MULTIPLE SCLEROSIS JOURNAL | MSJ

Original Research Paper

A multi-modal deep learning network for the classification of paramagnetic rim and remyelinated lesions in multiple sclerosis

Multiple Sclerosis Journal

2026, Vol. 32(6) 624–633

DOI: 10.1177/13524585261421491

Federico Spagnolo, Pedro M Gordaliza, Aarushi Bhardwaj, Po-Jui Lu, Mario Ocampo-Pineda, Meritxell Bach Cuadra, Matthias Weigel, Esther Ruberte, Xinjie Chen, Tommaso Sirito, Alessandro Cagol, Vincent Andrearczyk, Adrien Depeursinge and Cristina Granziera

Abstract

Objectives: Robust automated classification of paramagnetic rim lesions (PRLs) and remyelinated lesions based on iron and myelin content in people with multiple sclerosis (pwMS).

Materials and Methods: In this prospective study (2018–2022), three-dimensional (3D) brain magnetic resonance imaging (MRI) from 180 pwMS (mean age: 47 ± 14 years, 108 females) were acquired at 3T including fluid-attenuated inversion recovery, magnetization-prepared 2 rapid acquisition gradient echoes (MP2RAGE), and T2*-weighted segmented echo-planar imaging. Post-processed quantitative susceptibility mapping (QSM) and filtered phase unwrapped (PU) images were generated. Ground truth was established through expert manual rating. Performance was assessed using nested cross-validation (CV) including outer test sets. Three neural network configurations were evaluated: (1) PRL classification with QSM and MP2RAGE, (2) PRL classification with PU and MP2RAGE, and (3) multiple lesion phenotype (MLP) classification with QSM and MP2RAGE.

Results: For PRL classification, our network achieves a top mean validation F1 score of 0.737 ± 0.027 in the MP2RAGE-QSM configuration trained on PRLs, and a best test performance of 0.709 ± 0.040 when trained on MLP. The MP2RAGE-PU configuration yielded validation and test F1 scores of 0.733 ± 0.021 and 0.662 ± 0.037, respectively. The QSM-based MLP classification achieved a macro F1 test score of 0.728 ± 0.012.

Conclusion: Deep learning can automate the classification of PRLs and QSM lesion phenotypes with high accuracy, potentially facilitating the detection of PRLs (included in the new diagnostic criteria) and remyelinated lesions, guiding personalized treatment decisions in pwMS.

Keywords: Multiple sclerosis, deep learning, paramagnetic rim lesions, remyelination, MRI

Date received: 15 September 2025; revised: 19 December 2025; accepted: 20 January 2026

Correspondence to:
C Granziera
Translational Imaging in Neurology (ThINK) Basel, Department of Biomedical Engineering, Faculty of Medicine, University Hospital Basel and University of Basel, Basel, Switzerland.
Hegenheimermattweg 167B/C, 4123 Allschwil, Switzerland.
cristina.granziera@unibas.ch

Federico Spagnolo
Translational Imaging in Neurology (ThINK) Basel, Department of Biomedical Engineering, Faculty of Medicine, University Hospital Basel and University of Basel, Basel, Switzerland; Department of Neurology, University Hospital Basel, Basel, Switzerland
Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University Hospital Basel and University of Basel, Basel, Switzerland; MedGIFT, Institute of Informatics, School of Management, HES-SO Valais-Wallis University of Applied Sciences and Arts Western Switzerland, Sierre, Switzerland

Pedro M Gordaliza
CIBM, Center for Biomedical Imaging, Lausanne, Switzerland; Radiology Department, Lausanne University Hospital (CHUV) and University of Lausanne, Lausanne, Switzerland

Aarushi Bhardwaj
Translational Imaging in Neurology (ThINK) Basel, Department of Biomedical Engineering, Faculty of Medicine, University Hospital Basel and University of Basel, Basel, Switzerland; School of Engineering, University of Guelph, Guelph, ON, Canada

Po-Jui Lu
Translational Imaging

Introduction

Chronic active lesions (CALs), a subset of multiple sclerosis (MS) plaques in the white matter (WM), are marked by smoldering inflammation and are associated with more severe disease progression.[1–3] In contrast, chronic inactive lesions (CILs) are fully demyelinated and show inflammation and neurodegeneration levels similar to age-matched controls.[4]

CALs are definitively confirmed via histopathology only. However, imaging correlates have been identified,[5] including the paramagnetic rim seen on 7, 3, or 1.5T magnetic resonance imaging (MRI) using sequences like three-dimensional (3D) T2*-weighted GRE or echo-planar imaging (EPI). Reconstruction methods using magnitude and phase data[6] support the detection of paramagnetic rim lesions (PRLs) via susceptibility-weighted imaging (SWI), phase unwrapped (PU) images, or quantitative susceptibility mapping (QSM).

Manual PRL detection presents three key challenges: (1) confluent lesions—plaques that enlarge and merge with age[7]—must be separated to ensure clear

Table 1. Comparison of state-of-the-art methods: RimNet,[8] APRL,[9] QSMRim-Net,[10] and DeDA.[11]

Characteristics	APRL	RimNet	QSMRim-Net	DeDA
Resolution (mm^3)	0.65 iso	0.65 iso	$0.75 \times 0.93 \times 3$	$0.75 \times 0.75 \times 3$
Reconstruction	PU	PU	QSM	QSM
Radiomics	on PU	No	on QSM	No
Confluent lesions	Auto	Manual	Manual	Manual
Multi-labeled patches	No	Yes, exclusion	Yes, exclusion	Yes, none
External tests	None	One	None	None

The first and second rows report the resolution of T2*-w images in each MRI protocol and the adopted reconstruction method. In the work by Barquero et al.,[8] the separate hold-out test set had a resolution of 0.67 isotropic (iso). The third row reports whether the method included radiomic feature extraction (and on which MR contrast) or not. The fourth row refers to which approach was followed to split confluent lesions. Lou et al.[9] used the method described by Dworkin et al.[13] The fifth row reports whether the network was fed with patches containing multiple labels each, and which actions were taken: Barquero et al.[8] excluded patches of rim-negative lesions that also included PRLs; Zhang et al.[10] used manually annotated lesion masks to mask out non-lesion areas in input patches; Zhang et al.[11] used patches but did not report any actions taken. The last row shows how many external tests were conducted, after validation. Only Barquero et al.[8] reported test results, although only for a single hold-out test experiment.

identification; (2) it is extremely time-consuming; and (3) contrast variability impacts false positive/negative rates for both human raters and deep learning (DL).

