


Global and Regional Deep Learning Models for Multiple Sclerosis Stratification From MRI

Llucia Coll, MSc,^{1*}  Deborah Pareto, PhD,² Pere Carbonell-Mirabent, MSc,¹ Álvaro Cobo-Calvo, PhD,¹ Georgina Arrambide, PhD,¹ Ángela Vidal-Jordana, PhD,¹ Manuel Comabella, PhD,¹ Joaquín Castelló, PhD,¹ Breogán Rodríguez-Acevedo, MD,¹ Ana Zabalza, PhD,¹ Ingrid Galán, MD,¹ Luciana Midaglia, MD,¹ Carlos Nos, MD,¹ Cristina Auger, MD,² Manel Alberich, RT,² Jordi Río, PhD,¹ Jaume Sastre-Garriga, PhD,¹ Arnau Oliver, PhD,³ Xavier Montalban, PhD,¹ Àlex Rovira, MD,² Mar Tintoré, PhD,¹ Xavier Lladó, PhD,³ and Carmen Tur, PhD¹

Background: The combination of anatomical MRI and deep learning-based methods such as convolutional neural networks (CNNs) is a promising strategy to build predictive models of multiple sclerosis (MS) prognosis. However, studies assessing the effect of different input strategies on model's performance are lacking.

Purpose: To compare whole-brain input sampling strategies and regional/specific-tissue strategies, which focus on a priori known relevant areas for disability accrual, to stratify MS patients based on their disability level.

Study Type: Retrospective.

Subjects: Three hundred nineteen MS patients (382 brain MRI scans) with clinical assessment of disability level performed within the following 6 months (~70% training/~15% validation/~15% inference in-house dataset) and 440 MS patients from multiple centers (independent external validation cohort).

Field Strength/Sequence: Single vendor 1.5 T or 3.0 T. Magnetization-Prepared Rapid Gradient-Echo and Fluid-Attenuated Inversion Recovery sequences.

Assessment: A 7-fold patient cross validation strategy was used to train a 3D-CNN to classify patients into two groups, Expanded Disability Status Scale score (EDSS) ≥ 3.0 or EDSS < 3.0 . Two strategies were investigated: 1) a global approach, taking the whole brain volume as input and 2) regional approaches using five different regions-of-interest: white matter, gray matter, subcortical gray matter, ventricles, and brainstem structures. The performance of the models was assessed in the in-house and the independent external cohorts.

Statistical Tests: Balanced accuracy, sensitivity, specificity, area under receiver operating characteristic (ROC) curve (AUC).

Results: With the in-house dataset, the gray matter regional model showed the highest stratification accuracy (81%), followed by the global approach (79%). In the external dataset, without any further retraining, an accuracy of 72% was achieved for the white matter model and 71% for the global approach.

Data Conclusion: The global approach offered the best trade-off between internal performance and external validation to stratify MS patients based on accumulated disability.

Evidence Level: 4

Technical Efficacy: Stage 2

J. MAGN. RESON. IMAGING 2023.

View this article online at wileyonlinelibrary.com. DOI: 10.1002/jmri.29046

Received Apr 27, 2023, Accepted for publication Sep 18, 2023.

*Address reprint requests to: L.C., Passeig de Vall d'Hebron 119-129, 08035 Barcelona, Spain.

E-mail: llcoll@cem-cat.org

From the ¹Multiple Sclerosis Centre of Catalonia (Cemcat), Hospital Universitari Vall d'Hebron, Universitat Autònoma de Barcelona, Barcelona, Spain;

²Section of Neuroradiology, Department of Radiology, Hospital Universitari Vall d'Hebron, Universitat Autònoma de Barcelona, Barcelona, Spain; and

³Research Institute of Computer Vision and Robotics, University of Girona, Girona, Spain

Additional supporting information may be found in the online version of this article

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Multiple sclerosis (MS) is a chronic autoimmune disease of the central nervous system and is the main non-traumatic cause of irreversible disability in young adults.¹

MRI and associated biomarkers play an important role in the diagnosis, monitoring, and prognosis of MS, allowing quantification and tracking of patients' tissue structural damage over time.^{1,2} Biomarkers such as the number of lesions,¹ brain volume quantification³ or the combination of both⁴ have been associated with the accumulation of patient disability.^{5,6}

Deep learning-based models, especially convolutional neural networks (CNNs), have the ability to solve complex tasks by means of automatic feature extraction.⁷ In the last decade, deep learning-based models applied to medical imaging have been useful for the diagnosis or the prognostic classification of neurological conditions, as well as for brain lesion segmentation procedures.⁸ Deep learning based studies on MS have focused on the investigation of MRI biomarkers and their evolution over time, and have mostly resulted in new implementations for lesion segmentation and the detection of new lesions.⁹ However, deep learning-based methods have also been used in image pre-processing pipelines and dimensionality reduction strategies, which are mainly focused on diagnosis or prognosis predictions.¹⁰

Deep learning-based studies on MS classification or future prognosis of MS patients are limited and most have used a whole brain input strategy to build the models. Some approaches have used multiple MRI sequences,¹¹ the addition of pre-extracted biomarkers, such as brain lesion masks, or have used additional non-imaging data, such as clinical measures or patients' demographics.¹² Different sampling strategies, rather than using the whole brain, have been explored in other neurodegenerative diseases in which the use of brain MRI as input for classification tasks has been more widely studied (i.e., Alzheimer's Disease). Depending on the extent of the region-of-interest (ROI), studies can be classified in three categories: whole volume-level, regional-level, and patch-level. The whole volume strategy considers the whole volume of the analyzed structure (the whole brain in our case) as input. At the regional-level, the sampling is based on pre-segmented ROIs which correspond to structures that have been previously used as biomarkers in a given condition (eg, hippocampus, ventricles).^{13,14} This level would include masked regions of the whole-brain, such as segmented gray matter (GM) and white matter (WM), tissue probability maps^{15,16} or a variation of these, for example their modulation by the Jacobian of the deformation field.^{17,18} Finally, there would be the patch-level samplings, where the input would consist of several patches, whose size can be reduced as desired, without needing to contain a ROI in its entirety. Patch-level samplings, where patches are commonly randomly selected from abnormal tissue,¹⁹ are more widely used in segmentation tasks, where they were proposed in order to more effectively capture local structural changes.

The aim of this study was to investigate whether CNN approaches that focus on different brain regional structures perform better than a whole-brain CNN approach for the stratification of patients with MS based on their disability level, using a single brain MRI time-point.

