Received: 26 January 2022 | Revised: 2 June 2022 | Accepted: 24 September 2022

DOI: 10.1002/hbm.26106

RESEARCH ARTICLE

WILEY

Prediction of the information processing speed performance in multiple sclerosis using a machine learning approach in a large multicenter magnetic resonance imaging data set

Chiara Marzi[1,2] | Alessandro d'Ambrosio[1] | Stefano Diciotti[2,3] | Alvino Bisecco[1] | Manuela Altieri[1,4] | Massimo Filippi[5,6] | Maria Assunta Rocca[5,6] | Loredana Storelli[5] | Patrizia Pantano[7,8] | Silvia Tommasin[7] | Rosa Cortese[9] | Nicola De Stefano[9] | Gioacchino Tedeschi[1] | Antonio Gallo[1] | the INNI Network

[1]MS Center and 3T-MRI Research Unit, Department of Advanced Medical and Surgical Sciences (DAMSS), University of Campania "Luigi Vanvitelli", Napoli, Italy

[2]Department of Electrical, Electronic, and Information Engineering "Guglielmo Marconi" – DEI, Alma Mater Studiorum – University of Bologna, Bologna, Italy

[3]Alma Mater Research Institute for Human-Centered Artificial Intelligence, University of Bologna, Bologna, Italy

[4]Department of Psychology, University of Campania "Luigi Vanvitelli", Napoli, Italy

[5]Neuroimaging Research Unit, Division of Neuroscience, Vita-Salute San Raffaele University, IRCCS San Raffaele Scientific Institute, Milan, Italy

[6]Neurology and Neurophysiology Unit, Vita-Salute San Raffaele University, IRCCS San Raffaele Scientific Institute, Milan, Italy

[7]Department of Human Neurosciences, Sapienza University of Rome, Rome, Italy

[8]IRCCS Neuromed, Pozzilli, Italy

[9]Department of Medicine, Surgery and Neuroscience, University of Siena, Siena, Italy

Correspondence
Chiara Marzi, Nello CMS Center and 3T-MRI Research Unit, Department of Advanced Medical and Surgical Sciences (DAMSS), University of Campania "Luigi Vanvitelli,"

Abstract

Many patients with multiple sclerosis (MS) experience information processing speed (IPS) deficits, and the Symbol Digit Modalities Test (SDMT) has been recommended as a valid screening test. Magnetic resonance imaging (MRI) has markedly improved the understanding of the mechanisms associated with cognitive deficits in MS. However, which structural MRI markers are the most closely related to cognitive performance is still unclear. We used the multicenter 3T-MRI data set of the Italian Neuroimaging Network Initiative to extract multimodal data (i.e., demographic, clinical, neuropsychological, and structural MRIs) of 540 MS patients. We aimed to assess, through machine learning techniques, the contribution of brain MRI structural volumes in the prediction of IPS deficits when combined with demographic and clinical features. We trained and tested the eXtreme Gradient Boosting (XGBoost) model following a rigorous validation scheme to obtain reliable generalization performance. We carried out a classification and a regression task based on SDMT scores feeding each model with different combinations of features. For the classification task, the model trained with thalamus, cortical gray matter, hippocampus, and lesions volumes achieved an area under the receiver operating characteristic curve of 0.74. For the regression task, the model trained with cortical gray matter and thalamus volumes, EDSS, nucleus accumbens, lesions, and putamen volumes, and age reached a mean absolute error of 0.95. In conclusion, our results confirmed that damage to cortical gray matter and relevant deep and archaic gray matter structures, such as the

Chiara Marzi, Alessandro d'Ambrosio contributed equally to this work and share first authorship. Stefano Diciotti and Alvino Bisecco contributed equally to this study. Gioacchino Tedeschi and Antonio Gallo supervised equally.