Four automated approaches have tackled the first two issues.[8–11] While these methods show promise, each faces different limitations described in Table 1. In general, (1) acquiring images with sub-millimeter resolution would minimize partial volume and interpolation effects; (2) radiomic features, though informative, can hinder clinical adoption;[12] (3) patch-based methods may train the network with information partially coming from non-target lesions (e.g. non-PRLs can be in patches centered on PRLs and vice versa), which may result in the exclusion of selected patches,[8] or of perilesional signal[10] (masking non-lesion areas depends heavily on annotation accuracy); and (4) no robust estimate of methods' generalizability has been conducted beyond cross-validation (CV).

Another crucial unmet need is developing approaches that can help disentangle the heterogeneity of MS lesions, including remyelinated (REM) lesions. Manual classification in Rahmanzadeh et al.[13] identified CALs, CILs, and remyelinated lesions based on QSM signal patterns. Figure 1 illustrates the link between histological and imaging phenotypes. Automating this classification could aid in monitoring disease progression, for example, by quantifying rim lesions formation/disappearance and remyelination over time.

This paper aims to (1) explore the performance of PRLs classification; (2) compare QSM and PU inputs on the same images and lesions; and (3) expand on Rahmanzadeh et al.[13] to classify additional lesion phenotypes.

Materials and methods

Participants and imaging protocol

A total of 180 pwMS (mean age: 47 ± 14 years; 108 females; 58 secondary progressive multiple sclerosis (SPMS), 22 primary progressive multiple sclerosis (PPMS), 100 relapsing-remitting multiple sclerosis (RRMS)) were enrolled in a prospective study (NCT05177523) at the University Hospital of Basel. The advanced 3T brain MRI protocol included 3D fluid-attenuated inversion recovery (FLAIR), 3D magnetization-prepared 2 rapid gradient echo (MP2RAGE), and 3D T2*-weighted segmented EPI sequences. Conventional MRI with gadolinium was performed within 3 months to reasonably rule out acute lesions. Detailed acquisition parameters and inclusion/exclusion criteria are reported by Cagol et al.[3] The study was approved by the local ethics review committee, and all participants gave written consent prior to the study.

Image processing

PU and QSM maps were derived from T2* EPI images using the MEDI (1.0.376.GE) algorithm.[14] MP2RAGE skull-stripping employed FreeSurfer 6.0.0[15] and HD-BET.[16] FLAIR and MP2RAGE were registered to T2* space using FSL 6.0,[17] with MP2RAGE-based brain masks applied to FLAIR and PU images. QSM intensities were clipped between −200 and +200 parts per billion (*ppb*).

in Neurology (ThINK) Basel, Department of Biomedical Engineering, Faculty of Medicine, University Hospital Basel and University of Basel, Basel, Switzerland; Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University Hospital Basel and University of Basel, Basel, Switzerland

Meritxell Bach Cuadra
CIBM, Center for Biomedical Imaging, Lausanne, Switzerland; Radiology Department, Lausanne University Hospital (CHUV) and University of Lausanne, Lausanne, Switzerland

Mario Ocampo-Pineda
Esther Ruberte
Cristina Granziera
Translational Imaging in Neurology (ThINK) Basel, Department of Biomedical Engineering, Faculty of Medicine, University Hospital Basel and University of Basel, Basel, Switzerland; Department of Neurology, University Hospital Basel, Basel, Switzerland; Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University Hospital Basel and University of Basel, Basel, Switzerland

Matthias Weigel
Translational Imaging in Neurology (ThINK) Basel, Department of Biomedical Engineering, Faculty of Medicine, University Hospital Basel and University of Basel, Basel, Switzerland; Department of Neurology, University Hospital Basel, Basel, Switzerland; Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University Hospital Basel and University of Basel, Basel, Switzerland; Division of Radiological Physics, Department of Radiology, University Hospital Basel, Basel, Switzerland

Xinjie Chen
Translational Imaging in Neurology (ThINK) Basel, Department of Biomedical Engineering, Faculty of Medicine, University Hospital Basel and University of Basel, Basel, Switzerland; Department of Neurology, University Hospital Basel, Basel, Switzerland; Research Center for Clinical Neuroimmunology and

Neuroscience Basel (RC2NB), University Hospital Basel and University of Basel, Basel, Switzerland; Department of Pediatric Neurology and Developmental Medicine, University Children's Hospital Basel (UKBB), Basel, Switzerland

Tommaso Sirito
Alessandro Cagol
Translational Imaging in Neurology (ThINK) Basel, Department of Biomedical Engineering, Faculty of Medicine, University Hospital Basel and University of Basel, Basel, Switzerland; Department of Neurology, University Hospital Basel, Basel, Switzerland; Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University Hospital Basel and University of Basel, Basel, Switzerland; Dipartimento di Scienze della Salute, Università degli Studi di Genova, Genova, Italy

Vincent Andrearczyk
MedGIFT, Institute of Informatics, School of Management, HES-SO Valais-Wallis University of Applied Sciences and Arts Western Switzerland, Sierre, Switzerland

Adrien Depeursinge
MedGIFT, Institute of Informatics, School of Management, HES-SO Valais-Wallis University of Applied Sciences and Arts Western Switzerland, Sierre, Switzerland; Nuclear Medicine and Molecular Imaging Department, Lausanne University Hospital (CHUV) and University of Lausanne, Lausanne, Switzerland

Data Availability Statement included at the end of the article.