Materials and Methods

The study was approved by the Vall d'Hebron Institute of Research – Research and Ethics Committee (PR(AG)389/2021 and PR(AG)99/2017) and informed consent was obtained from each patient, for all data used during this research.

Datasets

In this study, we used data from two different cohorts (in-house and external) of MS patients. The inclusion criteria were the same for both datasets: 1) MRI scans available for image analysis and 2) the corresponding clinical examination.

MRI sequence acquisitions included sagittal 3D T1-weighted magnetization prepared rapid gradient-echo (MPRAGE) and transverse T2-weighted fluid-attenuated inversion recovery (T2-FLAIR). A clinical examination was performed within 6 months after the scan acquisition. Clinical examination included assessment of either the Expanded Disability Status Scale (EDSS) score (in-house cohort)²⁰ or the Patient Determined Disease Steps (PDDS) score (external cohort),²¹ depending on availability. The EDSS score ranges from 0 (no disability) to 10 (severe disability).^{2,22} The PDDS score ranges from 0 to 8 and has been shown to have a strong correlation with the EDSS score.²³ Figure 1 shows examples of MRI scans of MS patients with low and mild disability from the two cohorts.

IN-HOUSE DATASET. The in-house subjects were part of a larger cohort of patients, the Barcelona CIS cohort,² composed of consecutive patients from the Multiple Sclerosis center of Catalonia (Cemcat), Vall d'Hebron University Hospital (VHUh), prospectively followed over time after their first demyelinating attack. The selection of patients was based on the availability of MRI data that could be closely associated with the clinical evaluation, on patients that had their first demyelinating attack before the age of 50.

We included 319 unique patients, 215 with an EDSS score <3.0 and 104 with an EDSS score ≥3.0 (non-confirmed, i.e., EDSS ≥ 3.0 reached at least once, which might not be maintained over a minimum period of 6 months). Each MRI scan was associated with the closest available EDSS score (mean time difference = 37 days, range = [1–161]). Demographics are shown in Table 1. A total of 382 scans, acquired from 2010 to 2020, that included multiple scans at different time-points for 33 subjects.

The in-house dataset was acquired in a single center with five different Siemens (Heidelberg, Germany) scanner models at two different magnetic fields (1.5 T and 3.0 T), with standardized acquisition protocols for each scanner (see Table S1 in the Supplemental Material for specific acquisition parameters).

EXTERNAL VALIDATION (MS PATHS) DATASET. Multiple Sclerosis Partners Advancing Technology and Health Solutions (MS PATHS)²⁴ is a learning health system in MS, started in 2016, comprising a collaborative network of 10 healthcare centers, providing standardized routinely-acquired clinical and MRI data. From this

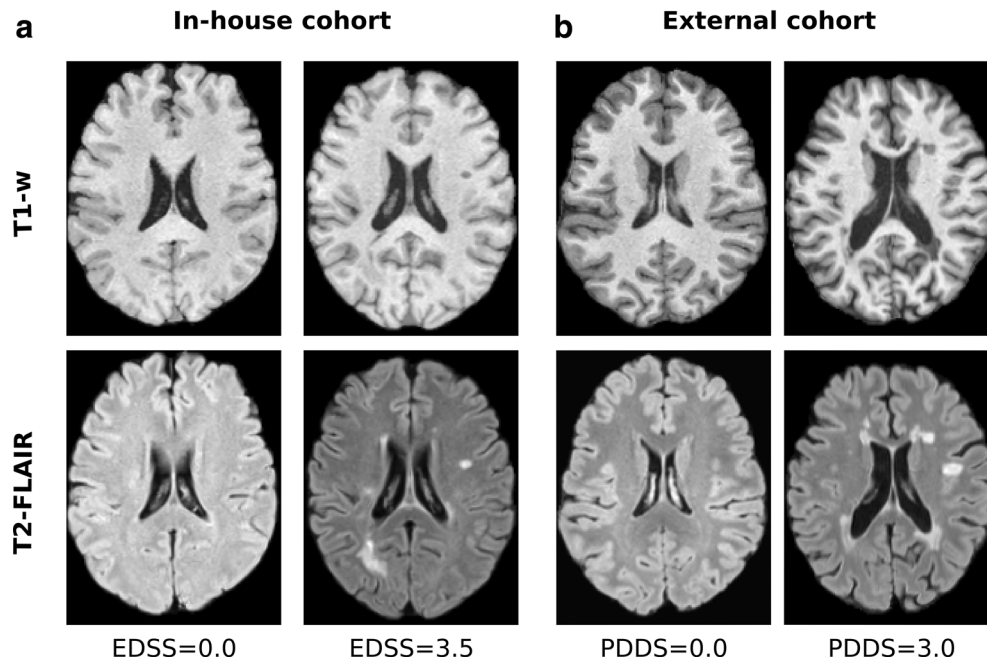


FIGURE 1: Example cases of low (EDSS/PDDS = 0.0) and mild (EDSS = 3.5, PDDS = 3.0) disability in (a) the in-house and (b) the external cohorts.

TABLE 1. Demographic and Clinical Data of Patients Included in the In-House Analysis

	Full Cohort N (Pts/Scans) = 319/382	EDSS < 3.0 N = 215/215	EDSS ≥ 3.0 N = 104/167
Female, N (%) ^a	207 (65)	147 (68)	60 (58)
Confirmed diagnosis, N (%) ^a	260 (82)	160 (74)	100 (96)
Age at diagnosis, years, mean [range]	33.2 [14–59]	33.5 [16–59]	32.7 [14–55]
DD, years, mean (SD)	10.4 (7.0)	7.6 (6.6)	14.0 (5.6)
EDSS, median [range]	2.0 [0.0–9.0]	1.5 [0.0–2.5]	5.0 [3.0–9.0]
Phenotype, N (%) ^b			
CIS	61 (16)	57 (26.5)	4 (2)
RRMS	232 (61)	157 (73)	75 (45)
SPMS	89 (23)	1 (0.5)	88 (53)
Scanner model, N (%) ^b			
Avanto	64 (17)	19 (9)	45 (27)
Avanto Fit	64 (17)	43 (20)	21 (13)
Symphony	10 (3)	7 (3)	3 (2)
Symphony Tim	51 (13)	13 (6)	38 (23)
Tim Trio	193 (50)	133 (62)	60 (36)

Pts = patients; EDSS = Expanded Disability Status Scale; DD = disease duration; GM = gray matter; WM = white matter; CIS = clinically isolated syndrome; RRMS = relapsing–remitting multiple sclerosis; SPMS = secondary progressive multiple sclerosis.

