. doi: 10.1016/j.nicl.2017.11.015

Schmidt, P., Pongratz, V., Küster, P., Meier, D., Wuerfel, J., Lukas, C., et al. (2019). Automated segmentation of changes in FLAIR-hyperintense white matter lesions in multiple sclerosis on serial magnetic resonance imaging. *NeuroImage* 23:101849. doi: 10.1016/j.nicl.2019.101849

Shaw, R., Sudre, C., Ourselin, S., and Cardoso, M. J. (2018). "MRI K-space motion artefact augmentation: model robustness and task-specific uncertainty," in *International Conference on Medical Imaging with Deep Learning-Full Paper Track*.

Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* 6, 1–48. doi: 10.1186/s40537-019-0197-0

Sudre, C. H., Cardoso, M. J., Ourselin, S., and Alzheimer's Disease Neuroimaging Initiative (2017). Longitudinal segmentation of age-related white matter hyperintensities. *Med. Image Anal.* 38, 50–64. doi: 10.1016/j.media.2017.02.007

Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., et al. (2018). "Training deep networks with synthetic data: bridging the reality gap by domain randomization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (IEEE)*, 969–977. doi: 10.1109/CVPRW.2018.00143

Tripathi, S., Chandra, S., Agrawal, A., Tyagi, A., Rehg, J. M., and Chari, V. (2019). "Learning to generate synthetic data *via* compositing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (IEEE)*, 461–470. doi: 10.1109/CVPR.2019.00055

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *J. Big Data* 3, 1–40. doi: 10.1186/s40537-016-0043-6

Zhang, H., Li, H., and Oguz, I. (2021). "Segmentation of new MS lesions with Tiramisu and 2.5 D stacked slices," in *MSSEG-2 Challenge Proceedings: Multiple Sclerosis New Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*, 61.

Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., et al. (2020). Generalizing deep learning for medical image segmentation to unseen domains *via* deep stacked transformation. *IEEE Trans. Med. Imaging* 39, 2531–2540. doi: 10.1109/TMI.2020.2973595

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 07 September 2022
DOI 10.3389/fnins.2022.981523

Check for updates

OPEN ACCESS

EDITED BY
Michel Dojat,
Institut National de la Santé et de la Recherche Médicale (INSERM), France

REVIEWED BY
Richard McKinley,
Bern University Hospital, Switzerland
Alejandro Veloz,
Universidad de Valparaiso, Chile

*CORRESPONDENCE
Julia Andresen
j.andresen@uni-luebeck.de

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 29 June 2022
ACCEPTED 18 August 2022
PUBLISHED 07 September 2022

CITATION
Andresen J, Uzunova H, Ehrhardt J, Kepp T and Handels H (2022) Image registration and appearance adaptation in non-correspondent image regions for new MS lesions detection. *Front. Neurosci.* 16:981523. doi: 10.3389/fnins.2022.981523

Image registration and appearance adaptation in non-correspondent image regions for new MS lesions detection

Julia Andresen[1]*, Hristina Uzunova[2], Jan Ehrhardt[1,2], Timo Kepp[1] and Heinz Handels[1,2]

[1]Institute of Medical Informatics, University of Lübeck, Lübeck, Germany, [2]German Research Center for Artificial Intelligence, Lübeck, Germany

Manual detection of newly formed lesions in multiple sclerosis is an important but tedious and difficult task. Several approaches for automating the detection of new lesions have recently been proposed, but they tend to either overestimate the actual amount of new lesions or to miss many lesions. In this paper, an image registration convolutional neural network (CNN) that adapts the baseline image to the follow-up image by spatial deformations and simulation of new lesions is proposed. Simultaneously, segmentations of new lesions are generated, which are shown to reliably estimate the real new lesion load and to separate stable and progressive patients. Several applications of the proposed network emerge: image registration, detection and segmentation of new lesions, and modeling of new MS lesions. The modeled lesions offer the possibility to investigate the intensity profile of new lesions.

KEYWORDS

convolutional neural networks, non-correspondences, image registration, shape and appearance adaptation, multiple sclerosis, new lesions

1. Introduction

Multiple sclerosis (MS) is a chronic inflammatory disease that progressively destroys the axons in the central nervous system. With an estimated number of more than 2 million affected, MS is the leading cause of neurological disability in young adults (WHO, 2008). The detection and quantification of new MS lesions based on magnetic resonance (MR) imaging is a crucial task in the monitoring of MS, since the presence of new lesions indicates drug inefficacy. The manual segmentation of MS lesions, however, is time-consuming and complex. In a postmortem study (Geurts et al., 2005), only 40% of lesions detected on histopathology were also found on FLAIR MR scans. The detection of new lesions is considered to be an even more challenging task, exhibiting high intra- and inter-rater variance. The automation of (new) MS lesion detection and segmentation has therefore attracted substantial attention recently, e.g., through several public challenges (Commowick et al., 2016, 2021; Carass et al., 2017).

Existing methods for automatic longitudinal examination of MS may be classified into lesion detection and change detection approaches (Lladó et al., 2012). Lesion detection approaches segment all lesions on MR volumes of single time points. For a longitudinal quantification of changes, a subsequent differentiation of static, dynamic and new lesions is needed. Köhler et al. for example use a semi-automatic segmentation approach to mark lesions in individual MR scans. Afterwards, they affinely register all images to a reference scan and finally distinguish between stable, dynamic and new lesions based on the intersection of lesion masks from all time points (Köhler et al., 2019).

Change-detection approaches on the other hand directly use both images from subsequent time points to detect changes between baseline and follow-up. These approaches can be subclassified into intensity- and deformation-based approaches (Salem et al., 2020). Intensity-based approaches compare pre-registered scans of subsequent time points on a voxel-by-voxel basis to segment new lesions, e.g., Moraal et al. (2010), Ganiler et al. (2014), and Battaglini et al. (2014); Jain et al. (2016); Fartaria et al. (2019). Deformation-based approaches, however, use non-rigid image registration and analyze the resulting deformation fields to find new or evolving lesions (Rey et al., 2002; Cabezas et al., 2016). Works combining ideas from intensity- and deformation-based approaches show improved performance compared to using intensity-based solutions alone (Cabezas et al., 2016; Salem et al., 2018).

The majority of recent methods for new MS lesion segmentation are based on deep learning (Krüger et al., 2020a; McKinley et al., 2020; Salem et al., 2020; Combès et al., 2021). A trend reflected in the submissions to the MICCAI 2021—Longitudinal Multiple Sclerosis Lesion Segmentation (MSSEG-2). Challenge (Commowick et al., 2021), most of which perform image registration as a pre-processing step and subsequently use a 2D or 3D U-Net-like architecture to segment new lesions. Especially promising segmentation results are achieved by Dalbis et al. (2021) and Zhang et al. (2021) that both use a 2.5D approach with image slices of all three directions as network input.

Salem et al. (2020) propose a fully convolutional network (FCN) that consists of four registration blocks followed by a segmentation block. Each registration block registers the baseline scan of a certain modality (T1, T2, PD, and FLAIR) to the respective follow-up scan. The resulting deformation fields are then fed to the segmentation part of the network (Salem et al., 2020). For the MSSEG-2 challenge, the authors adapt their approach to work with FLAIR images only.

Using image registration as a pre-processing step to lesion load change or new lesions detection may cause underestimation of changes, since not only geometrical distortions but also changes of interest are erroneously eliminated by the registration step. Joint image registration and non-correspondence estimation may overcome this problem (Dufresne et al., 2020). Classic, i.e., iterative approaches that estimate non-correspondences during the registration process can be found in (Ou et al., 2011; Chen et al., 2015; Dufresne et al., 2020; Krüger et al., 2020b). Ou et al. (2011) estimate the matching uniqueness between voxel pairs to weigh the image distance measure during the registration process. A similar approach is followed by Krüger et al. (2020b) who use probabilistic correspondences between sparse image representations to define the weight map. In Chen et al. (2015) and Dufresne et al. (2020), a segmentation mask of non-corresponding regions is generated during the registration process. This segmentation is used to mask out the image distance measure in non-corresponding image regions. Together with regularization of the segmentation, non-corresponding regions are thus found as outliers in the image distance and segmented directly. Following this approach, we propose in Andresen et al. (2022) what is, to the best of our knowledge, the first method that tackles joint image registration and non-correspondence segmentation with deep learning. For the MSSEG-2 challenge, we use this approach to register baseline and follow-up images of MS patients while simultaneously segmenting non-corresponding regions. The non-correspondence segmentation is then refined with a second FCN, resulting in a final segmentation of new MS lesions (Andresen et al., 2021).

While all these approaches handle non-correspondences by weighing them down during the registration process, other methods for image registration with non-correspondences directly model both spatial and intensity differences between images to make them look alike (Trouvé and Younes, 2005; Rekik et al., 2015; Wilms et al., 2017; Bône et al., 2020). Uzunova et al. propose the joint shape and appearance autoencoder (SAAE) that reconstructs images from a global template using spatial deformations and intensity transformations (Uzunova et al., 2021). This allows the reconstruction of different modalities within the same framework. To assure a proper disentanglement of shape and appearance, guided filtering (He et al., 2013) is used such that the appearance offsets do not change the shape of the template.

Inspired by Uzunova et al. (2021), we now extend our image registration CNN for new MS lesions detection (Andresen et al., 2021) to ANCR-Net (appearance adaptation in non-correspondent regions and image registration network). ANCR-Net not only spatially deforms the baseline image, but also changes its appearance in non-corresponding image areas to match the follow-up. The spatial displacement accounts for general misalignments between the baseline and the follow-up images, as well as for old lesions changing shapes and sizes. The intensity transformations, however, are not applied to the entire baseline images but only in non-corresponding areas, which allows us to directly model newly appearing MS lesions. Different from Andresen et al. (2021), we use only one CNN whose segmentation branch is trained in a supervised manner. The trained network offers several applications for MS lesion analysis: 1) detection and segmentation of new lesions, 2) registration of baseline to follow-up images and 3) modeling the appearance of new lesions.

2. Materials and methods

2.1. Training objective

As described in Andresen et al. (2022), CNN-based image registration of baseline image $\mathrm{B}:\Omega \rightarrow \mathbb{R}$ and follow-up image $\mathrm{F}:\Omega \rightarrow \mathbb{R}$ with simultaneous non-correspondence segmentation can be formulated with the following training objective (Andresen et al., 2022).

$$\mathcal{L}_{\mathrm{NCR}} = (1-N)\cdot \mathrm{D}\big(\mathrm{F}, \mathrm{B}\circ\varphi\big) + \alpha \mathrm{R}_{\varphi} + \beta \mathrm{R}_{N}, \qquad (1)$$

with image distance measure D and regularizers R_{φ} and R_{N}. The diffeomorphic deformation field $\varphi:\mathbb{R} \rightarrow \mathbb{R}^3$, with $\varphi = \exp(v)$ and the segmentation of non-correspondences $N:\Omega \rightarrow [0,1]$ are both network outputs. The regularizers $\mathrm{R}_{\varphi} = \|\nabla v\|_2^2$ and $\mathrm{R}_N = \sum_{x\in\Omega} N + \gamma \tanh\big(\|\nabla N\|_2\big)$ enforce smoothness of the velocity field v and small, regularly bordered segmentations N. The image distance measure D is evaluated only in corresponding image regions, while non-corresponding areas with large image distance are masked out. Based on outlier detection in the image distance measure, the network is able to simultaneously segment non-correspondences and to spatially align baseline and follow-up images.

Taking ideas from Uzunova et al. (2021), we now want to model new MS lesions as appearance offsets between baseline and follow-up images in non-corresponding image regions. This results in the new training objective

$$\mathcal{L} = \mathrm{D}\big(\mathrm{F}, (\mathrm{B} + N\cdot A)\circ\varphi\big) + \alpha \mathrm{R}_{\varphi} + \beta \mathcal{L}_{\mathrm{Dice}}(N\circ\varphi, \mathrm{S}). \qquad (2)$$

Appearance offsets $A:\Omega \rightarrow \mathbb{R}$ are masked with the non-correspondence segmentation N and added to the baseline image. The appearance adapted baseline is then spatially deformed to match the follow-up image. Normalized cross-correlation is used as an image distance measure. The regularizer R_{φ} is defined as in Eq. (1). Other than our previous approach, we now use the Dice loss between the network's non-correspondence segmentation N and the ground truth segmentation S, making the regularization of N obsolete.

The intuition behind this method is that only in the regions of new lesions, strong intensity changes are to be expected between the baseline and the follow-up. Thus, intensity transformations are only applied in the non-corresponding image regions in order to directly model the newly appearing MS lesions. The spatial displacement φ in turn accounts for old lesions changing shapes and sizes as well as for general misalignments between the baseline and the follow-up images, but not for newly appearing lesions.

2.2. Network architecture

Consistent with previous works, the proposed ANCR-Net consists of one encoder and two separate decoders whose exact architecture is shown in Figure 1. The encoder starts with two separate convolutional blocks that process input MR images and their subtraction image. The resulting feature maps are concatenated and passed through multiple max pooling and convolution operations, analogously to the U-Net (Ronneberger et al., 2015). Another common feature to the U-Net is that our network also has decoders connected to the encoder *via* skip connections. The first decoder outputs the diffeomorphic deformation φ and the other generates non-correspondence and appearance offset maps N and A. Outputs are generated on three levels of resolution to provide deep supervision on both branches (Hering et al., 2019; Andresen et al., 2022). The loss function is determined at all three levels of resolution and a weighted sum is calculated to give a final loss for backpropagation. The weighting factors are chosen to be 0.7, 0.2 and 0.1 for each level, respectively, giving the finest resolution level the highest weight. Input to the network are five stacked axial slices sampled to an isotropic resolution and image size of 368 × 512 pixels. To generate segmentation results for the entire image volume, we iterate slice-wise through the volume and keep the segmentation of the central slice of the stacked input patches.

2.3. Network training

For network training, we use the MSSEG-2 challenge dataset. It consists of 40 whole-head FLAIR MR image pairs. Baseline and follow-up images have been rigidly pre-aligned for each patient. New MS lesions—if present—were manually segmented in the pre-aligned images by four medical experts and combined to one ground truth label of new lesions, which are used for network training.

New MS lesions are rare and mostly small, resulting in lesions being severely underrepresented in the data. To account for the class imbalance problem, we pre-train the network by inserting simulated lesions into the images that do not have real new lesions and deforming them with random elastic deformations. The network is then trained in a supervised manner using Dice loss and mean squared error between predicted and ground truth deformations as loss function. For lesion simulation, we generate a mask indicating candidate locations of lesions as follows. First, brain extraction is performed on both time points separately and the union of the brain masks is defined as the final brain mask. Second, baseline and follow-up images are normalized to values between 0 and 1 and thresholded above 0.1 to exclude the ventricles from the final mask. The brain mask is then multiplied with the thresholded MR images. As the simulated lesions should not

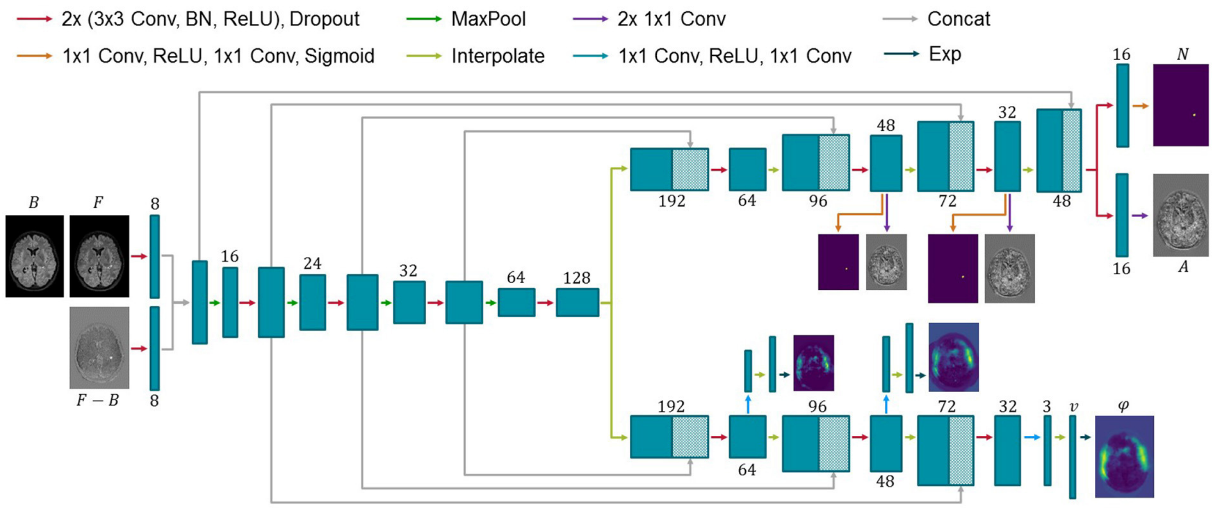


FIGURE 1
Architecture of the proposed network for new MS lesions modeling. Input to the network are baseline and follow-up MR images and their subtraction image. Two decoders generate a diffeomorphic deformation field φ for registration of baseline to follow-up and a segmentation N of new lesions together with appearance offsets A. Numbers below or above the blue boxes indicate the number of feature maps. Since the max pooling does not change the number of features, we omit the numbers before max pooling for visualization purposes.

protrude beyond the edge of the brain, the mask is subsequently shrunk using morphological erosion.

Artificial lesions are inserted on the fly during pre-training by first selecting a random number of new lesions (minimum one and maximum five) and randomly selecting locations from the candidate locations extracted before. At each selected location, we simulate a new lesion as a Gaussian ellipsoid whose values are added to the image intensities.

After lesion insertion, a random elastic deformation is applied to the image which then serves as fixed image whereas the original image is used as moving image. In addition, the following augmentation techniques are randomly applied to the moving and reference images during both the pre- and the final network training:

- Gaussian noise (inside brain region only)
- Rotation (±5°, performed on both images)
- Shift (±3 pixels in the axial plane, performed on both images)
- Brightness change (inside brain region only)
- Brightness gradient (inside brain region only)
- Adaptive histogram equalization

Pre-training is performed for 200 epochs, Adam optimization and a learning rate of $1e^{-4}$ that is decayed every 20th epoch with a factor of 0.8. After pre-training, ANCR-Net is trained with the loss function (2) using only image patches containing new lesions in the manual ground truth. Each of these patches is passed twice to the network, once with the original orientation and once flipped horizontally. Training is again performed with Adam optimization, exponentially decaying learning rate starting from $1e^{-4}$ and run for 400 epochs to assure full convergence. All code is made publicly available at https://github.com/juliaandresen/ANCRNet.git.

3. Experiments and results

The proposed method is validated on the test dataset of the MSSEG-2 challenge, consisting of 60 FLAIR MR image pairs. In our observations, the ground truth segmentation for one patient in the test data (ID 12) is not correct, thus we discard patient 12 from the test set and report results for the remaining 59 patients. For all experiments, we perform five-fold cross-validation on the training data, splitting the dataset into 32 training and 8 validation images per fold. The networks are ensembled and segmentations combined by majority vote. Each lesion in the resulting segmentations that is smaller than 3 mm^3 in volume is discarded. All metrics reported for new lesions detection and segmentation compare the manual consensus ground truth with the non-correspondence segmentations N. The non-correspondence segmentations are multiplied with brain masks generated by the default pre-processing pipeline[1] before metrics calculation.

[1] https://github.com/Inria-Empenn/lesion-segmentation-challenge-miccai21/

3.1. New lesions detection

New lesions detection performance is measured with several metrics. First, we report lesion sensitivity Sens_L, the proportion of detected new lesions in the ground truth. The lesion positive predictive value PPV_L gives the proportion of true positive lesions out of all lesions segmented by the network. Finally, the F_1-Score combines Sens_L and PPV_L as

$$\text{F}_1 = \frac{2 \cdot \text{Sens}_\text{L} \cdot \text{PPV}_\text{L}}{\text{Sens}_\text{L} + \text{PPV}_\text{L}}. \tag{3}$$

These metrics are not suitable for images that do not contain lesions in the ground truth. For these cases, we report average number and volume of erroneously detected lesions. We additionally give the proportion Det_p of patients correctly identified as progressing, i.e., at least one ground truth lesion is detected. For patients without new lesions we report Det_s, the proportion of patients correctly identified as stable, i.e., no segmentation is generated for these patients.

Results are summarized in Table 1 both for images with and without ground truth lesions. For comparison, we report the average performance of the four medical experts who segmented the MSSEG-2 challenge data and of the three teams achieving best results in the four metrics considered at the challenge: MedICL (Zhang et al., 2021) achieving the highest Dice score, Mediaire-B (Dalbis et al., 2021) achieving the best F_1-Score and LYLE (Ashtari et al., 2021) who performed best for number and volume of erroneously detected lesions. The results per patient can be found in the Supplementary material.

The PPV_L results show that most automated methods, including ours, tend to overestimate the number of new MS lesions and generate quite a lot of false positives. This is particularly true when the proposed pre-training is not used. In return, they are able to reliably detect real new lesions, even exceeding the average detection rate of medical experts. Despite the high proportion of false positives on the images with new MS lesions, ANCR-Net manages to correctly identify 89.3% of the 28 patients without a ground truth lesion in the test set as stable. At the same time, an average of 63.3% of ground truth lesions are correctly identified by our network. For 25 out of the 31 patients in the test set, our CNN manages to correctly detect at least one ground truth lesion. Considering not only correctly detected new lesions but all generated lesions, ANCR-Net identifies 29 patients as progressing. While the competitive methods achieve high detection rates either for stable or progressive patients, our method is the only one capable of reliably detecting new lesions and keeping the number of false positives low in stable patients, thus properly separating stable and progressing patients. In addition, our network also reliably estimates the real number of new lesions, with a mean error of only 1.322 lesions.

In Figure 2, contentious new lesions not included in the ground truth but segmented by at least one of the four experts and also by our proposed network are shown. The figure highlights the difficulty of the new lesions detection problem that is further aggravated by the changing size and shape of lesions. Automatic methods for new lesion detection inherently suffer from these difficulties, leading to the observed high proportion of false positives.

3.2. Segmentation of new lesions

To measure lesion segmentation performance, average Dice score, surface distance and Hausdorff distance are considered. Results are reported in Table 2 and again compared to experts' performance and best performing challenge submissions. Segmentation performances overall are quite low, which is reflected both in the Dice score and in the surface-based metrics. The average surface distance is comparable for almost all automatic methods with a value of just over 9 mm. Only LYLE achieves a mean surface distance of 7.209 mm. The results for the Hausdorff distance vary more. Here, too, LYLE performs best with 38.883 mm. Our methods achieves the second lowest value of 42.618 mm.

Considering Dice score, the best performing method (MedICL) achieves a value of 0.523. Our method scores second with 0.470. Even the experts only achieve an average Dice score of 0.573. This highlights the difficulty of the MS lesion segmentation task. Lesion borders often appear blurred, making their exact delineation difficult. Still, Dice scores do not take into account separate lesions, but only measure the overlap of all segmented pixels. We therefore also compute Dice scores for the test data on lesion-level and report scores averaged over 1) all lesions in ground truth and 2) all detected ground truth lesions. Lesion-wise Dice scores are even lower than the results in Table 2 with 0.412 for our method and 0.558 for the experts when averaging is performed over all ground truth lesions. For detected ground truth lesions, the average lesion-wise Dice score is 0.631, showing that lesion delineation works well in the case of identified lesions, but the gap to experts is still large (experts' average 0.817).

Finally, factors influencing the detection and segmentation quality of ANCR-Net are analyzed. For each lesion in the manual ground truth, volume, convexity, contrast to surrounding tissue and contrast to the baseline image are considered. For lesion volume, the cube root of the volume is used as a very rough estimate of lesion diameter. As described in Lian et al. (2012), the convexity is calculated as the quotient of the lesion volume and the volume of the convex hull of this lesion. To calculate the contrast to the surrounding tissue, we determine the mean intensities within the lesion and in a small area around the lesion (found by binary dilation of the lesion segmentation with a spherical structuring element). The contrast is then calculated as the difference in mean intensity divided by the average of the two mean intensities (Nabavizadeh et al., 2019). The contrast to the baseline image is determined analogously using the

TABLE 1 New lesion detection results for images with and without new lesions in ground truth.

Model	Data	With new lesions				Without new lesions		
		F_1	$Sens_L$	PPV_L	Det_p	Number	Volume	Det_s
Experts	Validation	0.732	0.697	0.772	0.871	0.000	0.000	1.000
Ours, w/o PT	Validation	0.591	0.634	0.624	0.828	0.545	6.959	0.636
Ours	Validation	0.622	0.666	0.623	0.862	0.455	6.948	0.636
Experts	Test	0.635	0.609	0.663	0.815	0.045	1.514	0.955
MedICL	Test	0.516	**0.760**	0.465	**0.871**	0.536	12.713	0.643
Mediaire-B	Test	0.559	*0.707*	0.507	0.806	0.536	29.235	0.643
LYLE	Test	0.455	0.431	0.522	0.742	**0.036**	**0.470**	**0.964**
Ours, w/o PT	Test	*0.566*	0.623	**0.612**	*0.839*	0.250	4.443	0.750
Ours	Test	**0.582**	0.633	*0.582*	0.806	*0.107*	*2.039*	*0.893*

Reported are average F_1 score, lesion sensitivity and positive predictive lesion value for images containing new lesions. For stable patients, the average number of erroneously detected lesions and their volume are reported. The results are given for the medical experts who generated the manual ground truth data as well as for our proposed method with and without pre-training (PT) and compared to the three pipelines that performed best in the MSSEG-2 challenge (Ashtari et al., 2021; Dalbis et al., 2021; Zhang et al., 2021). Best results are given in bold font and second best in italics. No method manages to significantly outperform all other methods (according to a Wilcoxon signed rank test with significance level 0.05).

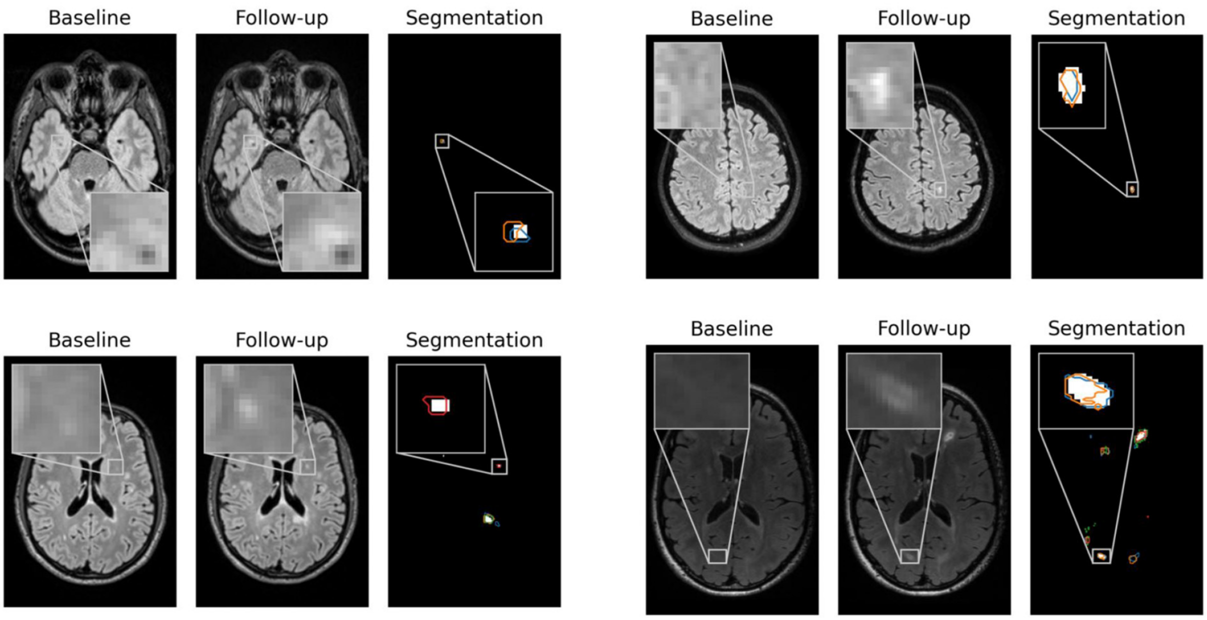


FIGURE 2
False positives generated by our CNN for patients 1, 23, 62, and 66 that were controversial among the experts. For each patient, baseline, follow-up and the segmentation of new lesions as generated by our network are shown. Segmentation contours of expert 1, 2, 3, and 4 are overlaid in blue, orange, green and red, respectively.

mean intensities within the lesion area in baseline and followup images.

Results are shown as scatter plots in Figure 3 where each point represents a ground truth lesion. It can be seen that lesion convexity does not seem to strongly influence the lesion detection performance. The pre-training on artificial lesions with an elliptical shape does not result in better detection of lesions with such a shape (as measured by convexity). The other considered metrics, however, have a greater impact on the detection performance of ANCR-Net. Larger lesions are detected with higher accuracy. Likewise, lesions that show a strong contrast to the background and especially to the baseline image are detected better than lesions with low contrast.

To analyze the influence of the considered lesion characteristics on the segmentation performance of ANCR-Net, linear regression is performed. For each lesion characteristic, we remove outliers biasing the regression results by discarding those lesions whose characteristic is smaller/larger than the 5 %-/95 % percentile of the respective characteristic. Also, we perform the regression once for all the remaining lesions and once for only those lesions that are detected by ANCR-Net. Each of the considered metrics shows a small positive correlation with Dice-score. A comparison of the regression results for all lesions and only for detected lesions shows again that the lesion volume and the contrast to the baseline image strongly influence the ANCR-Net detection rate. Interestingly, none of the metrics seem to have a very strong impact on segmentation performance when only looking at the detected lesions (red lines in Figure 3). Solely the contrast to the surrounding tissue gives a significant influence on the quality of the segmentation, with an R^2 of 0.085. Overall, lesion size and contrast to the baseline image are crucial for the detection of the lesions, but less so for their precise delineation, while contrast to the surrounding tissue is more critical for good segmentation.

TABLE 2 New lesion segmentation results for medical experts, our proposed method with and without pre-training (PT) as well as the three approaches performing best in the MSSEG-2 challenge (Ashtari et al., 2021; Dalbis et al., 2021; Zhang et al., 2021).

Model	Data	Dice	SD	HD
Experts	Validation	0.663	4.013	29.885
Ours, w/o PT	Validation	0.512	7.231*	41.502*
Ours	Validation	0.502	7.753*	37.688*
Experts	Test	0.573	6.211	32.639
MedICL	Test	**0.523**	9.352	61.835
Mediaire-B	Test	0.451	*9.010**	44.866*
LYLE	Test	0.422	**7.209***	**38.883***
Ours, w/o PT	Test	0.463	12.335	48.167
Ours	Test	*0.470*	9.053*	*42.618**

Reported are average Dice score, surface distance (SD) and Hausdorff distance (HD). Results marked with * are averaged over non-empty predicted segmentations only. For Mediaire-B two patients and for LYLE five patients are excluded from the distance calculation. With ANCR-Net, two patients from the validation data and two patients from the test data for the version trained with pre-training are excluded from the calculation. Significantly best results are presented in bold font.

3.3. Modeling of new lesions

Network outputs allow to not only spatially align baseline and follow-up, but also to model the appearance of newly formed lesions. To do so, the appearance offset map masked with the segmentation output is added to the baseline image and the adapted baseline is spatially deformed to match the follow-up image. In Figure 4 some exemplary results are shown for image registration and appearance adaptation between baseline and follow-up using new lesion modeling. For more examples refer to the Supplementary material.

The figure shows that the deformed and appearance adapted baseline images resemble the follow-up images well. The

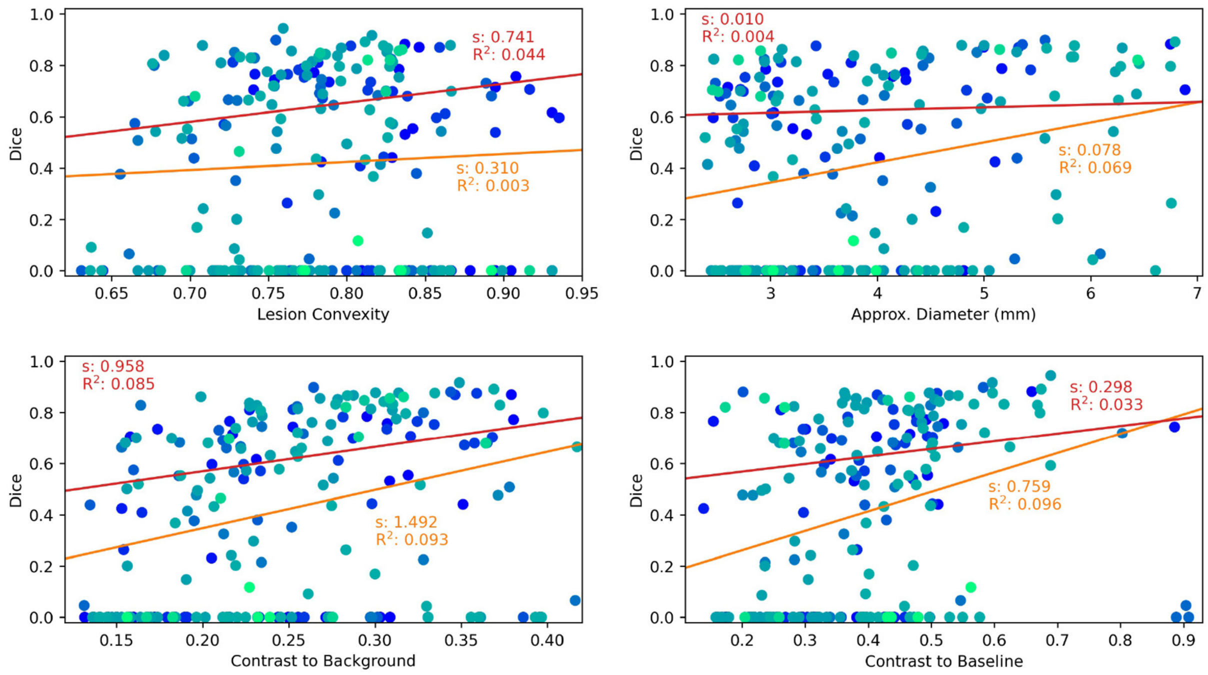


FIGURE 3
Lesion characteristics influencing the lesion detection and segmentation performance of ANCR-Net. Different colors represent different patients. The results of a linear regression measuring the influence of the respective lesion characteristic on Dice score are shown. Red lines show the results using only those lesions detected by ANCR-Net, whereas orange lines show the results considering all lesions. For each regression line, the slope s and the R^2 value are given.

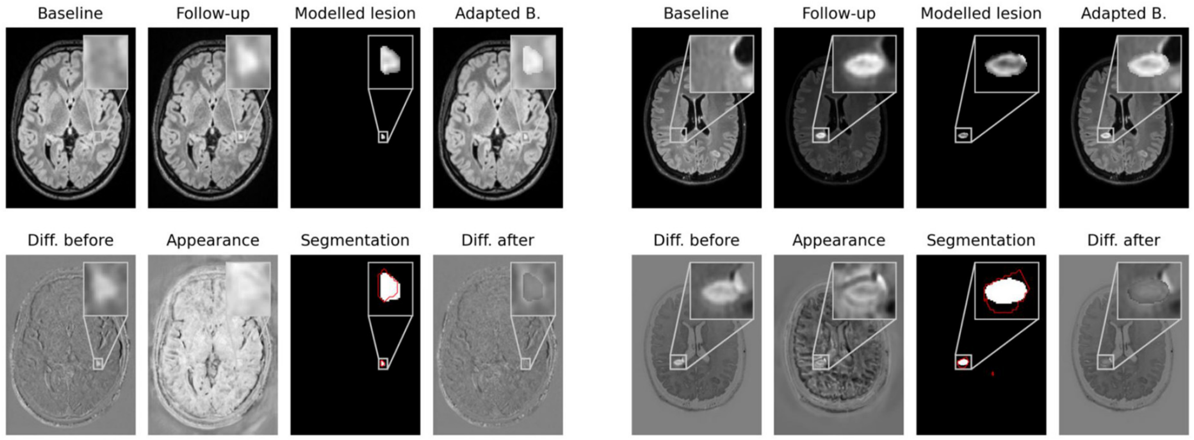


FIGURE 4
Modeling of new lesions. The appearance map is masked with the new lesions segmentation and added to the spatially deformed baseline image. The upper row shows baseline and follow-up images, the masked appearance map, deformed and appearance adapted baseline. In the lower row the difference image between follow-up and baseline, the appearance map, the segmentation of our CNN and the difference image between follow-up and adapted baseline are shown. The ground truth lesion segmentation is overlaid in red onto the network's segmentation.

modeled lesions do not overcompensate the overall intensity difference between baseline and follow-up images. Instead, the difference images show similar values inside and outside new lesions. The modeled lesions thus fit the intensity distribution of the baseline image. Investigating the modeled lesions, it can be seen that, even though MS lesions appear primarily as bright spots in FLAIR MR images, some of them still exhibit an irregular intensity profile. These irregular intensity profiles can be seen particularly well in the masked appearance maps (upper row in third and seventh columns in Figure 4), which might be used to analyze the morphology of newly forming MS lesions.

4. Discussion

We presented ANCR-Net, a CNN for the adaptation of baseline FLAIR MR images from MS patients to the respective follow-up images. Spatial deformations are applied to align baseline and follow-up structures, and new lesions are simulated in non-corresponding image areas. The trained network gives three outputs, namely a diffeomorphic deformation field to spatially align baseline and follow-up, a segmentation of new lesions and an appearance offset map that can be used to model newly appeared MS lesions.

New lesions detection and segmentation performances were compared to approaches scoring best in the MSSEG-2 challenge. The proposed CNN achieved highest lesion sensitivity (proportion of detected ground truth lesions) and F_1-Score. Most automatic methods for new MS lesions segmentation tend to produce quite a lot of false positives. ANCR-Net was the only method capable of keeping the number of such false positives comparably low while still detecting 63.3% of the new lesions on average. Thus, our method is the one best suited to separate stable and progressing patients.

Segmentation performances overall were quite low, but even the medical experts achieved an average Dice score of only 0.573. Our method achieved the second-best Dice score of all automatic methods, with a value of 0.470. Evaluations on lesion level showed that correctly detected lesions are indeed well delineated, a fact that the overall Dice score fails to reflect. Whether the exact delineation of the new lesions is actually crucial for MS monitoring, or rather their number and size, should be further investigated. Here, our network could be a valuable tool as it estimated the true number of new lesions very well, with a mean deviation of only 1.3 lesions.

The modeled new lesions were shown to fit well with the intensity profile of the baseline images and were able to match the baseline to the follow-up image. Some modeled lesions exhibit an irregular intensity profile that might give new insights into the morphology of MS lesions. The intensity profile of the lesions can be analyzed independently of the surrounding MR images using our masked appearance offsets maps. Distracting or influencing factors of the original images can thus be eliminated. Extensions to multimodal network inputs would also allow analyzing different types of MS lesions. Sheng et al. for example differentiate between hypo-, iso- and hyperintense lesions on susceptibility-weighted imaging (Sheng et al., 2019). Such a distinction could easily be made automatically based on our modeled lesions.

Network training using random intensity transformations makes the method robust to appearance variations between time points, as they might e.g., be introduced by imaging artifacts (see also Section 3 in Supplementary material). Still, the challenge training data is limited to 40 cases with high quality and well pre-registered images, thus performance may degrade in less controlled settings. The training dataset should therefore be extended with more data that reflects the natural variability of images in clinical practice. For example, the images could be noisier, or they could have been taken with different scanners at each visit. Also, the current method is designed for monomodal data. Extensions to multimodal inputs could be achieved by training ANCR-Net for each modality separately and then combining the results for the different modalities. How the method can be extended to take advantage of the different modalities in a single CNN will be the subject of future research.

Overall, the automatic analysis of new MS lesions remains a very difficult task. Our network achieves good values for all metrics considered, performing comparable to state-of-the-art methods for new MS lesions detection and segmentation. It is the only method capable of reliably separating stable and progressing patients, which additionally allows estimating the real new lesion load. Beyond that, the generated appearance offset maps offer the possibility to investigate morphology and intensity profile patterns of newly developed MS lesions. Our method is thus an important step toward automating the analysis of new MS lesions and achieving the performance of medical experts.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://shanoir.irisa.fr/shanoir-ng/challenge-request.

Author contributions

JA, HU, JE, and TK: methodology. JA and JE: software. JA: validation and writing—original draft preparation. JA, HU, JE, TK, and HH: writing—review and editing. HH: supervision. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2022.981523/full#supplementary-material

References

Andresen, J., Kepp, T., Ehrhardt, J., von der Burchard, C., Roider, J., and Handels, H. (2022). Deep learning-based simultaneous registration and unsupervised non-correspondence segmentation of medical images with pathologies. *Int. J. Comput. Assist. Radiol. Surg.* 17, 699–710. doi: 10.1007/s11548-022-02577-4

Andresen, J., Uzunova, H., Ehrhardt, J., and Handels, H. (2021). "New multiple sclerosis lesion detection with convolutional neural registration networks," in *MSSEG-2 Challenge Proceedings: Multiple Sclerosis New Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure* (Strasbourg), 111–114.

Ashtari, P., Barile, B., Van Huffel, S., and Sappey-Marinier, D. (2021). "Longitudinal multiple sclerosis lesion segmentation using pre-activation U-Net," in *MSSEG-2 Challenge Proceedings: Multiple Sclerosis New Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure* (Strasbourg), 61–64.

Battaglini, M., Rossi, F., Grove, R. A., Stromillo, M. L., Whitcher, B., Matthews, P. M., et al. (2014). Automated identification of brain new lesions in multiple sclerosis using subtraction images. *J. Magn. Reson. Imaging* 39, 1543–1549. doi: 10.1002/jmri.24293

Bône, A., Paul, V., Colliot, O., and Durrleman, S. (2020). "Learning joint shape and appearance representations with metamorphic auto-encoders," in *23rd International Conference on Image Computing and Computer Assisted Interventions-MICCAI 2020* (Lima), 202–211.

Cabezas, M., Corral, J. F., Oliver, A., Díez, Y., Tintoré, M., Auger, C., et al. (2016). Improved automatic detection of new T2 lesions in multiple sclerosis using deformation fields. *AJNR Am. J. Neuroradiol.* 37, 1816–1823. doi: 10.3174/ajnr.A4829

Carass, A., Roy, S., Jog, A., Cuzzocreo, J. L., Magrath, E., Gherman, A., et al. (2017). Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *Neuroimage* 148, 77–102. doi: 10.1016/j.neuroimage.2016.12.064

Chen, K., Derksen, A., Heldmann, S., Hallmann, M., and Berkels, B. (2015). "Deformable image registration with automatic non-correspondence detection," in *International Conference on Scale Space and Variational Methods in Computer Vision* (Lège-Cap Ferret), 360–371.

Combès, B., Kerbrat, A., Pasquier, G., Commowick, O., Le Bon, B., Galassi, F., et al. (2021). A clinically-compatible workflow for computer-aided assessment of brain disease activity in multiple sclerosis patients. *Front. Med.* 8, 740248. doi: 10.3389/fmed.2021.740248

Commowick, O., Cervenansky, F., and Ameli, R. (2016). "MSSEG challenge proceedings: multiple sclerosis lesions segmentation challenge using a data management and processing infrastructure," in *MICCAI* (Athènes).

Commowick, O., Cervenansky, F., Cotton, F., and Dojat, M. (2021). "MSSEG-2 challenge proceedings: multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure," in *24th International Conference on Medical Image Computing and Computer Assisted Intervention—MICCAI 2021* (Strasbourg).

Dalbis, T., Fritz, T., Grilo, J., Hitziger, S., and Ling, W. X. (2021). "Triplanar U-Net with orientation aggregation for new lesions segmentation," in *MSSEG-2 Challenge Proceedings: Multiple Sclerosis New Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure* (Strasbourg), 61–64.

Dufresne, E., Fortun, D., Kumar, B., Kremer, S., and Noblet, V. (2020). "Joint registration and change detection in longitudinal brain MRI," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* (Iowa City, IA: IEEE), 104–108.

Fartaria, M. J., Kober, T., Granziera, C., and Bach Cuadra, M. (2019). Longitudinal analysis of white matter and cortical lesions in multiple sclerosis. *Neuroimage Clin*. 23, 101938. doi: 10.1016/j.nicl.2019.101938

Ganiler, O., Oliver, A., Diez, Y., Freixenet, J., Vilanova, J. C., Beltran, B., et al. (2014). A subtraction pipeline for automatic detection of new appearing multiple sclerosis lesions in longitudinal studies. *Neuroradiology* 56, 363–374. doi: 10.1007/s00234-014-1343-1

Geurts, J. J., Bö, L., Pouwels, P. J., Castelijns, J. A., Polman, C. H., and Barkhof, F. (2005). Cortical lesions in multiple sclerosis: combined postmortem MR imaging and histopathology. *AJNR Am. J. Neuroradiol*. 26, 572–577.

He, K., Sun, J., and Tang, X. (2013). Guided image filtering. *IEEE Trans. Pattern Anal. Mach. Intell*. 35, 1397–1409. doi: 10.1109/TPAMI.2012.213

Hering, A., Kuckertz, S., Heldmann, S., and Heinrich, M. P. (2019). "Enhancing label-driven deep deformable image registration with local distance metrics for state-of-the-art cardiac motion tracking," in *Bildverarbeitung für die Medizin die Medizin 2019* (Lübeck), 309–314.

Jain, S., Ribbens, A., Sima, D. M., Cambron, M., De Keyser, J., Wang, C., et al. (2016). Two time point MS lesion segmentation in brain MRI: an expectation-maximization framework. *Front. Neurosci*. 10, 576. doi: 10.3389/fnins.2016.00576

Köhler, C., Wahl, H., Ziemssen, T., Linn, J., and Kitzler, H. H. (2019). Exploring individual multiple sclerosis lesion volume change over time: development of an algorithm for the analyses of longitudinal quantitative MRI measures. *Neuroimage Clin*. 21, 101623. doi: 10.1016/j.nicl.2018.101623

Krüger, J., Opfer, R., Gessert, N., Ostwaldt, A.-C., Manogaran, P., Kitzler, H. H., et al. (2020a). Fully automated longitudinal segmentation of new or enlarged multiple sclerosis lesions using 3D convolutional neural networks. *Neuroimage Clin*. 28, 102445. doi: 10.1016/j.nicl.2020.102445

Krüger, J., Schultz, S., Handels, H., and Ehrhardt, J. (2020b). Registration with probabilistic correspondences–Accurate and robust registration for pathological and inhomogeneous medical data. *Comput. Vis. Image Underst*. 190, 102839. doi: 10.1016/j.cviu.2019.102839

Lian, Z., Godil, A., Rosin, P., and Sun, X. (2012). "A new convexity measurement for 3D meshes," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (San Juan).

Lladó, X., Ganiler, O., Oliver, A., Martí, R., Freixenet, J., Valls, L., et al. (2012). Automated detection of multiple sclerosis lesions in serial brain MRI. *Neuroradiology* 54, 787–807. doi: 10.1007/s00234-011-0992-6

McKinley, R., Wepfer, R., Grunder, L., Aschwanden, F., Fischer, T., Friedli, C., et al. (2020). Automatic detection of lesion load change in Multiple Sclerosis using convolutional neural networks with segmentation confidence. *Neuroimage Clin*. 25, 102014. doi: 10.1016/j.nicl.2019.102104

Moraal, B., Wattjes, M. P., Geurts, J. J. G., Knol, D. L., van Schijndel, R. A., Pouwels, P. J. W., et al. (2010). Improved detection of active multiple sclerosis lesions: 3D subtraction imaging. *Radiology* 255, 154–163. doi: 10.1148/radiol.09090814

Nabavizadeh, A., Bayat, M., Kumar, V., Gregory, A., Webb, J., Alizad, A., et al. (2019). Viscoelastic biomarker for differentiation of benign and malignant breast lesion in ultra- low frequency range. *Sci. Rep*. 9, 5737. doi: 10.1038/s41598-019-41885-9

Ou, Y., Sotiras, A., Paragios, N., and Davatzikos, C. (2011). DRAMMS: deformable registration via attribute matching and mutual-saliency weighting. *Med. Image Anal*. 15, 622–639. doi: 10.1016/j.media.2010.07.002

Rekik, I., Li, G., Wu, G., Lin, W., and Shen, D. (2015). "Prediction of infant MRI appearance and anatomical structure evolution using sparse patch-based metamorphosis learning framework," in *Patch-Based Techniques in Medical Imaging: First International Workshop, Patch-MI 2015, Held in Conjunction With MICCAI 2015, Munich, Germany, October 9, 2015* (Munich), 197–204.