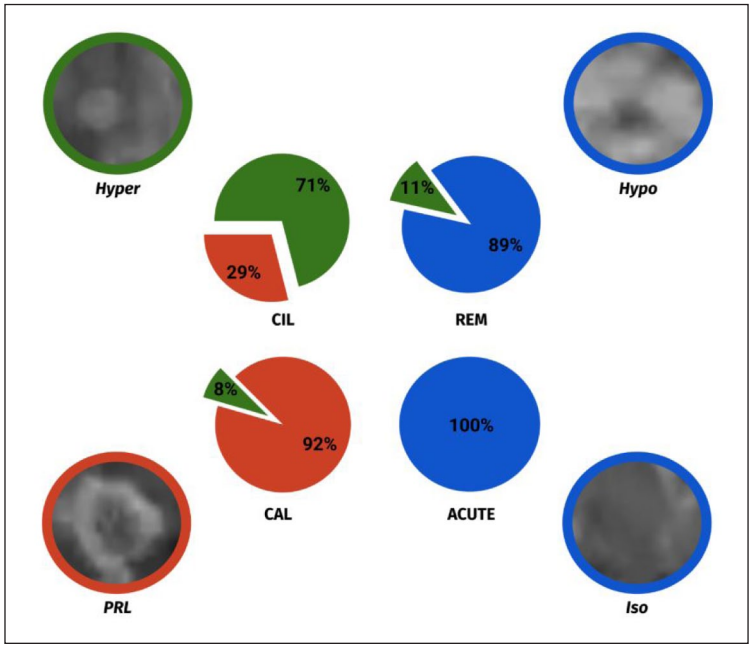

Figure 1. Illustration showing the relationships between histological phenotypes (CAL, CIL, REM, and acute) and radiological findings in QSM (PRL, hyper-intense, hypo-intense, and iso-intense).
The pie charts indicate the proportion of each histological phenotype that corresponds to specific radiological findings in QSM (defined by colors: red for PRL, green for hyper-intense, and blue for iso-/hypo-intense), as reported by Rahmanzadeh et al.[14] According to the study, 92% of chronic active lesions (CALs) were classified as paramagnetic rim lesions (PRLs), with the remaining 8% appearing hyper-intense. For chronic inactive lesions (CILs), 71% were classified as hyper-intense, while the rest showed a paramagnetic rim. In addition, 89% of remyelinated (REM) lesions appeared hypo- or iso-intense, with the remaining 11% classified as hyper-intense.

Lesion annotation

Lesion masks were initially generated via a 3D-UNet model[18] and manually corrected in FLAIR space by X.C. and E.R., a neurologist and a molecular biologist with 4 and >5 years of MS experience. Following registration to T2* space, lesions were reviewed, and confluent lesions were split by T.S., a neurology resident.

Lesions were classified twice using different contrasts. In the primary QSM-based assessment, T.S. categorized lesions into multiple lesion phenotypes (MLPs): (1) rim-positive (PRL), (2) hyper-intense, (3) iso-intense, (4) hypo-intense, and (5) unclassified (small lesions with QSM artifacts and confluent lesions with indistinguishable individual boundaries).

A.C., a neurologist with >5 years of MS experience, verified the classifications. Iso- and hypo-intense lesions, due to phenotype similarity (suggestive of remyelination),[13] were merged (hereafter "iso-intense").

In the secondary PU-based assessment, A.C. performed a binary PRL/non-PRL classification. All annotations were done using 3D Slicer 5.0.2.[19] An example of MR slice including the three different classes of the first assessment is shown in Figure 2.

Network architecture

Our multi-modal CNN builds on the architecture of Barquero et al.,[8] processing paired $28 \times 28 \times 28$ patches from MP2RAGE and QSM/PU through parallel branches (Figure 3). A third branch processes the dilated lesion mask (1 voxel in all directions) with a single convolutional block, before concatenating its features (early fusion) with those of the other two branches. This architectural choice aims to guide feature extraction within the lesional region, while preserving perilesional features and mitigating segmentation errors. Features from the remaining two branches are later concatenated (late fusion) and then fed through fully connected layers (128 and 32 neurons, dropout rate 0.5) to predict lesion classes.

Data set stratification

Data were split for nested CV: a threefold inner loop (model selection) and fivefold outer loop (performance evaluation). Patients were stratified