INNI Network: Federica Aprile[9], Marco Battaglini[9], Rocco Capuano[1], Alessandro De Rosa[1], Fabrizio Esposito[1], Daniele Gasparini[9], Costanza Gianni[7], Olga Marchesi[5], Damiano Mistri[5], Elisabetta Pagani[5], Nikolaos Petsas[7], Claudia Piervincenzi[7], Paolo Preziosa[5], Serena Ruggieri[7], Mauro Sibilia[5], Maria Laura Stromillo[9], Paola Valsasina[5].

Napoli, Italy.
Email: chiara.marzi@unicampania.it; chiara.marzi3@unibo.it

Funding information
Fondazione Italiana Sclerosi Multipla, Grant/Award Number: FISM2018/S/3

thalamus and hippocampus, is among the most relevant predictors of cognitive performance in MS.

KEYWORDS
artificial intelligence, cognitive performance, information processing speed, machine learning, MRI, multiple sclerosis, symbol digit modalities test

1 | INTRODUCTION

Multiple sclerosis (MS) is a chronic, inflammatory, demyelinating, and neurodegenerative disease of the central nervous system (Filippi et al., 2018) and is the commonest nontraumatic disabling disease affecting young adults (Dobson & Giovannoni, 2019). A large proportion of patients with MS, regardless of the clinical phenotype, experiences cognitive deficits (Benedict et al., 2020; Johnen et al., 2017; Ruano et al., 2017) with predominant involvement of information processing speed (IPS) and episodic memory domains (Benedict et al., 2020; Filippi et al., 2020). Cognitive impairment (CI), sometimes neglected, has a strong negative impact on social activities, employment status, and, more generally, on daily living and quality of life of these patients (Benedict et al., 2020). Therefore, there is an increased belief that monitoring the cognitive status of MS patients should be included in routine clinical assessment. To evaluate CI in MS, several neuropsychological (NP) batteries have been developed and licensed (Benedict et al., 2002; Langdon et al., 2012). Nevertheless, monitoring CI with such tools in routine clinical practice is hampered by the shortage of neuropsychologists and dedicated space and time. A solution would be to screen MS patients with short batteries, such as the Brief International Cognitive Assessment for Multiple Sclerosis (Langdon et al., 2012) or a single test with high sensitivity and predictive value. In this regard, the Symbol Digit Modalities Test (SDMT), which primarily assesses IPS (Smith, 1982), owing to its feasibility (a few minutes to administrate), reliability, sensitivity, ecology, and predictive value, has been recommended as a valid screening test for CI in MS (Benedict et al., 2020; Kalb et al., 2018; Parmenter et al., 2007; Van Schependom et al., 2014).

Magnetic resonance imaging (MRI) has markedly improved our understanding of the mechanisms associated with CI in MS patients (Benedict et al., 2020; Rocca et al., 2015), showing the relevant contribution of white matter (WM) lesion burden (Benedict, Weinstock-Guttman, et al., 2004; Foong et al., 2000; Rao et al., 1989; Stankiewicz et al., 2011), ventricular enlargement (Christodoulou et al., 2003; Rao et al., 1985) as well as whole brain and grey matter (GM) atrophy. As regards GM atrophy, in particular, the most relevant contribution to CI comes from global (Sanfilipo et al., 2006), cortical (Amato et al., 2004), and deep and archaic GM (Benedict et al., 2009; Benedict et al., 2013; Bisecco et al., 2015; Geurts et al., 2007; Houtchens et al., 2007; Sicotte et al., 2008) damage. However, which structural MRI markers are the most closely related to the cognitive performance of MS patients is still unclear. In fact, these studies have explored the contribution to CI in MS patients of just one or a limited number of specific brain structures. Since CI has been found to be related—as expected—to damage to many different brain regions, there is still a need to define, for monitoring and treatment implications, also at a single subject level, which brain regions are the most relevant or which combination of them is more predictive of CI in MS. An approach that integrates multiple MRI-derived metrics to infer brain damage patterns related to cognitive performance should better capture the complexity behind CI in MS, likely subtended by multiple biological processes acting together (Dolan, 2008; Van Schependom & Nagels, 2017). Moreover, the cognitive assessment, especially if repeated over time, is prone to some reliability concerns (Kalb et al., 2018). Thus, the selection of few and highly specific imaging and/or nonimaging features able to predict the cognitive status of an MS patient at a single subject level would also be extremely useful in a clinical setting.