Rey, D., Subsol, G., Delingette, H., and Ayache, N. (2002). Automatic detection and segmentation of evolving processes in 3D medical images: APP|lication to multiple sclerosis. *Med. Image Anal*. 6, 163–179. doi: 10.1016/S1361-8415(02)00056-7

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention-MICCAI 2015* (Munich), 234–241.

Salem, M., Cabezas, M., Valverde, S., Pareto, D., Oliver, A., Salvi, J., et al. (2018). A supervised framework with intensity subtraction and deformation field features for the detection of new T2-w lesions in multiple sclerosis. *Neuroimage Clin*. 17, 607–615. doi: 10.1016/j.nicl.2017.11.015

Salem, M., Valverde, S., Cabezas, M., Pareto, D., Oliver, A., Salvi, J., et al. (2020). A fully convolutional neural network for new T2-w lesion detection in multiple sclerosis. *Neuroimage Clin*. 25, 102149. doi: 10.1016/j.nicl.2019.102149

Sheng, H., Zhao, B., and Ge, Y. (2019). Blood perfusion and cellular microstructural changes associated with iron deposition in multiple sclerosis lesions. *Front. Neurol*. 10, 747. doi: 10.3389/fneur.2019.00747

Trouvé, A., and Younes, L. (2005). Metamorphoses through lie group action. *Found Comput. Math*. 5, 173–198. doi: 10.1007/s10208-004-0128-z

Uzunova, H., Handels, H., and Ehrhardt, J. (2021). "Guided filter regularization for improved disentanglement of shape and appearance in diffeomorphic autoencoders," in *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning, volume 143 of Proceedings of Machine Learning Research*, eds M. Heinrich, Q. Dou, M. de Bruijne, J. Lellmann, A. Schläfer, and F. Ernst (Lübeck: PMLR), 774–786.

WHO (2008). *Atlas: Multiple Sclerosis Resources in the World 2008*. Geneva: World Health Organization.

Wilms, M., Handels, H., and Ehrhardt, J. (2017). "Representative patch-based active appearance models generated from small training populations," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2017* (Quebec City, QC), 152–160.

Zhang, H., Li, H., and Oguz, I. (2021). "Segmentation of new MS lesions with tiramisu and 2.5D stacked slice," in *MSSEG-2 Challenge Proceedings: Multiple Sclerosis New Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure* (Strasbourg), 61–64.

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 29 September 2022
DOI 10.3389/fnins.2022.954662

Check for updates

OPEN ACCESS

EDITED BY
Benoit Combès,
Inria Rennes-Bretagne Atlantique
Research Centre, France

REVIEWED BY
Thomas Grenier,
Institut National des Sciences
Appliquées de Lyon, France
Cédric Meurée,
Inria Rennes-Bretagne Atlantique
Research Centre, France

*CORRESPONDENCE
Liliana Valencia
u1955564@campus.udg.edu

SPECIALTY SECTION
This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroscience

RECEIVED 27 May 2022
ACCEPTED 30 August 2022
PUBLISHED 29 September 2022

CITATION
Valencia L, Clèrigues A, Valverde S,
Salem M, Oliver A, Rovira À and Lladó X
(2022) Evaluating the use of synthetic
T1-w images in new T2 lesion
detection in multiple sclerosis.
Front. Neurosci. 16:954662.
doi: 10.3389/fnins.2022.954662

Evaluating the use of synthetic T1-w images in new T2 lesion detection in multiple sclerosis

Liliana Valencia[1]*, Albert Clèrigues[1], Sergi Valverde[2], Mostafa Salem[1,3], Arnau Oliver[1], Àlex Rovira[4] and Xavier Lladó[1]

[1]Research Institute of Computer Vision and Robotics, University of Girona, Girona, Spain, [2]Tensor Medical, Girona, Spain, [3]Department of Computer Science, Faculty of Computers and Information, Assiut University, Asyut, Egypt, [4]Magnetic Resonance Unit, Department of Radiology, Vall d'Hebron University Hospital, Barcelona, Spain

The assessment of disease activity using serial brain MRI scans is one of the most valuable strategies for monitoring treatment response in patients with multiple sclerosis (MS) receiving disease-modifying treatments. Recently, several deep learning approaches have been proposed to improve this analysis, obtaining a good trade-off between sensitivity and specificity, especially when using T1-w and T2-FLAIR images as inputs. However, the need to acquire two different types of images is time-consuming, costly and not always available in clinical practice. In this paper, we investigate an approach to generate synthetic T1-w images from T2-FLAIR images and subsequently analyse the impact of using original and synthetic T1-w images on the performance of a state-of-the-art approach for longitudinal MS lesion detection. We evaluate our approach on a dataset containing 136 images from MS patients, and 73 images with lesion activity (the appearance of new T2 lesions in follow-up scans). To evaluate the synthesis of the images, we analyse the structural similarity index metric and the median absolute error and obtain consistent results. To study the impact of synthetic T1-w images, we evaluate the performance of the new lesion detection approach when using (1) both T2-FLAIR and T1-w original images, (2) only T2-FLAIR images, and (3) both T2-FLAIR and synthetic T1-w images. Sensitivities of 0.75, 0.63, and 0.81, respectively, were obtained at the same false-positive rate (0.14) for all experiments. In addition, we also present the results obtained when using the data from the international MSSEG-2 challenge, showing also an improvement when including synthetic T1-w images. In conclusion, we show that the use of synthetic images can support the lack of data or even be used instead of the original image to homogenize the contrast of the different acquisitions in new T2 lesions detection algorithms.

KEYWORDS

brain, MRI, synthetic images, deep learning, multiple sclerosis

1. Introduction

Artificial intelligence, particularly deep learning (DL), is currently widely used in medical imaging applications (Zhou et al., 2021; Chen et al., 2022). Tasks such as processing images (Razzak et al., 2018), segmenting anatomical structures (Fritscher et al., 2016) or diagnosing diseases such as stroke (Feng et al., 2018), brain tumors (Işın et al., 2016), and multiple sclerosis (Nair et al., 2020), are subjects of numerous domains of research. DL has been demonstrated to be a revolutionary tool in the field, improving state-of-the-art results. However, the algorithms developed with DL techniques have the major drawback of needing a large amount of data to train the model. Traditional data augmentation approaches, such as geometric transformations, intensity operations, filtering (Shorten and Khoshgoftaar, 2019), and deformable techniques such as deformable image registration or randomized displacement field, have been used to overcome this inconvenience. Nevertheless, some of these techniques have their own limitation such as the case of the geometric transformations which do not account for variations resulting from different imaging protocols or sequences, sizes, shapes, locations and appearances of the specific pathology (Yi et al., 2019) and produce highly correlated images in the training set, which prevents model improvements. Therefore, novel ways to mitigate these limitations have been studied including the use of image synthesis with DL (Chlap et al., 2021) .

Image synthesis consists of the generation of new parametric images, including deriving more tissue contrast from a collection of image acquisitions (Lundervold and Lundervold, 2019). Image synthesis makes the synthesis of new medical images possible, including images that may not have been available in the original dataset. In medical imaging, image synthesis has been explored using different approaches, such as atlas based approaches (Burgos et al., 2015), machine learning approaches (Jog et al., 2017) and, lately, deep learning techniques (Pinaya et al., 2022), especially the use of generative adversarial networks (GANs) (Yi et al., 2019). This last method is currently widely used. The GAN framework was proposed by Goodfellow et al. (2014) and has lead to impressive results. Using GANs, it is possible to generate realistic- looking images from an implicit distribution that follows the real data distribution (Kazeminia et al., 2020). GAN approaches for synthesis can be either conditional, where an example of the desired output is specified and therefore labeled datasets are needed; or unconditional, where the output is a sample of a random class, using as unique input a noise vector. Unconditional strategies are less applied in the medical field. However, there were several studies, such as the one by Bermudez et al. (2018), where a deep convolutional GAN (DCGAN) learned to mimic the distribution of an entire high resolution magnetic resonance (MR) image, resulting in synthetic images that human observers could not reliably distinguish from the real images. From the conditional point of view, there are a large variety of works. For instance, in the image translation from computed tomography (CT) images to MR images, Wolterink et al. (2017) proposed a strategy using unpaired data of CT and MR cardiac images fed in a Cycle Consistency GAN (CycleGAN) (Zhu et al., 2017) for image translation and corresponding segmentation mask. The use of cross-modality in MR studies, such as the proposal by Lee et al. (2020), where a missing MR image (modality) can be inferred using its remaining contrast pairs with the application of collaGAN, an image imputation method (Lee et al., 2019). In Hi-Net (Zhou et al., 2020), the authors used different synthesis combinations, such as T1 and T2 sequences, to synthesize Fluid-attenuated inversion recovery (FLAIR) sequences, T1 and FLAIR sequences to synthesize T2 sequence, and T2 and FLAIR sequences to synthesize T1 sequences. Zhou et al. (2020) showed how their method outperformed state-of-the-art methods such as the pix2pix model (Isola et al., 2017) or CycleGAN (Zhu et al., 2017) by utilizing the correlation between different modalities for a modality-specific network that learns the representation of each individual modality and a fusion network dedicated to learn the common latent representation of the multimodal data.

Many medical image analysis approaches can take advantage of image synthesis as an strategy to overcome the lack of data or the necessity of several MR sequences. This is the case for multiple sclerosis (MS) which is a central nervous system inflammatory demyelinating disorder. MRI plays an essential role in establishing an accurate and early diagnosis of MS (Hemond and Bakshi, 2018), and monitoring treatment response, mainly by assessing new T2 lesion formations. There are several approaches of new T2 lesions detection pipelines using DL (McKinley et al., 2020; Salem et al., 2020). Two typical constraints in the pipelines are the lack of annotated data and the necessity of these models to use more than one MR image modality in order to determine the number, size and location of the lesion. Hence, some image synthesis proposals have been developed to overcome this drawback. For instance, Salem et al. (2019) proposed a model to generate synthetic MS lesions in MR images, while Wei et al. (2019) developed a model to synthesize the FLAIR modality by mapping multisequence source images.

We contribute to literature through the application of image synthesis to improve new T2 lesions detection for MS studies. To do so, synthetic T1-w MR images obtained of the original T2-FLAIR sequence are used in an algorithm for new T2 lesions detection. For the synthesis of the images, we propose an adversarial synthesis method based on the pix2pix approach (Isola et al., 2017). The performance of the synthetic images is evaluated when using them in the new T2 lesions detection pipeline from Salem et al. (2020). We also present the results of applying the proposed strategy to the MSSEG-2 challenge (Commowick et al., 2021). Our primary contribution is to demonstrate that the addition of synthetic T1-w images can contribute to the improvement of the sensitivity of the new T2 lesion detection algorithms when added to the original T2-FLAIR image as input to the detection models.

2. Materials and methods

In the development of this analysis, we used an in-house clinical dataset. The synthesis pipeline is based on 3D conditional GANs inspired by the pix2pix approach (Isola et al., 2017), while the recent proposal of Salem et al. (2020) is used for the detection of the new T2 lesions.

2.1. Dataset

The dataset used in this study contains 136 cases of MS patients with clinically isolated syndrome (CIS) where 73 cases had new T2 lesions in follow-up scans. The mean time between MR scans was 12 months (range 3–27 month). Basal and follow-up scans were obtained using a Siemens Tim Trio 3T with a 12-channel phased array coil. The MRI protocol included sagittal T1- weighted 3D magnetization-prepared rapid acquisition of gradient echo (MPRAGE) [repetition time (TR) = 2,300 ms, echo time (TE) = 2.98 ms, inversion time (TI) = 900 ms, voxel size = 1.0 x 1.0 x 1.2 mm^3] and transverse fast fluid-attenuated inversion recovery (FLAIR) (TR = 5,000 ms, TE = 394 ms, TI = 1,800 ms, flip angle = 120°, voxel size = 1.0 x 1.0 x 1.0 mm^3). The protocol was approved by the Vall d'Hebron Hospital (Barcelona, Spain) Research and Ethics Committee. Informed consent was obtained from each participant before enrolment in the study.

As the gold standard to evaluate the detection method, the number of new/enlarging T2 lesions was obtained after the review of the MRI images by an expert observer (a technician with more than 15 years of experience in assessing new T2 lesions for MS under neuroradiologist supervision) who was not blinded to the radiological report or clinical information.

In addition, we used the MSSEG-2 challenge dataset (Commowick et al., 2021) to extend the evaluation of our approach. A total of 100 MS patients were gathered where only 3D FLAIR sequences were acquired at a first and second timepoints (separated in from 1 to 3 years in time) using a total of 15 different MRI scanners (three GE scanners, six Philips scanners, and six Siemens scanners). The image characteristics vary with different resolutions and different voxel size (from 0.5 mm^3 to 1.2 mm^3). Data was separated according to 40 scans for training and 60 for testing. This database allows us to test the usefulness of our approach when missing T1 images in the training set.

2.2. Methodology

2.2.1. Preprocessing

The preprocessing done to all the images was the following. First, all images were registered to the MNI512 template. An affine transformation was applied to the follow-up image, while for the basal image, the concatenation between two affine transformations, one from basal to follow-up scans and the one from follow-up scans to the MNI512 template, was applied. ANTs (Avants et al., 2009) with default linear interpolation was used for this purpose. Later, skull stripping was applied with HD-BET (Isensee et al., 2019), and finally, the images were normalized in the range [0–1].

2.2.2. Proposed T1-w synthesis approach

The image generation architecture is based on the pix2pix architecture (Isola et al., 2017) which is a conditional GAN architecture where the network learn the mapping from the input to the output image as well as the loss function to train this mapping. Similarly to GANs, pix2pix architecture consists of a generator and a discriminator. During the training process, the generator tries to generate realistic samples in order to fool the discriminator while the discriminator tries to distinguish between real and synthetic samples (Xin et al., 2020).

A semantic image clustering of the T2-FLAIR image, which was obtained with the FSL FAST algorithm (Zhang et al., 2001; Jenkinson et al., 2012), together with its T1-w intensity pair as ground truth (Figure 1A) is used as input to the adversarial network. A different number of image clusters obtained using FSL FAST are considered in our experimental evaluation. We consider a minimum of 3 clusters corresponding to gray matter, white matter and cerebrospinal fluid (CSF), 5 clusters corresponding to gray matter, white matter, CSF and two partial volumes of the border between the tissues, and finally 7 and 9 clusters. These last clusterings of the image do not have a biological meaning but are considered here to study the impact on the synthesis model when smaller intensity clusters are used to perform the intensity mapping between modalities.

From each cluster volume and the T1-w image, patches of 32 x 32 x 32 are extracted and used as inputFrontiFron to the generator, which is a 3D ResUNet architecture of 8 blocks (Figure 2A), in essence a U-Net with residual layers. The UNet architecture (Ronneberger et al., 2015) is widely used in medical imaging due to its ability of capturing context through the extraction of high and low-level features and enable precise location. Adding residual connections allows merging feature maps from higher resolution layers with deconvolved maps to preserve localization details and improve back-propagation (He et al., 2016). Distinct from the original UNet architecture, which uses skip connections implemented with concatenations, we use summations to reduce the model complexity (Guerrero et al., 2018). After each residual layer in the downscaling path, pooling is applied. The discriminator is a ResNet with 4 blocks (Figure 2B), where the residual blocks are followed by pooling. Labels smoothing is used during the training of the model to improve the generalization and prevent the network to become over-confident about its

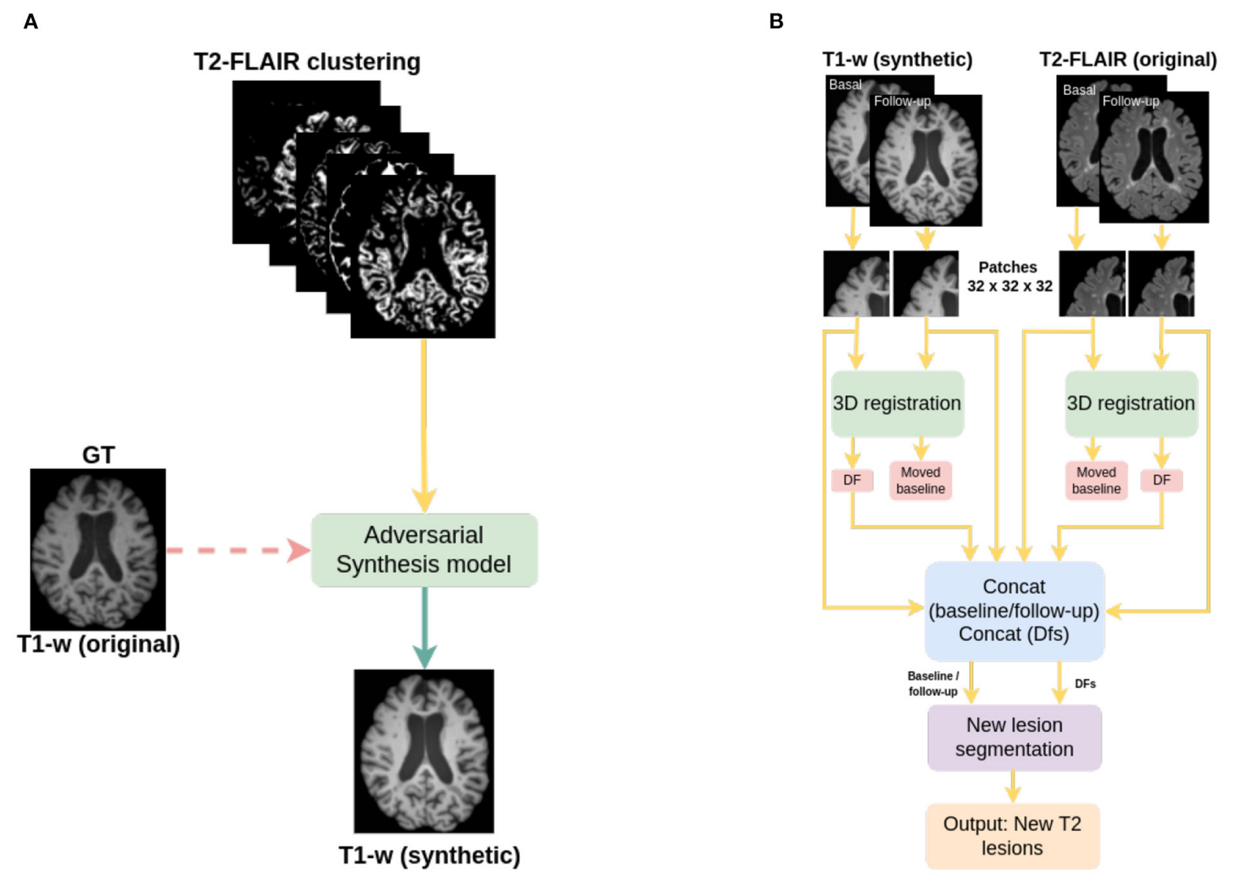


FIGURE 1
Pipelines used in this work; **(A)** pipeline for the synthesis of T1-w images, and **(B)** pipeline for new T2-w lesion detection in longitudinal analysis. The dashed line in **(A)** indicates that original T1-w images are used only in the training of the synthesizer.

prediction, therefore improving the accuracy (Müller et al., 2019).

Both the generator and discriminator have residual layers. Proposed by He et al. (2016), residual architectures facilitate the training of deeper networks, making them easier to optimize, and helping to improve the accuracy. Each block consists of two convolutions followed by batch normalization. The size of the kernel for the convolutions inside the residual blocks is 3 x 3 x 3. The pooling layers are implemented by striding with a kernel size of 2 x 2 x 2.

2.2.3. New T2 lesion detection algorithm

The detection of new T2 lesions in longitudinal images is performed using the approach of Salem et al. (2020). It consists of a fully convolutional network (FCNN) that accounts for two 3D architectures: first registration and then segmentation, which are trained end-to-end. The inputs to the FCNN are the basal and follow-up images, while the output is a new T2 lesion segmentation mask (Figure 1B).

The network consists of two architectures: the first one is a 3D U-Net for registration where for each input modality, the architecture learns the deformation fields and nonlinearly register the baseline image to the follow-up image. A second architecture, a 3D U-net, performs the final detection and segments the new T2-w lesions. Gradient descent is used as the optimizer and the network simultaneously learns both deformation fields and the new T2-w lesion segments. The loss function of the registration architecture is an unsupervised loss function (Balakrishnan et al., 2019) which has two components: one that penalizes differences in appearance and a second one that penalizes local spatial variation. For the segmentation architecture, the well known cross-entropy loss function is used. The network was trained using 3D patches of 32 x 32 x 32 with a step size of 16 x 16 x 16 extracted from both baseline and follow-up images. Adam was used as optimizer.

In the original work, Salem et al. (2020), the input modalities were T1-w, T2-w, PD-w, and T2-FLAIR. In this work, we modified them to be only T2-FLAIR (referred to FLAIR-only) or T2-FLAIR and T1-w images (referred to T2-FLAIR + T1). The

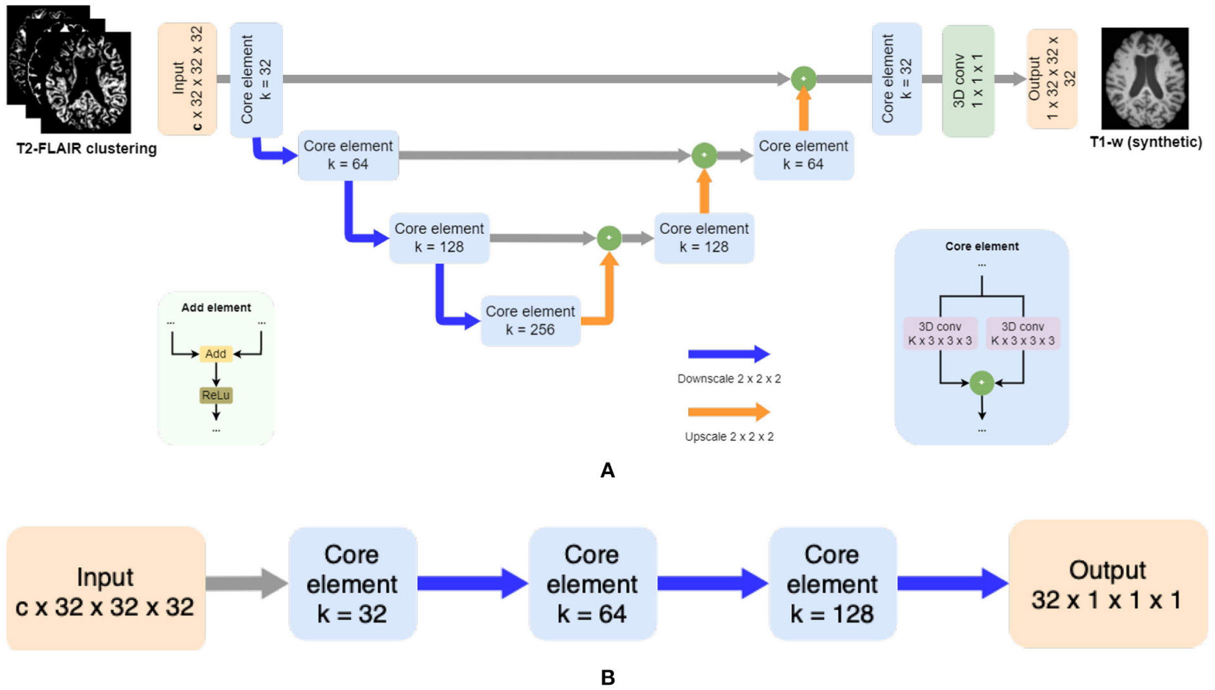


FIGURE 2
Adversarial synthesis model which takes as input the T2-FLAIR image clustering together with a T1-w intensity image as ground truth and generates a T1-w synthetic image as output. Core element modules in both architectures are composed of residual layers as described in He et al. (2016) with two convolutional layers ($k = 3 \times 3 \times 3$), followed by batch normalization, and finally a pooling layer by striding ($k = 2 \times 2 \times 2$). **(A)** The generator is a patchwise 3D encoder/decoder architecture where c is the number of clusters. Merged layers are implemented using summation instead of concatenation, using added element modules. **(B)** The discriminator is a patchwise 3D decoder architecture.

aim of this work is to evaluate the performance of the approach when using the synthetic T1 images generated as explained in the previous subsection.

2.3. Experimental evaluation

Three different experiments were performed in this study. First, we evaluated the image synthesis and determined which number of partial volumes improves the performance of the new T2 lesion detection algorithm.

Subsequently, using the in-house dataset, we compared the performance of using T1-w synthetic images for the lesion detection against two different models trained with original images, as shown in Figure 3, and described as:

- Baseline: model trained using original T2-FLAIR and T1-w images.
- FLAIR-only: model trained using only original T2-FLAIR images.
- Synthetic: model trained using original T2-FLAIR original images and synthetic T1-w images, obtained from the original T2-FLAIR images.

Finally, we also evaluated our image synthesis and lesion detection proposal using the data from the international MSSEG-2 challenge (Commowick et al., 2021), showing the obtained performance when using FLAIR-only and when adding the generated T1-w images.

2.3.1. Evaluation metrics for image quality

The quality of the images is evaluated locally measuring the voxel-wise intensity differences between a real image, y, and its approximation, $\bar{y}$, using the median absolute error (MAE) expressed as Equation (1). While the more similar images y and $\bar{y}$ are, the lower the MAE.

$$MAE(y, \bar{y}) = median\left|y - \bar{y}\right| \tag{1}$$

For a global evaluation, we use the structural similarity index metric (SSIM) proposed by Wang et al. (2004) and defined in Equation (2), which accounts for variations in luminance, contrast, and structure correlation, and has been found to correlate with the quality of perception of

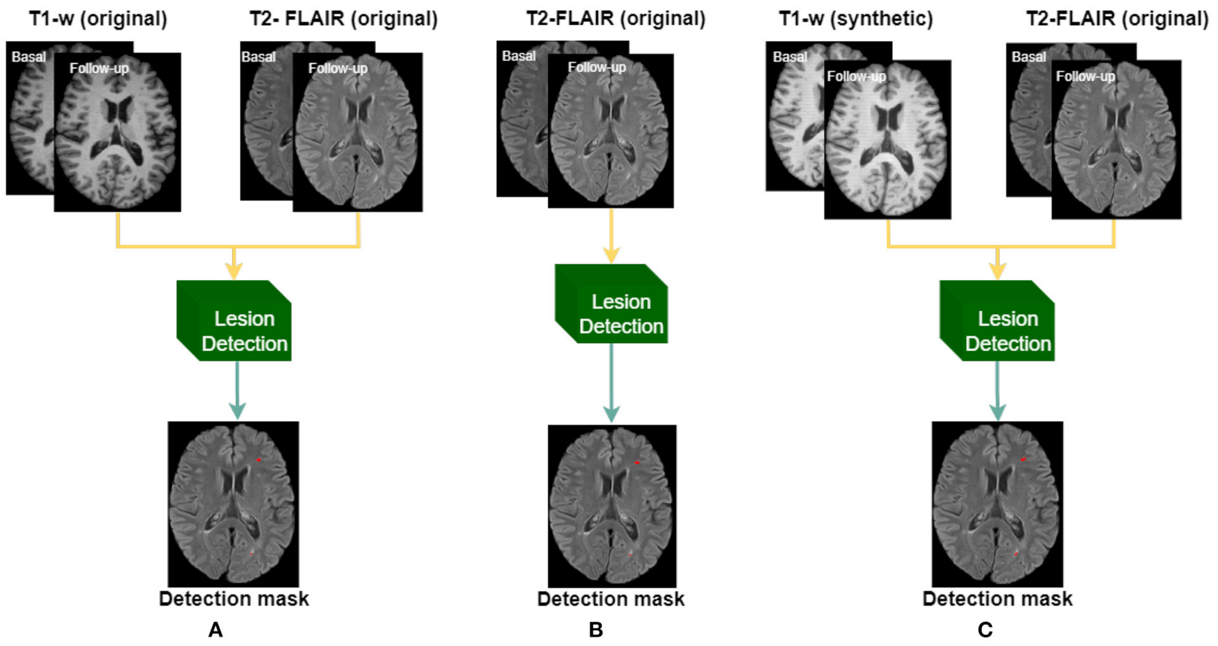


FIGURE 3
Three different models are trained according to the input images: **(A)** original T1-w and T2-FLAIR images, **(B)** using only original T2-FLAIR images, and **(C)** using synthetically obtained T1-w images (from the original T2-FLAIR images) along with the original T2-FLAIR images.

the human visual system (Hore and Ziou, 2010). It is defined as:

$$SSIM(y,\tilde{y}) = \frac{2\mu_y\mu_{\tilde{y}} + c_1}{\mu_y^2 + \mu_{\tilde{y}}^2 + c_1} \cdot \frac{2\sigma_y\sigma_{\tilde{y}} + c_2}{\sigma_y^2 + \sigma_{\tilde{y}}^2 + c_2} \cdot \frac{cov(y,\tilde{y}) + c_3}{\sigma_y\sigma_{\tilde{y}} + c_3}, \tag{2}$$

where μ denote the mean and σ is the standard deviation values of the luminance of the images, $cov(y,\tilde{y})$ is the covariance between y and $\tilde{y}$, and c_i is a constant that is used to avoid a null denominator (Hore and Ziou, 2010). The SSIM values range within zero and one, where zero indicates null similarity and one indicates total similarity.

2.3.2. Evaluation metrics for new T2 lesions detection performance

To evaluate the performance of the different trained models in the new T2 lesion detection algorithms, we use sensitivity, false discovery rate, and precision between the manual lesion annotation and the output segmentation mask. The sensitivity is defined as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

where TP and FN denote the number of correctly and missed lesion region candidates, respectively. In terms of detection, a lesion is considered TP if there is one voxel overlapping (Cabezas et al., 2016; Salem et al., 2018, 2020). The false discovery rate is:

$$FDR = \frac{FP}{FP + TP} \tag{4}$$

where FP denote the number of incorrectly classified lesion regions as positive. The precision is defined as:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

where TP and FP denote the numbers of correctly and miss classified lesion region candidates, respectively.

2.3.3. Statistical analysis

For each of the performance metrics of the detection of new T2 lesions, we applied the pairwise non-parametric Wilcoxon signed-rank test (two-sided) (Woolson, 2007), to assess the hypothesis of similar distributions between the different pairs of approaches. The results were considered significant for ($p < 0.05$).

3. Experimental results

To train and test the required models, we used the two subset configurations already available from the Vall d'Hebron

TABLE 1 Similarity between images and performance of the lesion detection algorithm when using 3, 5, 7, and 9 clusters in the synthesis of T1-w images.

	Similarity		Detection		
Modalities	**SSIM**	**MAE**	**Sensitivity**	**FDR**	**Precision**
T2-FLAIR + T1S (3c)	0.89 ± 0.07	0.11 ± 0.05	0.51 ± 0.38	0.07 ± 0.17	0.69 ± 0.42
T2-FLAIR + T1S (5c)	0.91 ± 0.07⋆	0.09 ± 0.05⋆	0.73 ± 0.31⋆	0.11 ± 0.20	0.83 ± 0.29⋆
T2-FLAIR + T1S (7c)	0.90 ± 0.07	0.10 ± 0.05	0.81 ± 0.23▽	0.14 ± 0.19	0.86 ± 0.19▽
T2-FLAIR + T1S (9c)	0.90 ± 0.07	0.09 ± 0.05 †	0.77 ± 0.30 †	0.25 ± 0.27 ⋄	0.73 ± 0.29

Significant differences in metrics between 5c and 3c are marked with ⋆, differences between 7c and 3c are marked with ▽, while differences between 9c and 3c are marked with †. Results of FDR for 9c are significantly lower with respect to the other 3 approaches (marked with ⋄).

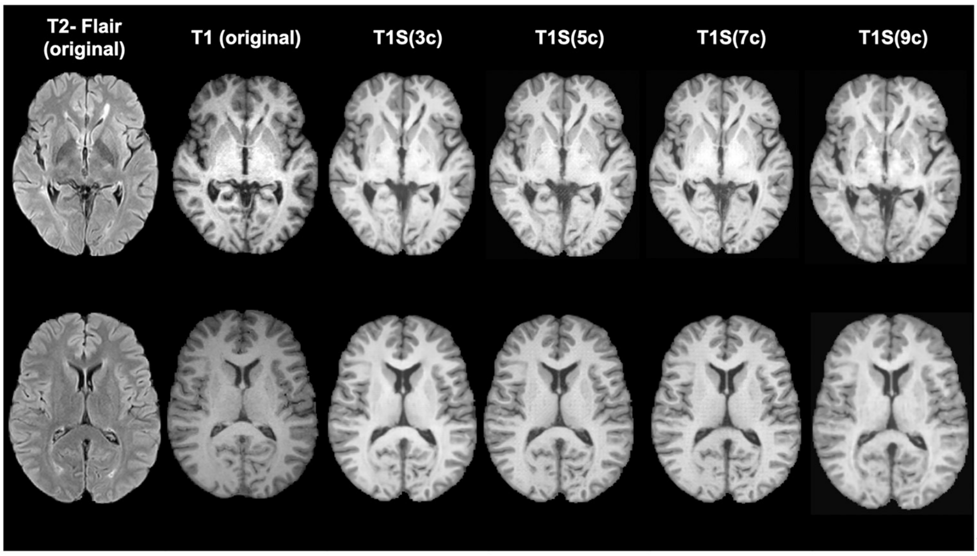


FIGURE 4
Examples of original and synthetically obtained images. The first column shows the original T2-FLAIR image, while the second column shows the original T1-w image. The following columns show the T1-w images obtained from the T2-FLAIR image using a different number of clusters (3, 5, 7, and 9).

Hospital. Set A included 101 patients, including 38 patients with new T2 lesions, and set B included 35 patients, all of whom had new T2 lesions. For the synthesis of T1-w images, set A was used for training, and set B was used for testing. Similarly, for the new T2 lesion detection models, the images from the 38 patients with new T2 lesions of set A were used for training, while the images from set B were used for testing (notice that for the model trained with synthetic images, the synthetic version of the images from set A were also computed).

We obtained the synthesized T1-w images using four different number of clusters of the T2-FLAIR image: 3, 5, 7, and 9 clusters. Table 1 shows the results of each case according to the similarity with the original image. For the inference of the new T2 lesion detection, voxels with ≥ 0.5 probability of being a lesion are taken as part of a lesion, while a lesion has a minimum of three neighboring voxels.

According to the similarity measures, the most similar image was obtained when using 5 clusters. Differences according to SSIM are small, while using MAE the performance of using 5 and 9 clusters are significantly different ($p < 0.05$) than when using 3 clusters. This difference in behavior of the measures shows the benefit of comparing the similarity between images both globally and locally. Figure 4 shows a qualitative example of each case, showing a high global similarity with respect to the ground truth, although there are discrepancies, mainly in the borders of the tissues, which are captured by the local similarity. Although the adversarial network exhibits common artifacts such as the intensity shift, they were more visible when using the approach with 3 clusters. On the contrary, when using 5, 7 and 9 clusters, axial slices generated tend to preserve better delineation of some structures.

Table 1 also shows the detection inferences computed with a fixed voxel probability threshold ≥ 0.5. Note that when using

T1S (7c), higher sensitivity and precision were obtained. To make these values more comparable, we inferred the detection using a threshold $\geq$ 0.3 for T1S (3c), T1S (5c) and T1S (9c) in an attempt to reach a similar operating point to that of the approach using T1S (7c). Under these conditions, T1S (3c) reached a sensitivity of 0.67 ± 0.33 with 0.14 ± 0.24 FDR, T1S (5c) increased the sensitivity to 0.78 ± 0.29 but with an FDR of 0.27±0.3, while T1S (9c) reached a sensitivity of 0.79±0.34 with 0.30 ± 0.26 FDR. Considering all these results, we can see that although all models were able to detect lesions, the best trade-off with the different detection measures was obtained when using 7 clusters in the synthesis of the T1-w sequence. Notice that images generated with 7c were not showing the best overall quality measurements but provided better feature information to improve the MS lesion detection.

Applying the synthesis based on 7 clusters (7c), we also evaluated the use of the synthetic T1-w images on the performance of the detection, using the 3 different approaches seen in Figure 3. Table 2 shows the obtained results. When using the original images, the sensitivity was 0.75 ± 0.29 at FDR of 0.09 ± 0.18, while when using only T2-FLAIR images as input the values were 0.63 ± 0.37 and 0.14 ± 0.24, respectively. When using the T2-FLAIR images along with the T1-w images synthesized from the same T2-FLAIR image, as an input, the sensitivity increased to 0.81 ± 0.23, without increasing the FDR with respect to the model using only T2-FLAIR images. The increase in sensitivity was significant with respect to the other models ($p < 0.05$). The precision between models showed that when using only T2-FLAIR images the performance was significantly lower ($p < 0.05$) than when using T1-w images, either real or synthesized. Comparing the use of both kinds of images, the results were similar.

3.1. Results using the MSSEG-2 dataset

In this experiment, we used our adversarial synthesis model trained with the in-house dataset to generate T1-w images for all the cases of the international MSSEG-2 challenge, where only T2-FLAIR images were available (Commowick et al., 2021).

TABLE 2 New T2 lesion detection performance evaluation using the models shown in Figure 3.

Modalities	Sensitivity	FDR	Precision
Results with original images			
T2-FLAIR + T1(Baseline)	0.75 ± 0.29	0.09 ± 0.18	0.85 ± 0.27
T2-FLAIR (FLAIR-only)	0.63 ± 0.37	0.14 ± 0.24	0.71 ± 0.38
Results with synthetic T1			
T2-FLAIR + T1S (7c)	0.81 ± 0.23 ⋆ ▽	0.14 ± 0.19	0.86 ± 0.19▽

Significant differences of the T2-FLAIR + T1S (7c) model w.r.t the Baseline and FLAIR-only models are marked with ⋆ and ▽, respectively.

We compared the performance of the MS lesion detection approach using only the T2-FLAIR images [original VICOROB submission to the challenge using Salem et al. (2020) with only T2-FLAIR images] vs. the model trained using both T2-FLAIR and T1-w synthetic images. Notice that all the MSSEG-2 training dataset was used to train both models, while the evaluation was done directly using the MSSEG-2 testing set, including both the active and stable cases.

The obtained results are illustrated in Table 3, where the two approaches are compared with some of the best pipelines participating in the challenge. Table 3 illustrates also the agreement of the approaches with the different expert raters. Interestingly, the performance of the model when using T1-w synthetic images was higher than the model using only T2-FLAIR images. For the active patients, we obtained an improvement in terms of sensitivity and precision of 0.12 and 0.2, respectively, while also reducing the FDR. Notice that the accuracy of the model was similar to that of some of the top participants in the challenge (MEDIARE$_B$, EMPENN and SNAC, see the MSSEG-2 challenge webpage for details of the participants), yielding also a performance that was comparable in terms of sensitivity to those of the human raters. Regarding the stable patients, where no new lesions were present, we observed a reduction in the total number of FP obtained and in the number of cases with FPs (11% of the 28 stable cases). Furthermore, it should be noted that our synthesis model was trained directly using the in-house dataset and only using images from a Siemens machine. This shows a capability of the model to adapt the source knowledge into the target domain of the challenge where data from different MRI scanners were available, producing T1-w images which indeed could be used to improve MS lesion detection.

4. Discussion

In this study, we investigated the usefulness of synthetic T1-w images in a longitudinal lesion detection pipeline. Starting from single T2-FLAIR images, we propose obtaining synthesized T1-w images that are subsequently used as an additional image modality to look for new abnormalities in the longitudinal analysis of the brain. Experiments show that although strong structural differences exist between T2-FLAIR and T1-w images, given the contrast difference between the two modalities, realistic T1-w images were able to be produced. In addition, the results show that adding the synthetic images to T2-FLAIR images in the detection pipeline provides new and reliable information that helps obtain better detection.

Our approach for generating T1-w images relies on intensity clustering of the T2-FLAIR images. The obtained clusters allow us to guide intensity information during the generation process. We have shown that images using more than 3 clusters are more similar to the original T1-w images. Most likely, the use of a few

TABLE 3 Results of the MSSEG-2 challenge 2021.

MSSEG-2 challenge	Active patients			Stable patients
	Sensitivity	FDR	Precision	N° of cases with FP (%)
Expert 1	0.71 ± 0.38	019 ± 0.31	0.72 ± 0.38	1 (4%)
Expert 2	0.61 ± 0.37	0.13 ± 0.21	0.68 ± 0.39	3 (11%)
Expert 3	0.61 ± 0.37	0.13 ± 0.23	0.69 ± 0.40	0 (0%)
Expert 4	0.47 ± 0.39	0.06 ± 0.19	0.66 ± 0.46	1 (4%)
MEDIARE$_B$	0.69 ± 0.40	0.39 ± 0.34	0.49 ± 0.36	10 (36%)
EMPENN	0.59 ± 0.32	0.33 ± 0.32	0.51 ± 0.36	8 (29%)
SNAC	0.66 ± 0.40	0.39 ± 0.33	0.49 ± 0.35	4 (14%)
VICOROB (FLAIR-only)	0.50 ± 0.39	0.43 ± 0.34	0.35 ± 0.32	6 (23%)
VICOROB (FLAIR + T1S)	0.62 ± 0.39	0.39 ± 0.38	0.55 ± 0.39	3 (11%)

Testing set composed by 60 patients: active patients $n = 32$, stable patients $n = 28$. Sensitivity, FPR and precision are shown for active patients, while the number of cases that presented FPs are provided for stable cases.

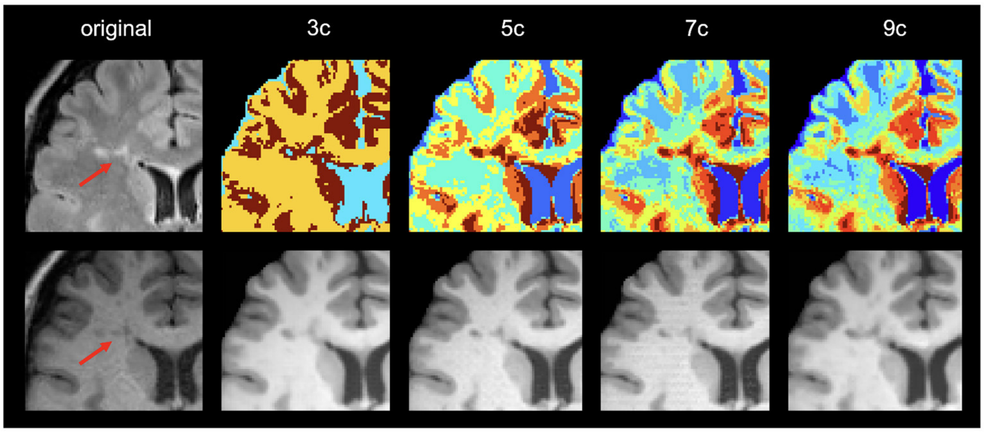


FIGURE 5
Example of image generations in a lesion area. First column shows the original T2-FLAIR and T1 image. The rest of the columns show the clustering result and the corresponding generated image using different numbers of clusters (3, 5, 7, and 9, respectively).

clusters does not account for the inherent partial volumes of MR images, while using more clusters allows better mapping of the partial volumes.

Regarding the lesion detection process, the best results were obtained when using 7 clusters. We observed that using more than 3 clusters allowed us to obtain additional information from the lesion areas that turns out to help in the lesion detection process. Note that the main goal of the synthesis is to provide images with complementary information to the network to improve lesion detection rather than produce high-quality synthetic images. Interestingly, we noticed that in the lesion areas, the model using 9 clusters tended to resemble too much the original T2-FLAIR cluster intensities in the generated T1 images, forcing an intensity mapping that deviates from the intensities present in the original T1. This can be seen in the example shown in Figure 5, where the generated image using 9 clusters produces more hypointense voxels in the lesion area than in the original T1 due to the larger number of clusters used and the intensity mapping learned from the model.

Comparing the detection performance when using only T2-FLAIR images vs. adding synthetic T1-w images, we found that there was a statistically significant difference in sensitivity between the two models. This indicates that the addition of T1-w synthetic images provides meaningful and additional information for the detection of the lesions. In contrast, the performance when using original T1-w images or synthetic images is similar, although we obtained slightly better results with the synthetic images. Our hypothesis is that in image synthesis, what is learned during training are the most predominant features of a T1-w image that can be extracted from a T2-FLAIR modality. These features may be related to the lesions, and therefore, the sensitivity during detection could improve. This may also be related to the number of clusters used. When using 5 clusters, we obtained more similar images than using 7 clusters, although the best performance for lesion detection was obtained when using the synthetic images from 7 clusters.

There is one limitation of this work that should be mentioned. All the images used in the study to train the synthesis model were taken from the same scanner, which was a Siemens Tim Trio 3T. Although the experiments done using the MSSEG-2 Challenge showed the capability of the synthesized images to improve the MS lesion detection even when using images from different MRI scanners (Siemens, Philips and GE), further investigations should be done in this line. As a future work, we plan to evaluate more exhaustively our synthesis approach when using images from different MRI scanners, analyzing not only the impact on the image generation and on the lesion detection performance, but also its applicability as an image standardization procedure. Furthermore, it could be very interesting to extend the study using more advanced synthesis models such as cycleGAN (Zhu et al., 2017) or Hi-Net (Zhou et al., 2020), which could in turn improve the generalization and the performance of the MS lesion detection approaches.

In conclusion, the results shown in this work demonstrate that the inclusion of synthetic images can support the lack of data. Specifically, we have seen how the inclusion of synthetic T1-w images on the lesion detection models helped to improve the overall performance. Our approach could benefit the clinical acquisition of MRI sequences, helping to reduce time and costs. Moreover, synthetic images could also be used instead of the original images to homogenize the contrast of the different acquisitions.

Data availability statement

The datasets presented in this article are not readily available because the dataset used in this work is an in-house dataset from the Vall d'Hebron Hospital (Barcelona, Spain) that includes T1-w and FLAIR images from 136 MS patients. Informed consent was obtained from each participant before enrolment in the study. The agreement done for sharing the data restricts the usability of the entities participating in this research study. Requests to access the datasets should be directed to xavier.llado@udg.edu.

Ethics statement

The studies involving human participants were reviewed and approved by Vall d'Hebron Hospital (Barcelona, Spain) Research and Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

Author contributions

LV, AC, MS, SV, AO, and XL contributed to the conception and design of the study. ÀR organized the database and provided clinical information. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

AC holds an FPI grant from the Ministerio de Ciencia, Innovación y Universidades with reference number PRE2018-083507. This work has been supported by DPI2020-114769RB-I00 from the Ministerio de Ciencia, Innovación y Universidades. The authors gratefully acknowledge the support of the NVIDIA Corporation with their donation of the TITAN X GPU used in this research. This work has been also supported by ICREA Academia Program.

Conflict of interest

Author SV was employed by company Tensor Medical.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Avants, B. B., Tustison, N., and Song, G. (2009). Advanced normalization tools (ants). *Insight J.* 2, 1–35. doi: 10.54294/uvnhin

Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., and Dalca, A. V. (2019). Voxelmorph: a learning framework for deformable medical image registration. *IEEE Trans. Med. Imaging* 38, 1788–1800. doi: 10.1109/TMI.2019.2897538

Bermudez, C., Plassard, A. J., Davis, L. T., Newton, A. T., Resnick, S. M., and Landman, B. A. (2018). "Learning implicit brain MRI manifolds with deep learning," in *Medical Imaging 2018: Image Processing, volume 10574* (Houston, TX: International Society for Optics and Photonics), 105741L.

Burgos, N., Cardoso, M. J., Guerreiro, F., Veiga, C., Modat, M., McClelland, J., et al. (2015). "Robust ct synthesis for radiotherapy planning: application to the head and neck region," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Munich: Springer), 476–484.