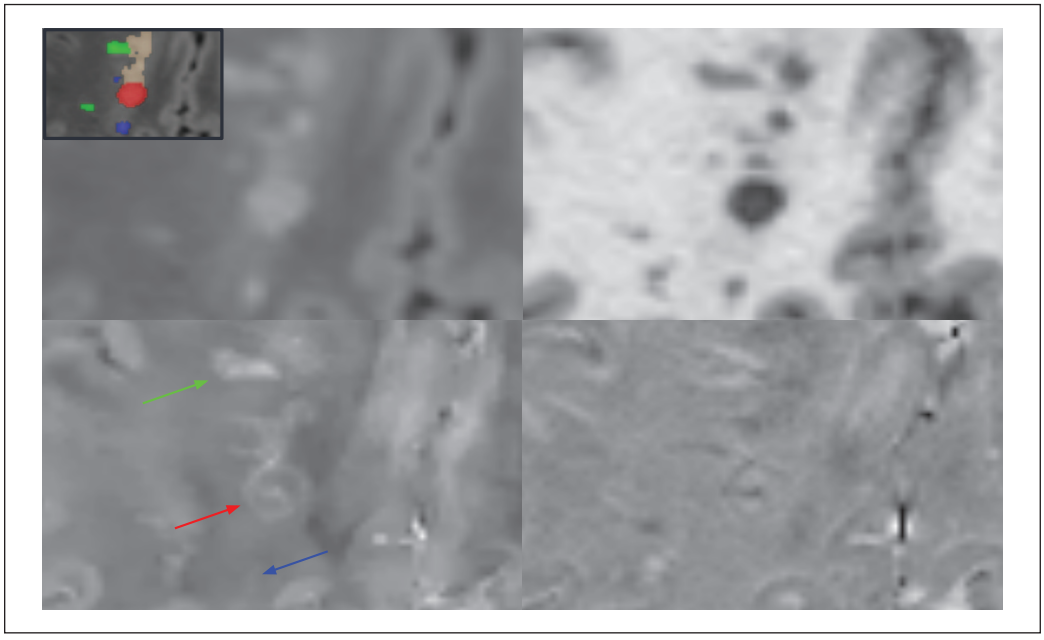

Figure 2. Axial slice of a FLAIR image (top left), MP2RAGE (top right), QSM (bottom left), and PU (bottom right). The same FLAIR image appears in the thumbnail with the manual lesion mask overlaid, presenting QSM phenotypes with different colors: red for PRL, green for hyper-intense, blue for iso-/hypo-intense, and light brown for unclassified components. The QSM image shows arrows pointing to an example of each class used during training.

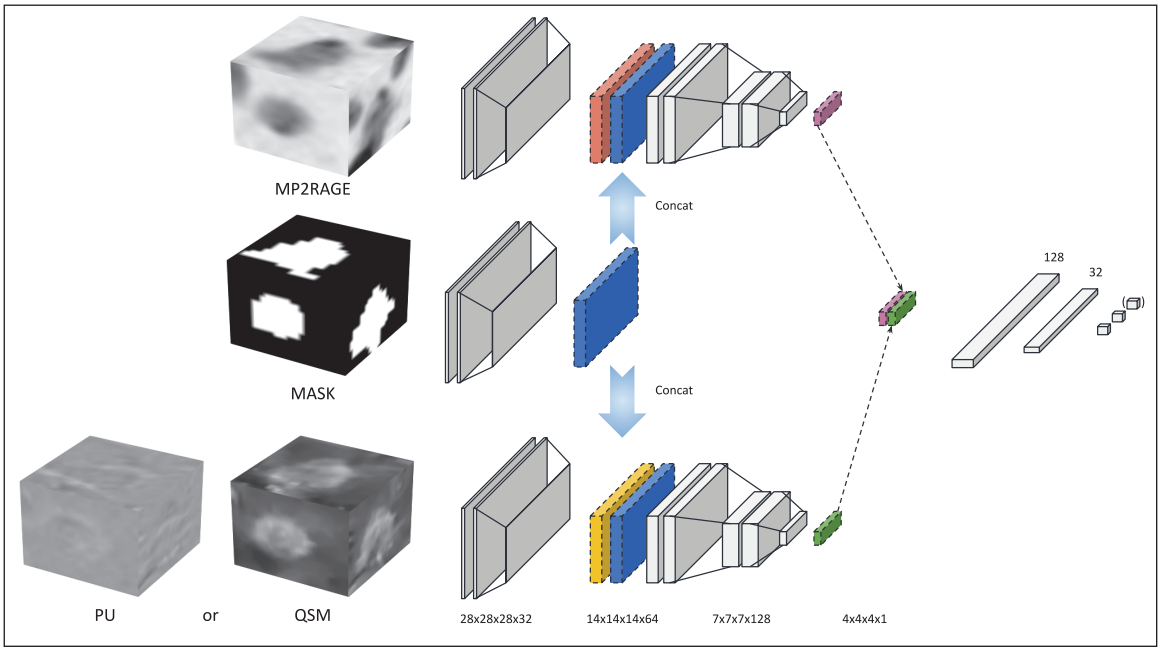

Figure 3. Our architecture leverages three inputs, MP2RAGE, the dilated binary lesion mask, and QSM or PU. The mask branch stops at the first convolutional block and fuses its feature maps with those of the other two parallel branches. After two more convolutional blocks, feature maps of the first and third branches are fused and passed through fully connected layers.

into four groups based on PRL counts (0, 1–3, 4–7, and >7) to balance folds. For each lesion, patches were cropped from MP2RAGE, PU/QSM, and lesion masks. MP2RAGE and PU were min–max normalized; QSM values scaled from (−200, 200) *ppb* to (0, 1).

Training

Patches of $28 \times 28 \times 28$ were first cropped around each lesion's center. These same patches were then used for all the following experiments. To address class imbalance, we augmented the training set by generating additional patches for underrepresented classes. For each PRL patch, we created seven additional patches by shifting the crop center by five voxels in various directions. For the MLP case, we similarly generated four extra patches for each hyper-intense lesion. We explored three configurations: (1) PRL classification with MP2RAGE and QSM, (2) MLP (PRL/hyper-/iso- intense) classification with MP2RAGE and QSM, and (3) PRL classification with MP2RAGE and PU. Best models were selected for testing based on macro F1 scores. An ablation study compared performance when removing branches or changing secondary inputs (e.g. FLAIR).

For a qualitative check of the network's attention (explainability), we used integrated gradients.[20] The resulting heatmap highlights the voxels that the model attributes the prediction to. Code was implemented in Python 3.8.0 and PyTorch 1.13.0, run on a NVIDIA A100 SXM4 GPU (40GB VRAM), and is publicly available at our *Github repository*.

Ablation study

In a first ablation study focusing on the role of the mask branch, we trained the network with only two inputs (MP2RAGE and QSM) for PRL classification under the same conditions described in the previous section, removing the guiding branch and retaining the early fusion by concatenating features of the MP2RAGE branch with those of the QSM branch (resulting in the QSM branch having double the feature maps of the MP2RAGE branch, reproducing the architecture in the work by Barquero et al.[8]).