In order to use MRI features to predict CI in individual patients, advanced statistical approaches are required (Bzdok et al., 2018). In the last few years, machine learning (ML) techniques have emerged as a very promising approach for studying high-dimensional data with a hidden complex pattern (Paulus et al., 2019). In neuroimaging research, the support of these advanced tools can help to understand how the biological system behaves and in forecasting unobserved outcomes or future behavior (Bzdok et al., 2018). So far, several studies have applied ML techniques to assist the diagnosis of MS (Bendfeldt et al., 2019; Mato-Abad et al., 2019; Neeb & Schenk, 2019; Wottschel et al., 2015; Wottschel et al., 2019; Zhang et al., 2019; Zurita et al., 2018), for classifying MS patients in the most common clinical phenotypes (Ion-Mărgineanu et al., 2017), or predicting physical disability (Tommasin et al., 2021). To our knowledge, only one recent work investigated the relationship between the cognitive status of MS patients and neuroimaging features using ML techniques (Buyukturkoglu et al., 2021). Due to the small sample size and some methodological limitations (i.e., feature selection not performed in the training/validation set only), previous studies may have shown overly optimistic results.

We hypothesized that ML techniques may identify the brain structural MRI volumes that, along with demographic and clinical data, are the best predictors of the cognitive status of patients with MS, as assessed by SDMT score. To investigate our hypotheses, we run a study with the following characteristics: (1) the use of a large multicenter multimodal data set containing high-quality clinical, NP, and 3T MRI data; (2) the application of appropriate and "state of the art" methodology for the harmonization of MRI data acquired in different centers; and (3) the implementation and use of ML algorithms following a rigorous validation scheme to obtain a robust, reliable, and generalizable prediction of the cognitive performance in MS.

2 | MATERIALS AND METHODS

2.1 | Participants

Five hundred and forty MS patients, whose NP and MRI examinations were included in the Italian Neuroimaging Network Initiative (INNI) repository (https://database.inni-ms.org) (Filippi et al., 2017) were included in the study. INNI is a multicenter multimodal repository financially supported by a special research grant from the Italian MS Foundation, where demographic, clinical, NP as well as 3T structural and functional MRI data sets are collected. Currently, the INNI project is run by the four founding centers (Milan, Neuroimaging Research Unit, IRCCS San Raffaele Scientific Institute; Rome, Department of Human Neurosciences, Sapienza University; Naples, Department of Advanced Medical and Surgical Sciences, University of Campania; Siena, Department of Medicine, Surgery and Neuroscience, University of Siena). Hereinafter, the centers are referred to as A, B, C, and D in any specific order, according to previous literature (Filippi et al., 2017; Storelli et al., 2019).

In the current cross-sectional study, MS patients were selected from the INNI repository based on the following inclusion criteria: (1) availability of complete demographic and clinical data, including sex, age, years of education, disease onset, disease course, and clinical disability, as assessed by the Expanded Disability Status Scale (EDSS) score (Kurtzke, 1983); (2) availability of axial T2-weighted (T2w) and anatomical, isotropic, 3D-T_1-weighted (3D-T1w) scans; and (3) collection of clinical and NP data within 180 days from the reference MRI scan.

2.1.1 | Neurological and NP evaluation

All enrolled MS patients underwent a neurological evaluation and an NP assessment performed at each participating site by experienced neurologists and neuropsychologists. The neurological evaluation included the main information about disease history/evolution and clinical disability scores. In particular, among the clinical data available in the INNI repository, we picked up the disease duration and the EDSS score.