Cabezas, M., Corral, J., Oliver, A., Díez, Y., Tintoré, M., Auger, C., et al. (2016). Improved automatic detection of new t2 lesions in multiple sclerosis using deformation fields. *Am. J. Neuroradiol.* 37, 1816–1823. doi: 10.3174/ajnr.A4829

Chen, X., Wang, X., Zhang, K., Fung, K.-M., Thai, T. C., Moore, K., et al. (2022). Recent advances and clinical applications of deep learning in medical image analysis. *Med. Image Anal.* 2022, 102444. doi: 10.1016/j.media.2022.102444

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., and Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. *J. Med. Imaging Radiat. Oncol.* 65, 545–563. doi: 10.1111/1754-9485.13261

Commowick, O., Cervenansky, F., Cotton, F., and Dojat, M. (2021). "Msseg-2 challenge proceedings: multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure," in *MICCAI 2021-24th International Conference on Medical Image Computing and Computer Assisted Intervention* (Strasburg), 1–118.

Feng, R., Badgeley, M., Mocco, J., and Oermann, E. K. (2018). Deep learning guided stroke management: a review of clinical applications. *J. Neurointerv. Surg.* 10, 358–362. doi: 10.1136/neurintsurg-2017-013355

Fritscher, K., Raudaschl, P., Zaffino, P., Spadea, M. F., Sharp, G. C., and Schubert, R. (2016). "Deep neural networks for fast segmentation of 3d medical images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Athens: Springer), 158–165.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial networks. *arXiv preprint arXiv:1406.2661*. doi: 10.48550/arXiv.1406.2661

Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joules, R., et al. (2018). White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. *Neuroimage Clin.* 17, 918–934. doi: 10.1016/j.nicl.2017.12.022

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV: IEEE), 770–778.

Hemond, C. C., and Bakshi, R. (2018). Magnetic resonance imaging in multiple sclerosis. *Cold Spring Harb. Perspect. Med.* 8, a028969. doi: 10.1101/cshperspect.a028969

Hore, A., and Ziou, D. (2010). "Image quality metrics: PSNR vs. SSIM," in *2010 20th International Conference on Pattern Recognition* (Istanbul: IEEE), 2366–2369.

Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., et al. (2019). Automated brain extraction of multisequence mri using artificial neural networks. *Hum. Brain Mapp.* 40, 4952–4964. doi: 10.1002/hbm.24750

Işın, A., Direkoğlu, C., and Şah, M. (2016). Review of mri-based brain tumor image segmentation using deep learning methods. *Procedia Comput. Sci.* 102, 317–324. doi: 10.1016/j.procs.2016.09.407

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 1125–1134.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). Fsl. *Neuroimage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015

Jog, A., Carass, A., Roy, S., Pham, D. L., and Prince, J. L. (2017). Random forest regression for magnetic resonance image synthesis. *Med. Image Anal.* 35, 475–488. doi: 10.1016/j.media.2016.08.009

Kazeminia, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., et al. (2020). Gans for medical image analysis. *Artif. Intell. Med. 109*, 101938. doi: 10.1016/j.artmed.2020.101938

Lee, D., Kim, J., Moon, W.-J., and Ye, J. C. (2019). "Collagan: Collaborative gan for missing image data imputation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 2487–2496.

Lee, D., Moon, W.-J., and Ye, J. C. (2020). Assessing the importance of magnetic resonance contrasts using collaborative generative adversarial networks. *Nat. Mach. Intell.* 2, 34–42. doi: 10.1038/s42256-019-0137-x

Lundervold, A. S., and Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik* 29, 102–127. doi: 10.1016/j.zemedi.2018.11.002

McKinley, R., Wepfer, R., Grunder, L., Aschwanden, F., Fischer, T., Friedli, C., et al. (2020). Automatic detection of lesion load change in multiple sclerosis using convolutional neural networks with segmentation confidence. *Neuroimage Clin.* 25, 102104. doi: 10.1016/j.nicl.2019.102104

Müller, R., Kornblith, S., and Hinton, G. (2019). When does label smoothing help? *arXiv preprint arXiv:1906.02629*. doi: 10.48550/arXiv.1906.02629

Nair, T., Precup, D., Arnold, D. L., and Arbel, T. (2020). Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Med. Image Anal.* 59, 101557. doi: 10.1016/j.media.2019.101557

Pinaya, W. H., Tudosiu, P.-D., Gray, R., Rees, G., Nachev, P., Ourselin, S., et al. (2022). Unsupervised brain imaging 3d anomaly detection and segmentation with transformers. *Med. Image Anal.* 79, 102475. doi: 10.1016/j.media.2022.102475

Razzak, M. I., Naz, S., and Zaib, A. (2018). "Deep learning for medical image processing: overview, challenges and the future," in *Classification in BioApps*, 323–350.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Munich: Springer), 234–241.

Salem, M., Cabezas, M., Valverde, S., Pareto, D., Oliver, A., Salvi, J., et al. (2018). A supervised framework with intensity subtraction and deformation field features for the detection of new t2-w lesions in multiple sclerosis. *Neuroimage Clin.* 17, 607–615. doi: 10.1016/j.nicl.2017.11.015

Salem, M., Valverde, S., Cabezas, M., Pareto, D., Oliver, A., Salvi, J., et al. (2019). Multiple sclerosis lesion synthesis in mri using an encoder-decoder u-net. *IEEE Access* 7, 25171–25184. doi: 10.1109/ACCESS.2019.2900198

Salem, M., Valverde, S., Cabezas, M., Pareto, D., Oliver, A., Salvi, J., et al. (2020). A fully convolutional neural network for new t2-w lesion detection in multiple sclerosis. *Neuroimage Clin.* 25, 102149. doi: 10.1016/j.nicl.2019.102149

Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* 6, 1–48. doi: 10.1186/s40537-019-0197-0

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 600–612. doi: 10.1109/TIP.2003.819861

Wei, W., Poirion, E., Bodini, B., Durrleman, S., Colliot, O., Stankoff, B., et al. (2019). Fluid-attenuated inversion recovery mri synthesis from multisequence mri using three-dimensional fully convolutional networks for multiple sclerosis. *J. Med. Imaging* 6, 014005. doi: 10.1117/1.JMI.6.1.014005

Wolterink, J. M., Dinkla, A. M., Savenije, M. H., Seevinck, P. R., van den Berg, C. A., and Išgum, I. (2017). "Deep mr to ct synthesis using unpaired data," in *International Workshop on Simulation and Synthesis in Medical Imaging* (Quebec City, QC: Springer), 14–23.

Woolson, R. (2007). "Wilcoxon signed-rank test," in *Wiley Encyclopedia of Clinical Trials*, 1–3.

Xin, B., Hu, Y., Zheng, Y., and Liao, H. (2020). "Multi-modality generative adversarial networks with tumor consistency loss for brain mr image synthesis," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* (Iowa City, IA: IEEE), 1803–1807.

Yi, X., Walia, E., and Babyn, P. (2019). Generative adversarial network in medical imaging: a review. *Med. Image Anal.* 58, 101552. doi: 10.1016/j.media.2019.101552

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* 20, 45–57. doi: 10.1109/42.906424

Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Van Ginneken, B., Madabhushi, A., et al. (2021). A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc. IEEE* 109, 820–838. doi: 10.1109/JPROC.2021.3054390

Zhou, T., Fu, H., Chen, G., Shen, J., and Shao, L. (2020). Hi-net: hybrid-fusion network for multi-modal mr image synthesis. *IEEE Trans. Med. Imaging* 39, 2772–2781. doi: 10.1109/TMI.2020.2975344

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice: IEEE), 2223–2232.

frontiers | Frontiers in Neuroscience

TYPE Methods
PUBLISHED 21 October 2022
DOI 10.3389/fnins.2022.1007453

OPEN ACCESS

EDITED BY
Olivier Commowick,
Inria Rennes-Bretagne Atlantique Research Centre, France

REVIEWED BY
Erin Savner Beck,
Icahn School of Medicine at Mount Sinai, United States
Paul Eichinger,
Helios Klinikum München West, Germany

*CORRESPONDENCE
Berke Doga Basaran
bdb19@ic.ac.uk

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 30 July 2022
ACCEPTED 06 October 2022
PUBLISHED 21 October 2022

CITATION
Basaran BD, Matthews PM and Bai W (2022) New lesion segmentation for multiple sclerosis brain images with imaging and lesion-aware augmentation.
Front. Neurosci. 16:1007453.
doi: 10.3389/fnins.2022.1007453

New lesion segmentation for multiple sclerosis brain images with imaging and lesion-aware augmentation

Berke Doga Basaran[1,2]*, Paul M. Matthews[3,4] and Wenjia Bai[1,2,3]

[1]Department of Computing, Imperial College London, London, United Kingdom, [2]Data Science Institute, Imperial College London, London, United Kingdom, [3]Department of Brain Sciences, Imperial College London, London, United Kingdom, [4]UK Dementia Research Institute, Imperial College London, London, United Kingdom

Multiple sclerosis (MS) is an inflammatory and demyelinating neurological disease of the central nervous system. Image-based biomarkers, such as lesions defined on magnetic resonance imaging (MRI), play an important role in MS diagnosis and patient monitoring. The detection of newly formed lesions provides crucial information for assessing disease progression and treatment outcome. Here, we propose a deep learning-based pipeline for new MS lesion detection and segmentation, which is built upon the nnU-Net framework. In addition to conventional data augmentation, we employ imaging and lesion-aware data augmentation methods, axial subsampling and CarveMix, to generate diverse samples and improve segmentation performance. The proposed pipeline is evaluated on the MICCAI 2021 MS new lesion segmentation challenge (MSSEG-2) dataset. It achieves an average Dice score of 0.510 and F_1 score of 0.552 on cases with new lesions, and an average false positive lesion number n_{FP} of 0.036 and false positive lesion volume V_{FP} of 0.192 mm^3 on cases with no new lesions. Our method outperforms other participating methods in the challenge and several state-of-the-art network architectures.

KEYWORDS

multiple sclerosis, new lesion detection, data augmentation, nnU-Net, MRI, longitudinal lesion segmentation, biomedical segmentation

1. Introduction

Multiple sclerosis (MS) is a chronic inflammatory neurological disease affecting the central nervous system (CNS). Generally detected in young adults, ages 20–40, demyelinated lesions in the CNS lead to cognitive and physical disabilities, affecting vision, learning and memory, musculoskeletal system, and internal organ dysfunctions (Ghasemi et al., 2017). While MS is not fatal, average life expectancy is 5–10 years lower than average. The McDonald diagnostic criteria (Thompson et al., 2018) for MS provides guidelines for diagnosing the patient based on the number of lesions, lesion size, and locations of lesions in the brain and spinal cord. Disease progression for MS patients is highly varied and unpredictable, therefore, identifying disease trajectories and closely following them are important for prognosis and treatment decisions.

Multiple sclerosis is typically diagnosed *via* the patient showing symptoms in combination with supporting medical imaging of the brain. Specifically, the presence of lesions on brain MRI scans is a predictive image-based biomarker for MS diagnosis. Common multi-modal brain MRI acquisitions are composed of T1, T2, fluid-attenuated inversion recovery (FLAIR) and proton-density modalities. Lesions in the periventricular, juxtacortical, and infratentorial regions are presented as hyperintensities on T2-weighted and FLAIR MRI, or hypointensities on T1-weighted MRI (Filippi et al., 2019).

To monitor the progression of the disease, patients may take multiple MRI scans at different time points, typically 6–12 months apart. The detection of newly formed lesions provides crucial information for assessing disease activity and treatment outcome. Formation of new lesions correlates with the progression and severity of the disease and is often complemented with increased symptoms (Weiner et al., 2000). Manual assessment of these imaging scans can be time consuming, especially when attempting to identify formations of new lesions compared to the baseline scan. Automated detection and segmentation of brain lesions substantially aid neuro-radiologists in tracking the progression of the disease. Additionally, state-of-the-art machine learning methods can provide fast and reliable quantitative information on detected abnormalities, such as lesion load, lesion number, or even patient outcome (Tousignant et al., 2019; McKinley et al., 2020).

Recent developments in convolutional neural networks (CNNs) have shown promising results for image segmentation tasks (Alzubaidi et al., 2021). The two-dimensional (2D) U-Net (Ronneberger et al., 2015) and three-dimensional (3D) U-Net (Çiçek et al., 2016) architectures have been widely adopted in biomedical image segmentation tasks due to their ability in incorporating multi-scale spatial context and generalisability across different biomedical domains. nnU-Net (Isensee et al., 2021a), a U-Net based medical image segmentation network which employs a self-adapting framework, has shown excellent performance in a number of organ segmentation tasks (Isensee et al., 2021a,b). nnU-Net stands for “no new U-Net.” Its strong performance across a variety of datasets is not due to a new network architecture, but rather to automating the process of manual configuration of setting up a neural network. nnU-Net configures its network and pipeline subject to dataset properties and available GPU memory budget, maximizing the training patch size which the GPU memory will allow.

Nevertheless, there are still several challenges in applying these methods to brain image segmentation tasks, such as for MS lesion segmentation. The first challenge is the scarcity of data and annotation. Most of the public MS lesion datasets, such as the 2016 MSSEG (Commowick et al., 2018) and the 2015 ISBI MS (Carass et al., 2017) datasets, only contain images from a dozen of subjects. In a field where data diversity is paramount, data augmentation methods become critical tools to boost model performance. The second challenge is the class imbalance problem. In MS lesion segmentation, almost all of the foreground voxels represent healthy brain tissues and the lesions only constitute for a minority of the voxels. This means that the deep learning models tend to learn from the healthy tissues instead of the lesions of interest. In an attempt to allow the network to learn features from underrepresented classes, patches which contain the underrepresented class are often oversampled (Rahman and Davis, 2013). Despite oversampling strategies, the class imbalance problem is amplified even more when working with longitudinal MS data, where the objective is to detect new lesions. New lesions to detect in follow-up scans can make up as little as 0.01% of the 3D image volume.

There is still room for improvement for current lesion segmentation methods in detecting small lesions and tracking their temporal trajectories in disease progression. Commowick et al. (2018) finds that lesion detection rates fall significantly as lesion volumes decrease, resulting in false negative results in automated segmentation. This forms a critical challenge when newly formed lesions need to be considered for MS progression monitoring, which these lesions are often small and hard to detect.

In this paper, we propose a deep learning pipeline for new MS lesion segmentation. The developed pipeline is built upon the nnU-Net framework and we incorporate multiple brain-image preprocessing steps as well as imaging and lesion-aware data augmentation techniques. We evaluate the pipeline on the MSSEG-2 challenge dataset (Commowick et al., 2021a), which demonstrates promising results for both new lesion cases and no new lesion cases.

2. Methods

2.1. Related works

2.1.1. Deep learning for MS lesion segmentation

There have been contributions to machine learning methods specifically for MS lesion segmentation. Numerous methods were developed following the 2015 ISBI Longitudinal Multiple Sclerosis Lesion Segmentation Challenge (Carass et al., 2017). Valverde et al. (2017) employed a cascade of two 3D patch-wise CNNs, where the first CNN proposed candidate lesion voxels and the second one reduced falsely classified voxels. Birenbaum and Greenspan (2017) developed a multi-view longitudinal CNN and utilized priors about lesion intensities and spatial distribution to extract candidate lesions. Similar to Valverde et al., and Birenbaum et al. also used 3D patches for model training. Contrary to patch-based training, Aslani et al. (2019) proposed a multi-branch CNN which takes whole slices of the brain as input. Three 2D ResNets were separately trained for the axial, sagittal, and coronnal planes, the outputs of which were fused to generate a final 3D segmentation. Zhang et al. (2019) developed a fully convolutional densely connected network

(Tiramisu) using a 2.5-dimensional input where slices were stacked from three anatomical planes, providing both global and local context in segmentation.

Transformer networks are now a widely adopted network model for both natural language processing and computer vision tasks due to their self-attention mechanisms. The Vision Transformer (ViT) (Dosovitskiy et al., 2021) showcased that a pure transformer applied on sequences of image patches can achieve competitive image classification performance. Consequently, a multitude of transformer-based frameworks for medical image segmentation have been proposed. The majority of these models utilize CNNs in conjunction with transformers, taking advantage of both local and global context information extraction. TransBTS performs 3D CNN encoding followed by a transformer for global feature modeling in multi-modal brain tumor segmentation (Wang et al., 2021). TransUNeT employs a hybrid CNN-Transformer architecture for multi-organ abdominal image segmentation (Chen et al., 2021). UNETR implements a pure transformer encoder based on ViT in combination with resolution-wise convolutions and a deconvolutional layer for decoding the image back into the original dimension (Hatamizadeh et al., 2022). It performs competitively with state-of-the-art methods in multi-organ CT and MRI brain tumor segmentation tasks.

2.1.2. Data augmentation

Data augmentation can be classified into four categories: affine transformations, elastic transformations, intensity alterations, and incorporation of synthetic data. Affine transformations include flipping, rotation, scaling, and shearing of the image. Affine transformations do not drastically change the shape characteristics of the abnormal region with respect to its surrounding tissue. Elastic transformations generate a displacement grid with random displacements, which is used to deform individual voxels of the input image (Çiçek et al., 2016). The non-linear transformations alter the boundaries of the abnormal region with respect to its surrounding tissue, producing diverse samples. Intensity alterations introduce Gaussian noise, Gaussian blurring, sharpening, salt and pepper noise, and gamma augmentation etc. to improve model robustness against intensity distribution shift, which concerns imaging scans acquired from different scanner models, scanner acquisition parameters, or scanner strengths. Synthetic data augmentation utilizes generative models or MixUp (Zhang et al., 2018) techniques to generate new samples. For example, generative adversarial networks (GANs) (Goodfellow et al., 2014) were introduced for data augmentation in biomedical image segmentation (Shin et al., 2018; Sandfort et al., 2019; Hong et al., 2021). MixUp (Zhang et al., 2018) and related methods such as MixMatch (Berthelot et al., 2019) and CutMix (Yun et al., 2019) designed specific operations on two or more images to generate new samples. For brain lesion images, a lesion-aware augmentation method, CarveMix (Zhang et al., 2021), was proposed to combine two brain MRI scans to increase training data diversity. CarveMix randomly extracts lesion regions on the sagittal plane from one image and overlays them onto a target image (Zhang et al., 2021).

2.2. Proposed pipeline

The proposed pipeline consists of a brain image preprocessing step, followed by nnU-Net (Isensee et al., 2021a) for lesion segmentation, which is trained with imaging and lesion-aware data augmentation. An overview of the pipeline is presented in Figure 1.

2.2.1. Preprocessing

Skull is stripped using an atlas-based brain extraction tool (Doshi et al., 2013) followed by N4 bias field correction (Tustison et al., 2010). This is implemented using the MSSEG-2 longitudinal preprocessing script on Anima[1] provided by the challenge organizers. In addition, as the segmentation problem concerns imaging scans taken at different time points, we also perform intra-subject image registration so that scans of the same subject can be aligned and new lesions can be better differentiated. Since new lesions are defined on the follow-up scan, we register the baseline scan to the space of the follow-up scan. Affine image registration is performed, followed by free-form deformation, implemented using the MIRTK toolbox (MIRTK, 2021) using normalized mutual information as the loss function. Free-form deformation assists lesion segmentation in two ways: (1) brain structures, such as gyri, ventricles etc., are better registered; (2) lesions which slightly grow between scans are elastically registered so that the subsequent segmentation network can focus more on newly formed lesions.

2.2.2. Segmentation network

We adopt nnU-Net (Isensee et al., 2021a) as the segmentation network, with a two-channel input: preprocessed baseline scan and preprocessed follow-up scan. The output of the network is a binary 3D prediction of new lesions which have formed in the follow-up scan. The network consists of six resolution levels, formed from contracting and expanding paths. On the contracting path, each resolution level consists of two convolutional layers, each with a 3 × 3 × 3 convolution kernel, followed by instance normalization and LeakyReLU operation. At the start of each resolution level, the first convolution has a stride of (2,2,2), which effectively downsamples the feature map. At the lowest resolution level, the first convolution has a stride of (2,1,2).

[1] Anima scripts: RRID:SCR_017072, https://anima.irisa.fr.

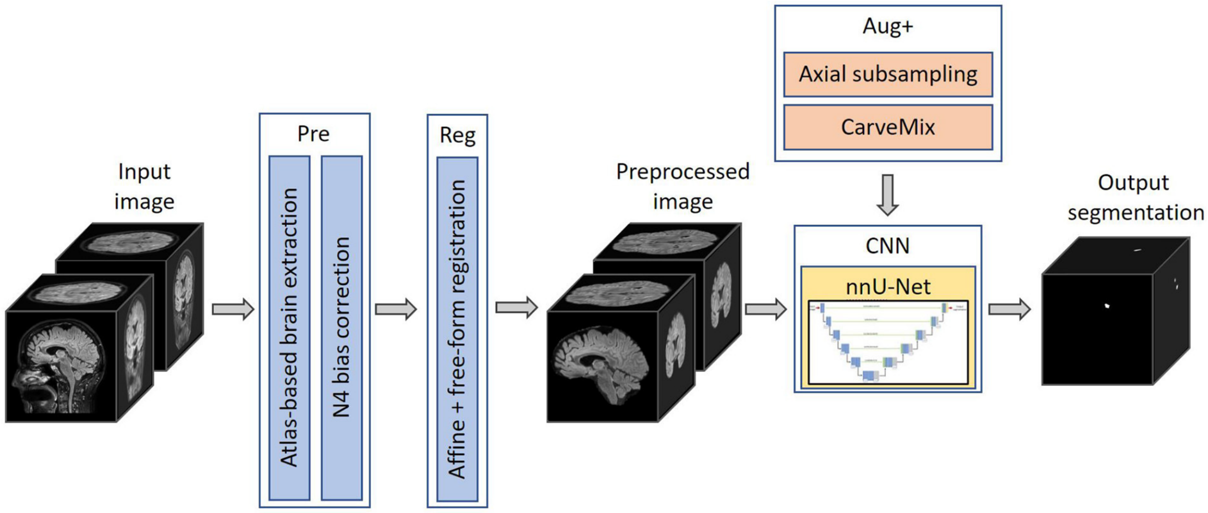


FIGURE 1
The proposed 3D framework for multiple sclerosis new lesion detection and segmentation. Framework comprises of brain extraction and N4 bias correction (Pre), intra-subject registration (Reg), imaging and lesion-aware data augmentations (Aug+), and nnU-Net. The pipeline takes input the baseline and follow-up FLAIR MRI scans, and outputs the proposed binary segmentation for new lesions.

On the expanding path, each resolution level consists of two convolutional layers with a 3×3×3 convolution kernel, followed by instance normalization and LeakyReLU operations, followed by an additional transposed 2 × 2 × 2 convolution operation. The transposed convolution has a stride of (2,1,2) at the lowest resolution level, and a stride of (2,2,2) at all other resolution levels. By utilizing skip connections, features extracted from the contracting path are concatenated with features at the expanding path at their respective resolution level. The network uses 32-dimensional features maps at the highest resolution layer, which is increased to 320 feature maps at the lowest resolution layer. Please refer to Figure 2 for a graphical representation of the architecture.

2.2.3. Hyperparameters and implementation details

We implement the 3D full-resolution U-Net model of nnU-Net, using the `3d_fullres` configuration, utilizing PyTorch. A single NVidia Tesla T4 GPU with 16GB RAM is used. Due to the GPU memory limit, 3D patches of size 128×112×160 are extracted from the original 3D images for model training. Patches are drawn randomly from the image with a 67% probability, and are ensured to include the lesion region with a 33% probability. The network is trained using a combination of Dice and cross-entropy loss, formulated as,

$$\mathcal{L} = -\frac{2}{|K|}\sum_{k\in K}\frac{\sum_{i\in I}\hat{y}_{i(k)}y_{i(k)}}{\sum_{i\in I}\hat{y}_{i(k)}+y_{i(k)}} - \sum_{i\in I}\sum_{k\in K} y_{i(k)}\log\hat{y}_{i(k)} \quad (1)$$

where k denotes the class, K denotes the number of classes ($K = 2$ in our method), i denotes a given voxel, I denotes the set of voxels over the image, $\hat{y}$ is the softmax output of the segmentation network, y is the one-hot encoding of the ground truth label for the new lesions, and subscript $i(k)$ is the number of voxels in the training patch for class k.

We use the stochastic gradient descent optimizer with Nesterov momentum of 0.99, an initial learning rate of 0.01, a polynomial learning rate decay and a batch size of 2 patches. When developing the model on the training data, five-fold cross-validation is used. Each model is trained for 1,000 epochs. After training, for each fold, we select the model which produces the highest Dice score. For inference, we ensemble segmentation outputs from the five models from each fold. No post-processing step is applied.

2.2.4. Imaging and lesion-aware data augmentation

Incorporation of data augmentation methods increases model generalizability and robustness, and decreases overfitting. We utilize multiple data augmentation techniques, including the default augmentations that nnU-Net provides in the batchgenerators data augmentation framework (Isensee et al., 2020). These augmentations include mirroring, rotating, scaling, channel translation to simulate registration errors, elastic deformations, linear downsampling, brightness and contrast augmentation, gamma augmentation, Gaussian and Rician noise augmentation, and random cropping.

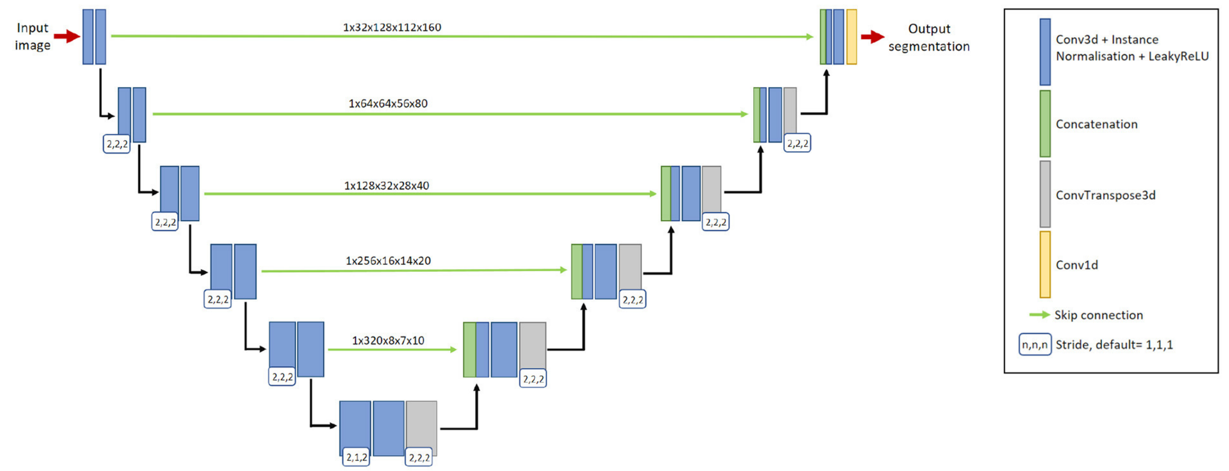


FIGURE 2
The 3D nnU-Net architecture for segmenting new lesions. It contains six resolution levels formed from contracting and expanding paths, and skip connections to recover fine-grain detail from the contracting path. The input to the network is a pair of baseline image and follow-up image. The output is the prediction of the new lesions.

In addition to these augmentations, inspired by Kamraoui et al. (2021, 2022), we introduce axial subsampling to simulate the image acquisition process on the axial plane. Brain MRI typically acquires a stack of 2D image slices in the axial plane to form a 3D volume, which can be of a high resolution within the axial plane but subject to low resolution across the plane (Chai et al., 2020). Axial subsampling augmentation is performed by applying a median filter of size $[1 \times 1 \times n]$ where $n \in 2, 3, 4$ to the axial image slices. This effectively blurs the image in the sagittal and coronal planes. Figure 3A illustrates an example of axial subsampling.

Finally, to increase the diversity of lesion images, a lesion-aware data augmentation method, CarveMix (Zhang et al., 2021), is used. CarveMix extracts a 3D region of interest (ROI) according to the lesion location and shape from one subject and mixes it with the brain image of another subject, thus creating augmented training samples. To increase diversity in augmentation, the lesion-aware ROI is generated by thresholding the distance transform of the lesion using a random threshold (Zhang et al., 2021). A synthetic image, **X**, and its label, **Y**, is generated by,

$$\mathbf{X} = \mathbf{X}_i \odot \mathbf{M}_i + \mathbf{X}_j \odot (1 - \mathbf{M}_i) \tag{2}$$

$$\mathbf{Y} = \mathbf{Y}_i \odot \mathbf{M}_i + \mathbf{Y}_j \odot (1 - \mathbf{M}_i) \tag{3}$$

where $\{\mathbf{X}_i, \mathbf{Y}_i\}$ denotes one pair of image and label, $\{\mathbf{X}_j, \mathbf{Y}_j\}$ denotes a second pair of image and label, $\mathbf{M}_i$ denotes the binary mask of the ROI, and $\odot$ represents voxel-wise multiplication. We randomly select two subjects for CarveMix augmentation when training. Incorporation of CarveMix data augmentation increases the total volume which the lesion class covers in an image, thus reducing the effect of class imbalance caused from the foreground class making up a small percentage of the overall image. Figure 3B illustrates an example of CarveMix augmentation.

3. Results

3.1. Data

We evaluate the pipeline on the *MICCAI 2021 MS new lesion segmentation challenge* dataset (MSSEG-2) (Commowick et al., 2021a), which provides 3D FLAIR images of 100 MS patients. The images were acquired from 15 different scanners, six of them 1.5T and nine of them 3T, including three GE scanners, six Philips scanners, and six Siemens scanners. Dataset scanner information can be found at the MICCAI 2021 MSSEG-2 challenge demographics data (Commowick et al., 2022). The images have varying image size and voxel spacing, which we resample to the median spacing of the dataset, $0.977 \times 0.977 \times 0.530mm^3$, before model training. Each patient was scanned twice, with 1–3 years between the two time points, constituting for a total of 200 images. Only new lesions at the second time point were annotated. Existing lesions, growing or shrinking lesions were not delineated. Each patient was annotated by four neuroradiologist and one consensus new lesion mask was provided. We use the consensus lesion masks as ground truth for model training and evaluation.

The dataset has been partitioned into 40 training and 60 test subjects by the challenge organizers. Of the 40 training subjects, 11 of them do not exhibit new lesions, which are referred to as "no-new lesion cases." We exclude these 11 no-new lesion cases

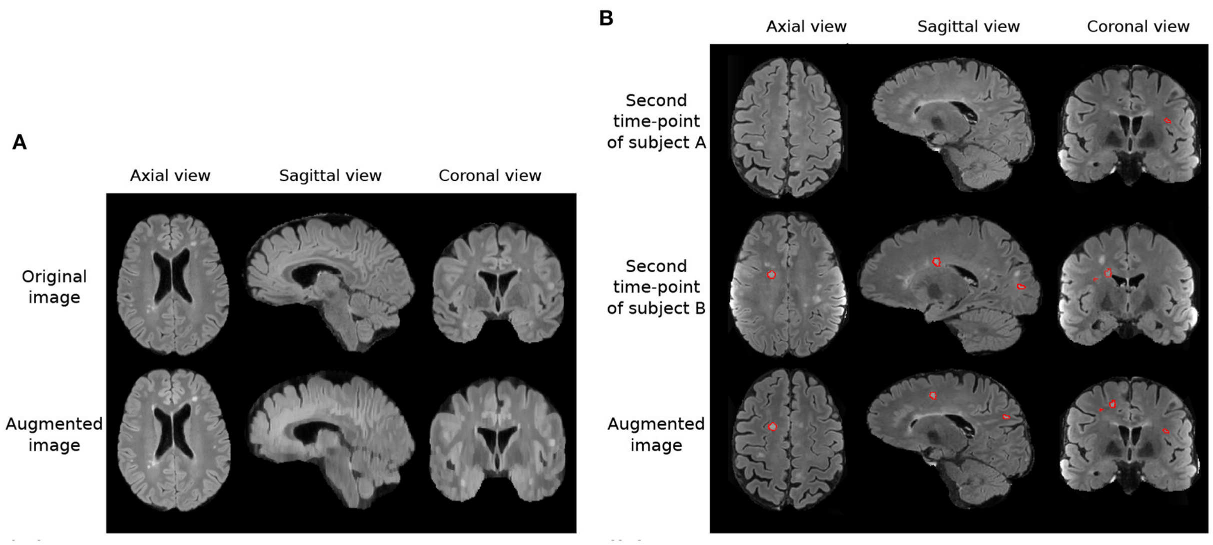


FIGURE 3
Imaging and lesion-aware data augmentations applied on the MSSEG-2 training set. **(A)** Example of axial subsampling ($n = 4$) to simulate the blurring in image acquisition. **(B)** Example of the CarveMix augmentation. Lesions from subject B are carved out and fused onto scan from subject A. Contours delineate lesion labels.

from the training set, utilizing the remaining 29 cases for model training. Of the 60 test subjects, 28 of them do not exhibit new lesions. We use all 60 subjects for testing.

3.2. Evaluation metrics

The method is evaluated using the Anima analyzer tool's `animaSegPerfAnalyzer`[2] function, provided by the MSSEG-2 challenge organizers in order to provide a fair comparison with other participating methods. In line with the MSSEG-2 evaluation, we use the default configuration of `animaSegPerfAnalyzer`, which excludes lesion volumes smaller than 3mm^3. The performance is evaluated separately for patients with new lesions and those with no-new lesions on the test set. For the new lesion cases, we report new lesion detection and segmentation performance, true positive lesion count, false positive lesion count, and false negative lesion count; for the no-new lesion cases, we calculate the number of new lesions detected (false positive lesions) and the volume of these false positive lesions.

3.2.1. Performance on new lesion cases

New lesion detection performance is evaluated using the F_1 score. The F_1 score measures how many lesions are correctly or incorrectly detected, regardless of the precision of its contours. It is formulated as,

$$F_1 = 2\frac{S_L \cdot P_L}{S_L + P_L}$$

where S_L denotes the lesion detection sensitivity (recall) and P_L denotes the positive predictive value (precision). The optimal F_1 score is 1. A lesion is considered as being detected or true positive if the automatic detection overlaps with at least 10% of the ground truth lesion volume and does not go outside by more than 70% of the volume (Commowick et al., 2018).

New lesion segmentation performance is evaluated using the Dice similarity coefficient, DSC, which measures spatial overlap. DSC is formulated as,

$$\text{DSC} = 2\frac{| A \cap G |}{| A | + | G |}$$

where A denotes the automatic segmentation and G denotes the ground truth. The optimal DSC is 1.

In addition to the metrics used in the MSSEG-2 challenge, we also present results for average true positive lesion count, n_{TP}, average false positive lesion count, n_{FP}, and average false negative lesion count, n_{FN}. Average true positive lesion count evaluates the average correctly detected lesions by the automated method. n_{FP} evaluates the average incorrectly detected lesions by the automated method. Finally, n_{FN} evaluates the lesions not detected by the automated method. These metrics are averaged over the 32 new lesion cases in the test set. The consensus ground truth segmentation contains a total of 224 new lesions, therefore the optimal average true positive lesion count, n_{TP}, is 7 ($\frac{224}{32}$). The optimal score for n_{FP} or n_{FN} is 0.

[2] Anima scripts: RRID:SCR_017072, https://anima.irisa.fr.

TABLE 1 Comparison of the proposed method to the challenge participating methods in terms of DSC, F_1 scores, the number of true positive lesions n_{TP}, the number of false positive lesions n_{FP}, the number of false negative lesions n_{FN}, and volume of false positive lesions V_{FP} (unit: mm^3), averaged across cases.

	New lesion cases (n = 32)					No-new lesion cases (n = 28)	
Method	**DSC**	F_1	n_{TP}	n_{FP}	n_{FN}	n_{FP}	V_{FP}
Expert 1	0.629	0.709	6.063	1.281	1.094	0.036	1.453
Expert 3	0.597	0.637	4.500	0.844	2.375	0.000	0.000
Expert 2	0.535	0.601	4.313	1.094	2.500	0.107	3.981
Expert 4	0.459	0.519	4.469	0.594	2.375	0.036	0.623
Proposed	**0.510**	**0.552**	4.969	2.031	2.281	**0.036**	**0.192**
MedICL	0.507	0.500	5.344	5.063††	1.875	0.536†††	12.713
LaBRI-IQDA	0.500	0.515	5.563	6.094††	1.656	1.143††	38.486*
SNAC	0.485	0.514	5.219	3.689†	2.031	0.321	5.726
LaBRI-D&E	0.472	0.496	5.500	9.156†††	1.750	1.964††	177.131
NVAUTO	0.469	0.464*	5.344	12.000†††	1.906	3.286†††	68.211*
LaBRI-Iw	0.453*	0.463*	5.000	6.719††	2.250	0.857†	27.761*
New Brain	0.451***	0.476***	4.032	2.903	3.355	0.786†	12.371
ITU	0.443	0.480	4.688	3.094	2.438	0.148	1.487
Mediaire-B	0.437**	0.541	**5.688**	4.469††	**1.500**	0.536†††	29.235*
Mediaire-A	0.432***	0.524	5.156	3.500	2.031	0.429†	15.908*
Empenn	0.424*	0.532	4.178	2.719	3.031	0.286††	4.258*
McEwan-IM	0.423***	0.453*	5.469	8.531†††	1.781	0.857†	16.504
PVG	0.414***	0.449*	4.032	2.903	3.355	0.107	1.031
Neuropoly-1	0.411***	0.425***	3.625	2.813	3.563	0.286†	8.615
IAMLAB	0.411***	0.412***	5.094	6.844†††	2.156	1.679†††	19.753*
LYLE	0.409***	0.443**	3.406	**1.250**	3.594	**0.036**	0.470
Neuropoly-2	0.409***	0.413***	3.656	1.906	3.469	0.107	0.498
SCAN	0.403***	0.431**	4.156	2.406	3.031	0.071	5.373
SCA-SimpleUNet	0.400***	0.448*	5.406	6.344†††	1.813	0.750†††	31.232*
I3M	0.398***	0.358***	4.250	4.313†	3.000	0.393	14.800
Neuropoly-3	0.379***	0.416***	3.719	2.625	3.500	0.321†	19.240
The NoCoDers	0.365***	0.381***	4.750	7.594†††	2.500	1.370†††	25.848*
Vicorob	0.357***	0.369***	3.906	4.094††	3.156	0.964†	88.402
HufsAIM	0.346***	0.407***	2.938	1.979	4.156	0.444†	17.128*
CMIC	0.330***	0.362***	3.906	6.094†††	3.344	4.714†††	123.442
MIAL	0.309***	0.332***	4.516	6.097†	2.774	1.464††	177.861
SCA-withPriors	0.223***	0.216***	2.750	6.719††	4.219	2.464†††	302.121
LIT	0.214***	0.242***	2.406	11.063	4.469	0.607†	35.404
IBBM	0.155***	0.145***	1.906††	7.625†††	5.188†	3.786†††	123.309***
Optimal score	1.000	1.000	7.000	0.000	0.000	0.000	0.000

The methods are sorted in the descending order of DSC. Best results are in bold. Asterisks indicate statistical significance ($^{*}p \leq 0.05$, $^{**}p \leq 0.01$, and $^{***}p \leq 0.005$) when using a paired Student's t-test compared to the proposed method. We implement the Mann-Whitney U-test for n_{TP}, n_{FP}, and n_{FN} metrics due to their non-normal distribution. The dagger symbols indicate statistical significance ($^{\dagger}p \leq 0.05$, $^{\dagger\dagger}p \leq 0.01$, and $^{\dagger\dagger\dagger}p \leq 0.005$) when using a Mann-Whitney U-test compared to the proposed method.

3.2.2. Performance on no-new lesion cases

For no-new lesion cases, the number and volume of falsely predicted lesions are evaluated. To count the number of false positive lesions, the Anima tool, `animaConnectedComponents`[3] function is used with default parameters. The volume of false positive lesions is

[3] Anima scripts: RRID:SCR_017072, https://anima.irisa.fr.

calculated by multiplying the number of lesion voxels by voxel spacing. We denote number of false positive lesions as n_{FP}, and volume of false positive lesions as V_{FP}. The optimal scores for both false positive lesion number and volume are 0.

3.3. Results

3.3.1. Comparison against participating methods in the challenge

The proposed pipeline is compared against MSSEG-2 participating methods and also four expert raters (Commowick et al., 2021b), reported in Table 1. The performance of MSSEG-2 participating methods and four expert raters is acquired from Commowick et al. (2021b). Table 1 shows that the proposed pipeline ranks competitively against methods submitted to the MSSEG-2 challenge. For the new lesion cases, it outperforms the other methods in terms of both the average DSC and the average F_1 scores. Also, our method outperforms three of the experts in n_{TP} and n_{FN} metrics. Our method correctly identifies 24 of the 32 new lesion cases as having new lesions. We achieve comparable performance to Experts 1, 2, 3 and 4, which correctly identify 26, 25, 27, and 22 of the 32 new lesion cases as having new lesions, respectively. A non-zero F_1 score is regarded as a method having correctly identified a new lesion case.

For the no-new lesion cases, the proposed pipeline achieves the lowest metrics for false positive lesions, including the average number n_{FP} and the average volume V_{FP}. It correctly identifies 27 out of 28 no-new lesion cases as subjects with no-new lesions. When comparing to expert raters, on the new lesion cases, the proposed pipeline outperforms Expert 4 in terms of DSC and F_1 scores and approaches the performance of Expert 2. On the no-new lesion cases, the proposed pipeline outperforms or achieves a comparable performance to Experts 1, 2 and 4.

3.3.2. Sensitivity vs. specificity analysis

Table 1 shows that there is a reverse correlation between the results for new lesion cases vs. no-new lesions cases, especially in the participating methods for the MSSEG-2 challenge. In Figure 4, we plot the average DSC and F_1 scores against the average n_{FP} and V_{FP} metrics. It shows that methods which perform well in new lesion cases do not perform as well in no-new lesion cases. Contrary to other methods, the proposed pipeline does not suffer from severe negative correlation, which performs well in both new lesion and no-new lesion cases.

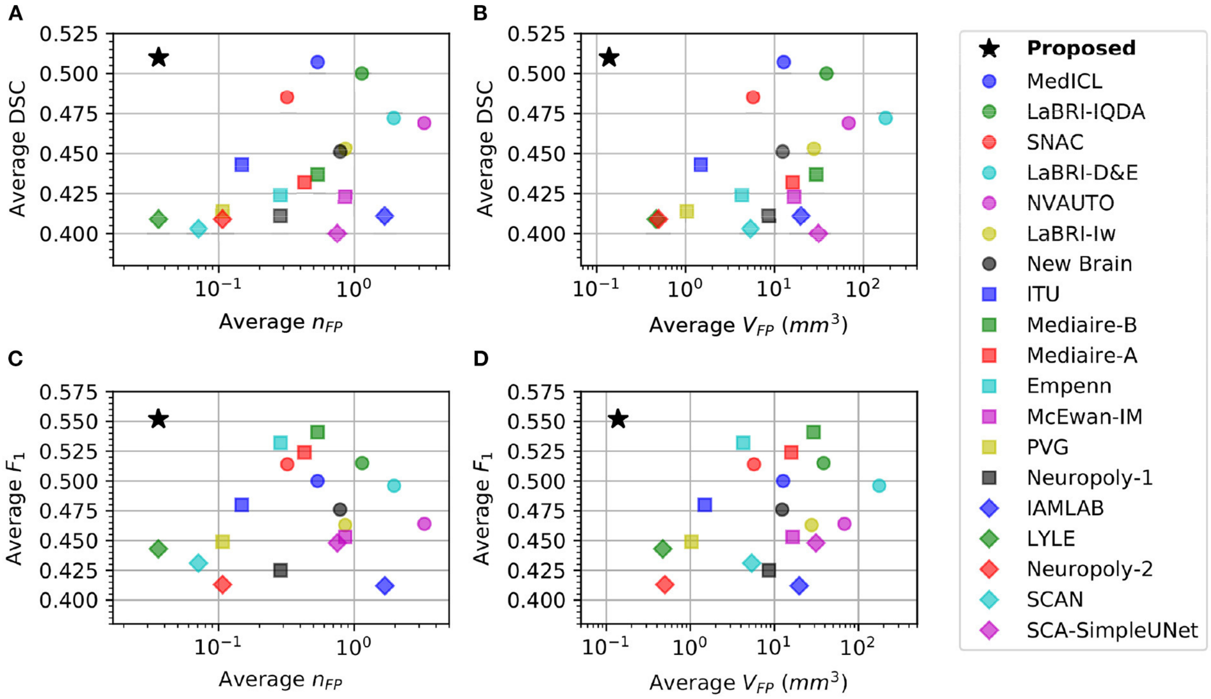


FIGURE 4
Comparison of different methods in new lesion metrics (DSC and F_1) vs. no-new lesion metrics (n_{FP} and V_{FP}). X-axis denotes one of the no-new lesion metrics in logarithmic scale and Y-axis denotes one of the new lesion metrics. Star denotes the proposed pipeline. **(A)** Plot of average DSC vs. average false positive lesion count n_{FP}. **(B)** Plot of average DSC vs. average false positive lesion volume V_{FP}. **(C)** Plot of average F_1 vs. average false positive lesion count n_{FP}. **(D)** Plot of average F_1 vs. average false positive lesion volume V_{FP}.

TABLE 2 Comparison of the proposed method to recent state-of-the-art deep learning architectures in terms of DSC and F1 scores, the number of false positive lesions n_{FP}, the number of true positive lesions n_{TP}, the number of false positive lesions n_{FP}, the number of false negative lesions n_{FN}, and volume of false positive lesions V_{FP} (unit: mm^3), averaged across cases.

	New lesion cases (n = 32)					No-new lesion cases (n = 28)	
Method	**DSC**	F_1	n_{TP}	n_{FP}	n_{FN}	n_{FP}	V_{FP}
Proposed	**0.510**	**0.552**	4.969	2.031	2.281	**0.036**	0.192
nnU-Net (Isensee et al., 2021a)	0.490	0.548	4.562	**1.281**	2.688	**0.036**	**0.138**
TransBTS (Wang et al., 2021)	0.477	0.470*	**5.492**	5.718††	**1.848**	0.939†	12.238
UNETR (Hatamizadeh et al., 2022)	0.462	0.468*	5.343	9.031†††	1.906	4.214†††	23.705*
TransUNet (Chen et al., 2021)	0.428**	0.434**	4.491	4.043††	2.102	1.081††	9.620
Tiramisu 2.5D (Zhang et al., 2019)	0.363***	0.365***	4.313	4.625††	2.938	1.384††	15.120
Optimal score	1.000	1.000	7.000	0.000	0.000	0.000	0.000

The methods are sorted in the descending order of DSC. Best results are in bold. Asterisks indicate statistical significance ($^{*}p \leq 0.05$, $^{**}p \leq 0.01$, and $^{***}p \leq 0.005$) when using a paired Student's t-test compared to our proposed method. We implement the Mann-Whitney U-test for n_{TP}, n_{FP}, and n_{FN} metrics due to their non-normal distribution. The dagger symbols indicate statistical significance ($^{\dagger}p \leq 0.05$, $^{\dagger\dagger}p \leq 0.01$, and $^{\dagger\dagger\dagger}p \leq 0.005$) when using a Mann-Whitney U-test compared to the proposed method.

3.3.3. Comparison against state-of-the-art architectures

We also compare the proposed pipeline to a number of state-of-the-art convolutional and transformer-based architectures, which have demonstrated excellent performance in biomedical image segmentation tasks. These architectures include the standard nnU-net (Isensee et al., 2021a), TransBTS (Wang et al., 2021), UNETR (Hatamizadeh et al., 2022), TransUNet (Chen et al., 2021), and Tiramisu 2.5D (Zhang et al., 2019). In order to evaluate methods fairly, we train these methods using the same preprocessed data, described in Section 2.2.1, which includes atlas-based brain extraction, N4 bias field correction, and free-form deformation registration, and use the standard data augmentation. The quantitative comparison results are reported in Table 2, and an example segmentation for visual comparison is provided in Figure 5. Table 2 shows that nnU-Net with standard data augmentations performs favorably against these state-of-the-art methods, and the proposed pipeline further improves performance possibly due to the additional data augmentation that we have introduced.