In a second experiment, we removed both the mask branch and the early fusion of features. To this end, we adapted the fully connected layers to the sizes of 64 and 16. Both this and the first study aim to show the impact of the guiding branch on the classification performance.

As a third ablation study aiming to quantify the impact of QSM signal on the classification of PRLs, we, in addition, report test results of the network removing the QSM branch from the previous architecture, leaving a single CNN branch using MP2RAGE as input.

A fourth experiment extended the third ablation study, with the addition of the guiding branch with early fusion of features from the lesion mask. This analysis further evaluates the possible advantage of a guiding branch, even when the only input is MP2RAGE.

A further analysis consisted in our best architecture, replacing MP2RAGE with FLAIR as secondary input for PRL classification.

Results

All patients enrolled in the study were included in the analyses. Table 1 in the Supplemental Material describes the number of patients and their MS phenotype for each training/testing partition.

In the primary assessment (on QSM), the lesion counts for the first four classes were as follows: 453 PRLs, 1212 hyper-intense, 4726 iso-intense, and 119 hypo-intense lesions. The fifth class, comprising 6654 lesions with artifacts or inseparable components, was excluded. In the secondary assessment (on PU), we identified 551 PRLs and 5959 non-PRLs. Lesions belonging to the fifth category in the primary assessment were also excluded from the secondary assessment to ensure that the same pool of lesions was considered in both PU and QSM analyses.

Table 2 summarizes the CV and test results of the following configurations: (1) multiclass classification network evaluated on MLP (reporting also results on PRLs only) with QSM, (2) binary classification network evaluated on PRLs with QSM, and (3) binary classification network evaluated on PRLs with PU. All results refer to performance at the lesion level. Performances of optimal models on all data partitions are reported in the Supplemental Material.

The per-class validation F1 scores (mean $\pm$ standard deviation) for the MLP classification case are 0.902 ± 0.011 for the iso-intense class, 0.728 ± 0.044 for PRLs, and 0.609 ± 0.028 for the hyper-intense class. Similarly, at test, we report 0.901 ± 0.011 for the iso-intense class, 0.709 ± 0.040 for PRLs, and 0.574 ± 0.042 for the hyper-intense class.

In Table 3, we report F1 score, precision, and recall of our methods to those of current state-of-the-art (SOTA) methods for PRL classification (each on its own data set), separating validation from testing results. Our test performance in PRL classification is also presented in Figure 4 where, for each configuration, boxplots report

Table 2. Validation and test metrics (mean ± standard deviation) for the following three classification cases: multiclass classification on multiple lesion phenotypes using MP2RAGE and QSM, binary classification on PRLs using MP2RAGE and QSM, and binary classification on PRLs using MP2RAGE and PU. Macro and weighted (W.) metrics are only evaluated for the multiclass case, which also reports metrics relative to the classification of PRLs only.

Results	MP2RAGE-QSM multiclass	MP2RAGE-QSM binary	MP2RAGE-PU binary
Validation			
F1 Macro	0.746 ± 0.020	–	–
ROC AUC Macro	0.907 ± 0.008	–	–
F1 PRL	0.728 ± 0.044	0.737 ± 0.027	0.733 ± 0.021
ROC AUC PRL	0.972 ± 0.008	0.968 ± 0.012	0.969 ± 0.008
Test			
F1 Macro	0.728 ± 0.012	0.688 ± 0.058	0.662 ± 0.037
Precision Macro	0.733 ± 0.026	0.700 ± 0.057	0.682 ± 0.041
Recall Macro	0.729 ± 0.017	0.682 ± 0.088	0.644 ± 0.051
F1 PRL	0.709 ± 0.040	0.688 ± 0.058	0.662 ± 0.037
Precision PRL	0.740 ± 0.072	0.700 ± 0.057	0.682 ± 0.041
Recall PRL	0.684 ± 0.042	0.682 ± 0.088	0.644 ± 0.051
F1 W.	0.827 ± 0.012	–	–
Precision W.	0.833 ± 0.012	–	–
Recall W.	0.824 ± 0.013	–	–
ROC AUC Macro	0.902 ± 0.010	0.964 ± 0.013	0.960 ± 0.013
ROC AUC PRL	0.968 ± 0.011	0.964 ± 0.013	0.960 ± 0.013
ROC AUC W.	0.896 ± 0.012	–	–

Table 3. Validation (Val) and test results of our network and (when available) other state-of-the-art methods for PRL classification, each on their own data: RimNet[8] and APRL[9] using PU images, QSMRim-Net,[10] and DA-TR-Net[11] using QSM. For our MLP configuration, we report the scores related to the PRL class.

Networks	F1		Precision		Recall		ROC	AUC
	Val	Test	Val	Test	Val	Test	Val	Test
APRL	0.552	–	0.436	–	0.750	–	0.820	–
RimNet	–	0.623	–	0.528	–	0.758	0.958	0.953
QSMRim-Net	0.703	–	0.732	–	0.680	–	–	–
DA-TR-Net	0.750	–	0.792	–	0.712	–	0.975	–
Ours (MLP)	0.728	0.709	0.749	0.740	0.719	0.684	0.971	0.968
Ours (PRL QSM)	0.737	0.688	0.748	0.700	0.731	0.682	0.968	0.964
Ours (PRL PU)	0.733	0.662	0.738	0.682	0.729	0.644	0.969	0.960

median, 25th and 75th percentile. In Figure S1 of the Supplemental Material, we show an example of explainable map computed for a patch that contains two confluent PRLs.

Ablation study

Table 4 provides an overview of the architectural changes in the ablation studies and the resulting performance.

The first ablation study removed the binary lesion mask as input (leaving MP2RAGE and QSM) and retained the fusion of low-level features. The test results showed a mean F1 score of 0.587 ± 0.060, which corresponds to a decrease of almost 15% from that reported in Tables 2 and 3 for the PRL classification task using QSM.