^a% calculated with the number of patients.

^b% calculated with the number of scans.

TABLE 2. Demographic and Clinical Data of Patients From MS PATHS Included in the Analysis

	Full Cohort N = 440	PDDS < 3.0 N = 220	PDDS ≥ 3.0 N = 220
Female, N (%)	310 (70)	170 (77)	140 (64)
Age at diagnosis, years, mean [range]	36.8 [19–69]	36.1 [19–62]	37.6 [19–69]
DD, years, mean (SD)	11.5 (9.1)	8.5 (7.6)	14.9 (9.5)
PDDS, median [range]	2.5 [0.0–7.0]	0.5 [0.0–2.0]	5.0 [3.0–7.0]

PDDS = Patient Determined Disease Steps; DD = disease duration.

large database, following the established criteria, we randomly selected a subset, with representation of all grades of disability and following the same distribution, along grades, than for the in-house subset. This independent set was used for external validation.

The resultant set was composed of 440 patients (and scans) imaged on four different Siemens 3 T scanner models (Biograph_mMR, Skyra, Trio Tim and Verio) from six different sites (excluding the provider of our in-house dataset). All scans were acquired using standardized image acquisition protocols, as in the Tim Trio scanner in Table S1 in the Supplemental Material. Each MRI scan was associated with the closest PDDS score (mean difference time 59 days, range = [1–185]). Demographic and clinical data are summarized in Table 2.

Proposed Method

We used a deep learning model to stratify patients based on their clinical evaluation score: moderate (EDSS ≥ 3.0) vs. mild (EDSS < 3.0) disability.² Different input strategies were studied on the same network architecture to investigate not only a global image-based approach but also different regional approaches as summarized in Fig. 2.

Our deep learning classifier was trained with brain T1-weighted (T1-w) and T2-FLAIR sequences from the in-house dataset, using a 7-fold cross validation strategy. The total available dataset was divided into 7 folds: four with 55 scans and three with 54 scans. In each iteration, the training set was composed of five folds (~70% of the dataset), the validation set of 1-fold (~15%), while the last one was used for inference (~15%). This process was repeated seven times until all folds were used for inference. We assured that patients with multiple scans were always in the same set (training, validation or inference) so as not to bias the model.

We considered the global approach, which was trained with the whole brain T1-w and T2-FLAIR sequences as input, as the reference model. In addition, we analyzed different regional approaches based on ROIs that have previously been shown to be associated with the prognosis of the disease, and which may serve as biomarkers (eg, localized atrophy measures) or reflect typical locations of white matter lesions^{3,5,6} in MS.

The selected regions were distinguished by their volume size, i.e., 1) small regions (the lateral ventricles, the subcortical GM and the brain stem and cerebellum [BSC]) and 2) large regions (WM and GM). Depending on the input region that was used for the model, different processing steps were applied to the images to highlight the input information.

IMAGE PRE-PROCESSING. The same fully automatic image pre-processing pipeline was applied to both VHUH (in-house) and MS PATHS (external validation) datasets. All T1-w and T2-FLAIR sequences were pre-processed with 1) bias correction,²⁵ 2) skull-stripping,²⁶ 3) registration²⁷ to MNI152 space, as well as co-registration of T2-FLAIR sequences to T1-w space, and 4) min–max voxel intensity normalization.

For the different input regional strategies, further processing was performed to calculate the tissue masks and ROI sizes. All ROIs and masks were obtained automatically and based on the average population of the study. First, we performed automatic lesion segmentation²⁸ to lesion fill the T1-w scans.²⁹ Afterwards, using the T1-w lesion filled scan, we extracted the whole brain parcellation with FastSurfer³⁰ to obtain the desired regions: subcortical GM structures (thalamus, putamen, caudate, and pallidum), lateral ventricles, brainstem, and cerebellum areas, as well as, WM and GM tissues. A graphical representation of the different samplings applied is shown in Fig. 2b.

For the GM regional input strategy, in addition to the T1-w and T2-FLAIR images, a third channel was incorporated as an input, the GM modulation. The steps required to extract the GM modulation are illustrated in Fig. 3. The GM modulation was used to preserve the GM volume of the native space, through the resulting Jacobian determinant from the nonlinear registration.³¹ This Jacobian determinant (Fig. 3d) contained the local volume change in each voxel when referred to the common MNI space.

NETWORK ARCHITECTURE. The proposed network was based on a modified ResNet CNN architecture,³² built with three-dimensional (3D) layers. The ResNet architecture has the capacity to achieve state-of-the-art results comparable to their deeper and more complex counterparts, with shorter training times, even when using limited hardware.³² To build on these advantages, modifications were included in order to reduce the number of parameters and the complexity of the final model. Each residual block was based on 3D convolutional layers that produce $3 \times 3 \times 3$ and $1 \times 1 \times 1$ kernel convolution layers, normalized with batch normalization and activated with leaky rectified linear unit (LeakyReLU). As shown in Fig. 4, the architecture was composed of four residual blocks with increasing numbers of kernels k (16, 32, 64, and 128), followed by a $2 \times 2 \times 2$ downscale max pooling operation. Afterwards, the feature map extracted was projected in a global adaptive max pooling layer to reduce feature dimensionality and allow independence of the