The INNI protocol includes a comprehensive NP evaluation based on the Brief Repeatable Battery of Neuropsychological Tests (BRB-N) (Rao, 1991) (Filippi et al., 2017). Among BRB-N tests, we selected the SDMT (Smith, 1982) in order to explore the cognitive domain that is most commonly affected by MS (Chiaravalloti & DeLuca, 2008), that is, the IPS. It consists of a symbol substitution task with a time limit, and the score is the number of correct answers (range 0–110) (Smith, 1982). Thus, higher SDMT scores represent better performance. In this study, we used the available normative data that are based on a sample of 200 healthy Italian adults to calculate demographic- and education-adjusted scores (Amato et al., 2006) and, successively, the Z-scores.

Descriptive statistics of clinical and NP data, along with demographic information of the MS patients included in this study, are reported in Table 1.

2.1.2 | MRI examination

All MS patients were scanned on the 3T MR system located in each INNI center. In this study, 3D-T1w and T2w images were utilized. All MRI data sets were acquired, at each center, on the same scanner with the same protocol, except for the 3D-T1w scans provided by center A, which were acquired with two different sequences. Thus, to adequately apply post-acquisition harmonization techniques (details in Section 2.2.3), we consider the images of center A as belonging to two different groups (A_0 and A_1). MRI acquisition parameters are detailed in Table 2.

2.2 | Methods overview

A schematic diagram of the data-analysis pipeline applied to each MS patient is shown in Figure 1. Briefly, after a preprocessing stage which

TABLE 1 Demographic, clinical, and neuropsychological information for each center participating in the INNI project

	Center A	Center B	Center C	Center D	Total
Demographic information					
# MS patients	279	151	83	27	540
Age, years mean (SD)	40.63 (12.16)	37.00 (10.62)	40.75 (10.59)	40.93 (7.91)	39.65 (11.43)
Sex, females/males	167/112	101/50	63/20	20/7	351/189
Education, years median (IQR)	13 (5)	13 (5)	13 (4)	13 (2.5)	13 (5)
Clinical evaluation					
Clinical phenotype, RR/PP/SP/CIS/BMS	182/18/54/1/24	127/1/6/17/0	75/2/3/3/0	25/1/0/0/1	410/22/63/2025
Disease duration, years mean (SD)	11.44 (7.98)	9.03 (8.95)	9.39 (8.73)	9.07 (6.89)	10.87 (8.75)
EDSS, median (IQR)	2 (3)	2 (1.5)	2 (1.5)	1.5 (0.5)	2 (2)
Neuropsychological assessment					
SDMT mean (SD)	43.14 (15.87)	40.31 (14.80)	46.18 (12.17)	41.85 (14.67)	42.75 (15.09)
SDMT z-scores mean (SD)	−0.90 (1.57)	−1.21 (1.44)	−0.71 (1.27)	−1.05 (1.52)	−0.96 (1.50)

Abbreviations: BMS, benign multiple sclerosis; CIS, clinically isolated syndrome; EDSS, Expanded Disability Status Scale; IQR, interquartile range; MS, multiple sclerosis; PP, primary progressive; RR, relapsing remitting; SD, standard deviation; SDMT, symbol digit modalities test.

TABLE 2 MRI acquisition parameters for each MR scanner participating in the INNI project