3.3.4. Ablation study

We carry out an ablation study to evaluate the impacts of different components of the pipeline, including brain extraction and N4 bias correction (Pre), affine and free-form image registration (Reg) and additional data augmentation methods including axial subsampling and CarveMix (Aug+). By default, standard data augmentation methods are used which come with nnU-Net, described in Section 2.2.4. The ablation study results are presented in Table 3.

Interestingly, adding pre-processing alone or registration alone does not drastically change performance metrics. However, when they are combined, for new lesion cases, the DSC score is increased from 0.476 to 0.490 and the F_1 score is increased from 0.533 to 0.548. When imaging-related and lesion-aware data augmentations (Aug+) are introduced, the DSC score is further increased to 0.510 and the F_1 score is increased to 0.552. This demonstrates that all the three components play an important role in the proposed pipeline. We also observe that when DSC and F_1 scores are increased, metrics concerning no-new lesion cases become poorer. The undesired increase in false positive lesion count and lesion volume is discussed in detail in Section 3.3.6.

3.3.5. Exclusion of no new lesion cases during training

The MSSEG-2 training dataset is composed of 40 subjects. 11 subjects do not exhibit new lesions in their follow-up scans. These subjects were removed from the training dataset, thus we only utilized 29 subjects. We carry out an additional study to investigate the impact of the exclusion of these images, by comparing the performance on the test set when utilizing

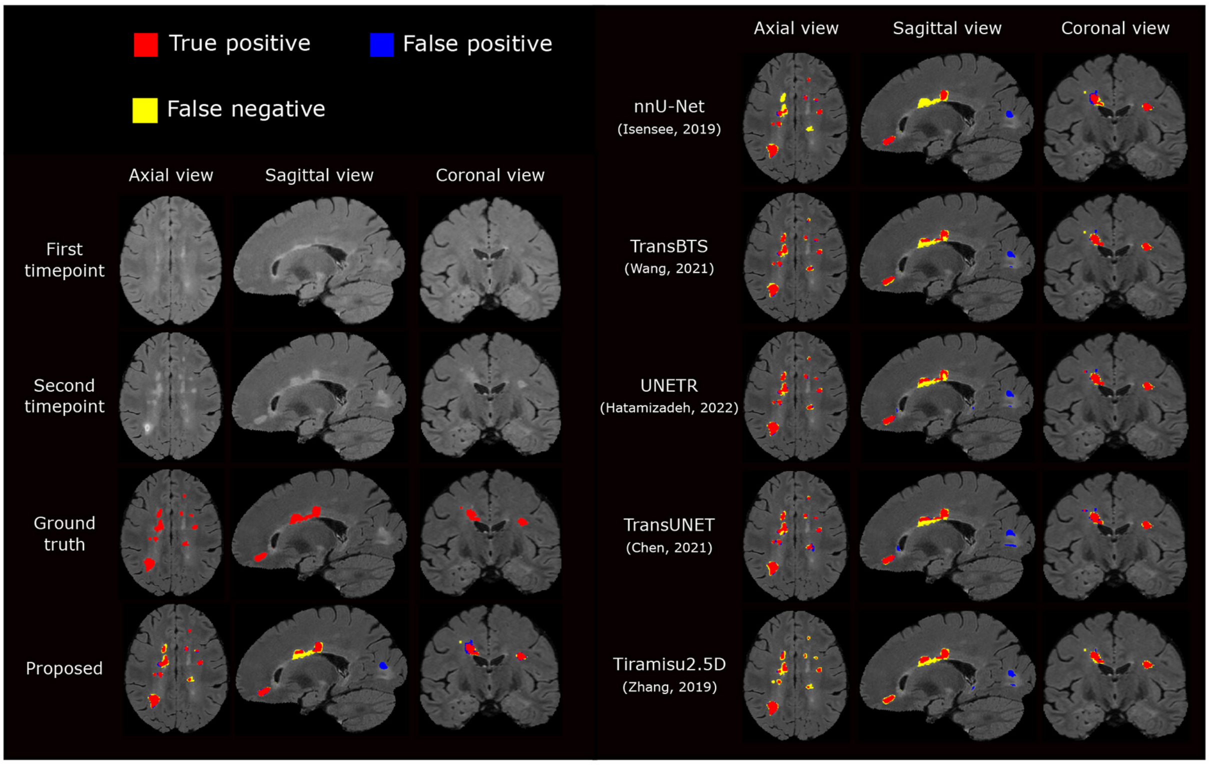


FIGURE 5
Visual comparison of the proposed segmentation pipeline to other methods. The proposed method produces a segmentation closest to the ground truth annotation.

TABLE 3 Results for the ablation study, presenting DSC and F_1 scores, the number of false positive lesions n_{FP}, the number of true positive lesions n_{TP}, the number of false positive lesions n_{FP}, the number of false negative lesions n_{FN}, and volume of false positive lesions n_{FP} (unit: mm^3), averaged across cases.

			New lesion cases (n = 32)					No-new lesion cases (n = 28)	
Pre	Reg	Aug+	DSC	F_1	n_{TP}	n_{FP}	n_{FN}	n_{FP}	V_{FP}
			0.476	0.533	4.250	**1.281**	3.000	**0.000**	**0.000**
✓			0.475	0.524	4.688	**1.281**	2.563	**0.000**	**0.000**
	✓		0.473	0.525	4.188	1.343	3.062	0.036	0.083
✓	✓		0.490	0.548	4.562	**1.281**	2.688	0.036	0.138
✓	✓	✓	**0.510**	**0.552**	**4.969**	2.031	**2.281**	0.036	0.192

Best results are in bold.

all 40 subjects for segmentation model training against using the 29 subjects with new lesions. Results are presented in Table 4. Interestingly, removing the no new lesion subjects result in slightly higher DSC and F_1 score, without compromising performance in the no new lesion cases. This is likely due to higher average representation of the foreground class (new lesions) in the altered training set. In addition, reducing the training set from 40 to 29 subjects decreased the model training time by 27.5%.

3.3.6. Sources of failure

We carry out a qualitative investigation on the test set to better understand where our method fails. In the no-new lesion cases, the proposed pipeline correctly classifies 27 out of the 28 subjects. The one misclassified (subject ID: 004) is incorrectly segmented to have 1 new lesion, which amount to 14 false positive voxels (3.875 mm^3), shown in Figure 6. The segmented region has a higher intensity compared to surrounding regions and we suspect that it is likely to be a new lesion. Two of four

TABLE 4 Comparison between using the complete MSSEG-2 training set (40 subjects) against using 29 subjects which excludes the no new lesion cases.

	New lesion cases ($n = 32$)		No-new lesion cases ($n = 28$)	
Training set	**DSC**	F_1	n_{FP}	V_{FP}
29 subjects with new lesions	**0.510**	**0.552**	**0.036**	**0.192**
All 40 subjects	0.502	0.530	**0.036**	**0.192**

We present the DSC and F_1 scores, the number of false positive lesions n_{FP} and their volume V_{FP} (unit: mm^3), averaged across cases. Best results are in bold.

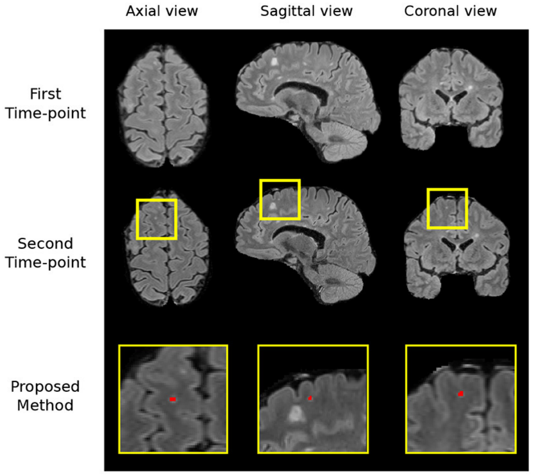


FIGURE 6
The region which we incorrectly classify as a lesion in the no-new lesion cases in the MSSEG-2 test set (subject ID: 004). We suspect the segmented region to be a new lesion, two of the human raters also classify this region as a new lesion.

expert raters also delineate this region as a new lesion, although the consensus segmentation does not regards this as a lesion, which leads to the misclassification of our method. This test set subject is the cause of the undesired increase in false positive lesion count and false positive lesion volume in Table 3.

In the new lesion cases, when assessing against the DSC and F_1 score, there is still room for improvement in performance. There are possibly three sources of failure that affect the DSC and F_1 scores. The first is the incorrect segmentation of growing lesions. The pipeline employs affine and non-rigid registration to align the baseline scan to the follow-up and thus suppresses the detection of growing lesions. However, the remaining mis-alignment for some growing lesions still leads to the boundary voxels, i.e., the grown regions of lesions, being incorrectly segmented as new lesions. Secondly, the proposed pipeline may miss some tiny and less apparent new lesions. In some cases, new lesions which form in the follow-up scan are very small and less hyperintense compared to large new lesions. This makes the detection of these lesions very difficult and leads to misclassifications. Finally, new lesion segmentation is a generally challenging task even for human raters and there are indiscrepancies between annotations from different human experts. The noise in the annotations may limit what an automated method can achieve (Zhang et al., 2020). We present examples of all three sources of failure in Figure 7.

4. Discussion and conclusion

Here we demonstrate that by incorporating appropriate preprocessing steps, an nnU-net segmentation network, imaging and lesion-aware data augmentation techniques, we can achieve promising performance in new MS lesion segmentation tasks. The proposed pipeline outperforms other challenge participating methods in both new lesion cases and no-new lesion cases, in terms of DSC, F_1, n_{FP} and V_{FP} scores. We also observe that in terms of network architecture, the recently popular transformer architectures may not necessarily outperform convolutional neural network architectures, such as nnU-net (Table 2). The design of proper pre-processing steps and problem-specific augmentations may play a more important role in this particular lesion segmentation task (Table 3).

In addition to the DSC and F_1 score used by the MSSEG-2 challenge, we introduce extra evaluation metrics, n_{TP}, n_{FP}, and n_{FN}, for the new lesion cases to understand the method performance. While many methods have a high n_{TP} and a lower n_{FN} score, the results suggest that a lower n_{FP} is what differentiates our method and the Experts to the other methods, thus providing a higher DSC and F_1 score. The n_{TP}, n_{FP}, and n_{FN} results also suggest that they should be analyzed with respect to each other, as evaluating a method solely with one of these metrics can be misleading. For example, the top performing method in correctly identified average true positive lesions, n_{TP}, ranks 10th in DSC score, and the top performing method in fewest average false positive lesion count ranks 17th in both DSC and F_1 score. Methods with higher n_{TP} score also have high n_{FP} scores, with respect to Experts' performance. The results on the new metrics show that methods differ on their approach to achieve optimal DSC and F_1 scores, and suggest that extra thought should be considered when evaluating a method solely on one metric.

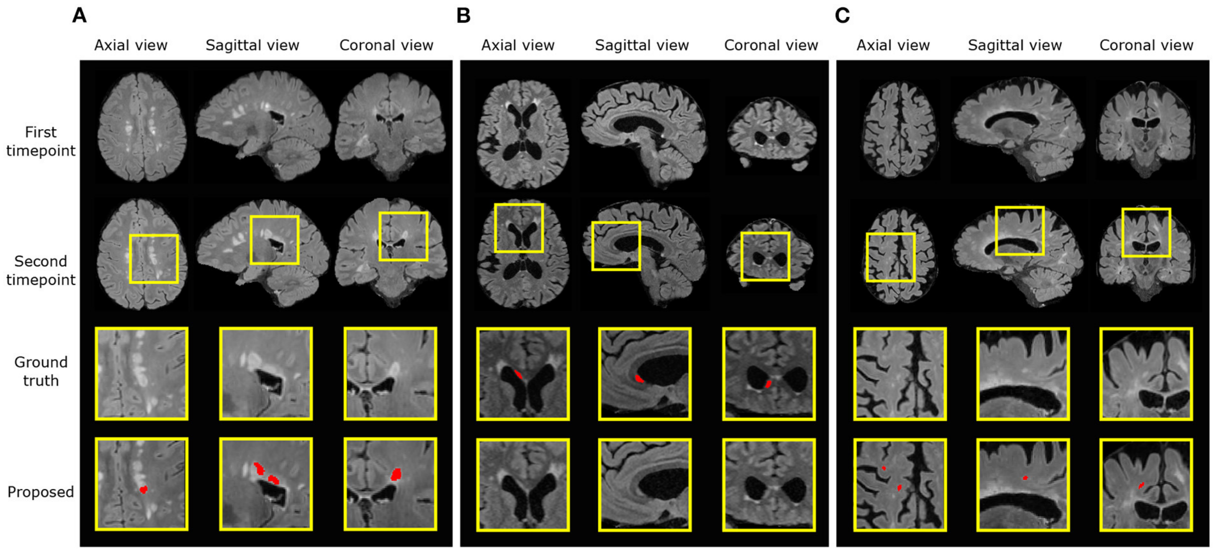


FIGURE 7
Examples of three different sources of error. Ground truth and predicted lesions are delineated in red. **(A)** (subject ID: 012) False positive segmentation of a growing lesion. **(B)** (subject ID: 078) False negative classification of a new lesion. **(C)** (subject ID: 036) False positive segmentation of a region classified as healthy/not-new in the consensus label, but annotated as a new lesion in two of the four provided expert annotations.

Future efforts to improve the proposed method include further investigation of the sources of failure described in Section 3.3.6 and bridging the gap between automatic segmentation and expert raters. The current MSSEG-2 challenge dataset only contains annotations of new lesions. To discriminate new lesions from growing lesions, future works may include curating a dataset of both lesion types and training automated methods for detecting and differentiating these lesions. Also, additional post-processing steps could be developed to inspect local neighborhoods of detected new lesions and check whether they are connected to existing lesions or not, thus decreasing false positives for new lesion detection. However, too large of a local context may come at the cost of decreasing true positives too. Furthermore, the proposed pipeline only focuses on lesions in the brain region and the pre-processing step removes the spinal cord region. Despite the MSSEG-2 testing dataset not featuring any new lesions in the spinal cord, MS lesions can form in this region. Inclusion of the spinal cord into the preprocessing step and training data will extend the application of the proposed pipeline.

In conclusion, we propose an nnU-Net-based pipeline for multiple sclerosis new lesion segmentation. A contribution of the pipeline is that it incorporates task-specific data augmentation methods, including axial subsampling, which simulates MRI acquisition-based image artifacts, and CarveMix, which increases the diversity of lesion images. When evaluating on the MSSEG-2 dataset, the proposed pipeline achieves excellent performance in evaluation metrics for both new lesion and no-new lesion cases.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

BB conducted all experiments and drafted the manuscript. WB provided significant assistance in technical issues and writing. PM provided guidance for the direction of research. All authors approved the final manuscript.

Funding

This work is supported by the UKRI CDT in AI for Healthcare http://ai4health.io (Grant No. EP/S023283/1).

Acknowledgments

The authors would like to acknowledge France Life Imaging (FLI), the National Institute for Research in Digital Science and Technology (Inria), and thank all individuals involved in organizing the MSSEG-2 challenge and preparing the MSSEG-2 dataset.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-dujaili, A., Duan, Y., Al-Shamma, O., et al. (2021). Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *J. Big Data* 8, 53. doi: 10.1186/s40537-021-00444-8

Aslani, S., Dayan, M., Storelli, L., Filippi, M., Murino, V., Rocca, M. A., et al. (2019). Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. *Neuroimage* 196, 1–15. doi: 10.1016/j.neuroimage.2019.03.068

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. (2019). Mixmatch: a holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*. doi: 10.48550/arXiv.1905.02249

Birenbaum, A., and Greenspan, H. (2017). Multi-view longitudinal CNN for multiple sclerosis lesion segmentation. *Eng. Appl. Artif. Intell.* 65, 111–118. doi: 10.1016/j.engappai.2017.06.006

Carass, A., Roy, S., Jog, A., Cuzzocreo, J. L., Magrath, E., Gherman, A., et al. (2017). Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *Neuroimage* 148, 77–102. doi: 10.1016/j.neuroimage.2016.12.064

Chai, Y., Xu, B., Zhang, K., Lepore, N., and Wood, J. C. (2020). Mri restoration using edge-guided adversarial learning. *IEEE Access* 8, 83858–83870. doi: 10.1109/ACCESS.2020.2992204

Chen, J., Lu, Y., Yu, Q., Luo, X., and Adeli, E. (2021). "Transunet: transformers make strong encoders for medical image segmentation," in *ICML 2021 Interpretable Machine Learning in Healthcare Workshop*.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3D U-net: learning dense volumetric segmentation from sparse annotation," in *MICCAI* (Athens).

Commowick, O., Cervenansky, F., Cotton, F., and Dojat, M. (2021a). "Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure," in *MICCAI 2021 MSSEG-2 Challenge Proceedings* (Strasbourg), 126.

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., et al. (2018). Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Sci. Rep.* 8, 13650. doi: 10.1038/s41598-018-31911-7

Commowick, O., Masson, A., Combes, B., Camarasu-Pop, S., Cervenansky, F., Kain, M., et al. (2021b). *MICCAI 2021 MSSEG-2 challenge quantitative results* [Data set]. Zenodo. doi: 10.5281/zenodo.5775523

Commowick, O., Masson, A., Combes, B., Camarasu-Pop, S., Cervenansky, F., Kain, M., et al. (2022). *MICCAI 2021 MSSEG-2 challenge demographics data* [Data set]. Zenodo. doi: 10.5281/zenodo.5824568

Doshi, J., Erus, G., Ou, Y., Gaonkar, B., and Davatzikos, C. (2013). Multi-atlas skull-stripping. *Acad. Radiol.* 20, 1566–1576. doi: 10.1016/j.acra.2013.09.010

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., et al. (2021). "An image is worth 16x16 words: transformers for image recognition at scale," in *International Conference on Learning Representations*.

Filippi, M., Preziosa, P., Banwell, B. L., Barkhof, F., Ciccarelli, O., De Stefano, N., et al. (2019). Assessment of lesions on magnetic resonance imaging in multiple sclerosis: practical guidelines. *Brain* 142, 1858–1875. doi: 10.1093/brain/awz144

Ghasemi, N., Razavi, S., and Nikzad, E. (2017). Multiple sclerosis: pathogenesis, symptoms, diagnoses and cell-based therapy citation: Ghasemi N, Razavi Sh, Nikzad E. Multiple sclerosis: pathogenesis, symptoms, diagnoses and cell-based therapy. *Cell J.* 19, 1–10. doi: 10.22074/cellj.2016.4867

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., et al. (2014). "Generative adversarial nets," in *Neural Information Processing Systems, Vol. 27*.

Hatamizadeh, A., Yang, D., Roth, H. R., and Xu, D. (2022). "Unetr: transformers for 3d medical image segmentation," in *Winter Conference on Applications of Computer Vision (WACV)* (Hawaii, HI), 1748–1758.

Hong, S., Marinescu, R., Dalca, A. V., Bonkhoff, A. K., Bretzner, M., Rost, N. S., et al. (2021). "3D-StyleGAN: a style-based generative adversarial network for generative modeling of three-dimensional medical images," in *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections*, eds S. Engelhardt, I. Oksuz, D. Zhu, Y. Yuan, A. Mukhopadhyay, N. Heller, S. X. Huang, H. Nguyen, R. Sznitman, and Y. Xue (Cham: Springer International Publishing), 24–34.

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., and Maier-Hein, K. H. (2021a). nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* 18, 203–211. doi: 10.1038/s41592-020-01008-z

Isensee, F., Jager, P., Wasserthal, J., Zimmerer, D., Petersen, J., Kohl, S., et al. (2020). *batchgenerators - a python framework for data augmentation (0.19.6).* Zenodo. doi: 10.5281/zenodo.3632567

Isensee, F., Jäger, P. F., Full, P. M., Vollmuth, P., and Maier-Hein, K. H. (2021b). "nnu-net for brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi and S. Bakas (Cham: Springer International Publishing), 118–132.

Kamraoui, R. A., Ta, V.-T., Manjon, J. V., and Pierrick, C. (2021). "Image quality data augmentation for new MS lesion segmentation," in *MICCAI 2021 MSSEG-2 Challenge Proceedings* (Strasbourg), 37.

Kamraoui, R. A., Ta, V.-T., Tourdias, T., Mansencal, B., Manjon, J. V., and Coupé, P. (2022). DeepLesionBrain: towards a broader deep-learning generalization for multiple sclerosis lesion segmentation. *Med. Image Anal.* 76, 102312. doi: 10.1016/j.media.2021.102312

McKinley, R., Wepfer, R., Grunder, L., Aschwanden, F., Fischer, T., et al. (2020). Automatic detection of lesion load change in Multiple Sclerosis using convolutional neural networks with segmentation confidence. *Neuroimage Clin.* 25, 102104. doi: 10.1016/j.nicl.2019.102104

MIRTK (2021). *Medical Image Registration ToolKit (MIRTK)*. Available online at: https://mirtk.github.io/.

Rahman, M. M., and Davis, D. N. (2013). Addressing the class imbalance problem in medical datasets. *Int. J. Mach. Learn. Comput.* 3, 224–228. doi: 10.7763/IJMLC.2013.V3.307

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Munich), 234–241.

Sandfort, V., Yan, K., Yan, K., Pickhardt, P. J., and Summers, R. M. (2019). Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. *Sci. Rep.* 9, 16884. doi: 10.1038/s41598-019-52737-x

Shin, H.-C., Tenenholtz, N. A., Rogers, J. K., Schwarz, C. G., Senjem, M. L., Gunter, J. L., et al. (2018). "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *Simulation and Synthesis in Medical Imaging*, eds A. Gooya, O. Goksel, I. Oguz, and N. Burgos (Cham: Springer International Publishing), 1–11.

Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., et al. (2018). Diagnosis of multiple sclerosis: 2017 revisions of the

McDonald criteria. *Lancet Neurol.* 17, 162–173. doi: 10.1016/S1474-4422(17)30470-2

Tousignant, A., Lemaître, P., Precup, D., Arnold, D. L., and Arbel, T. (2019). "Prediction of disease progression in multiple sclerosis patients using deep learning analysis of mri data," in *International Conference on Medical Imaging with Deep Learning, volume 102 of Proceedings of Machine Learning Research* (London), 483–492.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., et al. (2010). N4itk: improved n3 bias correction. *IEEE Trans. Med. Imaging* 29, 1310–1320. doi: 10.1109/TMI.2010.2046908

Valverde, S., Cabezas, M., Roura, E., González-Villà, S., Pareto, D., Vilanova, J. C., et al. (2017). Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *Neuroimage* 155, 159–168. doi: 10.1016/j.neuroimage.2017.04.034

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., and Li, J. (2021). "TransBTS: multimodal brain tumor segmentation using transformer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Strasbourg), 109–119.

Weiner, H. L., Guttmann, C. R., Khoury, S. J., Orav, E. J., Hohol, M. J., Kikinis, R., et al. (2000). Serial magnetic resonance imaging in multiple sclerosis: correlation with attacks, disability, and disease stage. *J. Neuroimmunol.* 104, 164–173. doi: 10.1016/S0165-5728(99)00273-8

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. (2019). "Cutmix: regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul: IEEE), 6023–6032.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2018). "MixUp: beyond empirical risk minimization," in *International Conference on Learning Representations* (Vancouver, BC).

Zhang, H., Valcarcel, A. M., Bakshi, R., Chu, R., Bagnato, F., Shinohara, R. T., et al. (2019). "Multiple sclerosis lesion segmentation with tiramisu and 2.5 d stacked slices," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Shenzhen), 338–346.

Zhang, L., Tanno, R., Xu, M.-C., Jin, C., and Jacob, J. (2020). "Disentangling human error from ground truth in segmentation of medical images," in *Neural Information Processing Systems, Vol.* 33, 15750–15762.

Zhang, X., Liu, C., Ou, N., Zeng, X., Xiong, X., Yu, Y., et al. (2021). "CarveMix: a simple data augmentation method for brain lesion segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Strasbourg), 196–205.

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 26 October 2022
DOI 10.3389/fnins.2022.975862

New multiple sclerosis lesion segmentation and detection using pre-activation U-Net

Pooya Ashtari[1,2]*, Berardino Barile[1,2], Sabine Van Huffel[1] and Dominique Sappey-Marinier[2]

[1]Department of Electrical Engineering (ESAT), STADIUS Centre for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven, Belgium, [2]CREATIS (UMR 5220 CNRS – U1294 INSERM), Université Claude Bernard Lyon 1, Université de Lyon, Villeurbanne, France

OPEN ACCESS

EDITED BY
Michel Dojat,
Institut National de la Santé et de la Recherche Médicale (INSERM), France

REVIEWED BY
Shen Zhao,
Sun Yat-sen University, China
Cédric Meurée,
Inria Rennes—Bretagne Atlantique Research Centre, France

*CORRESPONDENCE
Pooya Ashtari
pooya.ashtari@esat.kuleuven.be

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 22 June 2022
ACCEPTED 14 September 2022
PUBLISHED 26 October 2022

CITATION
Ashtari P, Barile B, Van Huffel S and Sappey-Marinier D (2022) New multiple sclerosis lesion segmentation and detection using pre-activation U-Net. *Front. Neurosci.* 16:975862. doi: 10.3389/fnins.2022.975862

Automated segmentation of new multiple sclerosis (MS) lesions in 3D MRI data is an essential prerequisite for monitoring and quantifying MS progression. Manual delineation of such lesions is time-consuming and expensive, especially because raters need to deal with 3D images and several modalities. In this paper, we propose Pre-U-Net, a 3D encoder-decoder architecture with pre-activation residual blocks, for the segmentation and detection of new MS lesions. Due to the limited training set and the class imbalance problem, we apply intensive data augmentation and use deep supervision to train our models effectively. Following the same U-shaped architecture but different blocks, Pre-U-Net outperforms U-Net and Res-U-Net on the MSSEG-2 dataset, achieving a Dice score of 40.3% on new lesion segmentation and an F_1 score of 48.1% on new lesion detection. The codes and trained models are publicly available at https://github.com/pashtari/xunet.

KEYWORDS

multiple sclerosis, new lesions, segmentation, U-Net, pre-activation

1. Introduction

Multiple sclerosis (MS) is a common chronic, autoimmune demyelinating disease of the central nervous system (CNS), which causes inflammatory lesions in the brain, particularly in white matter (WM). Multi-parametric MRI is widely used to diagnose and assess MS lesions in clinical practice. Particularly, FLuid Attenuated Inversion Recovery (FLAIR) images provide high contrast for white matter lesions appearing as high-intensity regions. It is highly relevant to monitor lesion activities, especially the appearance of new lesions and the enlargement of existing lesions, for several purposes, including prognosis and follow-up. More specifically, lesional changes between two longitudinal MRI scans from an MS patient are the most important markers for tracking disease progression and inflammatory changes. To this end, the accurate segmentation of new lesions is an essential prerequisite to quantifying lesional changes and measuring features, such as new lesion volumes and locations. However, manual delineation of such lesions is tedious, time-consuming, and expensive, especially because experts need to deal with 3D images and several modalities; therefore, accurate computer-assisted methods are needed to automatically perform this task.

Longitudinal MS lesion segmentation, however, remains very challenging since MS images often change subtly over time within a patient, and new lesions can be very small although they vary dramatically in shape, structure, and location across patients. The MSSEG-2 MICCAI 2021 challenge (Confavreux et al., 1992; Vukusic et al., 2020) aims to develop effective data-driven algorithms for the segmentation of new MS lesions by providing a dataset of 40 pairs of 3D FLAIR images acquired at two different time points (with varying intervals) and registered in the intermediate space between the two time points. For each pair, new lesions are manually annotated by multiple raters, and the consensus ground truths are obtained through a voxel-wise majority voting (see Figure 1).

Over the past decade, convolution neural networks (CNNs) with an encoder-decoder architecture, known as U-Net (Ronneberger et al., 2015), have dominated medical image segmentation. In contrast to a hand-crafted approach, U-Net can automatically learn high-level task-specific features for MS lesion segmentation. This work extends our previous effort (Ashtari et al., 2021a) in the MSSEG-2 and proposes Pre-U-Net, a 3D U-Net architecture with pre-activation residual blocks (He et al., 2016a,b), for segmenting new MS lesions. We use deep supervision (Lee et al., 2015) and perform intensive data augmentation to effectively train our models. In contrast to the existing methods, our models directly segment new MS lesions on longitudinal 3D FLAIR images in an end-to-end fashion in contrast to the common two-step approach, where cross-sectional segmentation is first performed individually for each time point, and new lesions are then extracted by comparing the longitudinal segmentation maps and applying further post-processing. Depending on the metric used, the MSSEG-2 challenge has four leaderboards. Our Pre-U-Net model achieved competitive scores, and our team, LYLE, was ranked first in two of the leaderboards among 30 participating teams in the challenge.

The rest of this paper is organized as follows: Section 2 briefly reviews relevant semantic segmentation techniques. Section 3 presents our approach to longitudinal MS lesion segmentation. Experiments are presented in Section 5. We conclude this paper in Section 6.

2. Related work

Over the past few years, considerable efforts have been made in the development of fully convolutional neural networks for semantic segmentation. Encoder-decoder architectures, in particular U-Net (Ronneberger et al., 2015) and its variants,

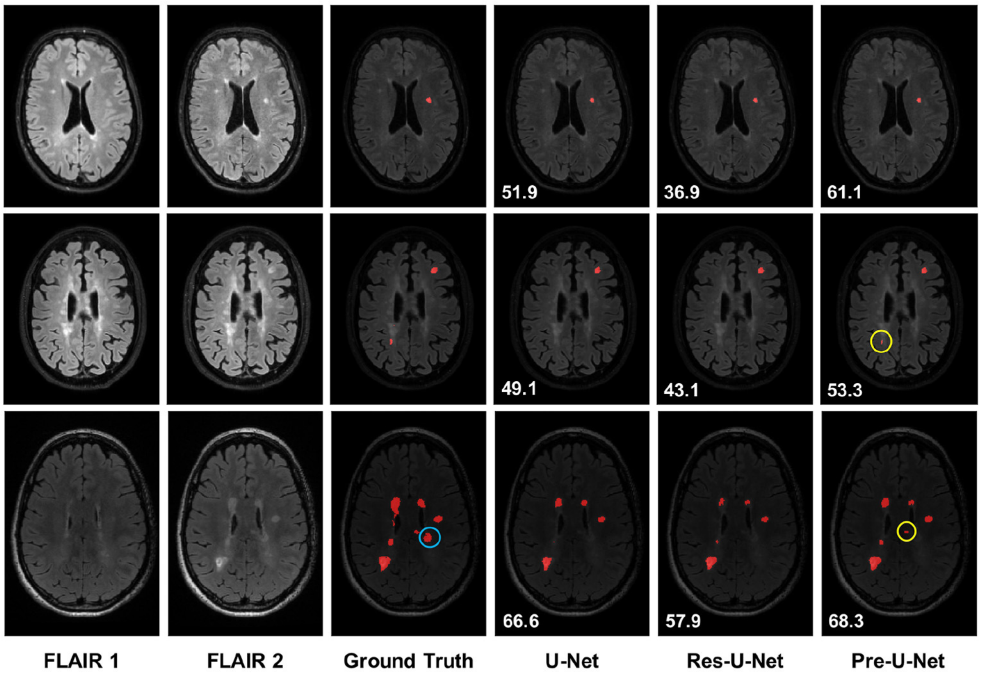


FIGURE 1
Qualitative results on new MS lesion segmentation. The three examples are from three different patients in the test set. The new lesions are shown in red in the segmentation maps. The new lesions circled in yellow (rows 2-3 and column 6) are successfully detected only by Pre-U-Net, while the new lesion circled in blue (row 3 and column 3) is not captured by any of the models, representing a very difficult case. The patient-wise Dice score for each example is displayed on the segmentation map.

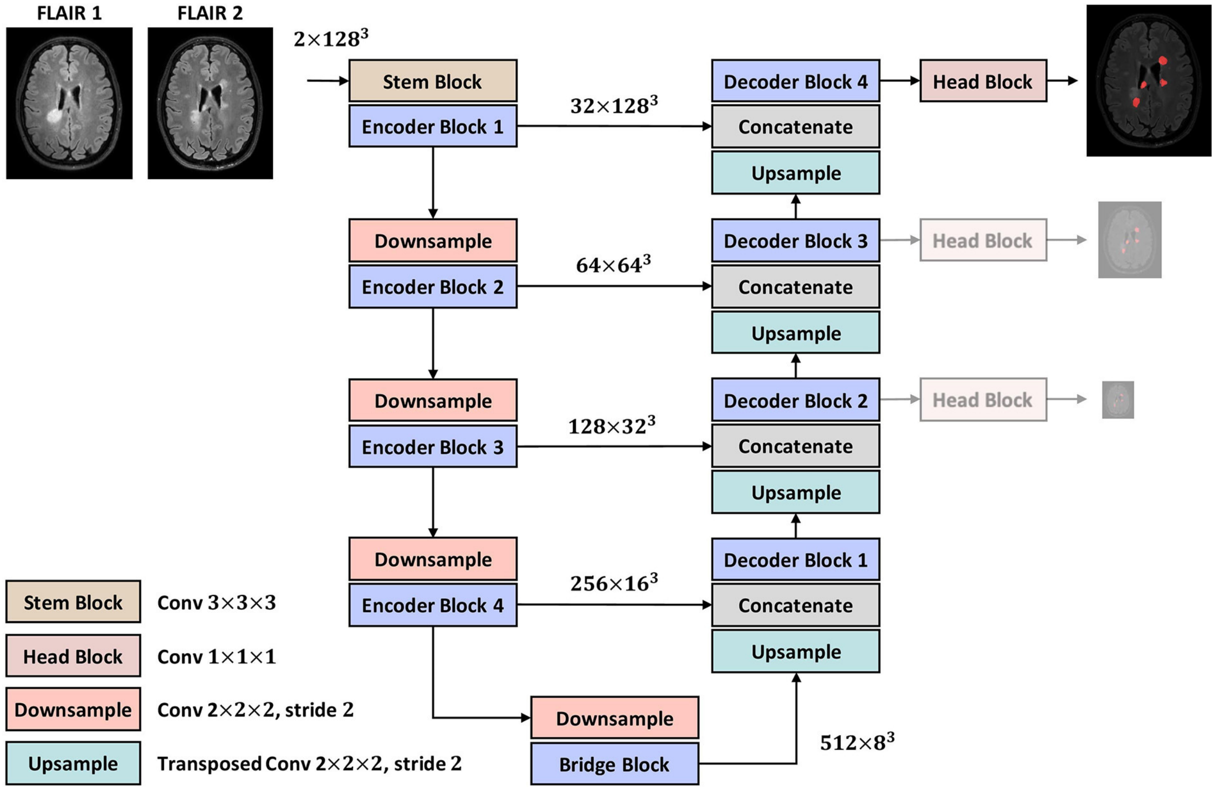


FIGURE 2
The proposed encoder-decoder architecture. The two lower-resolution auxiliary maps are only used in the training phase as deep supervisions.

are dominant in the segmentation of brain lesions. nnU-Net (Isensee et al., 2019) makes minor modifications to the standard 3D U-Net (Çiçek et al., 2016), automatically configuring the key design choices. It has been successfully applied to many medical image segmentation tasks, including longitudinal MS lesion segmentation (Isensee et al., 2020). McKinley et al. (2018) proposed an architecture, in which dense blocks (Huang et al., 2017) of dilated convolutions are embedded in a shallow encoder-decoder network. Myronenko (2019) proposed a U-Net-style architecture with a heavier encoder but a lighter decoder for brain tumor segmentation, taking a variational auto-encoder (VAE) approach by adding a branch to the encoder endpoint. Ashtari et al. (2021b) proposed a lightweight CNN for glioma segmentation, with low-rank constraints being imposed on the kernel weights of the convolutional layers in order to reduce overfitting. Aslani et al. (2019) proposed a deep architecture made up of multiple branches of convolutional encoder-decoder networks that perform slice-based MS lesion segmentation. La Rosa et al. (2020) proposed a U-Net-like model, to automatically segment cortical and white matter lesions based on 3D FLAIR and MP2RAGE images. These works and most of the MS research in medical imaging have focused on the cross-sectional segmentation of lesions, while only a few efforts have been made to detect and segment new lesions on longitudinal MRI scans. For example, Nills et al. (2020) proposed a two-path CNN jointly processing two FLAIR images from two time points to address longitudinal segmentation of new and enlarged lesions. In contrast, this paper proposes a single-path U-shaped architecture whose input is the 2-channel image constructed simply by concatenating two longitudinal FLAIR images which are co-registered.

3. Method

In this section, we describe the proposed encoder-decoder architecture, called Pre-U-Net, and its building blocks.

3.1. Overall architecture

The overall architecture, as shown in Figure 2, follows a U-Net-like style made up of encoder and decoder parts. A $3 \times 3 \times 3$ convolution is used as the stem layer. The network takes a 2-channel image of size $128 \times 128 \times 128$ and outputs a *probability map* with the same spatial size. The network has

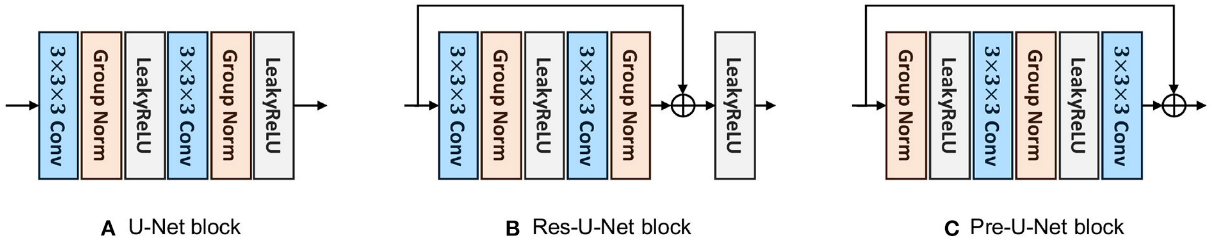


FIGURE 3
The proposed blocks. **(A)** U-Net block. **(B)** Res-U-Net block. **(C)** Pre-U-Net block.

4 levels, at each of which in the encoder (decoder), the input tensor is downsampled (upsampled) by a factor of two while the number of channels is doubled (halved). Downsampling and upsampling are performed *via* strided convolution and transposed convolution, respectively. The kernel size of all downsamplers and upsamplers is 2 × 2 × 2. We use deep supervision at the three highest resolutions in the decoder, applying pointwise convolutions (head blocks) to get three auxiliary logit tensors.

3.2. Baseline models

Depending on which block is used, we build and compare three baselines: (i) U-Net, (ii) Res-U-Net, and (iii) Pre-U-Net. All these variants follow the same overall architecture as explained in Section 3.1 but differ in their encoder/decoder blocks. The block for each model is detailed in the following.

3.2.1. U-Net block

The U-Net block used here is similar to that of nnU-Net (Isensee et al., 2019) except for some minor modifications. As shown in Figure 3A, this block is composed of two convolutional layers with kernel sizes of 3 × 3 × 3. A Group Normalization (Wu and He, 2018) layer (with a group size of 8) comes after each convolutional layer and before LeakyReLU activation.

3.2.2. Res-U-Net block

Inspired by the basic ResNet block (He et al., 2016a), a Res-U-Net block is, as shown in Figure 3B, similar to U-Net block except that a shortcut connection is used between the last Group Normalization layer and the last LeakyReLU activation. A pointwise convolution (i.e., a kernel size of 1 × 1 × 1) may be used in the shortcut connection to match the input dimension with the output dimension of the residual mapping. As investigated by He et al. (2016a), residual connections have been proven effective to avoid vanishing/exploding gradients and speed up the convergence, especially in very deep networks.

3.2.3. Pre-U-Net block

Similar to the pre-activation residual block (He et al., 2016b), a pre-U-Net block consists of two convolutional layers with kernel sizes of 3 × 3 × 3, with LeakyReLU activation coming before each convolutional layer and after Group Normalization (with a group size of 8). Note that the pre-U-Net block, in contrast to U-Net and Res-U-Net blocks, starts with normalization, applying convolution-activation-normalization in reverse order (see Figure 3C). He et al. (2016b) suggest that such a pre-activation design together with identity mappings as the shortcut connections makes information propagate more smoothly than the post-activation design (which is used in the basic ResNet block). Through ablation experiments, they show that the pre-activation design reduces overfitting more significantly, meaning that it leads to slightly higher training loss at convergence but lower test error compared to the post-activation design.

4. Experiments

All the models are implemented using PyTorch (Paszke et al., 2019) and PyTorch Lighting (Falcon, 2019) frameworks and trained on NVIDIA P100 GPUs. We evaluate the performance of Pre-U-Net for MS lesion segmentation on the MSSEG-2 dataset. We follow the same training workflow in all the experiments. In the following, we first provide the details of this workflow, then present the evaluation protocol and the results.

4.1. Setup

4.1.1. Data

A total of 40 and 60 MS patients are represented in the MSSEG-2 training and test set, respectively. For each patient, two longitudinal 3D FLAIR images are acquired at different time intervals (e.g., 1 year, 3 years) and registered in the intermediate space between the two time points. New lesions that a patient developed between the two time points were manually delineated by multiple raters, and the consensus

TABLE 1 An overview of the MSSEG-2 dataset.

Data	Modality	Median voxel size (mm)	Median shape	No. of total cases	No. with-new-lesion cases	No. without-new-lesion cases
Training	FLAIR	(0.53, 0.98, 0.98)	(320, 256, 256)	40	29	11
Test	FLAIR	(0.65, 0.98, 0.98)	(280, 256, 256)	60	32	28
All	FLAIR	(0.60, 0.98, 0.98)	(297, 256, 256)	100	61	39

The third column indicates the median value of voxel size for each axis. The fourth column indicates the median number of voxels along each axis.

ground truths were obtained through a voxel-wise majority voting (see Figure 1). The training (test) set includes images that have no new lesions since, in real clinical practice, many patients under treatment do not develop any new lesions during the time interval. Further details on the MSSEG-2 dataset are reported in Table 1. Note that both the training and test sets were fixed across our experiments as well as for all the challengers.

4.1.2. Preprocessing

For each case, we first concatenate the two FLAIR images to form a 2-channel 3D image as the input. This is valid since the two FLAIR images are co-registered, and therefore, spatially aligned. The resulting image and its ground truth are then cropped with a minimal box filtering out zero regions. MSSEG-2 data are heterogeneous in the sense that the images may be acquired with different protocols in multiple institutes using different scanners, making intensity values greatly vary across patients and even across time points within the same patient. Therefore, we normalize each image channel-wise using a z-score to have intensities with zero mean and unit variance. Moreover, all the images and their ground truths are then resampled to the same voxel size of 1 mm^3 using trilinear interpolation.

4.1.3. Data augmentation

To reduce overfitting caused by data insufficiency and heterogeneity, it is crucial to perform an effective data augmentation workflow before feeding the data into the network. ^{P3}During training, the data preprocessing and augmentation are integrated into a single pipeline operating on a batch of 2 samples at each step on the fly. From each sample, we first crop a random 128 × 128 × 128 patch whose center lies within the foreground (i.e., new lesions) with a probability of 66%. Such an oversampling technique ensures that at least 66% of the patches contain some lesion, which in turn alleviates the class imbalance problem caused by the relatively small size of new lesions. The patches then undergo spatial transforms, including random affine and random flip along each spatial dimension, and intensity transforms, including random additive Gaussian noise, random Gaussian smoothing, random intensity scaling and shifting, random bias field, and random contrast adjusting. All the preprocessing operations and augmentation transforms are computed on CPU using the MONAI library ().

4.1.4. Optimization

All networks are trained for 100,000 steps with a batch size of 2 (each patch is processed on one GPU) using AdamW optimizer with an initial learning rate of 1e−5, weight decay of 1e−2, and cosine annealing scheduler. ^{P3}Therefore, each network in training is fed by a total of 200,000 different patches of size 128 × 128 × 128. It is worth mentioning that since the training set consists of 40 subjects, there are 5,000 = 200,000/40 patches per subject, among which around 3,300 = 5,000 ×0.66 patches are expected to contain new lesions.

The loss $\mathcal{L}_{\text{total}}$ is computed by incorporating the three deep supervision outputs and the corresponding downsampled ground truths, according to

$$\mathcal{L}_{\text{total}} = \lambda_0 \mathcal{L}(\mathbf{G}_0, \mathbf{P}_0) + \lambda_1 \mathcal{L}(\mathbf{G}_1, \mathbf{P}_1) + \lambda_2 \mathcal{L}(\mathbf{G}_2, \mathbf{P}_2), \quad (1)$$

where $\lambda_0 = 1$, $\lambda_1 = 0.5$, and $\lambda_2 = 0.25$; $\mathbf{G}_i$ and $\mathbf{P}_i$ correspond to the deep supervision at resolution $[128/(2^i)]^3$; and the loss function $\mathcal{L}(\cdot,\cdot)$ is the sum of soft Dice (Milletari et al., 2016) and Focal loss (Lin et al., 2017), that is

$$\mathcal{L}(\mathbf{G}, \mathbf{P}) = \mathcal{L}_{\text{Dice}}(\mathbf{G}, \mathbf{P}) + \mathcal{L}_{\text{Focal}}(\mathbf{G}, \mathbf{P}), \quad (2)$$

where

$$\begin{aligned} \mathcal{L}_{\text{Dice}}(\mathbf{G}, \mathbf{P}) &= 1 - \frac{2\langle \mathbf{G}, \mathbf{P} \rangle + \epsilon}{\|\mathbf{G}\|^2 + \|\mathbf{P}\|^2 + \epsilon}, \\ \mathcal{L}_{\text{Focal}}(\mathbf{G}, \mathbf{P}) &= -\frac{1}{N} \langle \mathbf{G}, (\mathbf{1} - \mathbf{P})^{\gamma} \log(\mathbf{P}) \rangle, \end{aligned} \quad (3)$$

where $\mathbf{G} \in \{0,1\}^{J \times N}$ and $\mathbf{P} \in [0,1]^{J \times N}$ represent the one-hot encoded ground truth and the predicted probability map for each voxel, respectively, with J denoting the number of segmentation classes and N denoting the number of voxels in the patch. The small constant $\epsilon = 10^{-5}$ is commonly used to smooth the soft Dice loss and avoid division by zero.

The focusing parameter $\gamma = 2$ smoothly controls the rate at which well-classified voxels are suppressed in the Focal loss, and $\mathbf{1}$ denotes a $J \times N$ matrix of ones. The Focal loss has proved effective in tackling the class imbalance problem, which is present in the MSSEG-2 training set since the total volume of new lesions is generally much smaller than that of the background, and nearly one-third of the patients have no new lesions.

4.1.5. Inference

A test image in the inference is first subjected to z-score intensity normalization and resampled to a voxel size of 1 mm^3. The prediction is then made using a sliding window approach with a 50% overlap and a window size of 128 × 128 × 128 (which is equal to the patch size used in training). For a given voxel from overlapping windows, the mean of the predictions is simply taken as the final value (the `SlidingWindowInferer` module from MONAI was used to perform the sliding window inference). The resulting probability map is resampled back to the original voxel size and finally thresholded by 0.5 to obtain a binary segmentation map.

4.1.6. Evaluation

The Dice score and Hausdorff Distance (HD) are used as metrics to assess the performance of segmentation for the patients that have some new lesions in their ground truths. The Dice score measures the voxel-wise overlap between the ground truth and the prediction, defined as

$$\text{Dice}(\mathbf{g}, \mathbf{y}) = \frac{2\sum_{n=1}^{N} g_n y_n}{\sum_{n=1}^{N} g_n + \sum_{n=1}^{N} y_n} \tag{4}$$

where $g_n \in \{0, 1\}$ and $y_n \in \{0, 1\}$ represent the ground truth and the binary prediction for a voxel, respectively, and N is the number of voxels. Hausdorff Distance (HD) evaluates the distance between the boundaries of ground truth and prediction, computed according to:

$$\text{HD}(G, Y) = \max\{\max_{\mathbf{g}\in G} \min_{\mathbf{y}\in Y} \|\mathbf{g} - \mathbf{y}\|, \max_{\mathbf{y}\in Y} \min_{\mathbf{g}\in G} \|\mathbf{y} - \mathbf{g}\|\}, \tag{5}$$

where G and Y denote the set of all voxels on the surface of ground truth and prediction, respectively.