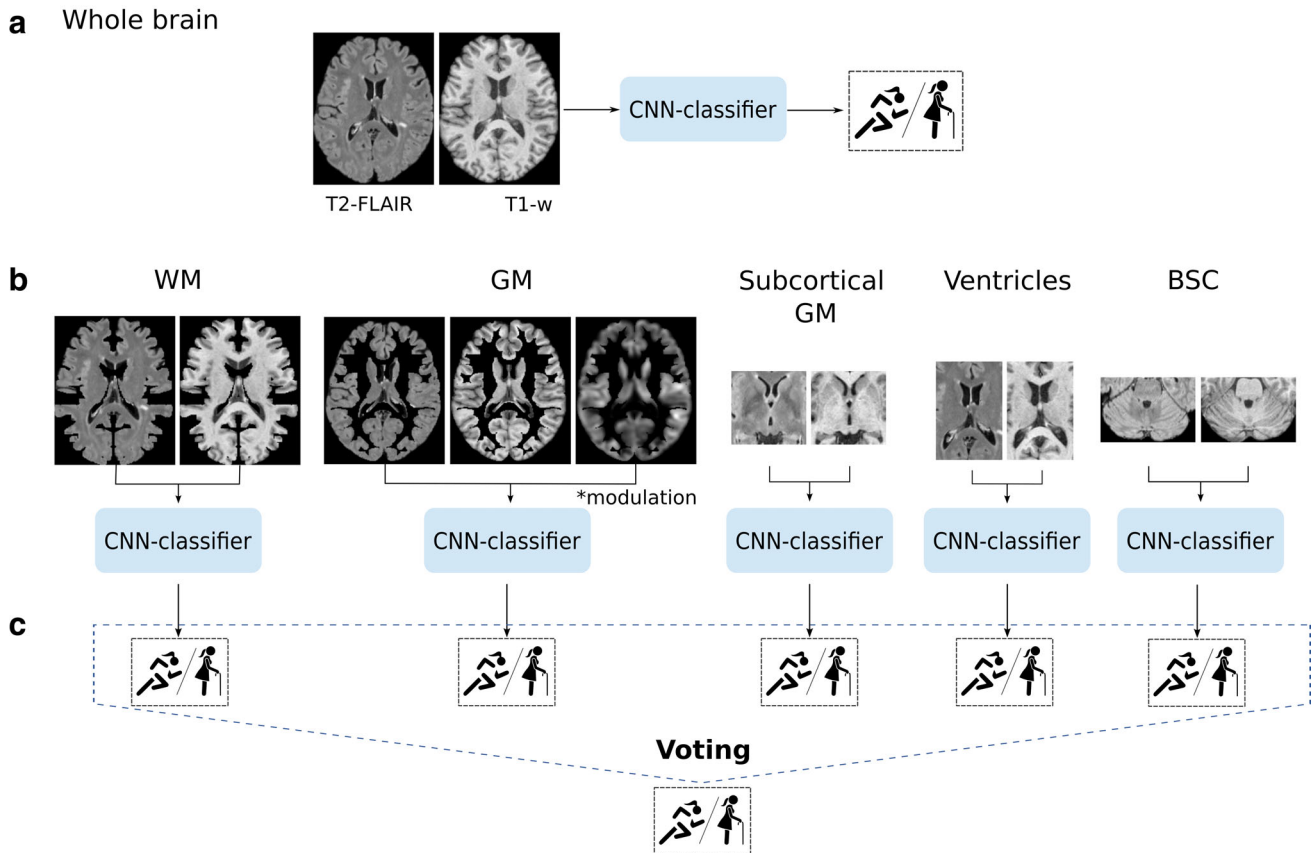


FIGURE 2: Classification strategies. (a) Whole brain and (b) individual regions were evaluated by the classifier model to predict the probability of belonging to $EDSS \geq 3.0$ or $EDSS < 3.0$. Large regions (WM and GM) were masked from the whole brain and, small regions (subcortical-GM, lateral ventricles, and BSC) are cropped to the ROI comprising the specific structure. (c) The single regional model predictions were combined in two voting fusion ensembles to predict the disability status of the patient: 1) based on the maximum probability of the five regional models (max) and 2) based on the mode of the final regional predictions (majority). WM = white matter; GM = gray matter; BSC = brain stem and cerebellum.

input patch size. The final classification layer (fully connected layer) was replaced by three successive $1 \times 1 \times 1$ 3D convolutional layers, with $k = 128, 64, 2$, where the first two were activated with ReLU and the last one with a Softmax, that outputs the probability to belonging to one or the other class.

TRAINING PROCEDURE. For training the different models, only the in-house dataset was used. A 7-fold patient cross validation strategy was used to train and test each model. For each sampling approach, we obtained seven different sets of parameters, i.e., seven different models, that were evaluated on the corresponding inference fold (~15% of patients). The folds were sampled to keep the same class distribution in each one, while also following the distribution present in the total dataset.

The sampling for each model was decided depending on the type of input region analyzed. For small input regions, i.e., lateral ventricles, subcortical GM and BSC, we used a square ROI-based patch,^{13,14} automatically delineated from the maximum average map of the region in question from the individual parcellations. The resultant intensity patches had a size of $75 \times 113 \times 41 \text{ mm}^3$ for the lateral ventricles, $80 \times 72 \times 55 \text{ mm}^3$ for subcortical GM and $123 \times 85 \times 61 \text{ mm}^3$ for the BSC, and were centered on each structure. For the large regions, WM and GM tissues, we took the whole

brain patch size ($144 \times 184 \times 152 \text{ mm}^3$), but only kept the intensities of the tissue we intended to use as input. This was done by masking the intensity patch with a dilated average mask of the studied tissue. Therefore, only the intensities inside the analyzed mask were considered¹⁸ (see Fig. 2b).

To mitigate the class imbalance, data augmentation was used. Depending on the input region size, we applied different strategies. For whole brain patches, the global approach and regional WM and GM approaches, an axial flip was applied to all subjects with $EDSS \geq 3.0$ and to a random 75% of the patients with $EDSS < 3.0$, trying to find an equilibrium between balancing the data and not letting the model to learn a non-characteristic feature as the axial flip. For ROI-based models, where the patches were smaller than the whole brain patch, a random voxel displacement in the three dimensions was used to generate additional patches of all subjects, considering the 1:3 proportion of patients with $EDSS \geq 3.0$ in the dataset.

Each model was trained using T1-w and T2-FLAIR scans (from the in-house dataset) and the corresponding EDSS-based class ($EDSS \geq$ or < 3.0). We trained the model for a maximum of 200 epochs, with an early stopping strategy to prevent overfitting. The model was optimized with a learning decay strategy depending also on the validation performance, and trained by minimizing a weighted cross entropy loss as cost function.