	A_1	A_2	B	C	D
MR scanner	**Philips Medical System Intera (A04051C)**	**Philips Medical System Intera (A04051C)**	**GE Healthcare Signa HDxt**	**Siemens Magnetom Verio**	**Philips Medical System Achieva**
Coil	**Eight-channel head coil**	**Eight-channel head coil**	**Eight-channel head coil**	**Twelve-channel head coil**	**Thirty two-channel head coil**
3D T_1-weighted imaging					
Sequence	FFE	TFE	IR-FSPGR	MPRAGE	FFE
Imaging plane	Axial	Sagittal	Sagittal	Sagittal	Axial
Matrix	256 × 256	256 × 240	256 × 256	256 × 256	256 × 256
FOV (mm^2)	230 × 230 × 176	256 × 240 × 192	256 × 256 × 199.2	256 × 256 × 176	256 × 256 × 192
Slice thickness (mm)	0.8	1	1.2	1	1
Number of slices	220	192	166	176	192
TR (ms)	25	7	6.988	1900	10
TE (ms)	4.6	3.2	2.85	2.9	3.9
TI (ms)	-	900	650	900	900
FA (°)	30	8	8	9	8
T_2-weighted imaging					
Sequence	Dual-echo	Dual-echo	Dual-echo	Dual-echo	Dual-echo
Imaging plane	Axial	Axial	Axial	Axial	Axial
Matrix	256 × 256	256 × 256	384 × 256	384 × 384	240 × 240 (recon 352 × 352)
FOV (mm^2)	240 × 240	240 × 240	240 × 240	220 × 220	240 × 240
Slice thickness (mm)	3	3	3	3	3
Number of slices	[44–50]	[44–50]	44	45	44
TR (ms)	[2599–2910]	[2599–2910]	3120	[3320–5310]	4000
TE (ms)	16/80	16/80	24/122	10/103	15/100
FA (°)	90	90	90	150	90
ETL	6	6	8	6	4

Abbreviations: ETL, echo train length; FA, flip angle; FFE, fast field echo; FOV, field of view; FSPGR, fast spoiled gradient echo; IR, inversion recovery; MPRAGE, magnetization prepared gradient echo; MRI, magnetic resonance imaging; TE, echo time; TFE, turbo field echo; TI, inversion time; TR, repetition time.

includes (i) focal MS lesions segmentation from T2w scans, (ii) lesion refilling and bias field correction of 3D-T1w images (Section 2.2.1), cortical, subcortical, and cerebellum tissues segmentation was performed (Section 2.2.2) to compute the volumes of several brain structures (Section 2.2.2). For handling nonbiological variance introduced by different MRI scanners and acquisition protocols, MRI-derived volumes were harmonized (Section 2.2.3). We then performed SDMT score prediction using different combinations of demographic, clinical, and MRI-derived volumes through an advanced ML approach (Section 2.2.4).

2.2.1 | MRI preprocessing

Focal WM hyperintensities of the whole brain were semi-automatically segmented in T2w images by experienced researchers at each of the participating centers using a local thresholding segmentation technique (Medical Image Processing, Analysis, and Visualization; v. 4.2.2; http://mipav.cit.nih.gov; Jim 8, Xinapse Systems Ltd, Northants, UK). For each subject, the total T2w lesion volume (T2LV) was then computed to be used as a predictor in the ML analysis.

All 3D-T1w MRI data went through two preprocessing stages. In the first stage, focal WM lesion masks were used to refill lesions in the 3D-T1w images using the *lesion_filling* tool (Battaglini et al., 2012) part of the FMRIB Software Library (FSL version 6.0.1; https://fsl.fmrib.ox.ac.uk/fsl/fslwiki). Refilling the lesions with intensities matching the surrounding normal-appearing WM ensured accurate tissue segmentation and measurement of brain subregional volumes. In the second stage, intensity inhomogeneity (bias field) in lesions-refilled 3D-T1w images was estimated and corrected by using the well-established N4 method from the Advanced Normalization Tools (ANTs) toolbox version 1.9 (Tustison et al., 2010).