The second ablation study removed both the binary lesion mask as input and the fusion of low-level

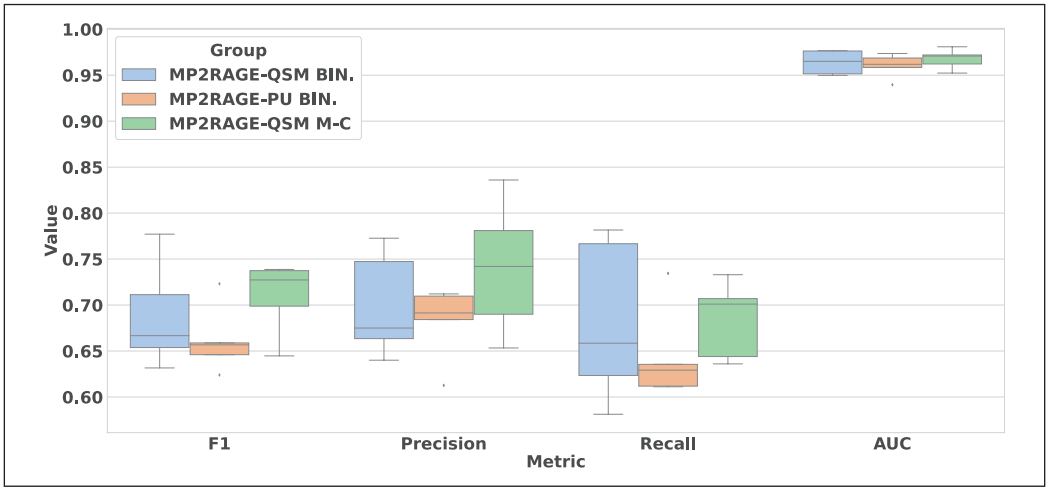

Figure 4. Boxplot of the test performance in the three configurations for PRL classification: (1) training on PRLs using MP2RAGE and QSM (blue), (2) training on PRLs using MP2RAGE and PU (orange), and (3) training on MLP using MP2RAGE and QSM (green). Metrics are reported in different columns: F1, precision, recall, and AUC.

Table 4. Description of the best performing architecture and the ablation studies (one for each row, counting the additional replacement of MP2RAGE with FLAIR).

Architectures	MR inputs	Early fusion	Total branches	Guiding branch	Test performance
Best QSM	MP2RAGE, QSM	Yes	3	Yes	0.688 ± 0.058
First	MP2RAGE, QSM	Yes	2	No	0.587 ± 0.060
Second	MP2RAGE, QSM	No	2	No	0.528 ± 0.047
Third	MP2RAGE	No	1	No	0.380 ± 0.016
Fourth	MP2RAGE	Yes	2	Yes	0.517 ± 0.076
Additional	FLAIR, QSM	Yes	3	Yes	0.642 ± 0.033

The first column reports the MR inputs to the network. The second specifies whether low-level features were concatenated. The third counts the total number of CNN branches, and the fourth shows if the mask was used to guide feature extraction. The last column reports the test performance (F1 score) of the different architectures.

features (again MP2RAGE and QSM as inputs). The test results showed a mean F1 score of 0.528 ± 0.047, which corresponds to a 10% decrease with respect to the first ablation study.

The third experiment consisted in a single CNN branch using MP2RAGE as the only input and reported a test F1 score of 0.380 ± 0.016. This shows a drop of 28% from the second ablation study.

In the last experiment, restoring the guiding branch (inputs were then MP2RAGE and mask), the test F1 score was 0.517 ± 0.076. This represents an increase of 36% from the third study.

In addition, the use of FLAIR as secondary input for the PRL classification yielded a mean test F1 score of 0.642 ± 0.033, which corresponds to 6.7% less than what we reported with MP2RAGE in Tables 2 and 3.

Discussion

In the present work, we developed a DL application which leverages multi-modal MRI to classify PRLs in pwMS. We performed a stratified nested CV to compare validation and test scores for different network's configurations: (1) classification of PRLs using MP2RAGE and QSM; (2) classification of PRLs using

MP2RAGE and PU; and (3) MLP classification of QSM lesion phenotypes based on QSM signal patterns.

The results obtained across all network configurations are competitive with SOTA methods, close but below those of Zhang et al.[11] As mentioned in the Introduction, their data and that used in this manuscript present some differences, for example, MRI resolution. From a methodological perspective, it is interesting to point out that their lesion masks were manually refined based on QSM intensities. This means that, for example, confluent lesions were furtherly separated to match QSM hyper-intense rims. This might indeed have helped the network achieve a better performance, at the expense of an additional manual intervention and a final discrepancy between the final annotations—with rim and negative lesions—and the initial masks annotated in FLAIR. Due to the unavailability of the code by Zhang et al.,[11] we were unable to train and validate their method on our data set. However, we were able to reproduce the architecture from Barquero et al.[8] (first ablation study), and we observed a significant drop in test performance compared with our network. Notably, our work is the only one to report multiple test results, providing confident estimates of the generalizability of our approach. Although we do report a performance decrease between validation and test, the validation loss did not increase during training, suggesting that the network was not overfitting the training set.

Our work offers two main contributions. First, we show that low-level features from lesion masks can significantly improve CNN performance by guiding the network's attention. Results from the second and fourth ablation studies confirm that the mask branch enhances feature extraction from both QSM and MP2RAGE. It also improves interpretability, helping the network handle patches containing multiple lesions (Figure 1 of the Supplemental Material).

Second, although mean test results of the MP2RAGE-QSM configuration were higher than those of the MP2RAGE-PU configuration, the difference in performance between the two was smaller than the standard deviation of both. To date, there is no consensus on the ideal MR sequence for PRL assessment.[6] Although QSM has shown stronger associations with myelin loss and disability,[21] it remains limited to research settings due to its high computational demands. Since our results do not clearly favor QSM, PU imaging may have an advantage for clinical integration of automated PRL classification.