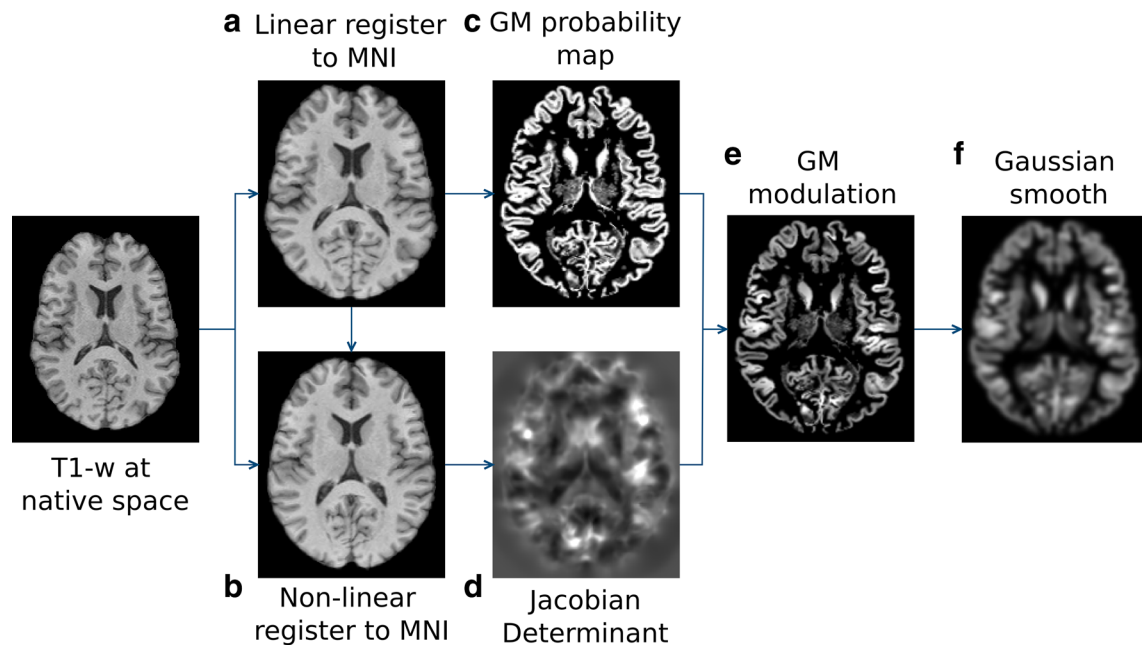


FIGURE 3: GM modulation. As part of the pre-processing, the T1-weighted scan in native space was registered to the MNI space (a) linearly and (b) non-linearly. The linearly registered scan was used to obtain the GM probability map (c) and the non-linearly registered scan was used to compute the Jacobian determinant (d). The GM modulation was obtained as the product of the GM probability map and the deformation (the Jacobian determinant) (e). A Gaussian kernel was then used to smooth the product at a FWHM of 4.7 mm (f). GM = gray matter; FWHM = full width at half maximum.

INFERENCE. For each individual CNN model, and following the same sampling procedure as described for model training, each model-specific patch, as described during training, was used as input through the trained model providing the output probabilities of belonging to one class or the other. The final classification was determined by the maximum of both probabilities, with a threshold set at 0.5.

In addition to the results of each individual model, we computed an ensemble of the regional models with a late fusion strategy.³³ As represented in Fig. 2c, the predictions obtained with each trained regional model were aggregated to make a final prediction based on two different voting strategies: 1) maximum and 2) majority voting. The maximum voting strategy was calculated as the prediction of the model with the highest probability, while the majority voting approach was calculated as the mode of the different predictions obtained after thresholding each model's probabilities. The calculation of the ensemble models provided individual

information of how subjects performed within the different models, presenting either a higher prediction probability (maximum voting) or full agreement across all different models (majority voting).

Without any retraining or fine-tuning of the different models trained with the in-house dataset, inference on the external validation set was also computed as described above. The final prediction per subject was obtained from the majority voting across the seven different cross-validation models, for each one of the sampling strategies. To evaluate the ensemble of regional models, a maximum voting was computed across the 7-folds of each regional model. Following this, using the winning fold, the specific voting strategy was computed, i.e., majority or maximum voting across regional models.

Evaluation Measures and Statistical Analysis

To evaluate and compare the performance of the models presented we used the following metrics:

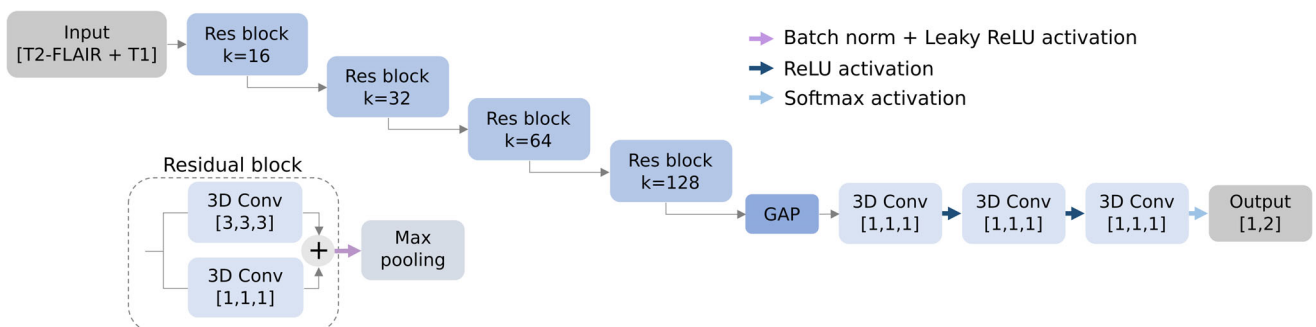


FIGURE 4: Residual convolutional neural network architecture. k = kernels; GAP = global adaptive max pooling; ReLU = rectified linear unit.

1. Sensitivity (SENS) of correctly classified subjects with EDSS ≥ 3.0 thresholded at 0.5, $\text{SENS} = \text{TP}/(\text{TP} + \text{FN})$
2. Specificity (SPEC) of correctly classified subjects with EDSS < 3.0 thresholded at 0.5, $\text{SPEC} = \text{TN}/(\text{TN} + \text{FP})$
3. Balanced accuracy (ACC) in correctly classifying each individual by means of their EDSS, $\text{ACC} = (\text{SENS} + \text{SPEC})/2$
4. ROC (receiver operating characteristic) curve showing the performance at all classification thresholds.
5. AUC, the area under the ROC curve, calculated based on all possible pairs of SENS and 1-SPEC obtained by changing the thresholds performed on the classification scores.

where TP (true positives) were the number of correctly classified patients with EDSS ≥ 3.0 , TN (true negatives) were the number of correctly classified patients with EDSS < 3.0 , FP (false positives) were the number of patients with EDSS < 3.0 classified as EDSS ≥ 3.0 , and FN (false negatives) were the number of patients with EDSS ≥ 3.0 classified as EDSS < 3.0 . The results were reported in terms of mean and standard deviation for the 7-fold cross validation computed on the in-house dataset and as majority voting of the seven models for the independent dataset.

Paired t -tests were used to compare the performance of each sampling strategy model against the others. T -tests were obtained using the output probabilities from each model to compare the performance on each individual patient in which they were evaluated. To compare the AUCs between models we used DeLong's test.³⁴ A P -value < 0.05 was considered statistically significant. The proposed method and analysis was entirely implemented in Python (<https://www.python.org/>), using the Pytorch library.³⁵ The model implementation and source code are publicly available at <https://github.com/suliciac/MStratification>. All the experiments were run on a GNU/Linux machine box running Ubuntu 20.04, with 125 GB RAM. For training the model, we used a single Quadro RTX 5000 GPU (NVIDIA Corp, USA) with 16 GB VRAM memory.