Lesion-wise sensitivity (SEN), positive predictive value (PPV), and F_1 score are used as metrics to quantify the detection rate of new lesions. Let $\mathbf{G}$ be the ground truth and $\mathbf{Y}$ be the prediction. To compute these lesion level metrics, we follow Commowick et al. (2018), according to which the connected components of $\mathbf{G}$ and $\mathbf{Y}$ (with a 18-connectivity kernel) are first extracted, and all new lesions smaller than 3 mm^3 in size are removed, yielding new tensors $\tilde{\mathbf{G}}$ and $\tilde{\mathbf{Y}}$. The metrics are then defined as

$$\begin{aligned} \text{SEN} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{PPV} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{F}_1 &= \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \end{aligned} \tag{6}$$

where TP, FP, and FN are the number of true positives, false positives, and false negatives, respectively, in the detection of new lesions (i.e., connected components). The rules by which a lesion is considered detected are explained in Commowick et al. (2018).

For cases without any new lesions in their ground truths, we use the following two metrics:

- The **N**umber of new **L**esions **P**redicted (NLP) by the algorithm. This is obtained by counting the number of connected components in the predicted segmentation.
- The **V**olume of new **L**esions **P**redicted (VLP) by the algorithm. This is obtained by simply multiplying the number of voxels in the predicted segmentation by the voxel volume.

All the metrics mentioned above were computed using `animaSegPerfAnalyzer` from the Anima toolbox (available at https://anima.irisa.fr/, RRID: SCR_017017 and RRID: SCR_01707).

5. Results and discussion

5.1. Quantitative evaluation

We performed five-fold cross-validation in all the experiments to estimate how capable our models are in generalizing to unseen data. The cross-validation results on the MSSEG-2 training set are reported in Table 2. For each network, we used an ensemble of the five models trained during the cross-validation on the training set for predicting the test set labels. The test results are reported in Table 3 and illustrated by notched box plots in Figure 4, where pairwise Wilcoxon signed-rank tests were used to identify the significant differences in the test scores of baselines.

Pre-U-Net was superior to all the other models in terms of both segmentation and detection performance for the test cases with some new lesions, achieving a Dice score of 40.3%, HD of 35.0, SEN of 47.5%, PPV of 53.6%, and F_1 score of 48.1%. While having almost the same number of parameters and the same computational complexity (FLOPS), Pre-U-Net outperformed U-Net, the second-best baseline, and significantly outperformed Res-U-Net, with p-value < 0.05 for the Dice score, p-value < 0.01 for the F_1 score, and p-value < 0.05 for HD.

TABLE 2 Results obtained by five-fold cross-validation on the MSSEG-2 training set.

Model	No. of params	FLOPs	With-new-lesion cases					Without-new-lesion cases	
			Dice (%)↑	HD (mm)↓	SEN (%)↑	PPV (%)↑	F_1 (%)↑	NLP↓	VLP (mm^3)↓
U-Net	28.7 M	1264.7 G	45.2 (5.5)	**39.0** (14.1)	51.0 (12.1)	52.9 (6.1)	48.9 (7.6)	0.1 (0.2)	9.2 (20.6)
Res-U-Net	28.9 M	1280.8 G	42.4 (11.4)	46.4 (15.9)	49.3 (22.9)	**60.6** (6.6)	49.9 (14.8)	0.2 (0.4)	4.0 (9.0)
Pre-U-Net	28.9 M	1280.8 G	**45.6** (9.5)	40.1 (13.2)	**54.5** (13.8)	53.8 (6.8)	**51.9** (11.3)	**0.0** (0.0)	**0.0** (0.0)

Symbols ↑ and ↓ indicate that a metric is desired to be higher and lower, respectively. The mean and standard deviation (SD) of a score across the folds are reported as "mean (SD)." The best results are in boldface.

TABLE 3 Results on the MSSEG-2 test set.

Model	No. of params	FLOPs	With-new-lesion cases					Without-new-lesion cases	
			Dice (%)↑	HD (mm)↓	SEN (%)↑	PPV (%)↑	F_1 (%)↑	NLP↓	VLP (mm^3)↓
U-Net	28.7 M	1264.7 G	38.9 (31.1)	43.1 (27.3)	45.2 (36.8)	51.2 (39.6)	45.3 (35.7)	0.0 (0.2)	0.4 (2.3)
Res-U-Net	28.9 M	1280.8 G	34.9 (29.5)	44.2 (29.0)	43.6 (38.4)	38.4 (38.5)	33.7 (33.1)	**0.0** (0.0)	**0.0** (0.0)
Pre-U-Net	28.9 M	1280.8 G	**40.3** (30.5)	**35.0** (22.3)	**47.5** (37.9)	**53.6** (38.3)	**48.1** (34.8)	0.0 (0.2)	0.5 (2.5)

Predictions were made using the five models from the cross-validation as an ensemble. Symbols ↑ and ↓ indicate that a metric is desired to be higher and lower, respectively. The mean and standard deviation (SD) of a score across patients are reported as "mean (SD)." The best results are in boldface.

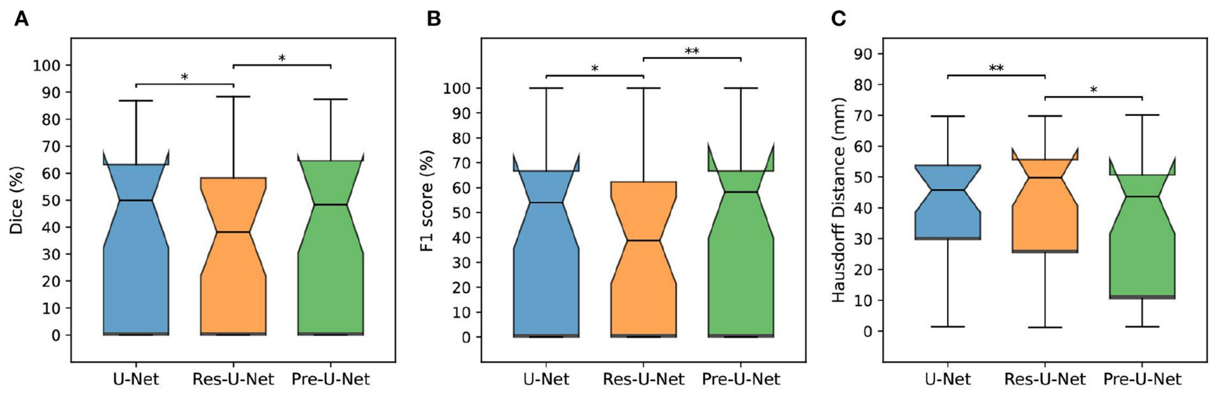


FIGURE 4
Comparison of different models on the MSSEG-2 test set. **(A–C)** Show box plots of Dice score (%), F1 score (%), and Hausdorff Distance (mm), respectively. The asterisks indicate how significantly a model score differs from those of the other baselines when using a pairwise Wilcoxon signed-rank test (*p-value < 0.05, **p-value < 0.01).

Overall, Pre-U-Net proved more effective than the other models at segmentation and detecting new lesions. Nevertheless, note that Pre-U-Net was only marginally superior to U-Net, and there was no statistically significant difference between the two models in terms of the segmentation or detection metrics.

Res-U-Net, with an NLP of 0.0 and VLP of 0.0, performed slightly better for the test cases that have no new lesions whereas Pre-U-Net is the winner in terms of validation scores. In fact, the differences in NLP and VLP scores are marginal, and all of our models are sufficiently accurate to detect no lesions (i.e., produce a segmentation map in which all elements are zero) for patients without any new lesions. Our team, LYLE, with the Pre-U-Net model (Ashtari et al., 2021a) was ranked first in the MSSEG-2 challenge in the two leaderboards based on the NLP and VLP metrics. All the four leaderboards (based on Dice, F_1 score, NLP, and VLP metrics) and the patient-wise scores for each participating team can be found on https://portal.fli-iam.irisa.fr/msseg-2/challenge-day/.

5.2. Qualitative evaluation

Figure 1 presents qualitative comparisons of baselines. The top row exemplifies a patient with a single lesion that is detected by all the models. However, Pre-U-Net, with a patient-wise Dice score of 61.1%, yields a lesion that overlaps most with the lesion in the ground truth compared to U-Net with a patient-wise Dice score of 51.9% and Res-U-Net with a patient-wise Dice score of 36.9%.

Moreover, Pre-U-Net demonstrates superior performance in detecting new lesions. This capability is evidenced in the middle and bottom rows, where Pre-U-Net detects the two new lesions circled in yellow whereas U-Net and Res-U-Net fail to capture them. Note that as observed in the bottom row, Pre-U-Net, with a patient-wise Dice score of 68.3%, shows only a slight improvement in the segmentation performance over U-Net, with a patient-wise Dice score of 66.6%; however, Pre-U-Net indeed outperforms U-Net significantly when it comes to new lesion detection. Nevertheless, some new lesions are extremely challenging to detect even for experts, and all the models fail to capture them. For example, the lesion circled in blue on the ground truth (row 3 and column 3 in Figure 1) is detected by none of the models including Pre-U-Net.

Future work aims at improving new MS lesion detection, especially in the presence of such difficult lesions. This might include, for instance, incorporating the individual delineations of raters into our models. Indeed, in cases where there is more uncertainty due to a weaker consensus among raters (e.g., three raters delineated a set of voxels differently than the other one), our models are also more likely to result in false predictions. Moreover, we will investigate the possibility of transfer learning from a simpler lesion segmentation task with a bigger dataset for further tackling the data insufficiency and class imbalance problems faced in this work.

6. Conclusion

We devised a U-Net-like architecture consisting of pre-activation blocks, called Pre-U-Net, for longitudinal MS lesion segmentation. We successfully trained our models by using data augmentation and deep supervision, alleviating the problem of data insufficiency and class imbalance. The effectiveness of Pre-U-Net was evaluated in segmenting and detecting new white matter lesions in 3D FLAIR images on the MSSEG-2 dataset. Pre-U-Net achieved a Dice score of 40.3% and F_1 score of 48.1%, outperforming the baselines, U-Net and Res-U-Net. In particular, Pre-U-Net is, as reflected by F_1 scores, more effective than the baselines at detecting new lesions, and it is competitive with U-Net in terms of segmentation performance, as evidenced by Dice and HD scores.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://portal.fli-iam.irisa.fr/msseg-2/data/.

Ethics statement

The studies involving human participants were reviewed and approved by MICCAI 2021 MSSEG-2 challenge. The patients/participants provided their written informed consent to participate in this study.

Author contributions

PA conceptualized, developed, validated the methodology of the study, wrote the first draft of the manuscript, and implemented the models. PA and BB contributed to the software development. SV acquired the funding. SV and DS-M supervised the work. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This research leading to these results has received funding from EU H2020 MSCA-ITN-2018: INtegrating Magnetic Resonance SPectroscopy and Multimodal Imaging for Research and Education in MEDicine (INSPiRE-MED), funded by the European Commission under Grant Agreement #813120. This research also received funding from the Flemish Government (AI Research Program). AI—KU Leuven Institute for AI, B-3000, Leuven, Belgium.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ashtari, P., Barile, B., Van Huffel, S., and Sappey-Marinier, D. (2021a). "Longitudinal multiple sclerosis lesion segmentation using pre-activation U-Net," in *MSSEG-2 Challenge Proceedings: Multiple Sclerosis New Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*, (Strasbourg), 45–51. Available online at: https://hal.inria.fr/hal-03358968v1/document#page=54

Ashtari, P., Maes, F., and Van Huffel, S. (2021b). "Low-rank convolutional networks for brain tumor segmentation," in *International MICCAI Brainlesion Workshop: BrainLes 2020. Lecture Notes in Computer Science, Vol. 12658*, eds A. Crimi and S. Bakas (Cham: Springer), 470–480. doi: 10.1007/978-3-030-72084-1_42

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

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016*, eds S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells (Cham: Springer), 424–432. doi: 10.1007/978-3-319-46723-8_49

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., et al. (2018). Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Sci. Rep.* 8, 1–17. doi: 10.1038/s41598-018-31911-7

Confavreux, C., Compston, D., Hommes, O., McDonald, W., and Thompson, A. (1992). EDMUS, a European database for multiple sclerosis. *J. Neurol. Neurosurg. Psychiatry* 55, 671–676. doi: 10.1136/jnnp.55.8.671

Falcon, W. (2019). *PyTorch Lightning*. GitHub. Available online at: https://github.com/PyTorchLightning/pytorch-lightning

He, K., Zhang, X., Ren, S., and Sun, J. (2016a). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (IEEE)*, 770–778. doi: 10.1109/CVPR.2016.90

He, K., Zhang, X., Ren, S., and Sun, J. (2016b). "Identity mappings in deep residual networks," in Computer Vision – ECCV 2016, eds B. Leibe, J. Matas, and M. Welling (Cham: Springer), 630–645.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (IEEE)*, 4700–4708. doi: 10.1109/CVPR.2017.243

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2020). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* 18, 203–211. doi: 10.1038/s41592-020-01008-z

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2019). "No New-Net," in *International MICCAI Brainlesion Workshop: BrainLes 2018. Lecture Notes in Computer Science*, eds A. Crimi and S. Bakas, and others (Cham: Springer), 234–244.

La Rosa, F., Abdulkadir, A., Fartaria, M. J., Rahmanzadeh, R., Lu, P.-J., Galbusera, R., et al. (2020). Multiple sclerosis cortical and WM lesion segmentation at 3T MRI: a deep learning method based on FLAIR and MP2RAGE. *NeuroImage. Clin.* 27, 102335. doi: 10.1016/j.nicl.2020.102335

Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., and Tu, Z. (2015). "Deeply-supervised nets," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, eds G. Lebanon and S. V. N. Vishwanaathan (San Diego, CA: PMLR), 562–570. Available online at: http://proceedings.mlr.press/v38/lee15a.pdf

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (IEEE)*, 2980–2988. doi: 10.1109/ICCV.2017.324

McKinley, R., Meier, R., and Wiest, R. (2018). "Ensembles of densely-connected CNNs with label-uncertainty for brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries,* eds A. Crimi, S. Bakas, H, Kuijf, F. Keyven, M. Reyes, and T. van Walsum (Cham: Springer), 456–465. doi: 10.1007/978-3-030-11726-9_40

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-Net: fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision*, 565–571. doi: 10.1109/3DV.2016.79

MONAI Consortium. (2020). MONAI: Medical Open Network for AI. *Zenodo*. doi: 10.5281/zenodo.6114127

Myronenko, A. (2019). "3D MRI brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop: BrainLes 2018. Lecture Notes in Computer Science, Vol. 11384*, eds A. Crimi, S. Bakas, et al. (Cham: Springer), 311–320. doi: 10.1007/978-3-030-11726-9_28

Nills. G., Kruger, J., Opfer, R., Ostwaldt, A. -C., Manogaran, P., Kitzler, H. H., Schippling, S., et al. (2020). Multiple sclerosis lesion activity segmentation with attention-guided two-path CNNs. *Computer. Med. Imag. Graph.* 84, 101772. doi: 10.1016/j.compmedimag.2020.101772

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, *Vol. 32*, eds H. Wallach, H. Larochelle, A. Beygelzimer, F. Alche-Buc, E. Fox, and R. Garnett (Curran Associates, Inc). Available online at: https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015*, eds N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi (Cham: Springer International Publishing), 234–241. doi: 10.1007/978-3-319-24574-4_28

Vukusic, S., Casey, R., Rollot, F., Brochet, B., Pelletier, J., Laplaud, D.-A., et al. (2020). Observatoire français de la sclérose en plaques (OFSEP): a unique multimodal nationwide MS registry in France. *Multip. Scler.* 26, 118–122. doi: 10.1177/1352458518815602

Wu, Y., and He, K. (2018). "Group normalization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–19. doi: 10.1007/978-3-030-01261-8_1

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 04 November 2022
DOI 10.3389/fnins.2022.1004050

Check for updates

OPEN ACCESS

EDITED BY
Michel Dojat,
Institut National de la Santé et de la Recherche Médicale (INSERM), France

REVIEWED BY
Richard McKinley,
Bern University Hospital, Switzerland
Govind Nair,
National Institutes of Health (NIH), United States
Robert Fekete,
New York Medical College, United States
Vijay Venkatraman,
The University of Melbourne, Australia

*CORRESPONDENCE
Olivier Colliot
olivier.colliot@cnrs.fr

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 26 July 2022
ACCEPTED 11 October 2022
PUBLISHED 04 November 2022

CITATION
Schmidt-Mengin M, Soulier T, Hamzaoui M, Yazdan-Panah A, Bodini B, Ayache N, Stankoff B and Colliot O (2022) Online hard example mining vs. fixed oversampling strategy for segmentation of new multiple sclerosis lesions from longitudinal FLAIR MRI. *Front. Neurosci.* 16:1004050. doi: 10.3389/fnins.2022.1004050

Online hard example mining vs. fixed oversampling strategy for segmentation of new multiple sclerosis lesions from longitudinal FLAIR MRI

Marius Schmidt-Mengin[1,2†], Théodore Soulier[2†], Mariem Hamzaoui[2], Arya Yazdan-Panah[1,2], Benedetta Bodini[2,3], Nicholas Ayache[4], Bruno Stankoff[2,3] and Olivier Colliot[1*]

[1]Institut du Cerveau-Paris Brain Institute, Centre National de la Recherche Scientifique, Inria, Inserm, Assistance Publique-Hôpitaux de Paris, Hôpital de la Pitié Salpêtrière, Sorbonne Université, Paris, France, [2]Institut du Cerveau-Paris Brain Institute, Centre National de la Recherche Scientifique, Inserm, Assistance Publique-Hôpitaux de Paris, Hôpital de la Pitié Salpêtrière, Sorbonne Université, Paris, France, [3]Department of Neurology, Assistance Publique-Hôpitaux de Paris, Hôpital Saint-Antoine, Paris, France, [4]Inria, Epione Project-Team, Sophia-Antipolis, France

Detecting new lesions is a key aspect of the radiological follow-up of patients with Multiple Sclerosis (MS), leading to eventual changes in their therapeutics. This paper presents our contribution to the MSSEG-2 MICCAI 2021 challenge. The challenge is focused on the segmentation of new MS lesions using two consecutive Fluid Attenuated Inversion Recovery (FLAIR) Magnetic Resonance Imaging (MRI). In other words, considering longitudinal data composed of two time points as input, the aim is to segment the lesional areas, which are present only in the follow-up scan and not in the baseline. The backbone of our segmentation method is a 3D UNet applied patch-wise to the images, and in which, to take into account both time points, we simply concatenate the baseline and follow-up images along the channel axis before passing them to the 3D UNet. Our key methodological contribution is the use of online hard example mining to address the challenge of class imbalance. Indeed, there are very few voxels belonging to new lesions which makes training deep-learning models difficult. Instead of using handcrafted priors like brain masks or multi-stage methods, we experiment with a novel modification to online hard example mining (OHEM), where we use an exponential moving average (i.e., its weights are updated with momentum) of the 3D UNet to mine hard examples. Using a moving average instead of the raw model should allow smoothing of its predictions and allow it to give more consistent feedback for OHEM.

KEYWORDS

segmentation, deep learning, hard example mining, multiple sclerosis, MRI

Introduction

Multiple Sclerosis (MS) is a chronic autoimmune demyelinating inflammatory disease of the central nervous system and represents the leading cause of non-traumatic motor disability of young people in Europe and North America (Howard et al., 2016). MS lesions, consisting of focal areas of demyelination, edema, and auto-immune inflammation, are visible on Magnetic Resonance Imaging (MRI), especially on Fluid Attenuated Inversion Recovery (FLAIR) as contiguous areas of hypersignal (Filippi et al., 2019). The decrease or absence of new FLAIR lesion formation over time is a key radiological endpoint in clinical trials assessing disease-modifying therapies in MS, and the absence of such radiological activity takes part in the "No Evidence of Disease Activity" score, used to monitor patient's disease control and to discuss potential therapeutic change at the individual level (Hegen et al., 2018). Novel lesion identification and segmentation is usually performed manually, or using semi-automated procedures, by radiologists or neurologists and is time-consuming and subject to intra- and inter-rater variability (Altay et al., 2013). The aim of the MICCAI MSSEG-2 challenge was to benchmark new automatic methods to segment new lesions based on two FLAIR MRIs from two longitudinal visits (baseline and follow-up) of the same patient. Already published methods for this task consists mostly of either non-deep learning methods (Cabezas et al., 2016) or deep learning methods using multiple MRI sequences (McKinley et al., 2020; Salem et al., 2020); there are very few deep learning methods for this precise task based uniquely on FLAIR sequences (Gessert et al., 2020). The present paper describes our contribution to the challenge. The backbone of our approach is a patch-wise 3D UNet (Çiçek et al., 2022). Our key methodological contribution is to introduce online hard example mining (Shrivastava et al., 2016) (OHEM) to tackle class imbalance. Indeed, one important characteristic of the dataset is that there are fewer voxels belonging to a new lesion (positive) than not belonging to a new lesion (negative), images comprise on average approximately 0.005% of positive voxels. Notably, we use a moving average of our 3D UNet to perform inference for hard example mining. Our goal is that, similar to He et al. (2020), doing so will provide more stable predictions as training progresses. The present paper extends that published in the proceedings of the MICCAI MSSEG-2 2021 workshop (Commowick et al., 2021) by providing a more extensive description of the methodology as well as more detailed experimental results including the testing of the algorithm on another cohort (Bodini et al., 2016) distinct from the MICCAI MSSEG-2 testing dataset.

Methods

Preprocessing

We resampled each FLAIR image to a voxel size of 0.5 mm as it is the highest resolution of the training dataset and applied a z-score normalization to each FLAIR individually. As the two consecutive FLAIR images (baseline and follow-up) of a patient have been aligned in the halfway space using a rigid transformation by the challenge providers, our method starts by concatenating them along the channel dimension, resulting in a tensor of shape 2*D*H*W, where D, H, and W are, respectively, the depth, height, and width of the resampled FLAIR image. This tensor is then subdivided into patches of shape 2*32*32*32, which are passed through a 3D UNet to obtain the segmentation.

Model

Our backbone model is a standard 3D UNet, which can be described by the following equations:

$$\mathrm{B}\,(\mathrm{n})\ :=2\mathrm{x}\ \left\{\mathrm{3DConvolution}\,(\mathrm{n})\rightarrow \mathrm{Group\ Normalization}\rightarrow \mathrm{ReLU}\right\}$$

$$\mathrm{3D\ UNet}\ :=\mathrm{B}\,(16)\downarrow\ \rightarrow\ \mathrm{B}\,(32)\downarrow\rightarrow\ \mathrm{B}\,(64)\ \rightarrow\ \uparrow\mathrm{B}\,(32)\ \rightarrow\uparrow\mathrm{B}\,(16)\ \rightarrow\ \mathrm{Conv}\,(1)$$

where the numbers in the parentheses are the number of filters, ↓ indicates max pooling and ↑ indicates trilinear upsampling. The model is trained on patches of size 32. For inference, we split the image into a grid of patches of size 32, with a stride of 24. This means that the patches have an overlap of 8 pixels. In these overlapping regions, we averaged all predictions and binarized the final output with a threshold of 0.5.

Dataset

We used the MICCAI MSSEG-2 datasets (Commowick et al., 2021) for training, validation, and the first testing set (see Appendix). We also used a second testing set consisting of a previously published MS cohort from our laboratory (Bodini et al., 2016). This cohort was constituted of 19 patients with active relapsing remitting MS (13 women, mean age 32.3 years sd 5.6) who underwent two MRIs with FLAIR spaced from minimum 31 days to maximum 120 days. Of those 19 patients, only 18 had available FLAIR MRIs for each visit. As only one of those 18 remaining patients had no new lesions at the second visit, we focused on the 17 patients that presented new lesions

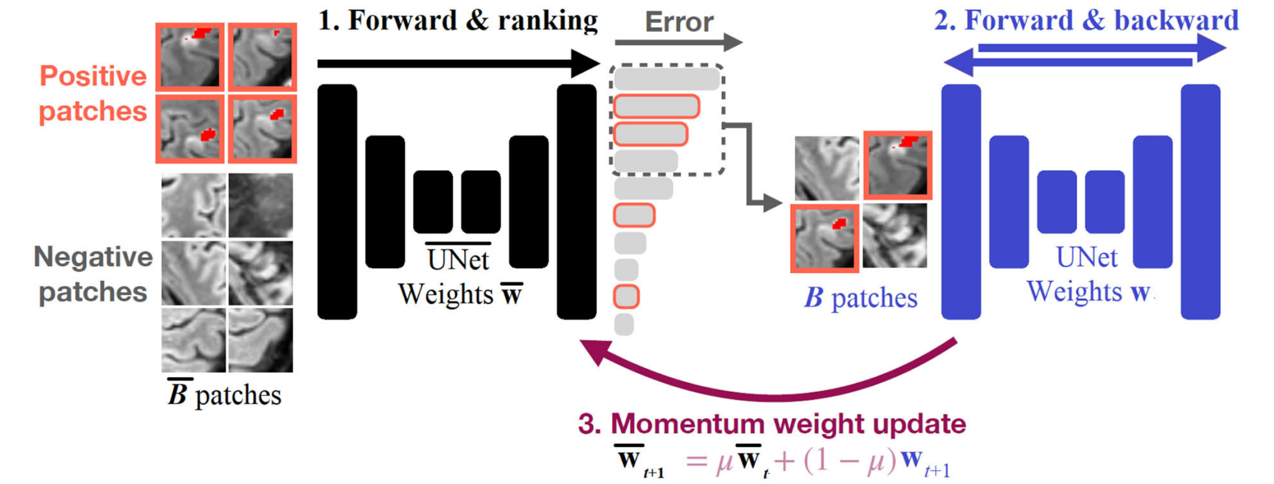

Positive patches
Negative patches
$\overline{B}$ patches
1. Forward & ranking
UNet
Weights $\overline{\mathbf{w}}$
Error
B patches
2. Forward & backward
UNet
Weights $\mathbf{w}$
3. Momentum weight update
$\overline{\mathbf{w}}_{t+1} = \mu\overline{\mathbf{w}}_t + (1-\mu)\mathbf{w}_{t+1}$

FIGURE 1
Illustration of our training strategy. $\overline{B}$ patches are fed to a first 3D $\overline{UNet}$ and the segmentation errors are computed for each patch. The patches are ranked according to their errors, and the top_B patches are selected to perform a training step with a second 3D *Unet*. The weights of the first 3D $\overline{UNet}$ are momentum-updated with the weights of the second 3D *Unet*.

at the second visit for the second testing dataset. For these 17 patients, the new lesions at the second visit were manually contoured in native space and verified by a senior neurologist. After rigid co-registration to halfway space (FLIRT, http://fsl.fmrib.ox.ac.uk/) (Jenkinson and Smith, 2001), we gave the baseline and the follow-up FLAIR as input to our algorithm, and the manually contoured lesion mask as ground truth to evaluate our algorithm performances. Acquisitions for our testing cohort were run on a 3 Tesla Siemens machine, with a 32-channel head coil (Repetition Time: 8.88 ms; Echo Time: 129 ms; Inversion Time: 2.5 ms; Flip Angle: 120°; Pixel size: 0.9 × 0.9 × 3 mm).

Training

As the images contain very few positive voxels, we do not sample the patches uniformly during training. One common strategy is to over-sample patches containing positive regions with a constant ratio. However, this ratio must be fine-tuned by hand. If it is too high, it can result in many false positives. Instead, our method uses a 3D UNet with momentum weight updates to perform hard example mining. A training iteration consists of three steps, illustrated in Figure 1 and described by Algorithm 1. In the first step, we select a batch of $\overline{B} = 128$ patches, which contains 30% of positive patches and 70% of uniformly sampled patches (i.e., mostly negatives due to the class imbalance). We then pass this batch through a 1st 3D UNet, denoted by $\overline{UNet}$, to obtain a prediction for each element of the batch and compute the segmentation errors with respect to the ground truth. Second, we select the $B = 32$ patches with the highest error and perform a training step on them with a second 3D UNet, denoted *Unet*. Last, we perform a momentum update of the weights of the 1st 3D $\overline{UNet}$, with the second 3D *Unet*. The use of momentum ensures that the predictions given by the 1st 3D $\overline{UNet}$ do not fluctuate too much during training and provide reliable samples for online hard example mining.

```
$\overline{B}$: batch size for hard example mining
$B < \overline{B}$: batch size for gradient descent
$\overline{\texttt{UNet}}$: momentum-updated 3D U-Net
$\mathbf{w}$: weights of $\texttt{UNet}$
$\overline{\mathbf{w}}$: weights of $\overline{\texttt{UNet}}$
Initialization: $\overline{\mathbf{w}}_0 = \mathbf{w}_0$
$\mu$: momentum coefficient
while training do
    $\mathbf{x}_1 \dots \mathbf{x}_{\overline{B}} = \texttt{sample\_patches(positive\_ratio=0.3)}$
    $e_1 \dots e_{\overline{B}} = \texttt{error}(\overline{\texttt{UNet}}(\mathbf{x}_1 \dots \mathbf{x}_{\overline{B}}),\ \mathbf{y}_1 \dots \mathbf{y}_{\overline{B}})$
    $i_1 \dots i_B = \texttt{top}_B(e_1 \dots e_{\overline{B}})$
    $\ell_1 \dots \ell_B = \texttt{loss}\ (\texttt{UNet}(\mathbf{x}_{i_1} \dots \mathbf{x}_{i_B}), \mathbf{y}_{i_1} \dots \mathbf{y}_{i_B})$
    $\mathbf{w}_{t+1} = \texttt{Adam}(\mathbf{w}_t, \ell_1 \dots \ell_B)$
    $\overline{\mathbf{w}}_{t+1} = \mu\overline{\mathbf{w}}_t + (1-\mu)\mathbf{w}_{t+1}$
end
```

Algorithm 1. The algorithm used for the training with OHEM and momentum update.

Training—OHEM vs. oversampling comparison

We optimized each network for 3 h on one NVIDIA Tesla P100 graphic card using Adam (Kingma and Ba, 2022). Note that for OHEM, the duration of one iteration is roughly 2 times longer. In the end, 3 h of training corresponds to about 30k iterations with OHEM and 64k without. The initial learning rate was set to 10^{-3} and decayed to 10^{-4} and 10^{-5} after, respectively, 50 and 80% of the training time. We split the dataset into 30 patients for training, and 10 for validation.

We compared the learning curves using the Dice score on the validation set for six training procedures: three with OHEM with a momentum of, respectively, 0, 0.9, and 0.99, and three without OHEM but with oversampling with a probability p of, respectively, 0 (uniform), 0.1, and 0.5. This oversampling probability meant that we sampled positive patches (i.e., with a new lesion at a second time point) with a probability p and other patches (that could be randomly positive or negative) with a probability 1-p for the training.

Training—Final approach provided for the MSSEG-2 challenge

We used the model described before, using OHEM with a momentum of 0.9, and trained the model on the whole MICCAI MSSEG-2 training dataset for 30k iteration. As above, the initial learning rate is set to 10^{-3} and decayed to 10^{-4} and 10^{-5} after, respectively, 50 and 80% of the training time.

Evaluation metrics for the testing dataset

The evaluation procedure was defined by the MICCAI MSSEG-2 committee (Commowick et al., 2021). We briefly recall this procedure in the following. The MICCAI MSSEG-2 testing dataset of 60 patients was divided into two subsets, according to the presence or absence of new lesions in patients: 28 patients without new lesions and 32 patients with new lesions. Those two datasets were evaluated differently.

All new lesions from the ground truth and our algorithm prediction were individualized by computing the connected components, and all lesions smaller than 3 mm^3 were removed (Commowick et al., 2018). The detection was defined at the lesion level using the algorithm described by Commowick et al. (2018) with the parameters $\alpha = 10\%$, $\beta = 65\%$, and $\gamma = 70\%$, which were set by the MICCAI MSSEG-2 committee.

For the 28 patients without new lesions, the following metrics are reported: the lesion volume prediction per patient in mm^3, and the new lesion detection rate per patient.

For the 32 patients with new lesions, the evaluation aimed at assessing both the quality of the detection and the segmentation. For evaluating the segmentation, the (voxel-level) Dice score per patient was reported. For evaluating the detection, the following metrics were used: the mean sensitivity *Sens* (=recall) at the lesion level per patient for detecting new lesions, and the mean positive predictive value *PPV* (=precision) at the lesion level per patient for detecting new lesions and the F_1 score at the lesion level (which combines lesion-level *Sens* and *PPV*) per patient (Commowick et al., 2018).

The calculation of those metrics is described below. True positives with respect to the ground truth TP_{gt} were defined as the number of new lesions from the ground truth that were correctly detected by our algorithm. True positives with respect to our prediction TP_{pred} correspond to the number of new lesions predicted by our algorithm that were correctly detected by the ground truth.

- $Dice = \frac{2\,|PRED \cap GT|}{|PRED| + |GT|}$, where PRED is the network prediction and GT the ground truth segmentation, $|PRED \cap GT|$ is the number of overlapping voxels between the prediction and the ground truth, $|PRED|$ is the number of voxels in the prediction and $|GT|$ the number of voxels in the ground truth.
- $Sens = \frac{TP_{gt}}{n_{new\ lesions_{gt}}}$ where TP_{gt} and $\text{n}_{\text{new lesions\_gt}}$ are, respectively, the true positives with respect to the ground truth and the number of new lesions in the ground truth.
- $PPV = \frac{TP_{pred}}{n_{new\ lesions_{pred}}}$ where TP_{pred} and $\text{n}_{\text{new lesions\_pred}}$ are, respectively, the true positives with respect to our prediction and the number of new lesions in our prediction.
- $F_1 = \frac{2^{*}Sens^{*}PPV}{Sens+PPV}$ where *Sens* and *PPV* are, respectively, the previously defined sensitivity and Positive Predictive Value.

All of those metrics were compared to zero for patients without new lesion, and to the ground truth segmentation of patients with new lesion, which is the consensual segmentation from four expert annotators (Commowick et al., 2021). All results are presented as mean, Standard Error to the Mean (SEM), and rank among other challenge pipelines when available.

For the second testing dataset, constituted by the 17 patients with new lesions in our cohort, we used exactly the same evaluation procedure that we described above for the patients with new lesions from the MICCAI MSSEG-2 testing dataset.

Implementation details

Our algorithms were implemented on PyTorch (Paszke et al., 2017) and written using TorchIO library (Pérez-García et al., 2021). The implementation was based on that of Wolny et al. (2020). Training was performed on an NVIDIA Tesla P100 graphic card.

Results

Results on the validation set: Impact of the OHEM procedure

The comparison of the learning curves for the proposed OHEM procedure and the forced oversampling procedure is

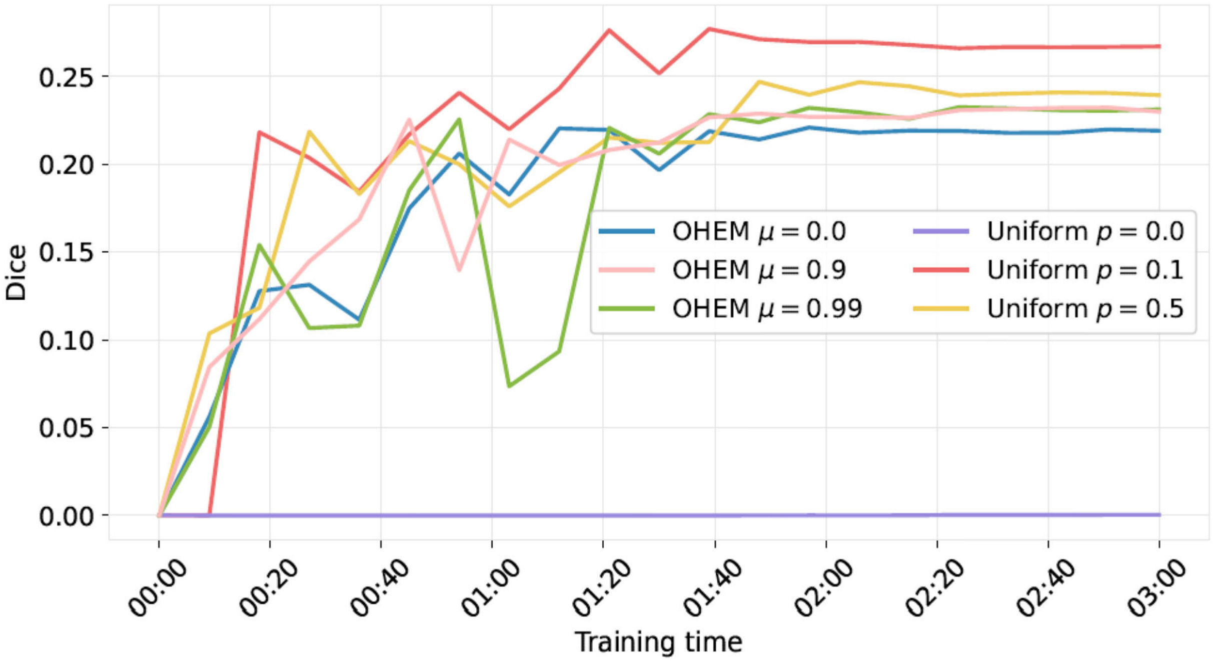


FIGURE 2
Evolution of the Dice score as a function of training time for the OHEM and the forced oversampling procedure (denoted as "Uniform"). For OHEM, μ is the momentum. For "Uniform", patches were sampled with respective probabilities, p for those with new lesions, and 1-p for the rest (not necessarily without new lesions). One can observe that the "Uniform" procedure with $p > 0$ ended up performing best and that, when using OHEM, choosing $\mu > 0$ seems to be beneficial.

shown in Figure 2. One can observe that, on this task, the OHEM procedure, even with increasing the momentum to 0.99, did not give better results in terms of training speed nor plateau of the Dice score. However, we observed that using OHEM gives a positive momentum that helped to reach a higher plateau of the Dice score on the validation set compared to a null one.

Results on the testing set

Results on the first testing set from MICCAI MSSEG-2 are shown in Table 1. In the 32 patients with new lesions, our network achieved a mean lesion-level F_1 score per patient of 0.446 (SEM 0.057), ranking 13th over 29 approaches for this metric. The mean Dice per patient was 0.400 (SEM 0.051), which ranked 18/29. Our mean sensitivity at the lesion level per patient was 0.616 (SEM 0.069) and our mean positive predictive value at the lesion level per patient was 0.383 (SEM 0.054). Concerning the 28 patients without new lesions, for whom any prediction is a pure false positive, on average, 0.75 (SEM 0.32) new lesions were predicted per patient (ranking our approach 15/29), with a mean lesion volume per patient among those 28 patients without new lesion of 31.2 mm^3 (SEM 13.0), which corresponded to a rank of 20/29.

On our second testing set from our laboratory, on the 17 patients with new lesions, the mean Dice per patient was 0.465 (SEM 0.046). At the lesion level, our network achieved a mean sensitivity per patient of 0.901 (SEM 0.043) and a mean positive predictive value per patient of 0.239 (SEM 0.030), resulting in a mean lesion-level F_1 score per patient of 0.365 (SEM 0.038).

Figure 3 shows an example of inference on a follow-up MRI from this second testing set from our laboratory.

Discussion

The main contribution of this work was the introduction of online hard example mining (OHEM) to deal with class imbalance. The rest of the approach is constituted of a standard 3D UNet. We first showed that the use of a non-negative momentum helped the training procedure. However, overall, OHEM did not perform better than a predefined fixed oversampling and especially performed worse when an oversampling probability of $p = 0.1$ was used for fixed oversampling.

On the MICCAI MSSEG-2 testing set, our approach ranked in the mid-class of the challenge (Dice score of 0.400, corresponding rank 18/29; lesion-level F_1 score of 0.446, rank 13/29). Interestingly, compared to other pipelines of the challenge, our worst performances were on the subset of patients

TABLE 1 Results on the testing set using MICCAI MSSEG-2 evaluation metrics, with the specific evaluation metrics from MICCAI MSSEG-2 testing dataset for the 32 patients with new lesions (a) as well as for the 28 patients without new lesions (b), and the 17 patients with new lesions from our second testing dataset (c).

(a) MICCAI MSSEG-2 testing dataset: patients with new lesions ($n = 32$)

Lesion-level F_1 score per patient, mean (SEM); rank	Dice score per patient, mean (SEM); rank	*Sens* at lesion level per patient, mean (SEM)	*PPV* at lesion level per patient, mean (SEM)
0.446 (0.057); 13th/29	0.400 (0.051); 18th/29	0.616 (0.069)	0.383 (0.054)

(b) MICCAI MSSEG-2 testing dataset: patients without new lesion ($n = 28$)

Number of new lesions predicted per patient, mean (SEM); rank	Lesion volume predicted in mm^3 per patient, mean (SEM); rank
0.750 (0.320); 15th/29	31.2 (13.0); 20th/29

(c) Second testing dataset: patients with new lesions ($n = 17$)

Lesion-level F_1 score per patient, mean (SEM)	Dice score per patient, mean (SEM)	*Sens* at lesion level per patient, mean (SEM)	*PPV* at lesion level per patient, mean (SEM)
0.365 (0.038)	0.465 (0.046)	0.901 (0.043)	0.239 (0.030)

without new lesions, where any prediction is a false positive. Together with the relatively high sensitivity but relatively low PPV, this could be explained by a bias in the OHEM training toward a high detection rate, resulting in a greater false positive rate. This trend was even stronger when we evaluated the algorithm performances on our second testing dataset, with a higher Dice score of 0.465, a higher sensitivity of 0.901 but a lower PPV of 0.239.

When compared to other pipelines of the challenge, the best pipeline in the subset of patients without new lesions, consisting of a 3D UNet with pre-activation block, also used an oversampling strategy for Regions of Interest with new lesions, but was also ranked in the mid-class of the challenge for the Dice score on the patients with new lesions (with a Dice score of 0.409). The most accurate pipeline in terms of Dice score (even better than several annotators), which did not use any oversampling strategy, was ranked in the mid-class of the challenge for the subset of patients without new lesions for the score of new lesions detection rate. This is consistent with the idea that dealing with the oversampling of positive examples is a key problem in the balance between false positive and false negative predictions in this new lesion segmentation task. We believe, given the medical utility of this task at the individual level for patient follow-up, that a compromise between sensitivity and PPV favoring sensitivity is clinically relevant if the algorithm is considered as an auxiliary to the neurologist or radiologist. Indeed, the interrater variability in manual new lesions detection is mainly explained by false negative rate (Altay et al., 2013), i.e., new lesions that were not detected by the rater. We believe that sensitive algorithms could help neurologists or radiologists to detect those overlooked new lesions. The clinicians could subsequently easily remove false positive predictions of the algorithm after visual checking. However, there is still a long way to go for clinical applications of algorithms for new lesion segmentation. This will require not only algorithm improvement but also prospective validation studies on larger and very diverse datasets.

There was only one pipeline in the challenge that did not use deep learning. Even if they outperformed four deep learning teams on average, their ranking was low on the MICCAI MSSEG-2 testing dataset, with a mean Dice of 0.309 for patients with new lesions, and a mean volume of new lesions detected of 177.9 mm^3 for patients without new lesions. This does not mean that non-deep-learning methods are not potentially useful for this task but this would require additional comparisons which are outside of the scope of the present work. To our knowledge, most of the previously published deep learning algorithms (McKinley et al., 2020; Salem et al., 2020) or recent non deep learning based on deformation field (Cabezas et al., 2016) used to segment new lesions on MS MRIs are based on multiple MRI sequences and not only on a single sequence. It is the same when looking at previously published deep learning algorithms used to segment the lesion load transversally (Valverde et al., 2019; Zeng et al., 2020). So, even if clinically relevant (Hegen et al., 2018), the challenge task allows neural network to learn less information for prediction than in most of the state of art methods, and it can partly explain the difficulty of the task. The previous work from Gessert et al. (2020) based on attention gated two

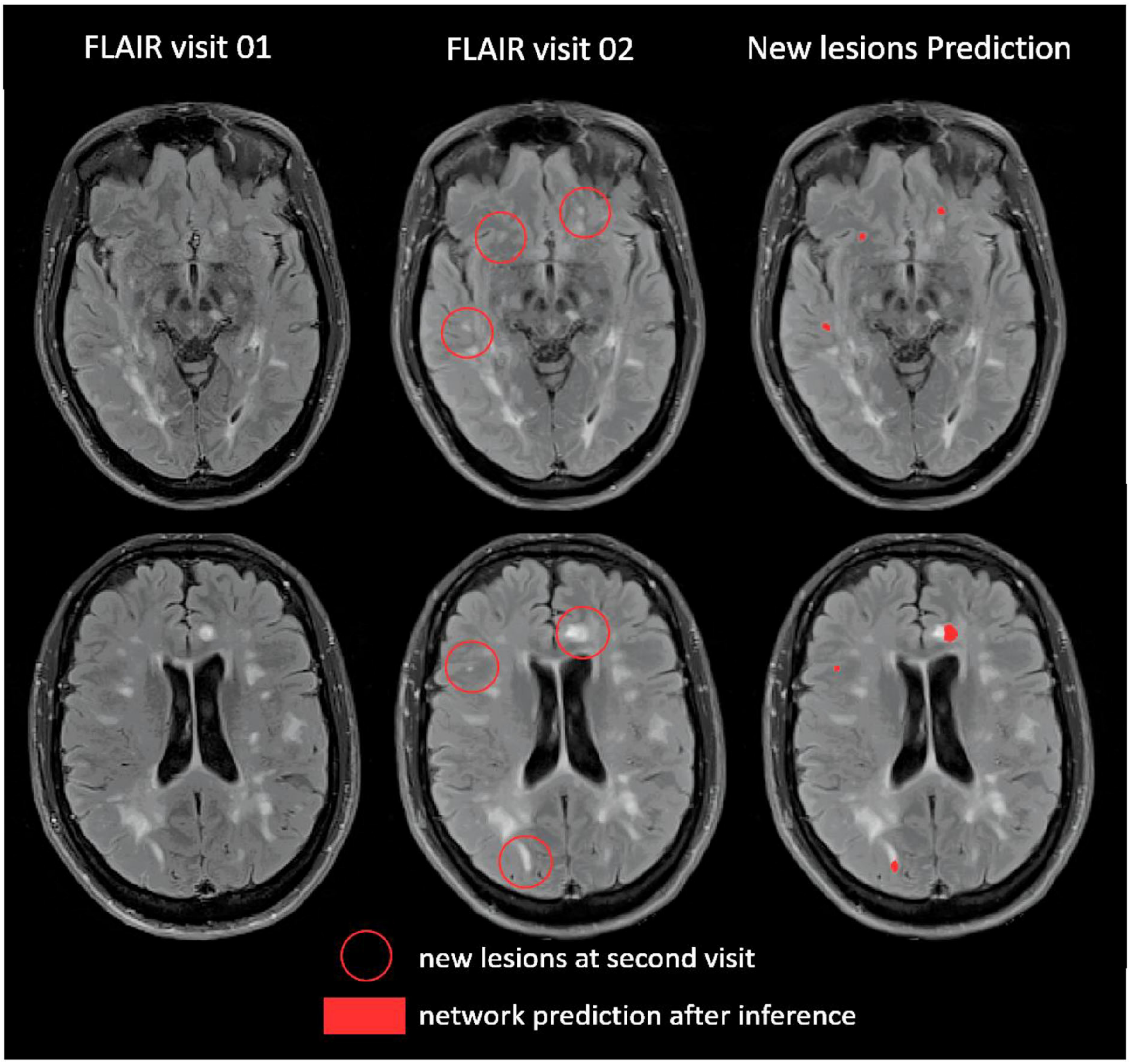

FLAIR visit 01
FLAIR visit 02
New lesions Prediction
new lesions at second visit
network prediction after inference

FIGURE 3
Example of prediction on one patient from our second testing dataset (Bodini et al., 2016).

paths convolutional neural networks was to our knowledge the most relevant deep learning work published on the task of segmenting MS new lesion based only on two follow up FLAIR sequences. They did not require an oversampling procedure to deal with class imbalance and had very good lesion-wise true positive rate and lesion-wise false positive rate. However, we could not compare methods since their proposed evaluation metrics differed from the ones provided by MICCAI MSSEG-2 (Commowick et al., 2018).