Our study differs from previous studies in key aspects. Lesions were segmented semi-automatically and split manually, unlike in Lou et al.[9] We evaluate both QSM and PU within the same architecture, unlike prior work that focused on a single MRI contrast.[8–11] Our imaging includes sub-millimeter isotropic resolution, reducing partial volume and interpolation artifacts compared with Zhang et al.[10,11] Finally, our study constitutes a first attempt on automated QSM multiple lesion phenotyping to identify remyelinated lesions, which may support prognosis, treatment response assessment, and patient stratification.[22]

We also acknowledge some limitations of this study. Partially relying on manual input (i.e. semi-automatic annotation of lesion masks) to extract patches hinders clinical applicability; yet integrating lesion segmentation tools—such as those described by Spagnolo et al.[23]—could help. However, confluent lesions remain challenging to manage. While Lou et al.[9] used the method proposed by Dworkin et al.[24] to split such lesions, it overlooks lesion size and can erroneously assign voxels to the closest center.[25] Second, our approach relied on the exclusion of a non-negligible pool of lesions, which were segmented on FLAIR. This methodological choice can be seen as a hurdle toward its implementation, but aimed at a more precise data set curation. The above issue is a well-known limitation in the field, and our approach is consistent with previous studies.[13] Third, our study was conducted in a single center. A multi-centric study[26] has shown that data heterogeneity reduces performance, although coordinated manual annotations across centers are difficult to obtain.

The MLP experiment also has limitations. Our iso-intense class groups both acute and remyelinated lesions, which are pathologically distinct. They appear similar on QSM; acute lesions reflect active inflammation, while remyelinated ones indicate repair. While we excluded acute lesions in this data set, contrast-enhanced MRI could improve future phenotyping. Furthermore, our network showed lower accuracy for the hyper-intense class, likely due to (1) lack of stratification for this class; (2) its intermediate characteristics between PRLs and iso-intense lesions; and (3) neglecting contralateral QSM intensity, which influenced manual annotation but was not included in our patch-based input.

It is also worth discussing the choice of MP2RAGE rather than FLAIR as the complementary contrast. While FLAIR helps delineate lesion boundaries, this role is partially covered by the annotated binary mask. Moreover, T1 hypointensity has been linked to greater tissue damage in PRLs,[27,28] and our results showed

better mean performance using MP2RAGE, supporting its relevance for PRL classification, although by a thin margin. Considered alongside the QSM–PU comparison, this pattern indicates that the method is likely to reach its best performance in research settings equipped with MP2RAGE and QSM, yet remains effective in clinical environments where FLAIR and SWI or EPI phase sequences are more commonly acquired.

In conclusion, the proposed approach can serve as a foundation for future studies aimed at identifying PRLs in clinical practice. In addition, the MLP configuration can aid in detecting remyelinated lesions and, therefore, be integrated into pipelines for the development of remyelinating drugs.

Acknowledgements
We acknowledge access to the expertise of the CIBM Center for Biomedical Imaging, a Swiss research center of excellence founded and supported by CHUV, UNIL, EPFL, UNIGE, and HUG.

Declaration of Conflicting Interests
The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The USB, as the employer of C. Granziera, has received the following fees which were used exclusively for research support: (1) advisory board and consultancy fees from Actelion, Genzyme-Sanofi, Novartis, GeNeuro, and Roche; (2) speaker fees from Genzyme-Sanofi, Novartis, GeNeuro, and Roche; and (3) research support from Siemens, GeNeuro, and Roche. C. Granziera was supported by the Swiss National Science Foundation grant PP00P3_176984, the Stiftung zur Förderung der gastroenterologischen und allgemeinen klinischen Forschung, and the EUROSTAR E!113682 HORIZON2020.
A.C. has received speaker honoraria from Novartis and Roche.
All other authors have nothing to disclose.

Ethical Considerations
The study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethic Review Committee of the University of Basel (no. NCT05177523), with the need for written informed consent waived.

Funding
The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Hasler Foundation with the project MSxplain number 21042, the Swiss National Science Foundation (SNSF) with the project 205320_219430, and the Swiss Cancer Research foundation with the project TARGET (KFS-5549-02-2022-R).

ORCID iDs
Federico Spagnolo https://orcid.org/0000-0001-9606-0400
Pedro M Gordaliza https://orcid.org/0000-0001-5305-3104
Aarushi Bhardwaj https://orcid.org/0009-0005-8077-1298
Po-Jui Lu https://orcid.org/0000-0001-6031-6865
Mario Ocampo-Pineda https://orcid.org/0000-0003-2239-7355
Meritxell Bach Cuadra https://orcid.org/0000-0003-2730-4285
Matthias Weigel https://orcid.org/0000-0002-4910-1414
Esther Ruberte https://orcid.org/0000-0002-0195-4602
Xinjie Chen https://orcid.org/0000-0002-3683-6489
Tommaso Sirito https://orcid.org/0000-0002-9016-8709
Alessandro Cagol https://orcid.org/0000-0002-7703-0553
Vincent Andrearczyk https://orcid.org/0000-0003-0793-5821

Data Availability
The data that support the findings of this study are available from the University Hospital of Basel, Switzerland, but restrictions apply to the availability of these data, which were used under license for this study, and so are not publicly available. Data are, however, available from the authors upon reasonable request and with permission of the University Hospital of Basel, Switzerland.

Supplemental Material
Supplemental material for this article is available online.

References
1. Luchetti S, Fransen NL, van Eden CG, et al. Progressive multiple sclerosis patients show substantial lesion activity that correlates with clinical disease severity and sex: A retrospective autopsy cohort analysis. *Acta Neuropathol* 2018; 135(4): 511–528.
2. Absinta M, Sati P, Masuzzo F, et al. Association of chronic active multiple sclerosis lesions with disability in vivo. *JAMA Neurol* 2019; 76: 1474–1483.