Results

Evaluation of Deep Learning Models: In-House Dataset

The global approach, using the whole brain patch as input, achieved a mean balanced accuracy of 79% (range across folds: [70–83]%), 77% sensitivity and 81% specificity, for classifying patients with an EDSS $<$ or ≥ 3.0 . The best performing individual regional model was the GM model, with a mean accuracy of 81% [74–87]% and the highest sensitivity of 79% (specificity 83%). The subcortical-GM (a subregion of the GM model) achieved a 78% [72–91]% accuracy, in line with the WM model which had the same accuracy and less variability [72–88]%. Both, subcortical-GM and WM models, also achieved similar sensitivity (77% and 75%, respectively) and specificity (79% and 81%). The other two regional models, ventricles and BSC, showed a similar but lower accuracy (76% [66–86]% and 76% [65–83]%, respectively), but with differentiated sensitivity (76% and 68%) and specificity (76% and 84%) at 0.5 operation point.

Figure 5a shows the ROC curves and AUC values for all the approaches. As observed with the thresholded values at 0.5 (sensitivity and specificity) the GM regional model had the highest AUC (0.87) and the BSC had the lowest (0.82).

When performing a t -test analysis between pairs of models, the BSC and the ventricles regional models had significantly poorer results compared to the other models, which were not significantly different from each other. All the combinations of t -tests between models are shown in Table S2 in the Supplemental Material.

When combining the final performance of the regional models in a voting ensemble approach (see Fig. 2c), we obtained accuracies of 81% [79–84]% and 80% [77–85]% using maximum and majority fusion strategies, respectively, with a lower variability across folds and a higher specificity (88% and 83%, respectively) than the individual models and a similar sensitivity (73% and 77%). In Fig. 5a, note that the highest AUC value is obtained with the ensemble model using majority voting (AUC 0.88). In general, the AUC results were slightly better in the voting ensemble approaches than in the best individual regional model (GM) in terms of accuracy and specificity. DeLong's test showed that most of the paired comparisons with the majority voting ensemble were statistically significant (see Table S3 in the Supplemental Material). In the majority voting ensemble, for most of patients, all regional models agreed on the same class attribution, being the WM the model most frequently contributing to the vote. In the maximum voting strategy, the models that contributed the most were the GM and BSC, reflected in the highest specificities of these models. The relation of all regional models which were winning in the maximum voting approach, i.e., having the highest output probabilities, can be found in Table S4 in the Supplemental Material.

Validation With the Independent MS PATHS Dataset

The overall performance was lower than with the in-house dataset, obtaining a balanced accuracy of 71% [68–72]%, 68% sensitivity and 75% specificity when using the whole brain approach. The WM regional model obtained the highest individual regional performance with 72% [69–72]% accuracy and the same sensitivity and specificity as the whole brain approach. GM and ventricles regional models presented a similar accuracy (70% [68–71]%). The subcortical-GM model achieved a similar accuracy (69%) with a much lower sensitivity (59%), and a higher specificity (80%).

When performing the statistical analysis on the so far presented models, there was no evidence of statistically significant differences between most of them (see Table S5 in the Supplemental Material for all paired t -test combinations). However, as in the in-house cohort, the BSC model performed significantly worse than the other models, with 67%

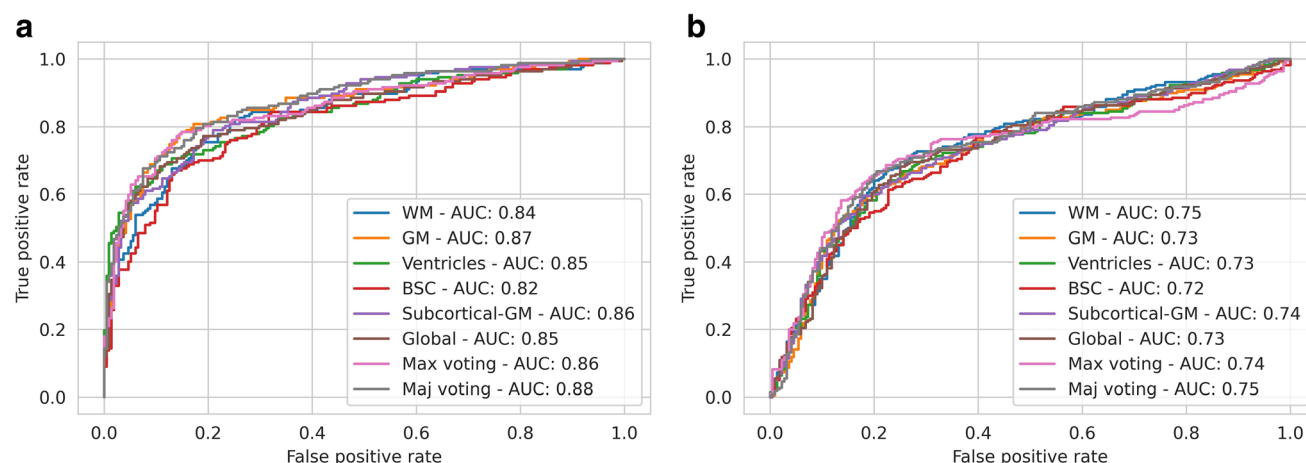


FIGURE 5: ROC curves and AUC values for each regional model, the global approach and the two ensemble strategies (maximum vote and majority vote) of the regional models for (a) the in-house dataset and (b) the external dataset. In (a), the mean ROC of all the in-house cases is represented, collected from the different inferred folds. In (b), the mean ROC is computed with each external cohort case majority voting result across the 7-folds evaluated.

accuracy, 48% sensitivity and 85% specificity at 0.5 operating point.

Figure 5b shows the mean across-folds ROC curves for all the models. As observed with the values obtained at 0.5 operating point, the WM model achieved the highest AUC (0.75) and the BSC model had the lowest (0.72). Delong's test showed that none of the AUCs results were statistically significant (see Table S6 in the Supplemental Material for all the models' combinations *P*-values).