This work has several limitations. First, concerning the OHEM training methodology (Shrivastava et al., 2016), it did not improve the training procedure on this task and did not outperform significantly other competing 3D UNets across the challenge. Despite being an interesting methodology to deal with class imbalance, we have to keep in mind that it has been developed for detection in 2D natural images (Shrivastava et al., 2016) using fast R-CNN (Wang et al., 2016). Even though it has shown promising results in Bian et al. (2022) work on heart MRI, unveiling its full potential for 3D medical image segmentation may require further adaptations and developments. Second, we chose to compare OHEM and fixed oversampling as a function of training time and not as a function of epochs. Training time could be influenced by many parameters like machine heat and GPU availability. However, we believe it was the fairest way to compare methods. Indeed, the unit cost of each epoch (or iteration) has no reason to be the same for the different techniques. Even worse, it can vary across epochs due to the nature of the OHEM method. Another limitation is that we used a single split into training and validation rather than a cross-validation strategy. Thus, we did not use all samples for testing and we did not assess their variability when varying the training set. We made this choice because we had to provide one single result for the challenge. We did not use data augmentation in our training strategy to be able to compare different oversampling strategies and momentum, but OHEM comportment should be explored with data augmentation in future work. Due to the short delay between baseline and follow-up MRIs in the MICCAI MSSEG-2 dataset (from 1 to 3 years) as well as in our second testing dataset (maximum 120 days), we could not explore the influence of severe atrophy in this task. An adjacent and clinically useful task for longitudinal follow-up of MS patients, that we could not assess here due to challenge constraints focusing on new lesions, is the detection of shrinking and enlarging lesions. Furthermore, it is likely that the use of multicontrast MRI could improve the results over the use of FLAIR alone. The aim of the MICCAI 2021 MSSEG-2 challenge was to develop an algorithm only based on two longitudinal FLAIRs. Thus, our present work only uses FLAIR as input and a comparison with a multicontrast input is left for future work. Another important aspect that remains to be studied is generalizability to other acquisition settings. In the MICCAI MSSEG-2 challenge, there was quite a variety of different MRI machines. Furthermore, it is interesting to note that the General Electric machines present in the MICCAI MSSEG-2 dataset were not present in the training dataset. However, further experiments, which could not be performed within the challenge setting, would be required to demonstrate generalizability across acquisition settings. Future work will be to go further into dealing with class imbalance during training with a fixed oversampling strategy, as it gave interesting results on the validation set and in other pipelines of the challenge. The difficulty with a fixed oversampling strategy is the arbitrary choice of the oversampling factor. Perhaps inserting neurological priors to guide the oversampling factors and adapting them to the anatomical region could be a promising idea, allowing to take into account the complexity of prediction in some brain areas and the variability of the lesion load over brain regions in MS to tune locally the probability of patches from those regions to be oversampled.

Conclusion

In this paper, we described our contribution to the MICCAI MSSEG-2 challenge (Commowick et al., 2021). The main new methodological component was the use of online hard example mining (OHEM) for handling class imbalance. Overall, on the challenge testing set, our pipeline ranked at the mid-class, with an average Dice of 0.400 and an average F_1 score of 0.446. For this specific application, on the validation set, OHEM did not provide any improvement over a standard fixed oversampling strategy. Nevertheless, such a strategy may deserve further investigation for medical imaging problems with class imbalance.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://portal.fli-iam.irisa.fr/msseg-2/data/.

Ethics statement

The studies involving human participants were reviewed and approved by OFSEP: https://www.ofsep.org/en. The patients/participants provided their written informed consent to participate in this study.

Author contributions

MS-M and TS contributed to the pipeline implementation and the manuscript redaction. MH and AY-P contributed to algorithm training, submission, and manuscript redaction. BB and BS contributed with clinical advice and revision. NA, BS, and OC contributed with implementation advice, work supervision, and manuscript revision. All authors contributed to the article and approved the submitted version.

Funding

The research leading to these results has received funding from the French government under management of Agence Nationale de la Recherche as part of the Investissements d'avenir program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), reference ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6), and reference number ANR-19-P3IA-0002 (3IA Côte d'Azur) and from ICM under the Big Brain Theory program (project IMAGIN-DEAL in MS-M). This work was supported by the Fondation pour la Recherche Médicale, Grant No. FDM202006011247 to TS and by the Fondation Sorbonne Université to MH.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Altay, E. E., Fisher, E., Jones, S. E., Hara-Cleaver, C., Lee, J. C., Rudick, R. A., et al. (2013). Reliability of classifying multiple sclerosis disease activity using magnetic resonance imaging in a multiple sclerosis clinic. *JAMA Neurol.* 70, 338. doi: 10.1001/2013.jamaneurol.211

Bian, C., Yang, X., Ma, J., Zheng, S., Liu, Y. A., Nezafat, R., et al. (2022). "Pyramid network with online hard example mining for accurate left atrium segmentation," in *International Workshop on Statistical Atlases and Computational Models of the Heart*, (Cham: Springer), 237–245. doi: 10.1007/978-3-030-12029-0_26

Bodini, B., Veronese, M., García-Lorenzo, D., Battaglini, M., Poirion, E., Chardain, A., et al. (2016). Dynamic I maging of I ndividual R emyelination P rofiles in M ultiple S clerosis. *Ann. Neurol.* 79, 726–738. doi: 10.1002/ana.24620

Cabezas, M., Corral, J. F., Oliver, A., Díez, Y., Tintoré, M., Auger, C., et al. (2016). Improved automatic detection of new t2 lesions in multiple sclerosis using deformation fields. *Am. J. Neuroradiol.* 37, 1816–1823. doi: 10.3174/ajnr.A4829

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2022). 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. Published online June 21, 2016. Available online at: http://arxiv.org/abs/1606.06650 (accessed June 28, 2022).

Commowick, O., Cervenansky, F., Cotton, F., and Dojat, M. (2021). "MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure," in *MICCAI 2021-24th International Conference on Medical Image Computing and Computer Assisted Intervention*, 1–118.

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., et al. (2018). Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Sci. Rep.* 8, 13650. doi: 10.1038/s41598-018-31911-7

Filippi, M., Brück, W., Chard, D., Fazekas, F., Geurts, J. J., Enzinger, C., et al. (2019). Association between pathological and MRI findings in multiple sclerosis. *Lancet Neurol.* 18, 198–210. doi: 10.1016/S1474-4422(18)30451-4

Gessert, N., Krüger, J., Opfer, R., Ostwaldt, A. C., Manogaran, P., Kitzler, H. H., et al. (2020). Multiple sclerosis lesion activity segmentation with attention-guided two-path CNNs. *Comput. Med. Imaging Graph.* 84, 101772. doi: 10.1016/j.compmedimag.2020.101772

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). "Momentum contrast for unsupervised visual representation learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE), 9726–9735. doi: 10.1109/CVPR42600.2020.00975

Hegen, H., Bsteh, G., and Berger, T. (2018). 'No evidence of disease activity' - is it an appropriate surrogate in multiple sclerosis? *Eur. J. Neurol.* 25, 1107–e101. doi: 10.1111/ene.13669

Howard, J., Trevick, S., and Younger, D. S. (2016). Epidemiology of multiple sclerosis. *Neurol. Clin.* 34, 919–939. doi: 10.1016/j.ncl.2016.06.016

Jenkinson, M., and Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* 5, 143–156. doi: 10.1016/S1361-8415(01)00036-6

Kingma, D. P., and Ba, J. (2022). Adam: A Method for Stochastic Optimization. Published online January 29, 2017. Available online at: http://arxiv.org/abs/1412.6980 (accessed June 28, 2022).

McKinley, R., Wepfer, R., Grunder, L., Aschwanden, F., Fischer, T., et al. (2020). Automatic detection of lesion load change in Multiple Sclerosis using convolutional neural networks with segmentation confidence. *NeuroImage Clin.* 25, 102104. doi: 10.1016/j.nicl.2019.102104

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). "Automatic differentiation in PyTorch," in *NIPS 2017 Workshop Autodiff Submission*.

Pérez-García, F., Sparks, R., and Ourselin, S. (2021). TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Comput. Methods Progr. Biomed.* 208, 106236. doi: 10.1016/j.cmpb.2021.106236

Salem, M., Valverde, S., Cabezas, M., Pareto, D., Oliver, A., et al. (2020). A fully convolutional neural network for new T2-w lesion detection in multiple sclerosis. *NeuroImage Clin.* 25, 102149. doi: 10.1016/j.nicl.2019.102149

Shrivastava, A., Gupta, A., and Girshick, R. (2016). "Training region-based object detectors with online hard example mining," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE), 761–769. doi: 10.1109/CVPR.2016.89

Valverde, S., Salem, M., Cabezas, M., Pareto, D., Vilanova, J. C., Ramió-Torrentà, L., et al. (2019). One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. *NeuroImage Clin.* 21, 101638. doi: 10.1016/j.nicl.2018.101638

Wang, X., Ma, H., and Chen, X. (2016). "Salient object detection via fast R-CNN and low-level cues," in *2016 IEEE International Conference on Image Processing (ICIP)* (IEEE), 1042–1046. doi: 10.1109/ICIP.2016.7532516

Wolny, A., Cerrone, L., Vijayan, A., Tofanelli, R., Barro, A. V., Louveaux, M., et al. (2020). Accurate and versatile 3D segmentation of plant tissues at cellular resolution. *eLife.* 9, e57613. doi: 10.7554/eLife.57613

Zeng, C., Gu, L., Liu, Z., and Zhao, S. (2020). Review of Deep Learning Approaches for the Segmentation of Multiple Sclerosis Lesions on Brain MRI. *Front Neuroinformatics.* 14, 610967. doi: 10.3389/fninf.2020.610967

Appendix

We used the MICCAI MSSEG-2 dataset (Commowick et al., 2021), consisting in 100 MS patients with two longitudinal FLAIR MRI spaced from 1 to 3 years, acquired with 6 Philips scanners (Ingenia 1.5T, 2 Ingenia 3T, 1 Achieva dStream 3T, 1 Achieva 1.5T, 1 Achieva 3T), 6 Siemens scanners (1 Aera 1.5T, 1 Skyra 3T, 1 Verio 3T, 1 Prisma 3T, 2 Avanto 1.5T), and 3 General Electrics (GE) scanners (Optima MR450w 1.5T, SIGNA HDx 3T, SIGNA HDxt 1.5T), with different voxel sizes (from 0.5 to 1.2 mm^3). Ground truth, consisting in new lesions on second time point, were delineated by 4 neuroradiologists from different centers manually on ITK-SNAP (http://www.itksnap.org/pmwiki/pmwiki.php), and consensus was obtained with the majority voting for each voxel. The whole dataset was divided by MSSEG-2 training committee into 40 patients available to challengers for training and validation, and 60 patients, not available to the challengers, for testing. All MRIs acquired with GE were only in the testing dataset.

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 24 November 2022
DOI 10.3389/fnins.2022.1007619

Check for updates

OPEN ACCESS

EDITED BY
Michel Dojat,
Institut National de la Santé et de la Recherche Médicale (INSERM), France

REVIEWED BY
Russell Ouellette,
Karolinska Institutet (KI), Sweden
Roland Opfer,
Jung Diagnostics GmbH, Germany

*CORRESPONDENCE
Mostafa Salem
mostafasalem@aun.edu.eg

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 30 July 2022
ACCEPTED 24 October 2022
PUBLISHED 24 November 2022

CITATION
Salem M, Ryan MA, Oliver A, Hussain KF and Lladó X (2022) Improving the detection of new lesions in multiple sclerosis with a cascaded 3D fully convolutional neural network approach. *Front. Neurosci.* 16:1007619. doi: 10.3389/fnins.2022.1007619

Improving the detection of new lesions in multiple sclerosis with a cascaded 3D fully convolutional neural network approach

Mostafa Salem[1,2]*, Marwa Ahmed Ryan[1,2], Arnau Oliver[1], Khaled Fathy Hussain[2] and Xavier Lladó[1]

[1]Research Institute of Computer Vision and Robotics, University of Girona, Girona, Spain,
[2]Department of Computer Science, Faculty of Computers and Information, Assiut University, Assiut, Egypt

Longitudinal magnetic resonance imaging (MRI) has an important role in multiple sclerosis (MS) diagnosis and follow-up. Specifically, the presence of new lesions on brain MRI scans is considered a robust predictive biomarker for the disease progression. New lesions are a high-impact prognostic factor to predict evolution to MS or risk of disability accumulation over time. However, the detection of this disease activity is performed visually by comparing the follow-up and baseline scans. Due to the presence of small lesions, misregistration, and high inter-/intra-observer variability, this detection of new lesions is prone to errors. In this direction, one of the last Medical Image Computing and Computer Assisted Intervention (MICCAI) challenges was dealing with this automatic new lesion quantification. The *MSSEG-2: MS new lesions segmentation challenge* offers an evaluation framework for this new lesion segmentation task with a large database (100 patients, each with two-time points) compiled from the OFSEP (Observatoire français de la sclérose en plaques) cohort, the French MS registry, including 3D T2-w fluid-attenuated inversion recovery (T2-FLAIR) images from different centers and scanners. Apart from a change in centers, MRI scanners, and acquisition protocols, there are more challenges that hinder the automated detection process of new lesions such as the need for large annotated datasets, which may be not easily available, or the fact that new lesions are small areas producing a class imbalance problem that could bias trained models toward the non-lesion class. In this article, we present a novel automated method for new lesion detection of MS patient images. Our approach is based on a cascade of two 3D patch-wise fully convolutional neural networks (FCNNs). The first FCNN is trained to be more sensitive revealing possible candidate new lesion voxels, while the second FCNN is trained to reduce the number of misclassified voxels coming from the first network. 3D T2-FLAIR images from the two-time points were pre-processed and linearly co-registered. Afterward, a fully CNN, where its inputs were only the baseline and follow-up images, was trained to detect new MS lesions. Our approach obtained a mean segmentation dice similarity coefficient of 0.42 with a detection F1-score of 0.5. Compared to the challenge participants, we obtained one of the highest precision scores (PPVL = 0.52), the best PPVL rate (0.53), and a lesion detection sensitivity (SensL of 0.53).

KEYWORDS

brain, MRI, multiple sclerosis, automatic new lesion detection, deep learning, learning-based registration, cascaded training

1. Introduction

Multiple sclerosis (MS) is an inflammatory disease of the central nervous system and spinal cord, with its etiology remains elusive. The progression of the disease starts almost in all cases with an inflammatory syndrome in the CNS, demyelination, and axonal loss when the immune system mistakenly starts to attack the protective myelin sheath in the brain. Due to the nature of the MS disease, no drugs offer neuroprotection when progression is observed (Ther et al., 2022), although they help to decrease the myelin loss ratio. MRI imaging techniques are one of the first choices to be used in clinical practice as reported in the 2017 revision of the McDonald criteria (McDonald et al., 2001; Thompson et al., 2018), because of their ability to detect the early stages of the disease. MS is detected in patients who have not developed clinically apparent neurological disabilities 5–10 times more frequently on conventional MRI than in the clinical assessment of relapses (Sahraian and Eshaghi, 2010). MS Lesion count and volume are very important indicators for MS diagnosis and progression and have been associated with the long-term outcome of the disease (Goodin et al., 2012; Uher et al., 2017; Ouellette et al., 2018). According to Rovira et al. (2015), patients with clinical and radiological MS findings that have not been diagnosed as patients with MS must undergo a follow-up brain MRI. On longitudinal analysis, new lesions are considered a high-impact prognostic factor for MS evolution prediction and risk of disability accumulation over time (Tintore et al., 2015). Furthermore, there is a need for a lesion quantification approach for the computation of the volumetric changes in each segmented lesion between two-time points for the MS lesion evolution (Köhler et al., 2019). Manual delineation of lesion load in brain volume should be the first choice during diagnosis, but a large number of MRI slices and different scanning modalities prevent it, due to being a time-consuming procedure with large intra- and inter-rater variability (Altay et al., 2013; Egger et al., 2017). Therefore, there is an increase in the demand for automatic methods to provide fast, more robust, and reliable results, specially for the computation of lesion volumetric changes between two-time points (Köhler et al., 2019)

Many methods were proposed to automatically detect the lesion load in MRI scans (Valverde et al., 2017b; Zhang et al., 2019) and even to review the improvements in the cross-sectional field (Lladó et al., 2012; Zeng et al., 2020; Shoeibi et al., 2021). Detecting changes in longitudinal analysis for new or enlarging lesions in the follow-up scan compared to the baseline was done initially with traditional image pre-processing tools. Based on the intensity subtraction between successive time points, Sweeney et al. (2013) used logistic regression coefficients to automatically model changes over time. Also, the work of Elliott et al. (2013) incorporated both spatial and temporal information in a two-stage classifier starting with the extraction of relevant features and brain tissues and used this information to finally segment lesions. In Battaglini et al. (2014) and Ganiler et al. (2014) authors relied on thresholding the subtraction of follow-up and baseline images. By taking the changes in surrounding tissue in mind and not depending only on the intensity change, deformation field-based methods were proposed to detect lesion change (Cabezas et al., 2016; Salem et al., 2018). Relying on segmenting both time points independently, Schmidt et al. (2019) extended their work on cross-sectional (Schmidt et al., 2012) in a new pipeline to provide lesion evolution patterns. Moreover, Jain et al. (2016), based on a joint expectation-maximization (EM) framework, used the subtraction of the two-time points and cross-sectional masks of follow-up and baseline to get the longitudinal changes. Krüger et al. (2020) used a shared encoder based on a 3D CNN to process both baseline and follow-up images. The outputs of the encoders were concatenated and passed to the decoder to detect the new or enlarged lesions that appear in the follow-up images. Most traditional methods depend on the manual threshold or mask subtraction which is affected by the required registration process and could not provide results comparable to those of human raters.

The recent advance in processing methods and shift made by artificial intelligence and deep learning methods, specially convolution neural networks (CNNs) and its ability to extract features, have made them one of the first choices to implement novel approaches. For instance, the first use of CNN in MS longitudinal data was proposed by Birenbaum and Greenspan (2016) to reduce false positives after candidate selection, obtaining segmentation accuracies near to a human rater. Inspired by the work of Balakrishnan et al. (2019) to compute the deformation field (DF), Salem et al. (2020) developed a new approach to simultaneously learn the nonlinear DF between follow-up and baseline and from the learned DF and input images learn the segmentation mask. Denner et al. (2021) used the same shared encoder and different decoders to learn the tasks of segmentation and non-rigid registration. To improve the lesion map segmentation, Gessert et al. (2020) extended the 4D context by adding a temporal history and adding convGRU to aggregate the 3D representations from encoders to be passed to the decoder for the final prediction map. Despite the increased demand for new lines in longitudinal studies, work was still hindered by no reference benchmark for proposed methods. Most methods mentioned previously were trained and evaluated on in-house data or no public code was available for comparisons among methods. To overcome this limitation, the *MICCAI Multiple Sclerosis new lesion segmentation (MSSEG-2) challenge* was proposed, offering a new opportunity to progress within this research and a public performance benchmark dataset.

In this article, we present a new pipeline for automated new lesion detection of MS patient images based on a cascade of two fully convolutional neural networks (FCNNs). The first FCNN, a filter for misclassified voxels, is used to discard the vast majority of negative voxels, while the second one is used to deal with more challenging voxels that were misclassified from the first FCNN and with the high unbalancing lesion voxels compared with background, specially hard in longitudinal data due to the few change in follow-up images (i.e., few lesions). The proposed architecture builds on an initial prototype that we presented at the MSSEG-2 challenge (Commowick et al., 2021). Other works exist either in other domains as coronary calcium segmentation (Wolterink et al., 2016), liver lesions in CT scans (Christ et al., 2016), or even based on CNN models in the MS domain such as the work of Valverde et al. (2017a), which used a cascaded CNN in cross-sectional lesion detection. The proposed pipeline was trained and tested with the MSSEG-2 challenge dataset. The results were obtained using the Anima[1] toolbox. The same measures for the challenge (detection/segmentation) are reported and compared with the rest of the participants.

2. Methods

The main basic block in our segmentation pipeline is the U-Net (Ronneberger et al., 2015; Çiçek et al., 2016), which proved its performance in segmentation tasks, especially in the medical area. One of the advantages the U-Net has provided to the medical community is the ability to use a small sample to create highly detailed segmentation maps, adopted in different medical applications and obtaining the best performance in medical challenges (Siddique et al., 2021). Due to its context-based learning in the two-path architecture of contracting and expansion paths, the network training is faster and provides more accurate results than other segmentation models. In this article, 3D patches were chosen to benefit from the spatial contextual information in 3D MRI and let the network deal with input of any size without the need to re-sample or resize images, which can suffer from information loss, or lesion deformation, especially in the smaller ones.

2.1. Cascade-based training

In general, training a model for the detection of small lesions, where the number of lesion voxels is much less than non-lesion voxels, makes the model biased to the non-lesion class. However, the problem is even more challenging in the new lesion change detection scenario, where the few changes in the follow-up images may be insufficient to train the model.

To tackle this class imbalance problem, we propose to perform the following patch extraction strategy around the lesion voxels (see Figure 1A):

1. Extract all lesion voxels in the training images,
2. Patches of size $32 \times 32 \times 32$ are extracted around every selected voxel in both baseline and follow-up images and stacked for the T2-w fluid-attenuated inversion recovery (T2-FLAIR) modality provided in the MSSEG-2 challenge.
3. FCNN_1 is trained with the selected patches (details of the model available in Section 2.2).
4. Overlapped patches are extracted and tested using the trained FCNN_1 to get the probability Y_1. The probability threshold (>0.5) is used to calculate the lesion map. Also, small lesions (<3 mm^3) are removed.
5. Based on the calculated lesion map, new patches are extracted with $32 \times 32 \times 32$ size and step $8 \times 8 \times 8$ around the lesion area and the misclassified lesion by FCNN_1.
6. The second network (FCNN_2) is trained from scratch with the newly extracted patches.
7. The output probability from the trained FCNN_1 (Y_1) is averaged with the output of the trained FCNN_2 (Y_2) to get the final lesion probability mask. To obtain the final segmentation mask, we threshold the voxel probability > 0.5 and remove the small lesions ($< 3mm^3$).

2.2. Network architecture

The FCNN used in our work for both FCNN_1 and FCNN_2 is shown in Figure 1B. It follows the most recent proposed architecture by Salem et al. (2020). The network is a fully CNN that takes the T2-FLAIR image modality in both baseline and follow-up as inputs and outputs of the new lesion segmentation mask. The network consists of two parts as shown in Figure 1C. The first part is a U-Net block that automatically learns the DF

[1] https://anima.irisa.fr/

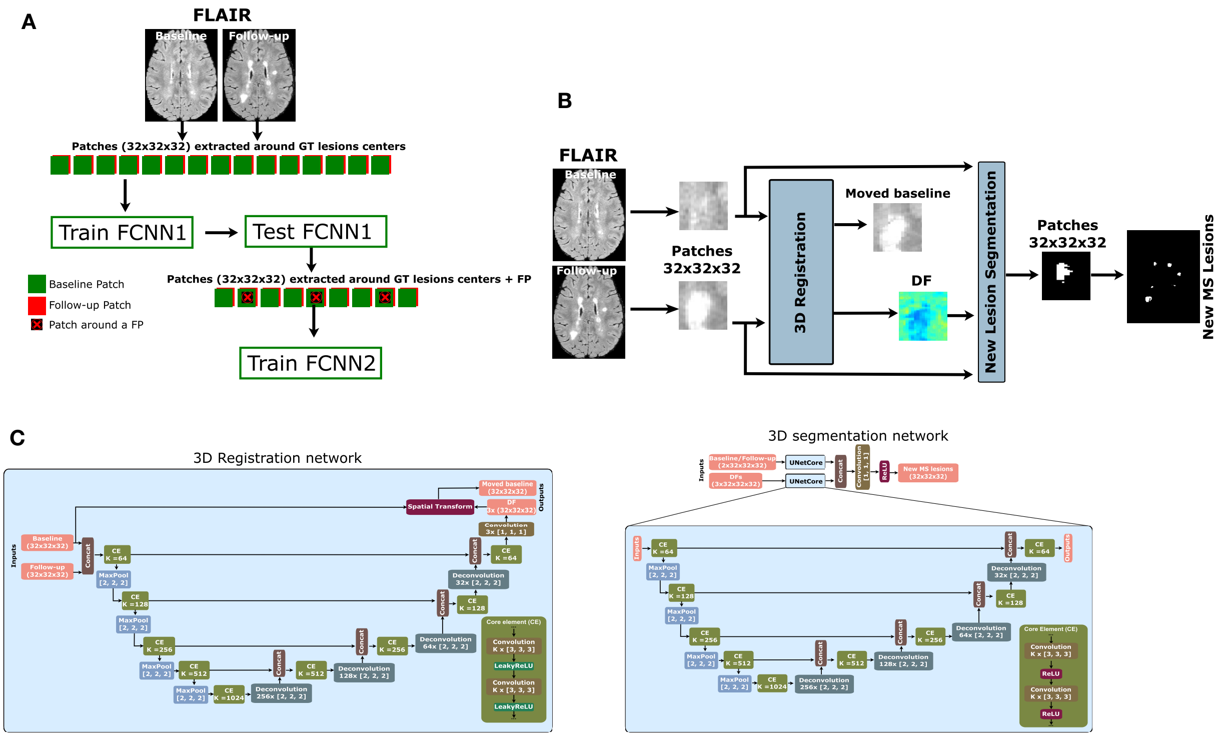


FIGURE 1
Proposed pipeline for new MS lesion detection. **(A)** Cascade-based pipeline, where the output of the first FCNN is used to select the input features of the second FCNN. **(B)** The proposed network consists of a 3D registration block and a 3D segmentation block. The inputs are baseline/follow-up images of the T2-FLAIR modality. The 3D registration block learns the deformation field (DF) and non-linearly registers the baseline image to the follow-up image. Afterward, the learned DF and the baseline and follow-up images are fed to the segmentation block, which performs the final detection and segmentation of the new lesions. The network is trained end-to-end using a combined loss function. **(C)** The 3D registration and segmentation architectures (see Salem et al., 2020 for more details).

that non-linearly registers the T2-FLAIR baseline image to the follow-up space. The learned DF and the baseline and follow-up images are then fed to a second part of the network, another U-Net that performs the detection and segments of the new lesions. The network is trained end-to-end with gradient descent and simultaneously learns both DF and new lesion segmentation. This model was updated for the MSSEG-2 challenge dataset and sent to the challenge (referred to as Vicorob).

3D registration architecture: A 3D registration block is built for the T2-FLAIR modality following the architecture explained in Salem et al. (2020). This block is inspired by VoxelMorph, a learning framework for deformable medical image registration (Balakrishnan et al., 2019). The registration block learns the DF that non-linearly registers the T2-FLAIR baseline image to the follow-up space. It is a fully convolutional network that follows a U-shaped architecture (Ronneberger et al., 2015). The U-Net architecture consists of four downsample (the contracting path) and upsample steps (the expansive path). The core element (CE) block is a two 3D convolution layer (kernel size = 3 and stride = 1) with K channels. Each convolution is followed by a LeakyReLU layer. The number of channels, K, of CE blocks is (64, 128, 256, and 512) and (512, 256, 128, and 64) for the contracting path and expansive path, respectively. The spatial transformation (Jaderberg et al., 2015; Balakrishnan et al., 2019) warps the baseline image to the follow-up image using the learned DF and enabling end-to-end training. The LeakyReLU activations are used instead of ReLU so that the learned DFs can have both positive and negative values (see Salem et al., 2020 for more details).

3D segmentation architecture: A 3D segmentation CNN is also used for segmenting the new lesions. It is a two-branch network where each branch is a U-Net following the architecture explained in Salem et al. (2020). The U-Net architecture is exactly the same as the U-Net used in the registration block, but uses a ReLU activation layer instead of the LeakyReLU layer. The inputs of the first branch are the T2-FLAIR image modality in both baseline and follow-up, while the second branch input is the DF learned from the first registration block. The outputs of the two branches are concatenated before the classification step.

2.3. Loss functions

The loss function used in this work consists of the summation of an unsupervised and a supervised loss functions. The unsupervised loss function controls the registration part of the network (Balakrishnan et al., 2019). It consists of two components: a similarity part that penalizes differences in appearance between the moved baseline and follow-up images combined with a regularization part that enforces a spatially smooth deformation and often is modeled as a linear operator on the spatial gradients of DF, as stated in Balakrishnan et al. (2019). The supervised function, $L_{CrossEntropy}$ (CrossEntropy), controls the segmentation part of the network and penalizes differences between the segmentation and ground truth. Therefore, the total loss function L_{Total} is:

$$L_{Total} = \underbrace{L_{CrossEntropy}(Seg, GT)}_{Segmentation\ loss\ function} + \underbrace{\sum_{m \in Modalities} \left(\overbrace{\frac{1}{N} \sum_{i=1}^{N} (F_{m_i} - B_m(DF_m)_i)^2}^{Similarity\ part} + \overbrace{\sum_{p \in DF} \| \triangledown DF_m(p) \|^2}^{Regularization\ part} \right)}_{Registration\ loss\ function} \tag{1}$$

where F_m, $B_m(DF_m)$, and DF_m are follow-up image, baseline image warped by DF (moved baseline), and DF for a modality m, respectively. *Seg* and *GT* are the automatic segmentation and the ground truth, respectively.

2.4. Model training

To adjust the weights of the cascaded pipeline, each network is trained individually. For FCNN_1 to be more sensitive with lesion voxels candidate, patches of size 32×32×32 are extracted around lesion voxels. For FCNN_2, the model is trained with more challenging voxels, which were wrongly classified with FCNN_1. Patches of size 32×32×32 and step size 8×8×8 are extracted in the area of lesion voxels and incorrectly predicted lesions from FCNN_1.

For training the pipeline, patches are extracted from the challenge's 40 patient volumes (the training set), with 25% of the selected patches used to validate the model after each epoch and to adjust the hyper-parameters. To adjust the pipeline weights, training is held for 100 epochs, with early stopping when no decrease was detected in the model validation loss after 10 epochs.

2.5. Model testing

When the pipeline training is completed, the weights can be used with the unseen data. The overlapped extracted patches from the T2-FLAIR modality in the baseline and follow-up images and the weights of FCNN_1 were used to get the probability P_1, then the same extracted patches are fed to FCNN_2 to get P_2. The average of the two probabilities is computed and threshold by $>$ 0.5 to get a binary mask. The final binary mask is obtained after removing the isolated voxels (region volume $< 3mm^3$). Figure 2 shows the cascade architecture for the testing procedure.

2.6. Implementation details

The proposed method has been implemented in Python[2], using Keras[3] with the TensorFlow[4] backend (Abadi et al., 2015). All experiments have been run on a GNU/Linux machine box running Ubuntu 18.04, with 128 GB RAM. The training was carried out on a single TITAN X GPU (NVIDIA Corp, United States) with 12 GB RAM. To promote the reproducibility and usability of our research, the proposed cascade new MS lesion detection pipeline will be available for downloading at our research website.

3. Experimental setup

3.1. Dataset

3.1.1. MSSEG-2

The database used in this article is the MSSEG-2 challenge dataset. A total of 100 patients with MS were gathered. Only a 3D T2-FLAIR sequence at the first timepoint and a 3D T2-FLAIR sequence at a second timepoint (from 1 to 3 years after the first one) are available. A total of 15 different MRI scanners are represented (nine scans from three GE scanners with field strength 1.5T and 3T, 63 scans from six Philips scanners with field strength 1.5T and 3T, and 28 scans from six Siemens scanners with field strength 1.5T and 3T). The image characteristics vary with different resolutions and different voxel sizes (from 0.5 mm^3 to 1.2 mm^3). The gathered data are separated according to 40 scans (11 scans with no new lesions detected in the second timepoint) for training and 60 (28 scans

[2] https://www.python.org

[3] https://keras.io

[4] https://www.tensorflow.org/

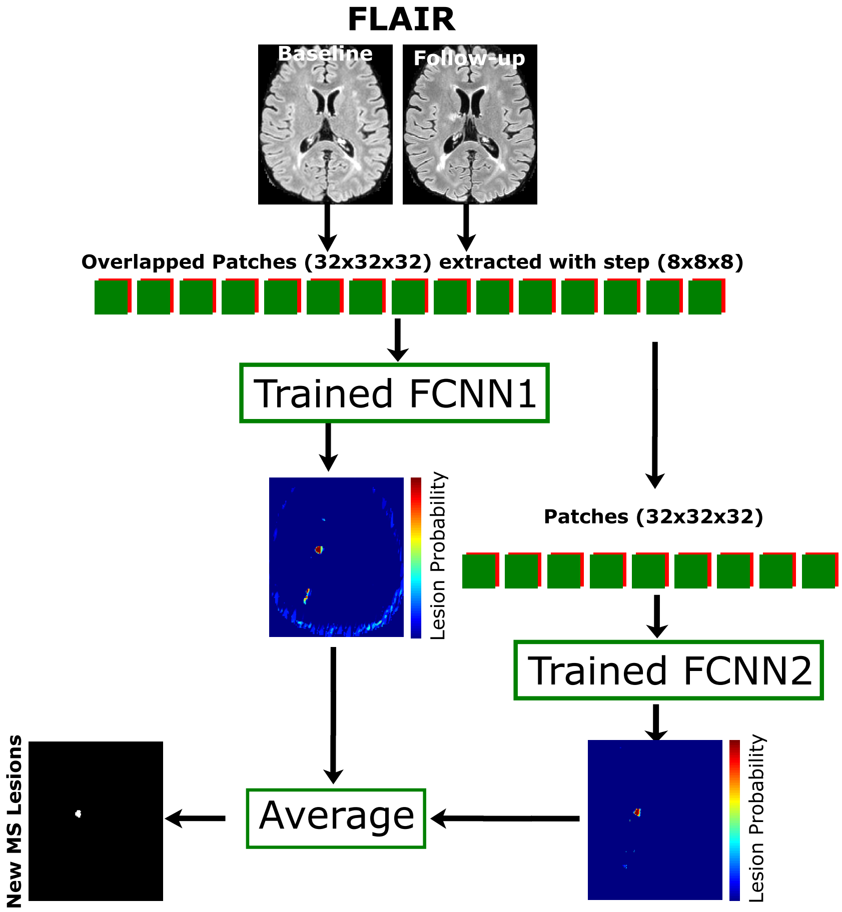


FIGURE 2
Proposed testing process. The cascade architecture of the trained network is used to segment the unseen data. Patches of size 32×32×32 are extracted from input modalities (baseline and follow-up) with step size 8×8×8 and fed to both FCNN_1 and FCNN_2. The average probability mask from both networks is thresholded with a minimum connected component (<3 mm^3) to get the final lesion mask.

with no new lesions detected in the second timepoint) for testing. All data from GE scanners have been excluded from the training set.

3.1.2. Pre-processing

The MSSEG-2 challenge dataset is available with a rigid registration already performed to bring the two-time points of each patient to a common middle point. For each patient, the same pre-processing steps were performed on both baseline and follow-up images. First, a brain mask was identified and delineated using the ROBEX Tool (Iglesias et al., 2011). Second, the T2-FLAIR images underwent a bias field correction step using the N4 algorithm from the ITK library. Finally, the baseline and follow-up intensity values from all the training sets were normalized using a histogram-matching approach based on Nyúl et al. (2000).

3.2. Evaluation

The MSSEG-2 challenge performance evaluation consists of two levels as follows:

- **New lesion detection**: how many individual new lesions in the ground truth were detected by the evaluated method, independently of the precision of their contours. F1-score was chosen for this criteria.
- **New lesion segmentation**: how well are the lesions in the ground truth overlapping with those of the evaluated method. Dice measure has been selected as a score in these criteria.

The Anima[5] toolbox, used by the challenge organizers for evaluation, is also used in all our evaluations

[5] https://anima.irisa.fr/

(animaSegPerfAnalyzer). Similar to the challenge, the evaluation of lesion detection and segmentation metrics were calculated using only 32 patients from the 60 scans provided for evaluation (only patients with at least one new lesion in the follow-up). The main metric for evaluating the detection of the new lesions is the F1-score, but we also computed the precision and recall, computed as follows:

$$\text{F1-score} = \frac{2 \cdot TP}{FN + FP + 2 \cdot TP}$$

$$\text{PPVL} = \frac{TP}{TP + FP}$$

$$\text{SensL} = \frac{TP}{TP + FN}$$

where PPVL denotes the model precision (the fraction of real lesions among the predicted ones) and SensL denotes model sensitivity or recall (the fraction of real lesions that were predicted). To evaluate the model performance in the cases with no new lesions detected at the follow-up image, the average volume (in mm^3) of incorrectly predicted lesions is added to the *VolTested* measure.

The main metric to evaluate the segmentation is the dice score (DSC), which is the equivalent of the F1-score on a voxel level, and is computed as follows:

$$DSC = \frac{2 \cdot TP_s}{FN_s + FP_s + 2 \cdot TP_s}$$

In segmentation, TP_s and FP_s denote the number of voxels correctly and incorrectly predicted as lesions, respectively, and FN_s represents the number of voxels incorrectly predicted as non-lesion.

To evaluate the significance of the obtained results, we used paired *t*-tests at a 5% level of confidence.

The following models were analyzed, aiming to show the benefits of the registration step:

- **VicorobCascade**: This is our main cascade-based model in which the registration block and segmentation block are trained simultaneously end-to-end using the loss function explained in Section 2.3. The T2-FLAIR image modality in both baseline and follow-up combined with the learned DF is fed to the segmentation block as first and second inputs, respectively.
- **DemonsDFCascade** (a.k.a. the proposed cascade-based network using the DF obtained from Demons Thirion, 1998): This model does not use the registration blocks of the proposed network shown in Figure 1B. It uses only the segmentation block with the T2-FLAIR image modality in both baseline and follow-up as the first input. The second input of the segmentation block is the DF directly computed by registering the baseline to the follow-up space for the T2-FLAIR modality using the multi-resolution Demons registration approach from ITK (Thirion, 1998). This model was used for comparison with the VicorobCascade model to highlight the impact of learned-based DF with end-to-end training over the DF from Demons.
- **NoDFCascade** (a.k.a. the proposed cascade-based network without DF): This model does not use the registration block of the proposed network shown in Figure 1B. It uses only the segmentation block with just the T2-FLAIR image modality in both baseline and follow-up as input. This model is used for comparison with the other two models to highlight the impact of the addition of the DF in increasing the detection of new lesions.

In addition to the above models, the **non-cascade** version of the three models was added to compare the normal 3D patch-based training with our proposed cascade-based training pipeline discussed in Section 2.1. Note that our original submission to the challenge is referred to here as Vicorob.

4. Results

Table 1 shows the F1-score, DSC, PPVL, and SensL of the proposed pipeline (VicorobCascade), the two variants (DemonsDFCascade, NoDfCascade), and the non-cascade version of each model. Results show the improvement achieved in evaluation metrics by using the cascaded-based pipeline over normal (no-cascade-based) training one. In addition, the results show the benefits of using DF and also the superiority of our cascade VicorobCascade model, where deformation fields are learned simultaneously with new lesion detection.

Figures 3, 4 show visual examples of the improvement of the VicorobCascade model with respect to the other evaluated models. In the figures, each column corresponds to the baseline T2-FLAIR image, the follow-up T2-FLAIR image, the NoDF, NoDFCascade, DemonsDF, DemonsDFCascade, Vicorob, and VicorobCascade prediction masks, and the ground truth mask. Figure 3 shows improvement in the sensitivity of the model, while Figure 4 shows improvement in precision.

Analyzing the results per patient, Figure 5 shows a box plot summarizing the performance of the VicorobCascade, the two variants (DemonsDFCascade, NoDFCascade), and the no-cascade-based version of the three models on the four metrics used in the evaluation (F1-score, DSC, PPVL, and SensL). The results show again the superiority of the VicorobCascade over the other methods.

TABLE 1 Lesion detection and segmentation results on the MSSEG-2 challenge test set: Comparison between the different models evaluated.

Method	F1-score	Dice	PPVL	SensL
Vicorob	36.88 ± 29.21	35.83 ± 30.53	34.28 ± 30.22	49.80 ± 39.49
VicorobCascade	**49.97** ± 36.75	**41.97** ± 31.51	**51.86** ± 39.31	52.74 ± 39.62
DemonsDF	31.21 ± 34.68	29.08 ± 29.16	35.20 ± 39.07	36.92 ± 38.58
DemonsDFCascade	45.59 ± 35.65	41.84 ± 30.98	46.51 ± 38.55	**55.70** ± 37.82
NoDF	23.97 ± 30.54	25.75 ± 28.58	34.12 ± 41.95	27.84 ± 34.81
NoDfCascade	43.30 ± 34.24	39.86 ± 29.19	46.12 ± 38.55	52.43 ± 40.58

The results represent the mean F1-score, DSC, PPVL, and SensL computed by the segmentation performance analyzer tool available in Anima (animaSegPerfAnalyzer). Best values are depicted in bold.

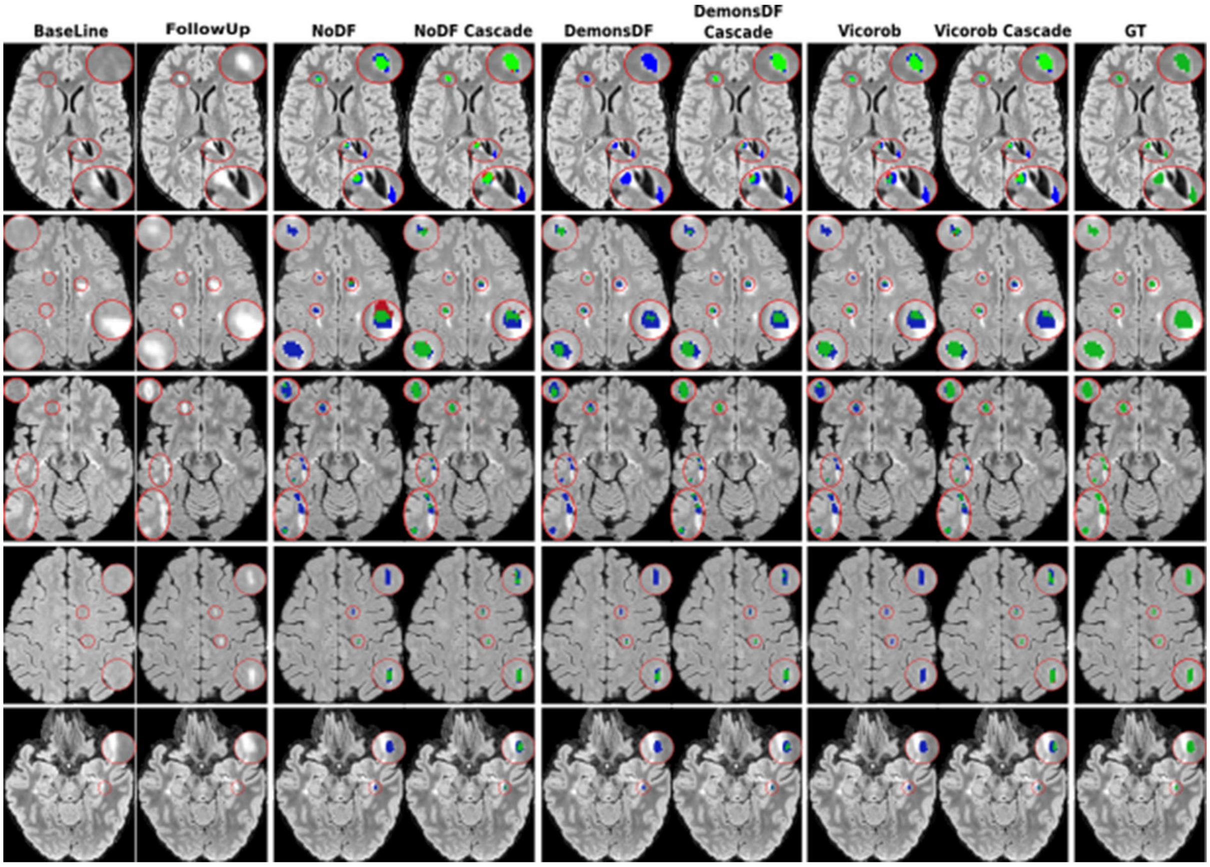


FIGURE 3
Examples of new lesion detection sensitivity improvement in axial slices. Columns correspond to baseline T2-FLAIR, follow-up T2-FLAIR and the predicted segmentation masks over follow-up T2-FLAIR for NoDF, NoDFCascade, DemonsDF, DemonsDFCascade, Vicorob, and VicorobCascade, respectively, along with the consensus ground truth (GT) mask, overlaid in green. For the predicted segmentation masks, green, red, and blue represent true positives, false positives, and false negatives, respectively.

Challenge results

The model previously submitted to the challenge under Vicorob team (referred to Vicorob) and our new cascade-based pipelines (VicorobCascade) are compared with the other challenge participants (29 pipelines for 24 teams submitted to the challenge). Figures 6, 7 show the boxplot summarizing the performance F1-score and PPVL per patient, respectively.

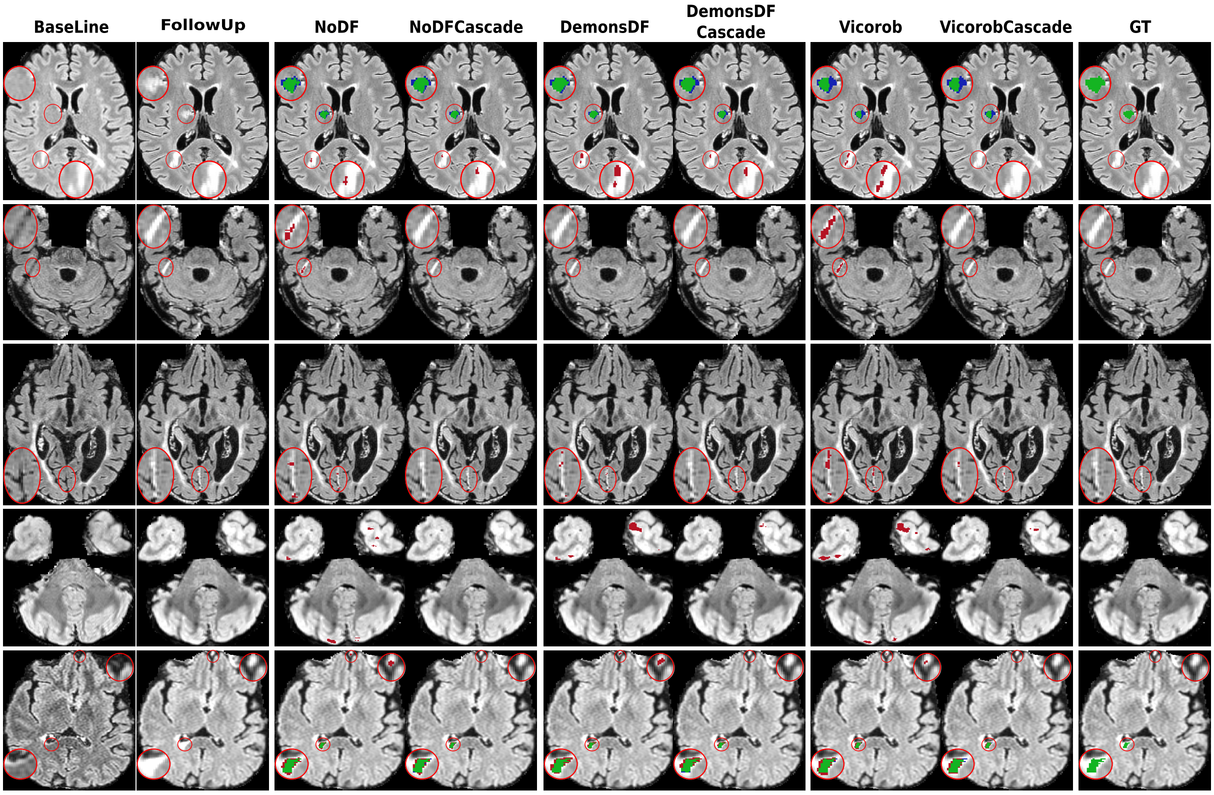


FIGURE 4
Examples of new lesion detection precision improvement in axial slices. Columns correspond to baseline T2-FLAIR, follow-up T2-FLAIR, and the predicted segmentation masks over follow-up T2-FLAIR for NoDF, NoDFCascade, DemonsDF, DemonsDFCascade, Vicorob, and VicorobCascade, respectively, along with the consensus ground truth (GT) mask, overlaid in green. For the predicted segmentation masks, green, red, and blue represent true positives, false positives, and false negatives, respectively.