3. Cagol A, Ocampo-Pineda M, Lu P-J, et al. Advanced quantitative MRI unveils microstructural thalamic changes reflecting disease progression in multiple sclerosis. *Neurol Neuroimmunol Neuroinflamm* 2024; 11: e200299.
4. Popescu BF, Pirko I and Lucchinetti CF. Pathology of multiple sclerosis: Where do we stand. *Continuum* 2013; 19(4): 901–921.
5. Dal-Bianco A, Oh J, Sati P, et al. Chronic active lesions in multiple sclerosis: Classification, terminology, and clinical significance. *Therap Adv Neurol Dis* 2024; 17: 1306684.
6. Bagnato F, Sati P, Hemond CC, et al. Imaging chronic active lesions in multiple sclerosis: A consensus statement. *Brain* 2024; 1479: 2913–2933.
7. Kim KW, MacFall J and Payne M. Classification of white matter lesions on magnetic resonance imaging in elderly persons. *Biol Psychiatry* 2008; 64: 273–280.
8. Barquero G, La Rosa F, Kebiri H, et al. RimNet: A deep 3D multimodal MRI architecture for paramagnetic rim lesion assessment in multiple sclerosis. *Neuroimage Clin* 2020; 28: 102412.
9. Lou C, Sati P, Absinta M, et al. Fully automated detection of paramagnetic rims in multiple sclerosis lesions on 3T susceptibility-based MR imaging. *Neuroimage Clin* 2021; 32: 102796.
10. Zhang H, Nguyen TD, Zhang J, et al. QSMRim-Net: Imbalance-aware learning for identification of chronic active multiple sclerosis lesions on quantitative susceptibility maps. *Neuroimage Clin* 2022; 34: 102979.
11. Zhang H, Wang R, Hu R, et al. DeDA: Deep directed accumulator. In: Greenspan H, Madabhushi A, Mousavi P, et al. (eds) *Medical image computing and computer assisted intervention—MICCAI 2023*, vol. 14221, Vancouver, BC, Canada: Springer Nature Switzerland, 2023, pp. 765–775.
12. Wang Y, Hu Z and Wang H. The clinical implications and interpretability of computational medical imaging (radiomics) in brain tumors. *Insights Imag* 2025; 16: 77.
13. Rahmanzadeh R, Galbusera R, Lu PJ, et al. A new advanced MRI biomarker for remyelinated lesions in multiple sclerosis. *Ann Neurol* 2022; 92(3): 486–502.
14. Liu T, Xu W, Spincemaille P, et al. Accuracy of the morphology enabled dipole inversion (MEDI) algorithm for quantitative susceptibility mapping in MRI. *IEEE Trans Med Imaging* 2012; 31(3): 816–824.
15. Fujimoto K, Polimeni J, Kouwe A, et al. Quantitative comparison of cortical surface reconstructions from MP2RAGE and Multi-Echo MPRAGE data at 3 and 7 Tesla. *Neuroimage* 2013; 90: 60–73.
16. Isensee F, Schell M, Pflueger I, et al. Automated brain ex- traction of multisequence MRI using artificial neural networks. *Human Brain Mapping* 2019; 40: 4592–4964.
17. Greve D and Fischl B. Accurate and robust brain image alignment using boundary-based registration. *Neuroimage* 2009; 48: 63–72.
18. La Rosa F, Abdulkadir A, Fartaria MJ, et al. Multiple sclerosis cortical and WM lesion segmentation at 3T MRI: A deep learning method based on FLAIR and MP2RAGE. *Neuroimage Clin* 2020; 27: 102335.
19. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D slicer as an image computing platform for the quantitative imaging network. *Magn Reson Imaging* 2012; 30(9): 1323–1341.
20. Sundararajan M, Taly A and Yan Q. Axiomatic attribution for deep networks. *arXiv*. 1703:01365, 2017, https://arxiv.org/abs/1703.01365
21. Huang W, Sweeney EM, Kaunzner UW, et al. Quantitative susceptibility mapping versus phase imaging to identify multiple sclerosis iron rim lesions with demyelination. *J Neuroimaging* 2022; 324: 667–675.
22. Khormi I, Al-iedani O, Abdulaziz A, et al. MR myelin imaging in multiple sclerosis: A scoping review. *J Neurol Sci* 2023; 455: 122807.
23. Spagnolo F, Depeursinge A, Schädelin S, et al. How far MS lesion detection and segmentation are integrated into the clinical workflow? A systematic review. *Neuroimage Clin* 2023; 39: 103491.
24. Dworkin JD, Linn KA, Oguz I, et al. An automated statistical technique for counting distinct multiple sclerosis lesions. *AJNR: Am J Neuroradiol* 2018; 39(4): 626–633.
25. Wynen M, Istasse M, Gordaliza PM, et al. ConfLUNet: Multiple sclerosis lesion instance segmentation in presence of confluent lesions. *Comput Biol Med* 2024; 197: 110919.
26. Chen L, Ren Z, Clark KA, et al. Multicenter validation of automated detection of paramagnetic rim lesions on brain MRI in multiple sclerosis. *J Neuroimaging* 2024; 34(6): 750–757.
27. Absinta M, Sati P, Schindler M, et al. Persistent 7-tesla phase rim predicts poor outcome in new multiple sclerosis patient lesions. *J Clin Invest* 2016; 126: 2597–2609.
28. Blindenbacher N, Brunner E, Asseyer S, et al. Evaluation of the "ring sign" and the "core sign" as a magnetic resonance imaging marker of disease activity and progression in clinically isolated syndrome and early multiple sclerosis. *Mult Scler J Exp Transl Clin* 2020; 6(1): 915480.