Evaluating the results obtained with the majority voting ensemble, in most cases there was an agreement between all regional models, with the WM model being the most frequent contributor, i.e., the model that correctly stratified more patients in agreement with the others, when calculating the majority voting. When analyzing the maximum voting strategy, we observed that the GM model provided the highest probabilities followed by the BSC model (see Table S7 in the Supplemental Material for the complete contributions in the maximum voting ensemble). As seen in the in-house dataset performance, the majority and maximum voting ensembles showed similar metrics to those of the individual models that best contributed to them (mean balanced accuracy of 73% and 72%, respectively).

Discussion

The current study investigated the ability of different input strategies (global, regional and the combination of these (ensemble)) to accurately classify MS patients based on their disability level through deep learning-based CNN models, using two sequences (T1-w and T2-FLAIR) of a single MRI time-point. The study was performed in a large cohort of patients with MS and validated in an external MS cohort.

Our findings showed that, in the in-house cohort (VHUH), the best accuracies were achieved with the regional

GM approach followed by the whole-brain approach, whereas the best performing models in the external dataset (MS PATHS) were from the regional WM approach, followed by the whole-brain approach. Thus, the global whole-brain approach offered the best trade-off between internal performance and external validation, although some regional models such as GM and WM models showed similar overall performances.

Among the different individual strategies presented, when evaluating the in-house dataset, the regional GM model achieved the best overall performance results. This may be explained by the fact that in this approach, a third input channel, the GM modulation, was incorporated. The GM modulation represents the deformation suffered by the image when registering the scans to a common space. Thus, it provides information about the native space, accounting for a possible effect of GM atrophy, which is known to be important for development of future disability in MS.^{36,37}

With the in-house dataset, the next best ranking models were the whole-brain approach and the regional WM and subcortical-GM models. The whole-brain and the WM models are those that may have a direct relationship with the WM lesion load. Indeed, a post-hoc analysis (see Supplemental Material for detailed information) showed that there was an association between WM lesion load and model output for the global and WM-regional models.

On the other hand, the subcortical-GM model performance was similar to that obtained with the GM tissue model. These results are in line with the strong correlation shown between GM subcortical volumes and disease severity.³⁸

In this study, the ventricles model performed acceptably well in both datasets. The presence (total or partial) of WM lesions together with the presence or absence of atrophy that can be measured by the ventricles size may have helped with the model accuracy.^{1,39} However, as for the BSC regional model, not having been given the whole brain image as input

may have constrained the model, leading to statistically significant poorer performance than the models with a larger regional input (see Table S2 in the Supplemental Material for absolute *P*-values). The BSC model had the poorest accuracy in both datasets. However, the BSC model was a relevant addition to the maximum voting ensemble model, due to it having the highest specificity that resulted in it being the model that contributed most when identifying patients with EDSS < 3.0. On the other hand, the GM model was the model that contributed most for the correct classification of patients with EDSS ≥ 3.0, due to it having the highest sensitivity when evaluated individually.

In general, the performance of the different models on the MS PATHS subset resulted in slightly lower performance than on the in-house cohort. However, considering that there was no additional training or fine-tuning to evaluate the unseen data, the accuracy of all models was satisfactory, suggesting the generalization of our models. However, using a different score (PDDS instead of the EDSS used for training) to classify this external validation dataset may be seen as a limitation of this evaluation, despite the strong correlation between both metrics.²³

Our results highlight the importance of considering whole-brain input sampling strategies to promote generalizability of CNN-based stratification models. Although this might be intuitive, this study quantitatively assessed the effect of the type of input that a CNN-based model must have in this MS stratification problem. Building accurate CNN-based models is key to predicting individual patients' disease course in order to achieve a personalized approach.¹ The methodology presented in this study, along with retraining or fine tuning, may have potential for diagnostic or progression prediction tasks.

Limitations

Apart from the relatively small sample size used for training the models from the in-house dataset, all five of the MRI scanners present in the study were from the same vendor. This can be seen as a limitation in terms of model generalization. However, the training cohort did include scans acquired at different strength fields (1.5 T and 3 T), with some variation in protocols between scanners. Approximately half of the patient data was acquired with the same acquisition protocol that was used in the external validation cohort (MS PATHS), where images were also acquired with different MRI scanners from the same vendor, three of which were not included in the in-house dataset. The external validation was also restricted by not using the same clinical score as used in the training set (PDDS instead of EDSS). Of note, although EDSS and PDDS are both nonlinear scales, the EDSS is obtained by a neurologist, after performing an anamnesis and a neurological examination, and the PDDS is instead reported by the patient, implying a strong subjective nature. Therefore, they reflect essentially different points of view of the disease.

However, they are highly correlated,²³ which is reassuring and suggests that they may be used for similar predictive and monitoring purposes. Indeed, the PDDS is frequently used in those clinical settings where the EDSS is not available. In our study, we considered an EDSS of 3.0 to be equivalent to a PDDS of 3.0. However, other equivalences were indeed possible and should be explored in further studies.

From the clinical point of view, setting a threshold at a certain EDSS may not involve all the factors that determine disability in MS patients at a cross-sectional point. Despite this, EDSS is the most used clinical score to quantify disability in MS clinical practice and clinical trials and reaching an EDSS ≥ 3.0 has been shown to be a relevant outcome when studying the disease progression course.² However, in this study, the EDSS ≥ 3.0 was not always confirmed in a follow-up visit, which means we may not have analyzed a clinically stable population. Also, we did not account for any disease-modifying treatment. Further studies taking these aspects into account are therefore necessary.

Conclusion

This study showed that CNN-based models were able to extract features from different input strategies and lead to a correct classification of MS patients based on their disability score. The global (whole-brain) and large ROI-input models (WM and GM) resulted in the highest classification accuracies. While their similar behavior suggests that the CNN is able to adapt to its inputs, this also indicates that focusing on specific regions, even if *a priori* important for MS, does not necessarily translate into better performance. Indeed, using global input approaches may result in a better generalization of such CNN models as it offered the best trade-off between internal performance and external validation.

Acknowledgments

This study has been possible thanks to a Junior Leader La Caixa Fellowship awarded to C. Tur (fellowship code is LCF/BQ/PI20/11760008) by “la Caixa” Foundation (ID 100010434). The salaries of C. Tur and Ll. Coll are covered by this award.

References

1. Thompson AJ, Baranzini SE, Geurts J, Hemmer B, Ciccarelli O. Multiple sclerosis. *Lancet* 2018;391(10130):1622-1636.
2. Tintore M, Rovira L, Rio J, et al. Defining high, medium and low impact prognostic factors for developing multiple sclerosis. *Brain* 2015;138(Pt 7):1863-1874.
3. Haider L, Chung K, Birch G, et al. Linear brain atrophy measures in multiple sclerosis and clinically isolated syndromes: A 30-year follow-up. *J Neurol Neurosurg Psychiatry* 2021;92(8):839-846.
4. Popescu V, Agosta F, Hulst HE, et al. Brain atrophy and lesion load predict long term disability in multiple sclerosis. *J Neurol Neurosurg Psychiatry* 2013;84(10):1082-1091.