5. Discussion and future work

In this article, we have proposed a novel automated new lesions detection approach in longitudinal brain MR images. The proposed patch-wise pipeline relies on a cascade of two identical FCNNs, where the first network is trained to be more sensitive revealing possible candidate lesion voxels, while the second network is trained to reduce the number of misclassified voxels coming from the first network output. As mentioned in Salem et al. (2020), the model is trained end-to-end and simultaneously learns both the DF and the appearance of new lesions. As the DF is learned inside the network and not computed separately using classic non-rigid registration methods, the execution time of the network on a testing image is reduced compared to the time required by the state-of-the-art methods (Cabezas et al., 2016; Salem et al., 2018) from 2 to 11 min according to the test image resolution.

Regarding the end-to-end training, we trained the proposed model (VicorobCascade), two other variants (DemonsDFCascade and NoDFCascade), and the no-cascade-based version of the three models. Regarding the results without cascading, in terms of F1-score, DSC, and SensL, the Vicorob model was significantly better than all the other methods ($p < 0.05$). The F1-score improved by 5.67% compared to the DemonsDF and by 12.91% with respect to the NoDF model. In terms of PPVL, however, the performance of the Vicorob model was similar to that of the DemonsDF, although both models provided better results than the NoDF model. Notice that the model trained without any DF (NoDF) detected new lesions with a sensitivity of 27.84% and an F1-score of 23.97%. This result shows, as previously discussed in Salem et al. (2020), that the addition of DF helps to increase the detection of new lesions. However, the results also show that training the model end-to-end, simultaneously learning both the DF and the new lesions (Vicorob pipeline), performs better than using DF computed by classic deformable registration methods such as Demons (Thirion, 1998).

Regarding the cascade-based training, the proposed pipeline using two FCNN outperforms the results obtained with the baseline (no-cascade-based) approaches. The reported results show that the cascaded proposed pipeline outperformed the baseline (no-cascade-based) pipeline in all the proposed

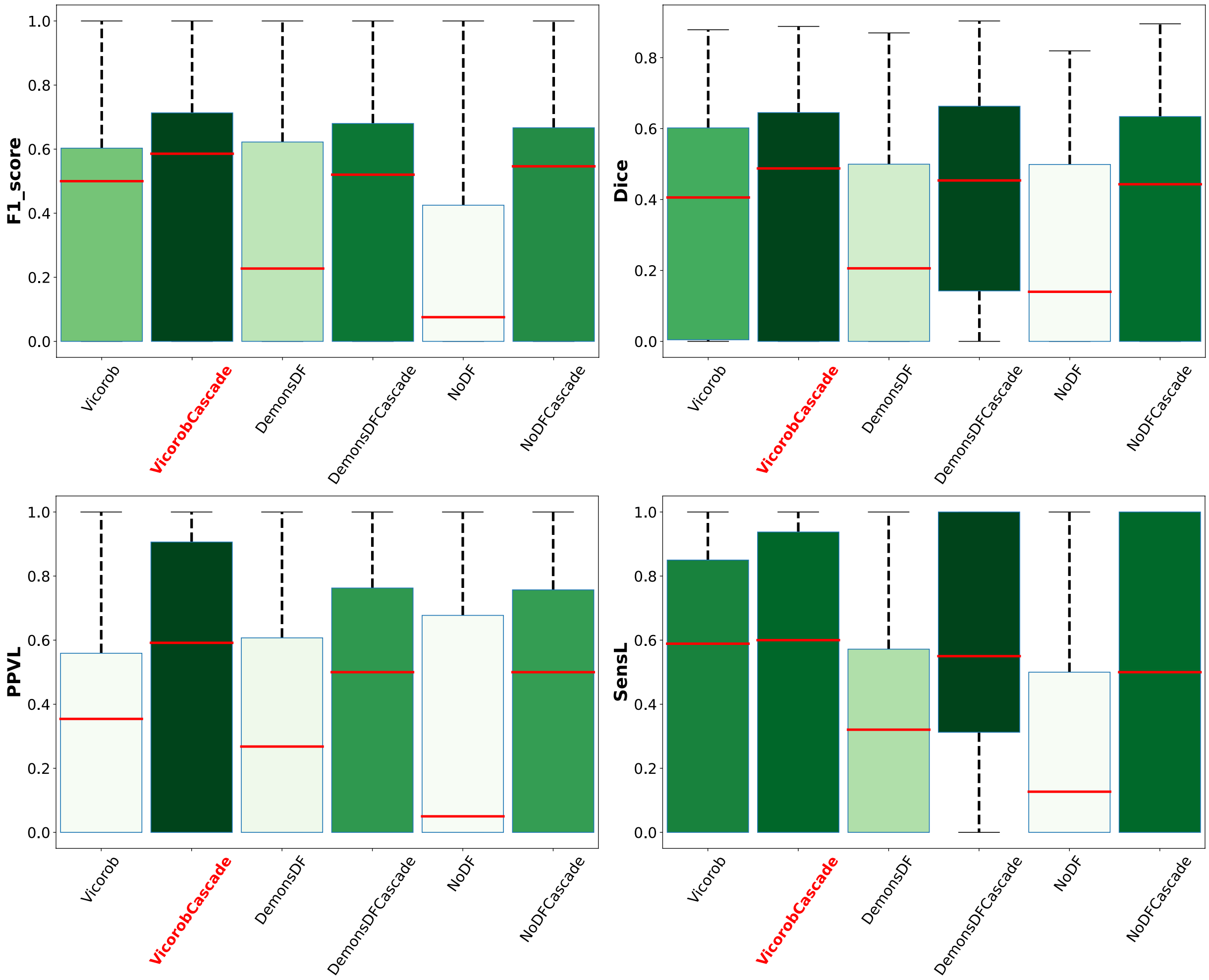


FIGURE 5
Box plot summarizing the per-patient performance of the VicorobCascade, the two variants (DemonsDFCascade, NoDFCascade), and the no-cascade-based version of the three models on the four metrics used in the evaluation (F1-score, DSC, PPVL, and SensL).

Vicorob, DemonsDF, and NoDF models for all the segmentation and detection metrics and showed also the superiority of our VicorobCascade model. The F1-score was significantly improved by 13.9%, 14.38%, and 20.85% for the Vicorob, DemonsDF, and NoDF models ($p < 0.05$), respectively. Moreover, Figure 3 shows a sensitivity improvement in the evaluated models. Notice that there is an increase in the number of true positive voxels (green ones) and decreasing in the number of false negative voxels (blue ones) between the non-cascaded and the cascaded-based models. Figure 4 shows a precision improvement for the VicorobCascade model. Notice also that there is a decrease in the number of false positive lesions compared to the other models. Regarding the cases with no new lesions, VolTested decreased from 88.40 mm^3 for the Vicorob model to 11.56 mm^3 for the VicorobCascade model.

Regarding the challenge results and compared to the challenge participants, our model (VicorobCascade) obtains one of the highest precision scores (PPVL = 0.52), the best PPVL rate (0.53), and a lesion detection sensitivity (SensL of 0.53) being superior to that of one of the challenge's human raters. Analyzing the results per scanner, the VicorobCascade model provided an F1-score of 0.22, 0.54, and 0.51 for GE, Philips, and Siemens scanners, respectively. Notice that the lower results for the GE scanner are due to the fact that data from this particular scanner were not available in the MSSEG-2 training set. Within this analysis, we also observed that the cascade-based approach obtained better results than the no-cascade one for the three scanners. Notice that there is a limitation in dealing with different image domains when data are not available. Furthermore, a clinical correlation with disability measurements could enrich the clinical evaluation

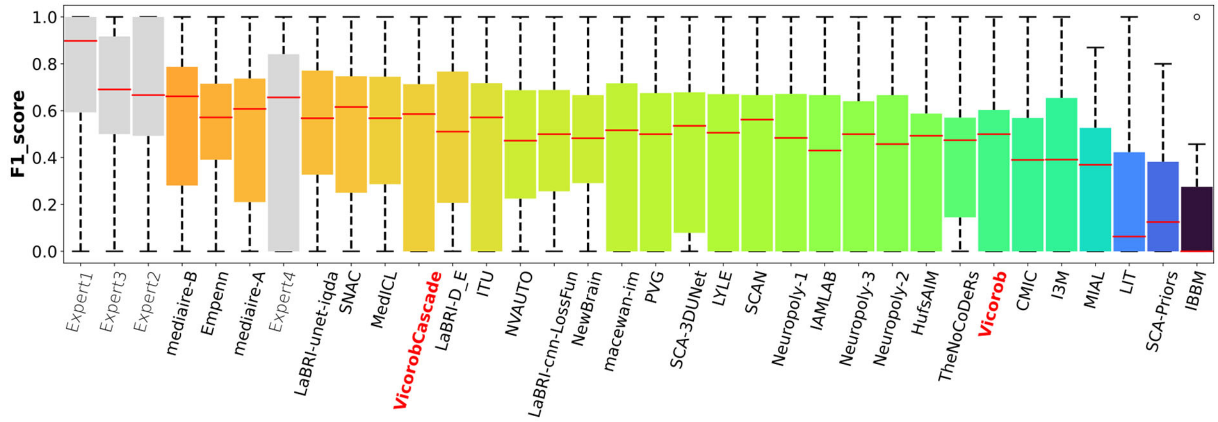


FIGURE 6
F1-score per-patient analysis. F1-score for the MSSEG-2 challenge experts, challenge teams' results, and our cascade-based pipeline (VicorobCascade).

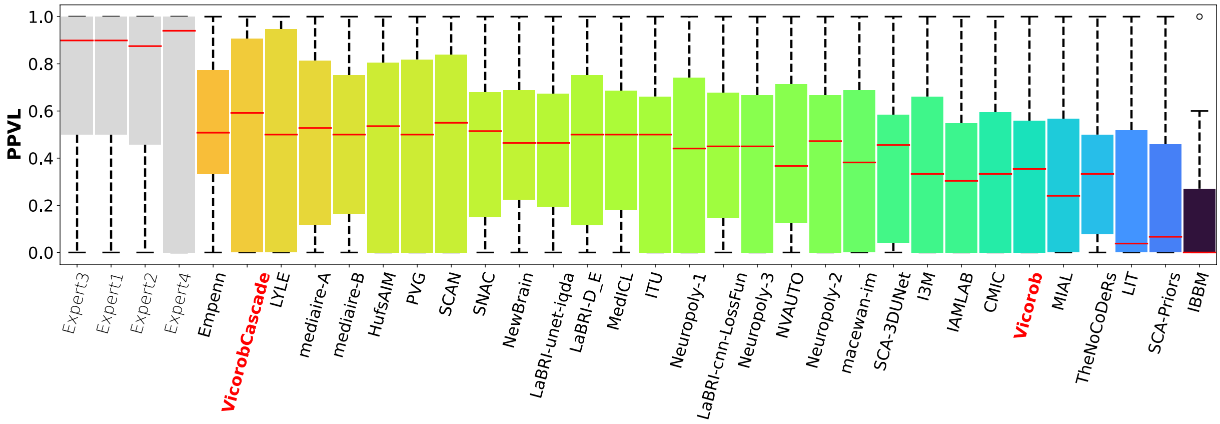


FIGURE 7
PPVL per-patient analysis. PPVL for the MSSEG-2 challenge experts, challenge teams' results, and our cascade-based pipeline (VicorobCascade). The VicorobCascade model got one of the best PPVL values between teams after the Empenn team.

of the automated segmentation results. Unfortunately, the MSSEG-2 challenge dataset does not include these clinical disability metrics. This will be taken into account in our future research work.

In conclusion, we have presented a novel approach for longitudinal analysis in patients with MS based on a cascade of two FCNNs, where the first one is able to find the potential candidates and the second one is optimized to detect new lesions and reduce the number of false positives. The obtained results indicate that the proposed end-to-end training model of the deformation fields along with the detection of new lesions combined within the cascade-based training pipeline increases the accuracy of the pipeline. Given the sensitivity and limited number of false positives, we strongly believe that the proposed method has the potential to be used in clinical studies in order to monitor the progression of the disease. We plan to release the proposed method for downloading at our research website.

Data availability statement

The data for training and testing all the presented pipelines was obtained as part of the MICCAI 2021 MSSEG-2 challenge (https://portal.fli-iam.irisa.fr/msseg-2/data/). Access was restricted to challenge participants. Requests to access these datasets should be directed to challenges-iam@inria.fr.

Ethics statement

The studies involving human participants were reviewed and approved by MICCAI 2021 MSSEG-2 challenge. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

MS: thinking about the main idea, writing code, running experiments, writing the manuscript, and revision. MR: writing code, running experiments, and writing the manuscript. KH: revision. AO and XL: write the manuscript and revision. All authors contributed to the article and approved the submitted version.

Funding

This work has been supported by DPI2020-114769RB-I00 from the Ministerio de Ciencia, Innovación y Universidades. The authors gratefully acknowledge the support of the NVIDIA Corporation with their donation of the TITAN X GPU used in this research. This work has been also supported by ICREA Academia program.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from https://tensorflow.org.

Altay, E. E., Fisher, E., Jones, S. E., Hara-Cleaver, C., Lee, J.-C., and Rudick, R. A. (2013). Reliability of classifying multiple sclerosis disease activity using magnetic resonance imaging in a multiple sclerosis clinic. *JAMA Neurol.* 70, 338–344. doi: 10.1001/2013.jamaneurol.211

Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., and Dalca, A. V. (2019). Voxelmorph: a learning framework for deformable medical image registration. *IEEE Trans. Med. Imaging* 38, 1788–1800. doi: 10.1109/TMI.2019.2897538

Battaglini, M., Rossi, F., Grove, R. A., Stromillo, M. L., Whitcher, B., Matthews, P. M., et al. (2014). Automated identification of brain new lesions in multiple sclerosis using subtraction images. *J. Magn. Reson. Imaging* 39, 1543–1549. doi: 10.1002/jmri.24293

Birenbaum, A., and Greenspan, H. (2016). "Longitudinal multiple sclerosis lesion segmentation using multi-view convolutional neural networks," in *2nd International Workshop on Deep Learning in Medical Image Analysis, DLMIA 2016* (Athens), 58–67.

Cabezas, M., Corral, J., Oliver, A., Díez, Y., Tintoré, M., Auger, C., et al. (2016). Improved automatic detection of new t2 lesions in multiple sclerosis using deformation fields. *Am. J. Neuroradiol.* 37, 1816–1823. doi: 10.3174/ajnr.A4829

Christ, P. F., Elshaer, M. E. A., Ettlinger, F., Tatavarty, S., Bickel, M., Bilic, P., et al. (2016). Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields. *Lecture Notes Comput. Sci.* 9901, 415–423. doi: 10.1007/978-3-319-46723-8_48

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3D U-net: learning dense volumetric segmentation from sparse annotation. *Lecture Notes Comput. Sci.* 9901, 424–432. doi: 10.1007/978-3-319-46723-8_49

Commowick, O., Cervenansky, F., Cotton, F., and Dojat, M. (2021). "MSSEG-2 challenge proceedings: multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure," in *MICCAI 2021-24th International Conference on Medical Image Computing and Computer Assisted Intervention* (Strasbourg), 126.

Denner, S., Khakzar, A., Sajid, M., Saleh, M., Spiclin, Z., Kim, S. T., et al. (2021). Spatio-temporal learning from longitudinal data for multiple sclerosis lesion segmentation. *Lecture Notes Comput. Sci.* 12658, 111–121. doi: 10.1007/978-3-030-72084-1_11

Egger, C., Opfer, R., Wang, C., Kepp, T., Sormani, M. P., Spies, L., et al. (2017). MRI FLAIR lesion segmentation in multiple sclerosis: Does automated segmentation hold up with manual annotation? *Neuroimage Clin.* 13, 264–270. doi: 10.1016/j.nicl.2016.11.020

Elliott, C., Arnold, D. L., Collins, D. L., and Arbel, T. (2013). Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain MRI. *IEEE Trans. Med. Imaging* 32, 1490–1503. doi: 10.1109/TMI.2013.2258403

Ganiler, O., Oliver, A., Diez, Y., Freixenet, J., Vilanova, J. C., Beltran, B., et al. (2014). A subtraction pipeline for automatic detection of new appearing multiple sclerosis lesions in longitudinal studies. *Neuroradiology* 56, 363–374. doi: 10.1007/s00234-014-1343-1

Gessert, N., Bengs, M., Krüger, J., Opfer, R., Ostwaldt, A.-C. C., Manogaran, P., et al. (2020). 4D deep learning for multiple sclerosis lesion activity segmentation. *arXiv* 1–5. doi: 10.48550/arXiv.2004.09216

Goodin, D. S., Traboulsee, A., Knappertz, V., Reder, A. T., Li, D., Langdon, D., et al. (2012). Relationship between early clinical characteristics and long term disability outcomes: 16 year cohort study (follow-up) of the pivotal interferon β-1b trial in multiple sclerosis. *J. Neurol. Neurosurg. Psychiatry* 83, 282–287. doi: 10.1136/jnnp-2011-301178

Iglesias, J. E., Liu, C.-Y., Thompson, P. M., and Tu, Z. (2011). Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Trans. Med. Imaging* 30, 1617–1634. doi: 10.1109/TMI.2011.2138152

Jaderberg, M., Simonyan, K., and Zisserman, A. (2015). "Spatial transformer networks," in *Advances in Neural Information Processing Systems* (Montreal, QC), 2017–2025.

Jain, S., Ribbens, A., Sima, D. M., Cambron, M., De Keyser, J., Wang, C., et al. (2016). Two time point MS lesion segmentation in brain MRI: an expectation-maximization framework. *Front. Neurosci.* 10, 576. doi: 10.3389/fnins.2016.00576

Köhler, C., Wahl, H., Ziemssen, T., Linn, J., and Kitzler, H. H. (2019). Exploring individual multiple sclerosis lesion volume change over time: development of an algorithm for the analyses of longitudinal quantitative mri measures. *Neuroimage Clin.* 21, 101623. doi: 10.1016/j.nicl.2018.101623

Krüger, J., Opfer, R., Gessert, N., Ostwaldt, A.-C., Manogaran, P., Kitzler, H. H., et al. (2020). Fully automated longitudinal segmentation of new or enlarged multiple sclerosis lesions using 3d convolutional neural networks. *Neuroimage Clin.* 28, 102445. doi: 10.1016/j.nicl.2020.102445

Lladó, X., Ganiler, O., Oliver, A., Martí, R., Freixenet, J., Valls, L., et al. (2012). Automated detection of multiple sclerosis lesions in serial brain MRI. *Neuroradiology* 54, 787–807. doi: 10.1007/s00234-011-0992-6

McDonald, W. I., Compston, A., Edan, G., Goodkin, D., Hartung, H.-P., Lublin, F. D., et al. (2001). Recommended diagnostic criteria for multiple sclerosis: guidelines from the international panel on the diagnosis of multiple sclerosis. *Ann. Neurol.* 50, 121–127. doi: 10.1002/ana.1032

Nyúl, L. G., Udupa, J. K., and Zhang, X. (2000). New variants of a method of MRI scale standardization. *IEEE Trans. Med. Imaging* 19, 143–150. doi: 10.1109/42.836373

Ouellette, R., Bergendal, Å. A., Shams, S., Martola, J., Mainero, C., Wiberg, M. K., et al. (2018). Lesion accumulation is predictive of long-term cognitive decline in multiple sclerosis. *Mult. Scler. Relat. Disord.* 21, 110–116. doi: 10.1016/j.msard.2018.03.002

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. *Med. Image Comput. Comput. Assisted Intervent.* 2015, 234–241. doi: 10.1007/978-3-319-24574-4_28

Rovira, À., Wattjes, M. P., Tintoré, M., Tur, C., Yousry, T. A., Sormani, M. P., et al. (2015). Magnims consensus guidelines on the use of mri in multiple sclerosis–clinical implementation in the diagnostic process. *Nat. Rev. Neurol.* 11, 471–482. doi: 10.1038/nrneurol.2015.106

Sahraian, M. A., and Eshaghi, A. (2010). Role of MRI in diagnosis and treatment of multiple sclerosis. *Clin. Neurol Neurosurg.* 112, 609–615. doi: 10.1016/j.clineuro.2010.03.022

Salem, M., Cabezas, M., Valverde, S., Pareto, D., Oliver, A., Salvi, J., et al. (2018). A supervised framework with intensity subtraction and deformation field features for the detection of new t2-w lesions in multiple sclerosis. *Neuroimage Clin.* 17, 607–615. doi: 10.1016/j.nicl.2017.11.015

Salem, M., Valverde, S., Cabezas, M., Pareto, D., Oliver, A., Salvi, J., et al. (2020). A fully convolutional neural network for new T2-w lesion detection in multiple sclerosis. *Neuroimage Clin.* 25, 102149. doi: 10.1016/j.nicl.2019.102149

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., et al. (2012). An automated tool for detection of flair-hyperintense white-matter lesions in multiple sclerosis. *Neuroimage* 59, 3774–3783. doi: 10.1016/j.neuroimage.2011.11.032

Schmidt, P., Pongratz, V., Küster, P., Meier, D., Wuerfel, J., Lukas, C., et al. (2019). Automated segmentation of changes in flair-hyperintense white matter lesions in multiple sclerosis on serial magnetic resonance imaging. *Neuroimage Clin.* 23, 101849. doi: 10.1016/j.nicl.2019.101849

Shoeibi, A., Khodatars, M., Jafari, M., and Moridian, P. (2021). Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: a review. *arXiv[Preprint].arXiv:2105.04881.* doi: 10.1016/j.compbiomed.2021.104697

Siddique, N., Paheding, S., Elkin, C. P., and Devabhaktuni, V. (2021). U-net and its variants for medical image segmentation: a review of theory and applications. *IEEE Access* 9, 82031–82057. doi: 10.1109/ACCESS.2021.3086020

Sweeney, E., Shinohara, R., Shea, C., Reich, D. S., and Crainiceanu, C. M. (2013). Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal MRI. *Am. J. Neuroradiol.* 34, 68–73. doi: 10.3174/ajnr.A3172

Ther, N., Collongues, N., Becker, G., Biname, F., Ayme-dietrich, E., Patte-mensah, C., et al. (2022). A narrative review on axonal neuroprotection in multiple sclerosis. *Neurol. Therapy* 11, 981–1042. doi: 10.1007/s40120-022-00363-7

Thirion, J.-P. (1998). Image matching as a diffusion process: an analogy with maxwell's demons. *Med. Image Anal.* 2, 243–260. doi: 10.1016/S1361-8415(98)80022-4

Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., et al. (2018). Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol.* 17, 162–173. doi: 10.1016/S1474-4422(17)30470-2

Tintore, M., Rovira, À., Río, J., Otero-Romero, S., Arrambide, G., Tur, C., et al. (2015). Defining high, medium and low impact prognostic factors for developing multiple sclerosis. *Brain* 138, 1863–1874. doi: 10.1093/brain/awv105

Uher, T., Vaneckova, M., Sobisek, L., Tyblova, M., Seidl, Z., Krasensky, J., et al. (2017). Combining clinical and magnetic resonance imaging markers enhances prediction of 12-year disability in multiple sclerosis. *Multiple Sclerosis J.* 23, 51–61. doi: 10.1177/1352458516642314

Valverde, S., Cabezas, M., Roura, E., González-Villà, S., Pareto, D., Vilanova, J. C., et al. (2017a). Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *Neuroimage* 155, 159–168. doi: 10.1016/j.neuroimage.2017.04.034

Valverde, S., Oliver, A., Roura, E., González-Villà, S., Pareto, D., Vilanova, J. C., et al. (2017b). Automated tissue segmentation of MR brain images in the presence of white matter lesions. *Med. Image Anal.* 35, 446–457. doi: 10.1016/j.media.2016.08.014

Wolterink, J. M., Leiner, T., de Vos, B. D., van Hamersvelt, R. W., Viergever, M. A., and Išgum, I. (2016). Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks. *Med. Image Anal.* 34, 123–136. doi: 10.1016/j.media.2016.04.004

Zeng, C., Gu, L., Liu, Z., and Zhao, S. (2020). Review of deep learning approaches for the segmentation of multiple sclerosis lesions on brain MRI. *Front. Neuroinf.* 14, 610967. doi: 10.3389/fninf.2020.610967

Zhang, H., Valcarcel, A. M., Bakshi, R., Chu, R., Bagnato, F., Shinohara, R. T., et al. (2019). Multiple sclerosis lesion segmentation with tiramisu and 2.5D stacked slices. In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science, vol 11766.* Shenzhen: Springer.

frontiers | Frontiers in Neuroimaging

TYPE Original Research
PUBLISHED 22 December 2022
DOI 10.3389/fnimg.2022.1008128

Check for updates

OPEN ACCESS

EDITED BY
Benoit Combès,
Inria Rennes-Bretagne Atlantique Research Centre, France

REVIEWED BY
Stefano Cerri,
Technical University of Denmark, Denmark
Renaud Hedouin,
Inria Rennes-Bretagne Atlantique Research Centre, France

*CORRESPONDENCE
Denis Fortun
✉ dfortun@unistra.fr

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroimaging

RECEIVED 31 July 2022
ACCEPTED 06 December 2022
PUBLISHED 22 December 2022

CITATION
Dufresne E, Fortun D, Kremer S and Noblet V (2022) A unified framework for focal intensity change detection and deformable image registration. Application to the monitoring of multiple sclerosis lesions in longitudinal 3D brain MRI. *Front. Neuroimaging* 1:1008128. doi: 10.3389/fnimg.2022.1008128

A unified framework for focal intensity change detection and deformable image registration. Application to the monitoring of multiple sclerosis lesions in longitudinal 3D brain MRI

Eléonore Dufresne[1], Denis Fortun[1]*, Stéphane Kremer[1,2] and Vincent Noblet[1]

[1]ICube UMR 7357, Université de Strasbourg, CNRS, Strasbourg, France, [2]Hôpitaux Universitaires de Strasbourg, Strasbourg, France

Registration is a crucial step in the design of automatic change detection methods dedicated to longitudinal brain MRI. Even small registration inaccuracies can significantly deteriorate the detection performance by introducing numerous spurious detections. Rigid or affine registration are usually considered to align baseline and follow-up scans, as a pre-processing step before applying a change detection method. In the context of multiple sclerosis, using deformable registration can be required to capture the complex deformations due to brain atrophy. However, non-rigid registration can alter the shape of appearing and evolving lesions while minimizing the dissimilarity between the two images. To overcome this issue, we consider registration and change detection as intertwined problems that should be solved jointly. To this end, we formulate these two separate tasks as a single optimization problem involving a unique energy that models their coupling. We focus on intensity-based change detection and registration, but the approach is versatile and could be extended to other modeling choices. We show experimentally on synthetic and real data that the proposed joint approach overcomes the limitations of the sequential scheme.

KEYWORDS

deformable 3D registration, change detection, longitudinal analysis, multiple sclerosis, joint minimization, alternating direction method of multipliers (ADMM)

1. Introduction

Multiple sclerosis (MS) is an auto-immune neurodegenerative disease characterized by the inflammation of the myelin coating that surrounds the nerves. As a consequence, the transmission of nervous impulses is impaired, causing motor, cognitive and sensorial disabilities. The evolution of MS is characterized by the apparition of focal lesions in the brain and in the spinal cord, and by a progressive atrophy of brain tissues. Both phenomena can be monitored thanks to Magnetic Resonance Imaging (MRI) (Kaunzner and Gauthier, 2017). In the clinical routine, the evolution of the lesion load and of the brain atrophy is generally assessed qualitatively. However, the precise quantification of lesion changes over time may be of great interest to finely characterize the course of the pathology and to evaluate at the early stage the effect of a therapeutic strategy (McNamara et al., 2017). Since the manual delineation of lesion changes in MRI is a tedious and time consuming task, which is prone to both intra- and inter-observer variability, there is a great need for efficient and reliable automated tools (Altay et al., 2013).

Most change detection methods dedicated to lesion monitoring rely on a sequential scheme that first consists in removing all changes that are not of interest in order to detect in a second step only the evolution of lesions (Radke et al., 2005). To correct for global intensity changes induced by the difference of MRI acquisition setups, algorithms for bias field inhomogeneity correction (Song et al., 2017) and histogram-based intensity normalization procedure (Shinohara et al., 2014) are generally considered. Then, geometrical discrepancies due to variation in patient positioning, acquisition-related geometrical distortion and brain atrophy are corrected thanks to registration algorithms involving either rigid, affine or deformable transformations. Finally, the remaining changes corresponding to the evolution of lesions are detected. This final step generally consists in thresholding an intensity-based (Bosc et al., 2003) or deformation-based (Rey et al., 2002) feature map. The threshold can be chosen according to some statistical modeling in order to control the expected number of false positive detections (Rousseau et al., 2007).

The main flaw of the sequential procedure is that it implicitly assumes that each correction step can be performed while not being influenced by the changes that remain to be corrected in the next steps. As a consequence, the sequence order of the correction procedures should be carefully chosen. Moreover, for each correction step, a trade-off should be found between its performance (i.e., the ability of the method to accurately and specifically correct a given kind of change) and its robustness (i.e., the ability not to be biased by another kind of remaining changes). This observation advocates for a unified formulation of the change detection problem allowing to estimate all the different kinds of changes jointly.

In this paper, we address more specifically the interplay between deformable image registration and focal intensity change detection. When deformable registration is performed in the presence of appearing lesions, the estimated transformation tends to make these new lesions disappear in order to minimize the dissimilarity between the two images. This is the reason why the most common practice is to consider only rigid or affine registration in order not to alter lesion shape. However, such linear transforms can only compensate for difference in patient positioning but are not able to capture the complex deformations induced by brain atrophy, which typically occurs in MS. These remaining deformations may yield to spurious detections in atrophied areas, especially in the cortex and around the ventricles.

We propose to account for the intertwining of deformable registration and focal intensity change detection by estimating them jointly. To this end, we show that these two separate tasks can be formulated as a single optimization problem involving a unique energy that models their coupling. Basically, areas corresponding to detected changes are ignored in the registration similarity criterion, which prevents the lesion elimination effect described above. Solving this issue allows us to use of deformable registration, which in turn prevents from detecting spurious changes in atrophied areas. We propose an efficient alternating optimization scheme to solve this unified optimization problem. We focus on demonstrating the benefits of this joint formulation in the particular case of a standard intensity-based data similarity criterion. Nevertheless, the proposed approach is versatile and could easily be extended to more elaborated modeling choices. Experimental analysis is performed on the BrainWeb synthetic dataset and on two annotated real datasets. We first demonstrate in each case the benefits of considering a deformable registration as compared to an affine registration only in order to reduce the number of false detections. Then, we highlight the benefit of the proposed joint formulation as compared to the standard sequential scheme in terms of change detection accuracy. A preliminary version of this work has been published as a conference paper in Dufresne et al. (2020). In this paper, we provide a more extensive experimental analysis, which helps to better characterize the behavior of the proposed method and better understand why it outperforms the sequential approach.

The paper is organized as follows. In Section 2, the sequential approach, which will be considered as the reference baseline method of this work, is described and its limitations are discussed. In Section 3, we introduce the proposed joint model and the optmization strategies that has been set up estimate both change detection map and deformation field. In Section 4, we give implementation details. Finally, we present and discuss experimental results in Section 5.2.

2. The sequential approach

The conventional sequential approach consists of three main steps. First, the images are corrected for global intensity variations, then they are spatially registered, and finally the focal intensity changes due to the evolution of lesions are detected. In this section, we give a brief overview of the common practices in the registration and change detection steps. Our goal is not to cover a comprehensive scope of the field but to formulate the general principles underlying existing methods. In the remainder of this article, we will denote $I_1, I_2 : \Omega \rightarrow \mathbb{R}$ the baseline and follow-up MRI images, where $\Omega \subset \mathbb{R}^3$ is the image domain.

2.1. Registration

The registration problem can be formulated as:

$$\widehat{\mathrm{w}} = \underset{\mathrm{w}}{\operatorname{argmin}} \sum_{\mathrm{x}\in\Omega} \rho(I_1, I_2, \mathrm{w}, \mathrm{x}) + \lambda_1 \Psi(\mathrm{w}), \tag{1}$$

where $\mathrm{w} : \Omega \to \mathbb{R}^3$ represents the transformation, $\rho(\cdot)$ is a data similarity term, and $\Psi(\cdot)$ is a regularizer weighted by a scalar $\lambda_1 > 0$. An overview of deformable registration methods in medical imaging can be found in Sotiras et al. (2013).

Several transformation models can be considered relying either on a parametric representation (e.g., rigid, affine, polynomial, or Bspline-based) or on a non-parametric deformable mapping (i.e., a displacement vector is estimated for each voxel).

The role of the data term is to penalize dissimilarity between I_1 and I_2 warped with the estimated transformation. For monomodal registration, it is common to use intensity-based measures such as the sum of squared intensity differences or the cross correlation. In the multimodal case, mutual information is one of the most widely used similarity metric (Kaunzner and Gauthier, 2017).

In the context of deformable registration, considering the data term only can lead to an ill-posed problem. To overcome this issue, the data term has to be balanced with an additional regularization term $\Psi(\mathrm{w})$ that enforces some constraints on the deformation field. For instance, penalizing the ℓ_2 or ℓ_1 norm of the gradient of w helps to promote smooth solutions.

2.2. Change detection

The change detection step generally consists in thresholding a map of feature differences between registered baseline and follow-up images (see the survey Lladó et al., 2012). These maps can be calculated directly from either intensity-based (Sweeney et al., 2013; Ganiler et al., 2014; Cabezas et al., 2016), or deformation-based (Rey et al., 2002; Cabezas et al., 2016; Salem et al., 2018) features, and sometimes integrate other kind of information (Elliott et al., 2013; Sweeney et al., 2013) .

In the perspective of integrating both registration and change detection in a single joint optimization problem, we advocate that they should both rely on the same data similarity. Consequently, we model the binary change map $c : \Omega \to \{0, 1\}$ defined at each voxel x as follow:

$$c(\mathrm{x}) = \begin{cases} 0 & \text{if } \rho(I_1, I_2, \mathrm{w}, \mathrm{x}) \le \lambda_2 \\ 1 & \text{otherwise,} \end{cases} \tag{2}$$

$\lambda_2 \in \mathbb{R}^{+*}$ being the detection threshold. The thresholding scheme (Equation 2) can be reformulated as the following optimization problem:

$$\widehat{c}(\mathrm{x}) = \underset{c\,:\,\Omega\to\{0,1\}}{\operatorname{argmin}} \sum_{\mathrm{x}\in\Omega} (1 - c(\mathrm{x}))\, \rho(I_1, I_2, \mathrm{w}, \mathrm{x}) + \lambda_2 c(\mathrm{x}). \tag{3}$$

Since simple thresholding can yield noisy results, most MS lesion change detection methods also integrate a denoising step in post-processing to obtain the final change map. The denoising can be realized jointly with the change detection by integrating a regularization term $\Phi(\cdot)$ in Equation (3):

$$\widehat{c}(\mathrm{x}) = \underset{c\,:\,\Omega\to\{0,1\}}{\operatorname{argmin}} \sum_{\mathrm{x}\in\Omega} (1 - c(\mathrm{x}))\, \rho(I_1, I_2, \mathrm{w}, \mathrm{x}) + \lambda_2 c(\mathrm{x}) + \lambda_3 \Phi(c), \tag{4}$$

Where λ_3 weights the regularization term.

2.3. Limitation of the sequential approach

The main limitation of the sequential approach is illustrated in Figure 1 involving the baseline (Figure 1) and follow-up (Figure 1) MRI acquisitions of a patient suffering from MS. One can observe in the follow-up scan the apparition of a new lesion and a slight enlargement of the ventricle reflecting the brain atrophy process. In the case of affine registration (Figures 1), we can see on the subtraction image that the lesion is well detected, but that spurious detection occur around the ventricles and in the cortical regions due to brain tissue atrophy. Using a deformable registration (Figures 1) helps to remove these spurious detection by compensating the ventricles enlargement and the cortical atrophy. However, it also tends to make the new lesion disappear (Figure 1), thus altering the shape of the corresponding detection (Figure 1). The goal of the proposed joint approach (Figures 1) is to perform an accurate atrophy correction while preserving the shape of appearing and evolving lesions, even when the lesion-to-tissue contrast is quite low.

3. Joint approach

3.1. General formulation

To overcome the limitations of the sequential approach, we advocate a joint modeling of registration and change detection. The two steps are fundamentally intertwined, since registration aims at finding correspondences between images, while change detection determines regions that does not admit correspondences. Therefore, both tasks should be defined with the same objective function to work in synergy. We formulate

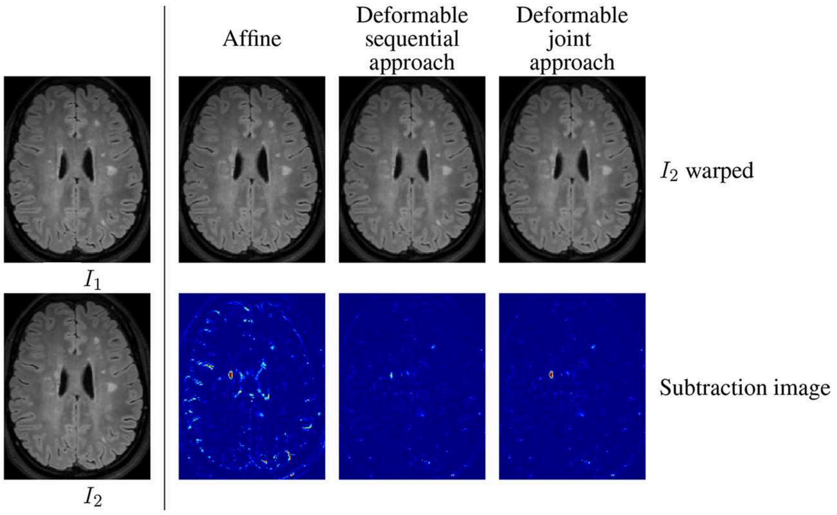


FIGURE 1
Warped follow-up images and subtraction images obtained with the sequential affine and deformable registration pipelines and the proposed joint deformable approach, on an example of the dataset described in COM (2021).

the following joint minimization problem that achieves this goal by unifying the principles described previously:

$$\widehat{w},\widehat{c} = \underset{w,c}{\text{argmin}} \sum_{x\in\Omega} \left[(1-c(x))\,\rho(I_1, I_2, w, x) + \lambda_2 c(x)\right] + \lambda_1\Psi(w) + \lambda_3\Phi(c). \quad (5)$$

With this model, the data term is cancelled in change regions (where $c(x) = 1$), so that the estimation of the transformation is only driven by the regularization term, thus producing smoothed deformation field in these areas. By this way, it prevents the lesion disappearing effect observed in Figure 1.

3.2. Modeling choices

The formulation (Equation 5) is versatile and could be instantiated with a variety of data and regularization terms. In this paper, the goal is to demonstrate the superiority of the joint formulation over the sequential approach under standard modeling choices, which are detailed in the sequel.

First, we assume that, thanks to the intensity normalization step done as a preprocessing, intensities of both images are comparable, thus allowing us to consider a data term based on intensity difference. Consequently, we consider the following standard similarity measure:

$$\rho(I_1, I_2, w, x) = \frac{1}{\sigma^2}\, \|I_2(x - w(x)) - I_1(x)\|_2^2, \quad (6)$$

Where σ is a normalization constant defined by the median absolute deviation of the intensity differences between I_1 and I_2. This data term has been used for motion estimation (Bruhn et al., 2005) and is representative of intensity-based features commonly used in change detection methods (Sweeney et al., 2013; Ganiler et al., 2014).

Secondly, we assume that the deformations induced by brain tissue atrophy are complex but still locally smooth. This is why we consider a non-parametric representation of the transformation field w while introducing a first order Tikhonov regularization term:

$$\Psi(w) = \sum_{x\in\Omega} \|\nabla w(x)\|_2^2, \quad (7)$$

Where $\nabla\cdot$ is the gradient operator.

Finally, we assume that the detected changes should be spatially coherent. Consequently, the change map c is regularized with a standard binary Potts model:

$$\Phi(c) = \sum_{x\in\Omega} \sum_{y\in\mathcal{N}(x)} (1 - \delta\left(c(x), c(y)\right)), \quad (8)$$

Where δ is the Kronecker function equal to 1 if its argument is true and $\mathcal{N}(x)$ is the 6-neighborhood of x.

3.3. Optimization

To solve the optimization problem (Equation 5), we rely on an alternating minimization strategy: at each iteration, we successively minimize with respect to (w.r.t.) each variable, while keeping the other fixed. We detail in this section the optimization strategies dedicated to each of the two subproblems.

Notice that, since the problem is nonconvex, convergence toward the global minimum cannot unfortunately be guaranteed. However, the results presented in Section 5.2 on several datasets suggest that the optimization process converges in practice toward a satisfying solution.

Minimization w.r.t. w To make the problem tractable, we consider the linearized version of the data term (Equation 6) obtained by replacing $I_2(\mathrm{y} - \mathrm{w}(\mathrm{x}))$ with its Taylor expansion around y:

$$\rho_l(I_1, I_2, \mathrm{w}, \mathrm{x}) = \frac{1}{\sigma^2}\,\|\nabla^\top I_2(\mathrm{x})\,\mathrm{w}(\mathrm{x}) + I_t(\mathrm{x})\|_2^2, \tag{9}$$

Where $I_t(\mathrm{y}) = I_2(\mathrm{y}) - I_1(\mathrm{y})$ is the temporal derivative. Since the Taylor expansion is valid only for small deformations, we embed the estimation in a coarse-to-fine scheme, which is a common practice in registration and motion estimation (Hill et al., 2001).

After Taylor development of $I_2(\mathrm{y} - \mathrm{w}(\mathrm{x}))$, the new data term ρ_l in Equation (9) is the composition of a quadratic function with a linear function of w, which yields a convex term. Since the regularization term (Equation 7) is also convex, the whole optimization problem is convex. By substituting ρ by ρ_l in Equation (5), the optimisation problem can be addressed with a variety of efficient optimization methods.

We chose to consider the alternated direction method of multipliers (ADMM) framework (Boyd et al., 2011). To this end, we introduce a splitting variable z that decouples the two terms of Equation (5) that depend on w, and we formulate the problem in the constrained form:

$$\min_{\mathrm{w}} \sum_{\mathrm{x}\in\Omega} (1 - c(\mathrm{x}))\,\rho_l(I_1, I_2, \mathrm{w}, \mathrm{x}) + \lambda_1 \Psi(\mathrm{z}) \quad \text{s.t}\ \ \mathrm{w} = \mathrm{z} \tag{10}$$

The ADMM algorithm is based on the minimization of the augmented Lagrangian associated with Equation (10) w.r.t. w and z, and a gradient ascent on the dual variable (Boyd et al., 2011). It leads to the following iterative updates of w and z (see Fortun et al., 2018 for a similar derivation with different data and regularization terms):

$$\mathrm{w}^{k+1} = \mathrm{prox}_{\sum_x (1-c(\mathrm{x}))\rho_l(I_1, I_2, \cdot, \mathrm{x})}\left(\mathrm{z}^k - \frac{\alpha^k}{\mu}\right) \tag{11}$$

$$\mathrm{z}^{k+1} = \mathrm{prox}_{\lambda_1 \Psi}\left(\mathrm{w}^{k+1} + \frac{\alpha^k}{\mu}\right) \tag{12}$$

$$\alpha^{k+1} = \alpha^k + \mu(\mathrm{w}^{k+1} - \mathrm{z}^{k+1}) \tag{13}$$

Where $\mathrm{prox}_f(\mathrm{x}) = \underset{\mathrm{y}}{\mathrm{argmin}} \frac{1}{2}\|\mathrm{x} - \mathrm{y}\|_2^2 + f(\mathrm{y})$ denotes the proximity operator of a function f. The subproblem (Equation 11) is voxel-wise and quadratic, and it admits a simple closed form solution. The subproblem (Equation 12) is equivalent to a denoising operation with the regularizer $\Psi(\cdot)$, and it also has a closed form linear solution that can be computed efficiently in the Fourier domain. μ is the parameter associated with the quadratic penalty in the Augmented Lagrangian associated with Equation (10). The ADMM algorithm is derived from this Augmented Lagrangian and the update (Equations 11, 12) are its minimization w.r.t. w and z. Intuitively, μ controls how fast the constraint $\mathrm{w} = \mathrm{z}$ is imposed through the optimization process. Thus, even if it is not strictly speaking a step size, it has a similar impact on the convergence speed.

Notice that the ADMM framework is flexible enough to cope with different data and regularization terms with low computational cost. The requirement is to be able to design a splitting of the cost function such that the proximity operators of each iteration have computationally efficient solutions. Examples of admissible models comprise data terms based on the ℓ_1 penalty function or cross-correlation (Vogel et al., 2013), and regularizations by total variation or Nuclear norm of the Jacobian (Bostan et al., 2014).

Minimization w.r.t. *c* When w is fixed, the estimation of c amounts to a binary segmentation problem with Potts regularization:

$$\widehat{c} = \underset{c}{\mathrm{argmin}} \sum_{\mathrm{x}\in\Omega} \big[\lambda_2 - \rho(I_1, I_2, \mathrm{w}, \mathrm{x})\big]\, c(\mathrm{x}) + \lambda_3 \Phi(c). \tag{14}$$

We solve it with a graph-cut method (Boykov et al., 2001), which is able to find an exact solution with very low computational cost.

4. Implementation details

4.1. Pre-processing

Before applying the change detection framework, the input images require to be pre-processed as follows. First, images are corrected for bias field inhomogeneity using the N4 algorithm (Tustison et al., 2010). Then, a global scaling of the intensity is performed in order to enforce the median value of the intensities inside the brainmask to be equal to 100. Images are then resampled to 1 mm isotropic resolution. These two steps aim at harmonizing all input images in terms of intensity range and spatial resolution. The follow-up scan is then rigidly registered on the baseline image using ANTs library (Avants et al., 2011)[1] with default parameters and mutual information metric. Differential bias field inhomogeneity is corrected thanks to the method described in Lewis and Fox (2004), while considering a 21 × 21 × 21 median filter size.

4.2. Post-processing

The detection maps are post-processed by discarding the connected components smaller than $3\mathrm{mm}^3$ and the detections

[1] https://github.com/ANTsX/ANTs

outside brain parenchyma (i.e., the union of gray and white matter). Notice that the binary change detection map can also be computed to reflect either positive or negative intensity changes only.

Brain parenchyma masks are computed from T1-weighted images using the FAST method provided in the FSL library (Zhang et al., 2001) or alternatively from FLAIR images using SAMSEG[2] brain parcellation tool (Cerri et al., 2021), in the case where no T1-weighted image is available.

4.3. Hyperparameter setting

Three hyperparameters have to be set in the proposed joint formulation (Equation 5): λ_1, controlling the spatial regularization of the deformation field, λ_2 acting as a threshold for the map of intensity differences, and λ_3, controlling the spatial regularization of the change map. Here, we suggest strategies to find out relevant parameters settings.

The value of λ_1 should be chosen to optimally estimate longitudinal brain atrophy, since it is the main source of brain deformation in MS. Thus, we consider a subset of 21 images from the dataset OASIS-3 (LaMontagne et al., 2019) that contains longitudinal Alzheimer and normal aging MRI data that exhibit various pattern of longitudinal brain atrophy. We determine the optimal value of λ_1 by selecting the one that leads to the best registration performance on this subset of OASIS-3. To this end, we derive a registration quality metric from the provided segmentation maps of brain structures obtained with Freesurfer. Concretely, the Dice score is computed for each structure between the segmentation maps of the baseline image and of the registered follow-up image. The global registration quality metric is then computed as the sum over all the regions of the median Dice score observed for each region. This procedure leads us to find $\lambda_1 = 70$ as an optimal value.