5. Cappelle S, Pareto D, Vidal-Jordana A, et al. A validation study of manual atrophy measures in patients with Multiple Sclerosis. *Neuroradiology* 2020;62(8):955-964.
6. Bonacchi R, Meani A, Pagani E, Marchesi O, Filippi M, Rocca MA. The role of cerebellar damage in explaining disability and cognition in multiple sclerosis phenotypes: A multiparametric MRI study. *J Neurol* 2022;269(7):3841-3857.
7. Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521(7553):436-444.
8. Bernal J, Kushibar K, Asfaw DS, et al. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: A review. *Artif Intell Med* 2019;95:64-81.
9. Commowick O, Cervenansky F, Cotton F, Dojat M. MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure. MICCAI 2021 – 24th Int Conf Med Image Comput Comput Assist Interv 2021;126-2021.
10. Shoeibi A, Khodatars M, Jafari M, et al. Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: A review. *Comput Biol Med* 2021;136:104697.
11. Tousignant A, Lemaître P, Doina P, Arnold DL, Arbel T. Prediction of disease progression in multiple sclerosis patients using deep learning analysis of MRI data. *Proc Mach Learn Res* 2019;102:483-492.
12. Roca P, Attye A, Colas L, et al. Artificial intelligence to predict clinical disability in patients with multiple sclerosis using FLAIR MRI. *Diagn Interv Imaging* 2020;101(12):795-802.
13. Ahmed S, Kim BC, Lee KH, Yub H, Jung HY. Ensemble of ROI-based convolutional neural network classifiers for staging the Alzheimer disease spectrum from magnetic resonance imaging. *PLoS One* 2020;15(12):e0242712.
14. Kwak K, Niethammer M, Giovanello KS, Styner M, Dayan E. Differential role for hippocampal subfields in Alzheimer's disease progression revealed with deep learning. *Cereb Cortex* 2021;32(3):467-478.
15. Basheera S, Ram MSS. Deep learning based Alzheimer's disease early diagnosis using T2w segmented gray matter MRI. *Int J Imaging Syst Technol* 2021;31(3):1692-1710.
16. Mehmood A, Yang S, Feng Z, et al. A transfer learning approach for early diagnosis of Alzheimer's disease on MRI images. *Neuroscience* 2021;460:43-52.
17. Cao P, Gao J, Zhang Z. Multi-view based multi-model learning for MCI diagnosis. *Brain Sci* 2020;10(3):181.
18. Zhou P, Jiang S, Yu L, et al. Use of a sparse-response deep belief network and extreme learning machine to discriminate Alzheimer's disease, mild cognitive impairment, and normal controls based on amyloid PET/MRI images. *Front Med* 2021;7:621204.
19. Zhu W, Sun L, Huang J, Han L, Zhang D. Dual attention multi-instance deep learning for Alzheimer's disease diagnosis with structural MRI. *IEEE Trans Med Imaging* 2021;40(9):2354-2366.
20. Kurtzke JF. Rating neurologic impairment in multiple sclerosis. *Neurology* 1983;33(11):1444-1452.
21. Rizzo MA, Hadjimichael OC, Preiningerova J, Vollmer TL. Prevalence and treatment of spasticity reported by multiple sclerosis patients. *Mult Scler J* 2004;10(5):589-595.
22. Tintoré M, Rovira A, Río J, et al. Baseline MRI predicts future attacks and disability in clinically isolated syndromes. *Neurology* 2006;67(6):968-972.
23. Learmonth YC, Motl RW, Sandroff BM, Pula JH, Cadavid D. Validation of patient determined disease steps (PDDS) scale scores in persons with multiple sclerosis. *BMC Neurol* 2013;13:37.
24. Mowry EM, Bermel RA, Williams JR, et al. Harnessing real-world data to inform decision-making: Multiple sclerosis partners advancing technology and health solutions (MS PATHS). *Front Neurol* 2020;11:632.
25. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: Improved N3 bias correction. *IEEE Trans Med Imaging* 2010;29(6):1310-1320.
26. Isensee F, Schell M, Pflueger I, et al. Automated brain extraction of multisequence MRI using artificial neural networks. *Hum Brain Mapp* 2019;40(17):4952-4964.
27. Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. FSL. *Neuroimage* 2012;62(2):782-790.
28. Schmidt P, Gaser C, Arsic M, et al. An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. *Neuroimage* 2012;59(4):3774-3783.
29. Prados F, Cardoso MJ, Kanber B, et al. A multi-time-point modality-agnostic patch-based method for lesion filling in multiple sclerosis. *Neuroimage* 2016;139:376-384.
30. Henschel L, Conjeti S, Estrada S, Diers K, Fischl B, Reuter M. FastSurfer – a fast and accurate deep learning based neuroimaging pipeline. *Neuroimage* 2020;219:117012.
31. Lungu O, Pantano P, Kumfor F, et al. Impaired self-other distinction and subcortical gray-matter alterations characterize socio-cognitive disturbances in multiple sclerosis. *Front Neurol* 2019;10:525.
32. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 2016;770-778.
33. Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *NPJ Digit Med* 2020;3:136.
34. Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process Lett* 2014;21(11):1389-1393.
35. Paszke A, Gross S, Massa F, et al. PyTorch: An imperative style, high-performance deep learning library. *Proceedings of the 33rd international conference on neural information processing systems*. Red Hook, NY, USA: Curran Associates Inc.; 2019. p 8026-8037.
36. Amiri H, de Sitter A, Bendfeldt K, et al. Urgent challenges in quantification and interpretation of brain grey matter atrophy in individual MS patients using MRI. *NeuroImage Clin* 2018;19:466-475.
37. Eshaghi A, Marinescu RV, Young AL, et al. Progression of regional grey matter atrophy in multiple sclerosis on behalf of the MAGNIMS study group*. *Brain* 2018;141:1665-1677.
38. Eshaghi A, Prados F, Brownlee WJ, et al. Deep gray matter volume loss drives disability worsening in multiple sclerosis. *Ann Neurol* 2018;83(2):210-222.
39. Brown JWL, Pardini M, Brownlee WJ, et al. An abnormal periventricular magnetization transfer ratio gradient occurs early in multiple sclerosis. *Brain* 2017;140(2):387-398.