The values of λ_2 and λ_3 have to be set to find the best compromise regarding: (i) The expected intensity difference, (ii) the noise level that corrupts the images, and (iii) the spatial extent of the changes. Here, we suggest an approach to find out optimized setting for each of the two considered databases (see Sections 5.1.2, 3.2). In practice, λ_2 and λ_3 have been fixed to maximize the overall performance of the *affine sequential* approach (see Section 5.1.5) in terms of local Dice Similarity Coefficient (*local* DSC, see Section 5.1.4) for each dataset. Considering the *local* DSC ensures to focus on the ability of the detection scheme (Equation 14) to recover the detected changes while not being influenced by false positive detections that can occur in other parts of the brain. Considering the affine transformation model ensures that the registration step does not to alter the geometry of evolving regions. This procedure leads us to find $\lambda_2 = 16$ and $\lambda_3 = 5$ as optimal setting for LesjakDB dataset and $\lambda_2 = 25$ and $\lambda_3 = 3$ for MSSEG-2 dataset.

4.4. Convergence and stopping criteria

The iterations of the alternated minimization of Equation (5) and of the ADMM algorithm (Equations 11–13) are stopped when a stopping criterion is verified or when a maximum number of iterations is reached. The stopping criterion is a threshold on the norm of the relative changes between two consecutive iterations, and is set to 10^{-3} for the alternated minimization and 2.10^{-3} for ADMM. The maximum number of iterations is set to 5 for the alternated minimization and 300 for ADMM.

5. Experimental evaluation

5.1. Evaluation framework

In this section, we report results obtained on one synthetic dataset and two publicly available real patients datasets. The synthetic dataset offer the advantage to have an unambiguously defined ground truth change detection map, while controlling the amount of noise, bias field inhomogeneity and brain atrophy that corrupt the images. The real datasets are used to evaluate the proposed approach in conditions that are closer to the clinical routine, while considering different acquisition conditions and various pathological evolution. The first real patients dataset, denoted in the sequel as *LesjakDB* (Lesjak et al., 2016), is dedicated to assess the ability of methods to detect every kinds of MS lesion evolutions (shrinkage, growth, new and disappearing), whereas the second dataset, denoted in the sequel as *MSSEG-2* (COM, 2021), only focus on the ability to detect new appearing lesions.

5.1.1. Synthetic dataset

We evaluate the proposed method on T2-weighted synthetic volumes generated with the Brainweb simulator (Cocosco et al., 1997), while considering the *normal* anatomical model (i.e., without lesion) and two multiple sclerosis anatomical models with *moderate* and *severe* lesion load. The images are simulated at a 1 mm^3 isotropic resolution (image size: 181 x 217 x 181) with bias field inhomogeneity (20%). To simulate realistic brain atrophy for the follow-up image, we applied a deformation field that has been estimated using a deformable registration (Avants et al., 2008) from two T1-weighted MRI scans acquired 4 years apart of a patient suffering from MS that exhibits a significant brain atrophy evolution (in-house dataset). It should be noted that these real data images were first affinely registered onto the brainweb image to ensure the estimated deformable

[2] https://surfer.nmr.mgh.harvard.edu/fswiki/Samseg

TABLE 1 Simulated longitudinal acquisitions.

Scenario	Baseline image	Follow-up image	Simulated atrophy
Lesion appearance without atrophy	Normal	Moderate	No
Lesion growth without atrophy	Moderate	Severe	No
Lesion appearance with atrophy	Normal	Moderate	Yes
Lesion growth with atrophy	Moderate	Severe	Yes

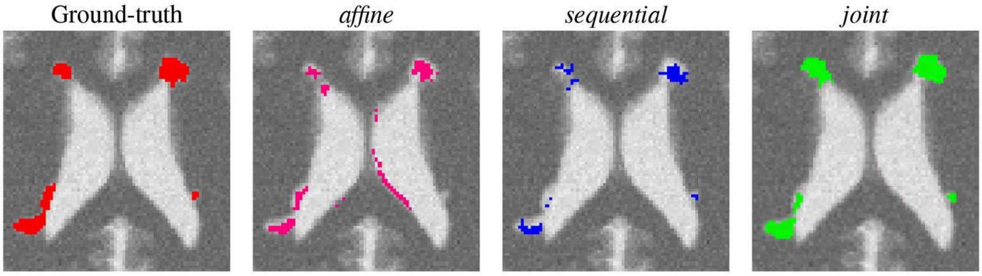


FIGURE 2
Qualitative comparison of the binary change detection maps obtained with the three methods on the synthetic dataset (lesion appearance with atrophy scenario).

registration to be consistent with the underlying anatomy. The visual inspection confirms that the simulated image exhibits a realistic atrophy pattern. Gaussian additive noise was added with a standard deviation fixed at 5% of the mean intensity in the brightest tissue (cerebrospinal fluid in the T2-weighted simulation). We consider several scenarios of simulated longitudinal acquisition that are summarized in Table 1.

5.1.2. Real dataset LesjakDB: All kinds of lesion evolution

LesjakDB dataset (Lesjak et al., 2016) is composed of 20 longitudinal MRI acquisitions of MS patients with two timepoints. The median time between the baseline and follow-up studies was 311 days, ranging from 81 to 723 days. Each MRI acquisition consists in a 2D T1-weighted, a 2D T2-weighted and 2D-FLAIR sequences. Change detection was conducted on the FLAIR images only. The FLAIR image size is 256 × 256 × 49 with an anisotropic spatial resolution of 0.9 × 0.9 × 3 mm. Ground truth change detection maps are also provided, which were obtained from manual annotations done by two expert raters. We adjusted some of the ground truth annotations that did not match the real lesion changes. The annotated changes include appearing, growing, shrinking and disappearing lesions. Ground-truth detection maps are compared to binary detection maps that include both positive and negative intensity changes.

5.1.3. Real dataset MSSEG-2: Only appearing lesions

MSSEG-2 dataset (COM, 2021) is composed of 100 pairs of FLAIR MRI scans from MS patients acquired on various MR scanners. The provided ground-truth is limited to new appearing lesions, and was build from the consensus of manual annotations delineated by four experts. The dataset is separated into training (40 patients) and testing (60 patients) sets. Since the proposed approach does not require any training step, we consider the whole dataset for testing. However, we distinguish two subgroups of data, namely *MSSEG-2-Change* corresponding the 61 subjects that exhibit at least one new appearing lesion and *MSSEG-2-NoChange* corresponding the 39 subjects that do not exhibit any new appearing lesion. Since the provided ground-truth is limited to new appearing lesions, they are compared only to the positive binary change detection maps obtained with the different methods. Notice, that the proposed method framework does not discriminate appearing from evolving lesion. Consequently, lesion evolutions, which are not labeled in the ground truth detection maps, are erroneously considered as false positive detection, thus introducing a bias in some of the evaluation metrics.

5.1.4. Metrics

We report four metrics to evaluate the performance of the methods to detect changes, namely the Dice Similarity Coefficient (DSC), the Positive Predictive Value (PPV), the True Positive Ratio (TPR) and the *local* DSC. Let *TP*, *TN*, *FP*, and *FN* be the number of voxels from estimated change detection map

TABLE 2 Results computed on the synthetic dataset.

Scenario	Method	DSC	PPV	TPR	Local DSC
Lesion appearance no atrophy	*affine*	**0.830**	0.782	**0.885**	**0.830**
	sequential	0.684	**0.921**	0.544	0.684
	joint	0.814	0.887	0.751	0.814
Lesion growth no atrophy	*affine*	0.766	0.635	**0.964**	0.770
	sequential	0.685	**0.734**	0.641	0.688
	joint	**0.806**	0.726	0.902	**0.808**
Lesion appearance simulated atrophy	*affine*	0.460	0.329	**0.767**	**0.810**
	sequential	0.626	**0.960**	0.465	0.627
	joint	**0.743**	0.925	0.621	0.744
Lesion growth simulated atrophy	*affine*	0.652	0.505	**0.919**	0.827
	sequential	0.753	**0.869**	0.664	0.754
	joint	**0.847**	0.833	0.861	**0.848**

The bold values indicate the highest scores among the four methods.

that correspond to *True Positive*, *True Negative*, *False Positive* and *False Negative*, respectively.

The DSC is defined as:

$$DSC = 2TP/(2TP + FP + FN)$$

and reflects the overall good overlap between the detection map and the ground truth.

The PPV is defined as:

$$PPV = TP/(TP + FP)$$

and reflects the proportion of relevant detections among all the detected changes.

The TPR is defined as:

$$TPR = TP/(TP + FN)$$

and reflects the proportion of the ground-truth changes that have been detected.

The *local* DSC correspond the DSC computed on a restricted area defined as the dilation with a 4-voxel radius spherical structuring element to the ground truth. This metric enables us to focus the evaluation on the local spatial accuracy of the detection method

In addition to the voxel-wise metrics, we also report lesion-wise metrics, namely the Lesion True Positive Ratio (L-TPR) and the Lesion Positive Predictive Value (L-PPV). These metrics have been evaluated thanks to the *animaSegPerfAnalyzer* validation tool while considering the same hyperparameters as in Commowick et al. (2018).

Since all these metrics are not relevant for data that do not exhibit any changes, we consider in that specific case the number of detected connected components as well as the volume of detected changes to characterize the false positive detections.

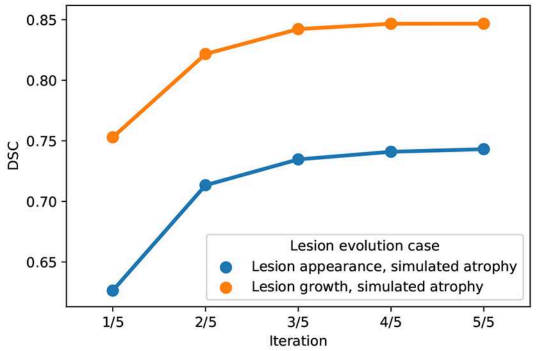


FIGURE 3
Evolution of DSC across iterations of the alternating optimization scheme of the *joint* approach on the synthetic dataset (blue: Lesion appearance, simulated atrophy, orange: Lesion growth, simulated atrophy).

5.1.5. Variants used for comparison

To demonstrate the benefits of the proposed joint modeling, we consider three variants of the change detection framework:

- *joint*: the proposed joint change detection and registration method described in Section 3.
- *sequential*: The sequential counterpart of the proposed method, which successively performs deformable

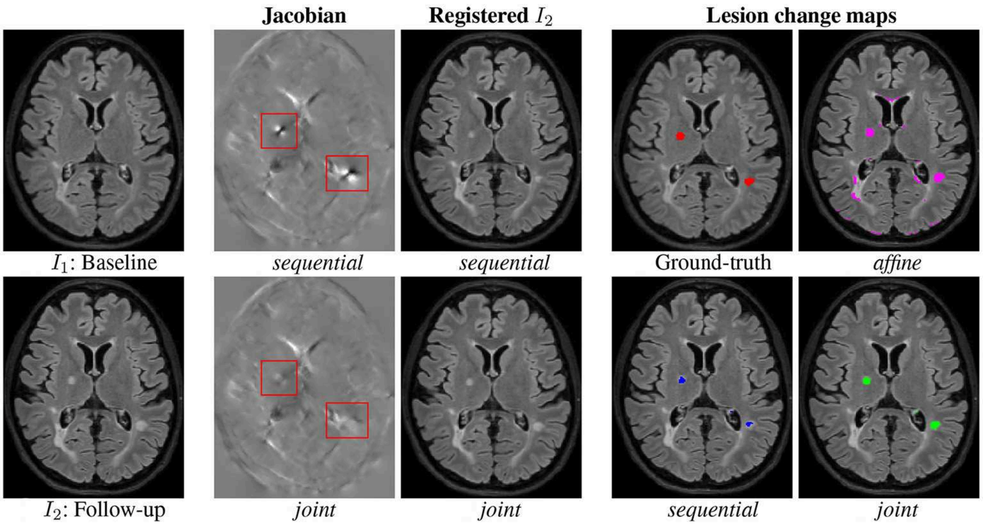


FIGURE 4
Qualitative comparison of the binary change detection maps obtained with the three methods and of the jacobian of the deformation field estimated with the *sequential* and *joint* approaches on one selected subject from the MSSEG-2 dataset. Hyperparameters: $\lambda_1 = 70$ (for *sequential* and *joint* methods), $\lambda_2 = 25$, $\lambda_3 = 3$. I_1: Baseline, I_2: Follow-up, *sequential*, *joint*, Ground-truth, and *affine*.

TABLE 3 Results computed on LesjakDB and MSSEG-2-Change datasets.

DataSet	Method	Local DSC	DSC	PPV	TPR	L-PPV	L-TPR
LesjakDB	*affine*	**0.539 ± 0.174**	0.152 ± 0.087	0.086 ± 0.060	**0.603 ± 0.223**	0.022 ± 0.022	0.576 ± 0.195
	sequential	0.424 ± 0.139	0.317 ± 0.117	0.293 ± 0.152	0.323 ± 0.125	0.088 ± 0.051	**0.577 ± 0.197**
	joint	0.501 ± 0.179	**0.353 ± 0.144**	**0.323 ± 0.148**	0.447 ± 0.207	**0.103 ± 0.061**	0.574 ± 0.208
MSSEG-2-Change	*affine*	**0.626 ± 0.224**	0.142 ± 0.165	0.081 ± 0.139	**0.633 ± 0.269**	0.015 ± 0.042	0.840 ± 0.271
	sequential	0.520 ± 0.196	0.310 ± 0.178	0.298 ± 0.264	0.379 ± 0.176	0.095 ± 0.150	**0.872 ± 0.307**
	joint	0.579 ± 0.219	**0.356 ± 0.208**	**0.336 ± 0.254**	0.474 ± 0.222	**0.111 ± 0.155**	**0.872 ± 0.304**
MSSEG-2-Change inverse	*affine*	0.626 ± 0.224	0.142 ± 0.165	0.081 ± 0.139	**0.633 ± 0.269**	0.015 ± 0.042	0.840 ± 0.271
	sequential	0.625 ± 0.217	0.348 ± 0.216	0.290 ± 0.244	0.550 ± 0.243	**0.094 ± 0.151**	**0.977 ± 0.292**
	joint	**0.655 ± 0.237**	**0.378 ± 0.233**	**0.312 ± 0.250**	0.619 ± 0.266	0.091 ± 0.164	0.947 ± 0.304

The median ± the median absolute deviation (MAD) computed over all subjects are reported for each metric. The MSSEG-2-Change inverse experiment consist in swapping the baseline and follow-up images to evaluate the ability of the methods to detect disappearing lesions. The bold values indicate the highest scores among the three methods.

registration and change detection. For the two steps, we use the same model and optimization algorithms as in substeps of the *joint* approach described in Section 3.3.

- *affine*: The *sequential* approach where the deformable registration has been replaced by affine registration, which corresponds to the most common case. The affine registration was estimated using ANTs library (Avants et al., 2011) [3] with default parameters and mutual information metric. Then, the thresholding and smoothing of the change map routine follows model (Equation 14).

5.2. Results

5.2.1. Synthetic dataset

First, a qualitative visual comparison of the three methods is provided in Figure 2 for the lesion appearance with atrophy scenario. The *affine* method succeeds to detect almost all the lesion areas, but it suffers from false positive detection around the ventricles due to brain atrophy. Both the *sequential* and the *joint* methods compensated for brain atrophy deformation since none of them exhibit false

[3] https://github.com/ANTsX/ANTs

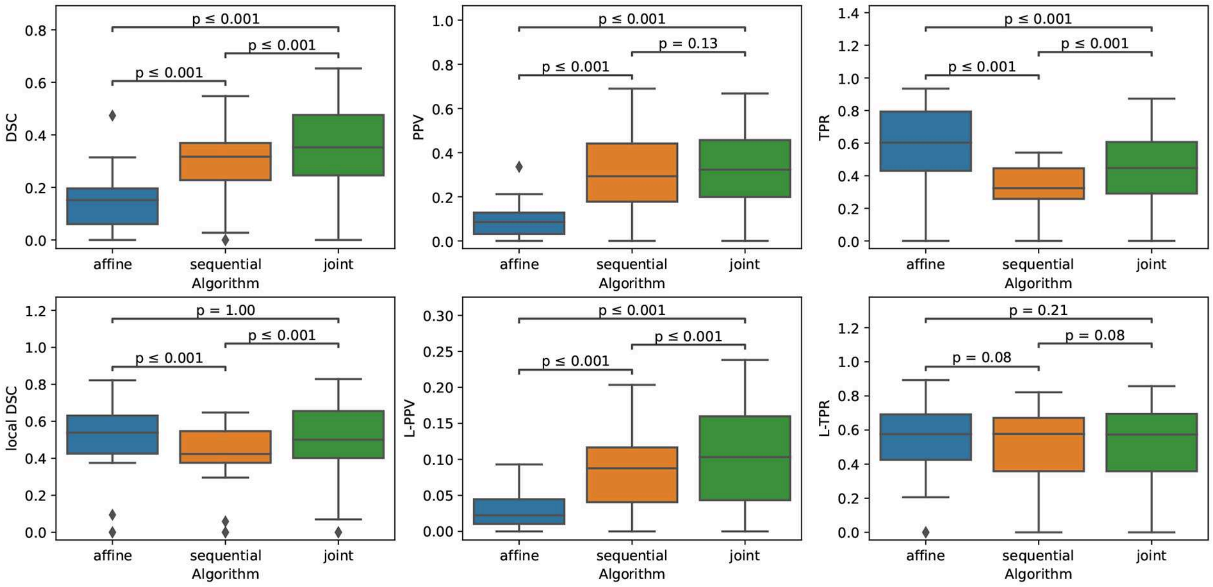


FIGURE 5
Boxplots corresponding to the results summarized in Table 3 for LesjakDB ($N_{Subject} = 20$). Statistical significancy is evaluated thanks to the Wilcoxon signed-rank test between each pair of methods while applying Benjamini/Hochberg FDR correction.

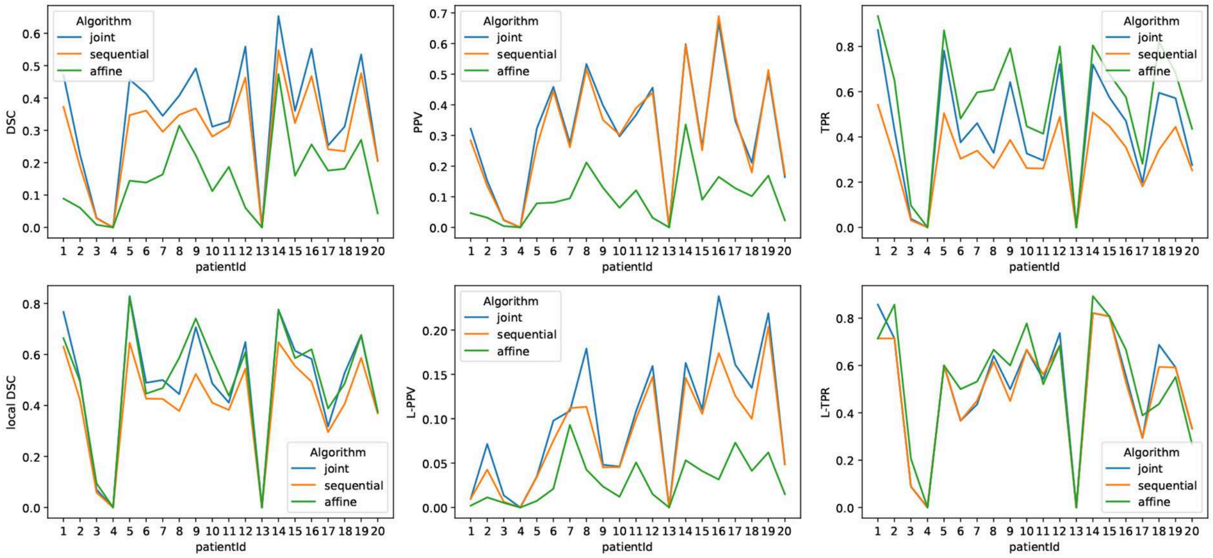


FIGURE 6
Metrics reporting the performance of the three methods for each subject of LesjakDB.

detections around the ventricles. However, the *sequential* method failed to detect the whole lesion areas due to the over-compensation of lesion changes. This limitation is overcome by the *joint* approach that succeeds to detect the entire lesion areas.

A quantitative comparison of the three methods under four scenarios is provided in Table 2. First, we consider the appearance of lesion without atrophy. Unsurprisingly, this scenario is the most favorable for the *affine* method since there is no geometric difference to compensate. The *sequential* approach

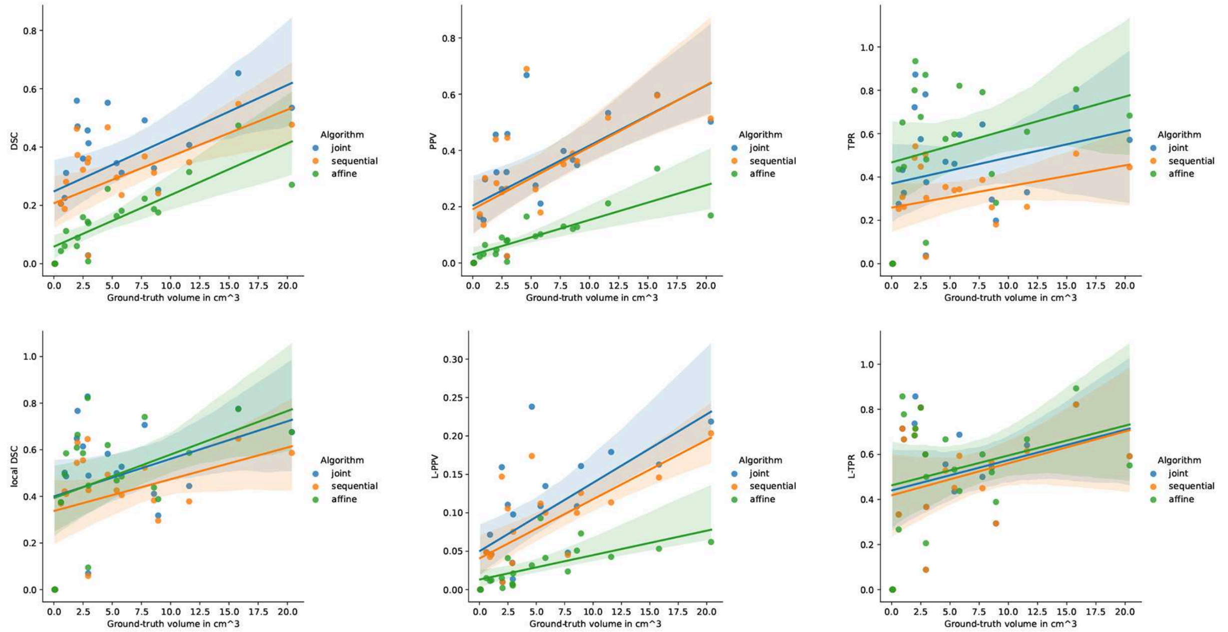


FIGURE 7
Correlation between the performance of the three methods and the ground-truth evolving lesion load (cm^3) of the subjects of LesjakDB.

yields to significantly lower values of DSC and local DSC. This is due to the lesion over-compensation effect, as supported by the observed low TPR value (i.e., lack of sensitivity) and high PPV value (i.e., high specificity). Finally, the *joint* approach overcomes the shortcoming of the *sequential* approach and have performances similar to the *affine* method, with a slight tendency to underestimate the detected area. Similar observations can be made about the second scenario involving lesion growth without atrophy.

The conclusions are drastically different for the two scenarios involving simulated atrophy. The performance of the *affine* method is significantly hampered by the numerous false detections due to the atrophy. This is illustrated by the significant decrease of the DSC and PPV values compared to the cases without atrophy, while the TPR and local DSC values are less modified. The *sequential* approach succeeds to compensate for the simulated brain atrophy, as highlighted by the high PPV value, but still underestimates the changes to detect, as indicated by the low TPR value. The *joint* approach clearly outperforms the two previous approaches in terms of detection accuracy, as objectified by the significantly higher DSC value.

The behavior of the *joint* approach can be monitored through the iterations of the alternating optimization scheme (see Figure 3). We can see that the DSC increases across the iterations, and the convergence is reached in a few iterations. Concerning the computational burden of the joint approach, it is about 24min on one single core (Intel(R) Xeon(R) Gold 6130 CPU @ 2.10GHz) for an experiment on the synthetic dataset (image size: 181 x 217 x 181).

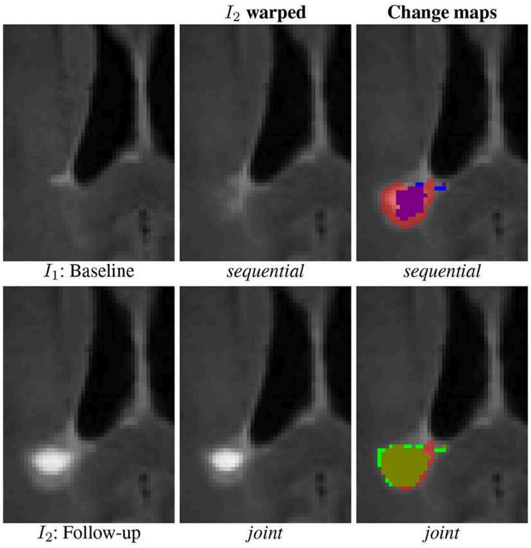


FIGURE 8
Qualitative comparison of the binary change detection maps obtained with the *sequential* (purple) and *joint* (green) approaches as compared to the ground truth (underlying transparent red) on one selected subject from the MSSEG-2 dataset.

5.2.2. LesjakDB

First, a qualitative visual comparison of the three methods is provided in Figure 4. We can draw similar conclusions as

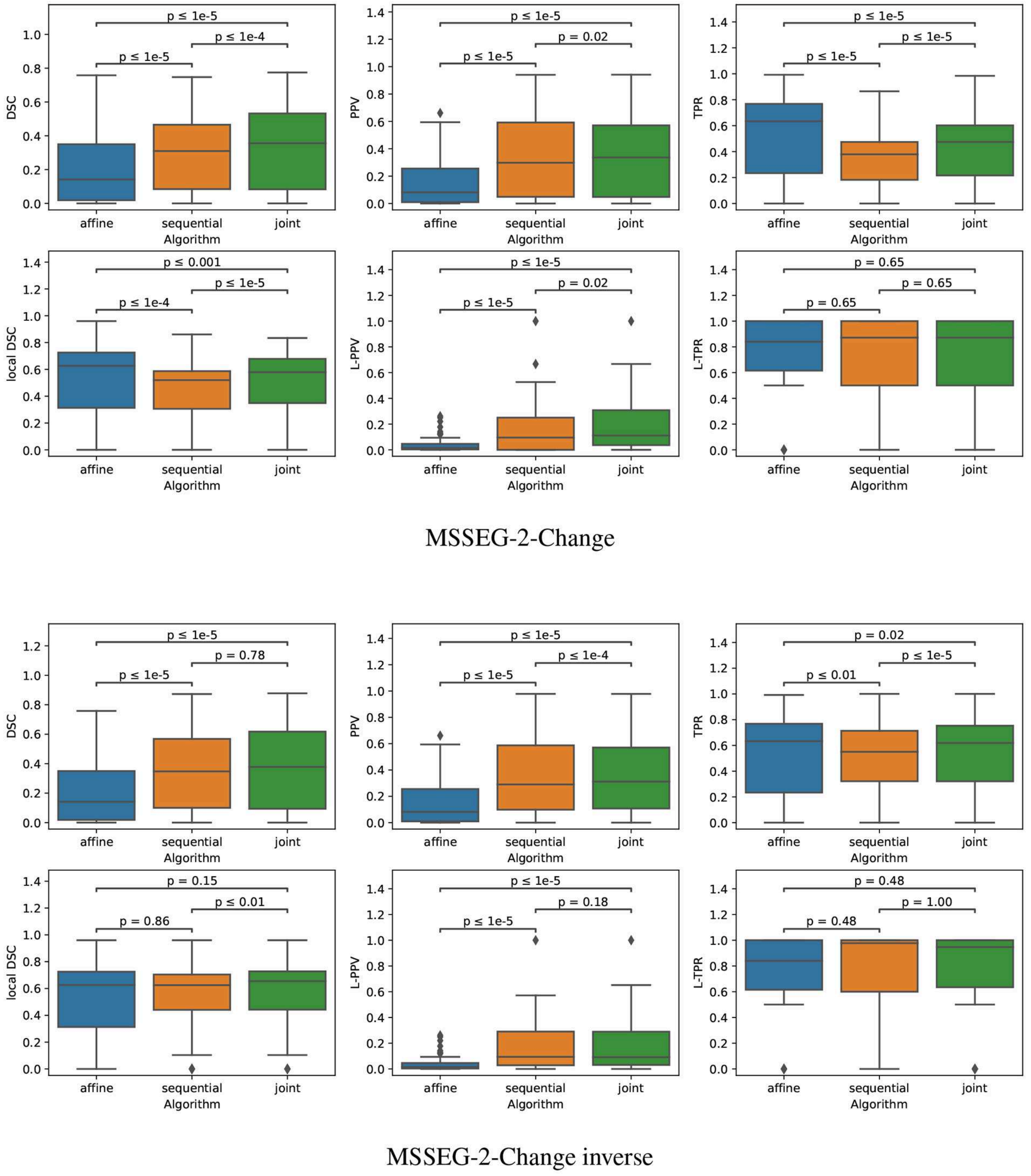


FIGURE 9
Boxplots corresponding to the results summarized in Table 3 for both MSSEG-2-Change and MSSEG-2-Change inverse ($N_{Subject} = 61$). Statistical significancy is evaluated thanks to the Wilcoxon signed-rank test between each pair of methods while applying Benjamini/Hochberg FDR correction.

for the synthetic dataset (see Figure 2). The *affine* method demonstrates a high sensitivity (i.e., the lesion evolution is well detected) but a lack of specificity (i.e., numerous false positive detections are detected around the ventricles and in the posterior part of the cortex). Conversely, the *sequential* method has high specificity but lacks sensibility. The *joint* approach provides the best visual results, thus illustrating its ability to achieve both high sensitivity and high specificity.

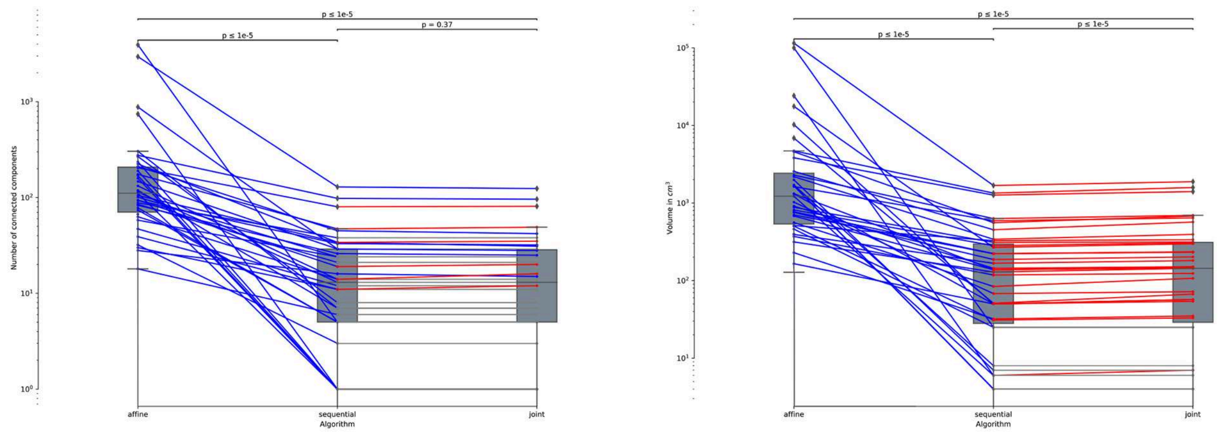


FIGURE 10
Number of connected components **(left)** and volume in cm^3 **(right)** of the detected changes computed for the 39 subjects of MSSEG-2-NoChange datasets (i.e., subjects without new appearing lesion). Statistical significancy is evaluated thanks to the Wilcoxon signed-rank test between each pair of methods while appling Benjamini/Hochberg FDR correction.

Figures 4 show the Jacobian of the deformation fields obtained by the *sequential* and *joint* methods, respectively. The specific pattern characterized by the alternance of both high and low values of the jacobian (see areas highlighted by the red squares in Figure 4) reflects the high local contraction and dilation induced by the deformation field to make the lesion disappear, thus explaining the lack of sensitivity of the detection results.

The quantitative evaluation shown in Table 3 (first row) and in Figure 5 confirms the conclusions of the visual analysis. The high sensitivity of the *affine* method is objectified at the voxel level by a statistically significantly higher TPR than the two other methods. At the lesion level, all the three methods exhibit similar L-TPR values, thus emphasizing their ability to detect the same amount of changing areas. Both the *sequential* and *joint* methods yield significantly higher PPV and L-PPV compared to the *affine* method, which illustrates their ability to reduce the number of false detections induced by brain atrophy at both voxel and lesion levels. This result highlights the benefit of using deformable registration in the context of MS lesion monitoring. The significantly lower TPR achieved by the *sequential* method compared to the *joint* method is the consequence of the lesion over-compensation effect. Finally, the *joint* approach significantly outperforms the two other approaches in term of voxel-wise global accuracy (see DSC).

Figure 6 highlights the variability of the performance of the methods across the subjects. It is interesting to notice that, although the performance of the detection methods greatly varies from one subject to the other, the ranking among the three methods appears to be highly consistent across the subjects. When investigating for the factors that may explain the observed variability, it appears that the volume of the ground-truth seems to play a prominent role: the larger is the volume to detect, the better is the performance of the change detection algorithm (see Figure 7).

5.2.3. MSSEG-2

Similarly as for the synthetic and LesjakDB datasets, the qualitative visual comparison of the two approaches based on deformable registration highlights the lack of sensitivity of the *sequential* method due the lesion over-compensation effect (see Figure 8). The resulting change detection map (purple) is too small compared to the ground-truth (underlying transparent red) due to the deformable registration that significantly shrinks the lesion. With the *joint* approach, the shape of the lesion is almost preserved in the warped follow-up image and the change detection map (green) matches almost perfectly the ground-truth.

The quantitative evaluation on the subset MSSEG-2-Change is reported in the second row of Table 3 and in the upper part of Figure 9.

The fact that both the *sequential* and *joint* approaches lead to significantly higher PPV values as compared to the *affine* approach advocates the use of deformable registration to reduce the number of false detections. The benefit of considering the *joint* over the *sequential* approach to overcome the lesion overcompensation effect is clearly demonstrated by the significantly higher TPR and local DSC values obtained with *joint* method.

It is also interesting to notice that the lesion over-compensation effect does not affect the special case of disappearing lesion. Indeed, when registering an image without lesion on a image with a lesion, the dissimilarity in the area of the disappearing lesion cannot be corrected by the transport of intensity of the registration (this is in fact only the case for non symmetric image registration method, see Noblet et al. (2004) for further explanations). To illustrate this phenomenon, we consider the MSSEG-2-Change inverse experiment (see the third row of both Table 3 and the bottom part of Figure 9) that consist in swapping the baseline and the follow-up image, so that the ground-truth now correspond to disappearing lesions. The same conclusion can be drawn from the DSC, PPV, and TPR as compared to the MSSEG-2-Change experiment. The most interesting point concern the local DSC that focuses the evaluation on the disappearing lesion. In that case, there is no significant difference any more between *sequential* and *joint* approaches contrary to the MSSEG-2-Change experiment, showing the absence of lesion over-compensation effect in the specific scenario of detecting disappearing lesions.

Note that all the results presented above in this section are evaluated on the 61 subjects of MSSEG-2-Change (i.e., subject presenting at least one new appearing lesion). Indeed, the presented metrics cannot be computed anymore for the 39 subjects of MSSEG-2-NoChange since the ground-truth change detection map is empty. This is why we only report the volume of detected changes for this subset of MSSEG-2 (see Figure 10). We can notice that both *sequential* and *joint* approaches lead to significantly lower volume of detected changes as compared to the *affine*, which appears in line with previous findings that support the use of deformable registration to reduce the number of false detections. Also note that the *joint* method yields consistently to slightly higher volume of detected changes as compared to the *sequential* method. This is also the consequence of the lesion over-compensation effect that affects the *sequential* appproach.

6. Conclusion and perspectives

We have presented a method that unifies registration and change detection for the analysis of longitudinal brain MRI. It is based on the joint modeling of these two tasks as the minimization of a single objective function, for which we have developed an efficient alternating optimization method. The proposed approach has been evaluated in the context of the follow-up of multiple sclerosis lesion, which requires deformable registration to capture characteristic brain atrophy, but also with the potential caveat to shrink appearing lesions. In this context, the conventional sequential detection pipeline leads to large detection inaccuracies around new appearing lesions. We have demonstrated on simulated and real data that the proposed joint approach is able to combine the ability of deformable registration to correct brain atrophy, and a good preservation of the lesions shape to ensure accurate change detection.

The implementation presented in this paper of the proposed joint model relies in fact on quite simple modeling assumptions. The versatility of the optimization approach opens the way for more sophisticated models that could be handled in the same framework. Alternative data fidelity terms such as cross correlation or mutual information, and regularizers such as total variation could be considered to potentially improve the performance of the method. Another perspective is to improve the convergence of the alternating optimization strategy to ensure a better robustness to local minima. To this end we could consider a fuzzy change detection map to turn its estimation into a continuous optimization problem. This would allow us to use more robust optimization approaches such as Proximal Alternating Linearized Minimization (Bolte et al., 2014). Finally, while we have addressed in this paper the registration problem, the unification principle could be extended to other steps of the change detection pipeline. In particular, the intensity normalization and the bias field inhomogeneity correction of the MRI acquisitions are crucial pre-processing tasks that are impacted by the presence of evolving lesions. Therefore, the integration of these two tasks in a single unified model would be a natural extension of the proposed framework.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://brainweb.bic.mni.mcgill.ca/selection_ms.html; http://lit.fe.uni-lj.si/tools.php?lang=eng; https://portal.fli-iam.irisa.fr/msseg-2/.

Author contributions

The proposed method was coded by ED. ED, DF, and VN equally contributed to the writing of the article and the interpretation of the results. SK provided a clinical expertise for the analysis of the results and the real annotated databases. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by the Région Grand Est and Philips Healthcare.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Altay, E. E., Fisher, E., Jones, S. E., Hara-Cleaver, C., Lee, J.-C., and Rudick, R. A. (2013). Reliability of classifying multiple sclerosis disease activity using magnetic resonance imaging in a multiple sclerosis clinic. *JAMA Neurol.* 70, 338–344. doi: 10.1001/2013.jamaneurol.211

Avants, B., Epstein, C., Grossman, M., and Gee, J. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Analysis* 12, 26–41. doi: 10.1016/j.media.2007.06.004

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., and Gee, J. C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage* 54, 2033–2044. doi: 10.1016/j.neuroimage.2010.09.025

Bolte, J., Sabach, S., and Teboulle, M. (2014). Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.* 146, 459–494. doi: 10.1007/s10107-013-0701-9

Bosc, M., Heitz, F., Armspach, J.-P., Namer, I., Gounot, D., and Rumbach, L. (2003). Automatic change detection in multimodal serial MRI: application to multiple sclerosis lesion evolution. *Neuroimage* 20, 643–656. doi: 10.1016/S1053-8119(03)00406-3

Bostan, E., Lefkimmiatis, S., Vardoulis, O., Stergiopulos, N., and Unser, M. (2014). Improved variational denoising of flow fields with application to phase-contrast MRI data. *IEEE Signal Process. Lett.* 22, 762–766. doi: 10.1109/LSP.2014.2369212

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundat. Trends Mach. Learn.* 3, 1–122. doi: 10.1561/2200000016

Boykov, Y., Veksler, O., and Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* 23, 1222–1239. doi: 10.1109/34.969114

Bruhn, A., Weickert, J., and Schnörr, C. (2005). Lucas/kanade meets horn/schunck: combining local and global optic flow methods. *Int. J. Comput. Vis.* 61, 1–21. doi: 10.1023/B:VISI.0000045324.43199.43

Cabezas, M., Corral, J., Oliver, A., Díez, Y., Tintor,é, M., Auger, C., et al. (2016). Improved automatic detection of new T2 lesions in multiple sclerosis using deformation fields. *Am. J. Neuroradiol.* 37, 1816–1823. doi: 10.3174/ajnr.A4829

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

Cocosco, C. A., Kollokian, V., Kwan, R. K.-S., Pike, G. B., and Evans, A. C. (1997). Brainweb: online interface to a 3D mri simulated brain database. *Neuroimage* 5, 425.

COM (2021). *MSSEG-2 Challenge Proceedings: Multiple Sclerosis New Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*, COM: Strasbourg.

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., et al. (2018). Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Sci. Rep.* 8, 13650. doi: 10.1038/s41598-018-31911-7

Dufresne, E., Fortun, D., Kumar, B., Kremer, S., and Noblet, V. (2020). “Joint registration and change detection in longitudinal brain MRI,” in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* (Iowa City, IA: IEEE).

Elliott, C., Arnold, D. L., Collins, D. L., and Arbel, T. (2013). Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain MRI. *IEEE Trans. Med. Imaging* 32, 1490–1503. doi: 10.1109/TMI.2013.2258403

Fortun, D., Storath, M., Rickert, D., Weinmann, A., and Unser, M. (2018). Fast piecewise-affine motion estimation without segmentation. *IEEE Trans. Image Process.* 27, 5612–5624. doi: 10.1109/TIP.2018.2856399

Ganiler, O., Oliver, A., Diez, Y., Freixenet, J., Vilanova, J. C., Beltran, B., et al. (2014). A subtraction pipeline for automatic detection of new appearing multiple sclerosis lesions in longitudinal studies. *Neuroradiology* 56, 363–374. doi: 10.1007/s00234-014-1343-1

Hill, D. L., Batchelor, P. G., Holden, M., and Hawkes, D. J. (2001). Medical image registration. *Phys. Med. Biol.* 46, R1. doi: 10.1088/0031-9155/46/3/201

Kaunzner, U. W., and Gauthier, S. A. (2017). MRI in the assessment and monitoring of multiple sclerosis: an update on best practice. *Therapeut. Adv. Neurol. Disord.* 10, 247–261. doi: 10.1177/1756285617708911

LaMontagne, P. J., Benzinger, T. L., Morris, J. C., Keefe, S., Hornbeck, R., Xiong, C., et al. (2019). *OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease*. Cold Spring Harbor Laboratory Press. Available online at: https://www.medrxiv.org/content/early/2019/12/15/2019.12.13.19014902.full.pdf

Lesjak, Ž, Z., Pernuš, F., Likar, B., and Špiclin, I. (2016). Validation of white-matter lesion change detection methods on a novel publicly available MRI image database. *Neuroinformatics* 14, 403–420. doi: 10.1007/s12021-016-9301-1

Lewis, E. B., and Fox, N. C. (2004). Correction of differential intensity inhomogeneity in longitudinal MR images. *Neuroimage* 23, 75–83. doi: 10.1016/j.neuroimage.2004.04.030

Llad,ó, X., Ganiler, O., Oliver, A., Mart,í, R., Freixenet, J., Valls, L., et al. (2012). Automated detection of multiple sclerosis lesions in serial brain MRI. *Neuroradiology* 54, 787–807. doi: 10.1007/s00234-011-0992-6

McNamara, C., Sugrue, G., Murray, B., and MacMahon, P. (2017). Current and emerging therapies in multiple sclerosis: implications for the radiologist, part 1—mechanisms, efficacy, and safety. *Am. J. Neuroradiol.* 38, 1664–1671. doi: 10.3174/ajnr.A5147

Noblet, V., Heinrich, C., Heitz, F., and Armspach, J.-P. (2004). “A topology preserving non-rigid registration method using a symmetric similarity function - application to 3-D brain images,” in *European Conference on Computer Vision (ECCV)* (Prague), 546-557.

Radke, R., Andra, S., Al-Kofahi, O., and Roysam, B. (2005). Image change detection algorithms: a systematic survey. *IEEE Trans. Image Process.* 14, 294–307. doi: 10.1109/TIP.2004.838698

Rey, D., Subsol, G., Delingette, H., and Ayache, N. (2002). Automatic detection and segmentation of evolving processes in 3D medical images: application to multiple sclerosis. *Med. Image Anal.* 6, 163–179. doi: 10.1016/S1361-8415(02)00056-7

Rousseau, F., Faisan, S., Heitz, F., Armspach, J.-P., Chevalier, Y., Blanc, F., et al. (2007). “An a contrario approach for change detection in 3D multimodal images: application to multiple sclerosis in MRI,” in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Lyon: IEEE), 2069–2072.

Salem, M., Cabezas, M., Valverde, S., Pareto, D., Oliver, A., Salvi, J., et al. (2018). A supervised framework with intensity subtraction and deformation field features for the detection of new T2-w lesions in multiple sclerosis. *Neuroimage Clin.* 17, 607–615. doi: 10.1016/j.nicl.2017.11.015

Shinohara, R. T., Sweeney, E. M., Goldsmith, J., Shiee, N., Mateen, F. J., Calabresi, P. A., et al. (2014). Statistical normalization techniques for magnetic resonance imaging. *Neuroimage Clin.* 6, 9–19. doi: 10.1016/j.nicl.2014.08.008

Song, S., Zheng, Y., and He, Y. (2017). A review of methods for bias correction in medical images. *Biomed. Eng. Rev.* 3, 1550. doi: 10.18103/bme.v3i1.1550

Sotiras, A., Davatzikos, C., and Paragios, N. (2013). Deformable medical image registration: a survey. *IEEE Trans. Med. Imaging* 32, 1153–1190. doi: 10.1109/TMI.2013.2265603

Sweeney, E., Shinohara, R., Shea, C., Reich, D., and Crainiceanu, C. (2013). Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal MRI. *Am. J. Neuroradiol.* 34, 68–73. doi: 10.3174/ajnr.A3172

Tustison, N. J., Avants, B. B., Cook, P. A., and Gee, J. C. (2010). N4ITK: improved N3 bias correction with robust B-spline approximation. *IEEE Trans. Med. Imaging* 29, 1310–1320. doi: 10.1109/ISBI.2010.5490078

Vogel, C., Roth, S., and Schindler, K. (2013). “An evaluation of data costs for optical flow,” in *DAGM Symposium on Pattern Recognition* (Berlin), 343–353.

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* 20, 45–57. doi: 10.1109/42.906424

Frontiers in Neuroscience

Provides a holistic understanding of brain function from genes to behavior

Part of the most cited neuroscience journal series which explores the brain - from the new eras of causation and anatomical neurosciences to neuroeconomics and neuroenergetics.

Discover the latest Research Topics

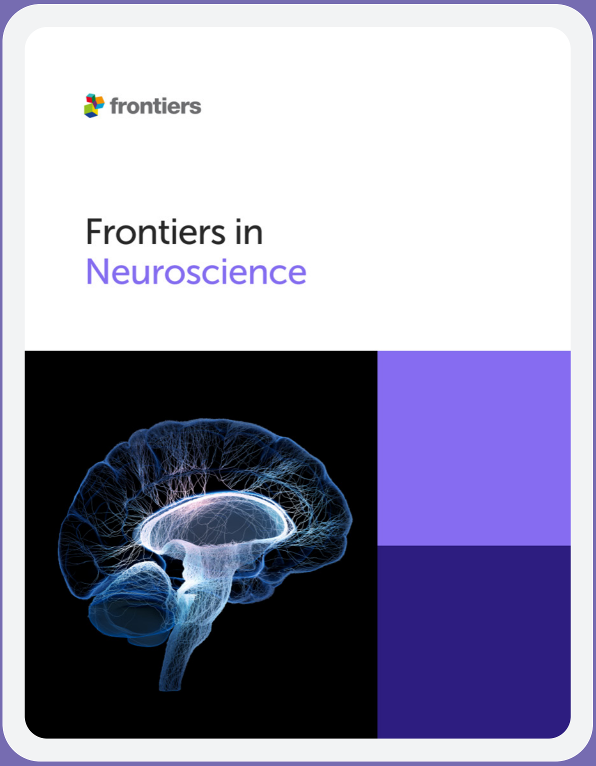
See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact



frontiers

Frontiers in
Neuroscience