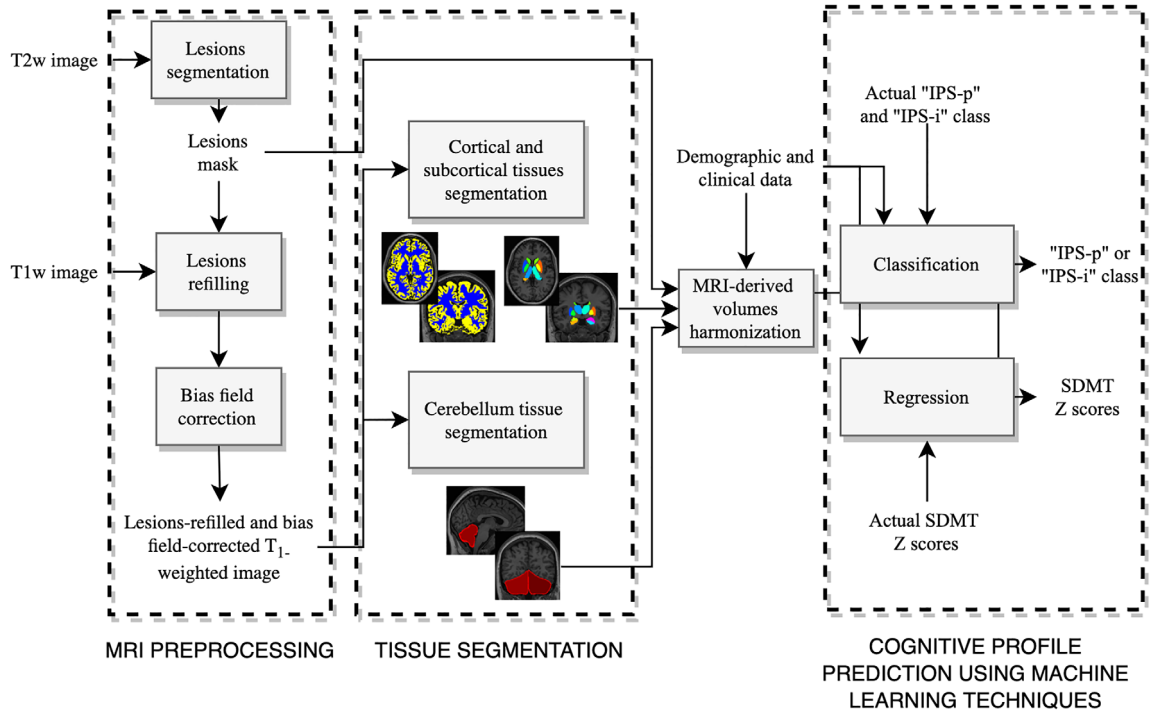

FIGURE 1 Overview of the entire magnetic resonance imaging (MRI) processing and machine learning (ML) analysis

2.2.2 | Tissue segmentation

Brain tissue segmentation was carried out applying FSL v.6.0.3 scripts (Jenkinson et al., 2012) to lesions-refilled and bias field-corrected 3D-T1w images. In particular, we used (i) the cross-sectional pipeline included in the SIENA-XL package (Battaglini et al., 2018) for computing whole brain, cortical GM; WM; thalamus; basal ganglia (i.e., putamen, caudate nucleus, nucleus accumbens, globus pallidus); hippocampus; and amygdala volumes, and (ii) FIRST scripts for cerebellar segmentation (Patenaude et al., 2011) with specific options (*-cort* option in the *first_flirt* registration tool and *-intref* option in the *run_first* tool).

All segmented volumes, including the T2LV obtained during the preprocessing (see Section 2.2.1) were computed in cm^3 and multiplied by the SIENAX scaling factor (which estimates the scaling between each subject's naïve image and standard space) to reduce head-size-related variability between subjects. For subcortical and cerebellar volumes, we considered the average volume between left and right structures.

2.2.3 | MRI-derived volumes harmonization

The success of pooling multicenter MR scans and MRI-derived metrics, for example, cortical and subcortical volumes, critically depends on the comparability of the images across centers, scanners, and imaging sequences. Indeed, MR images are subject to a large variability across scans due to differences in scanner manufacturers and heterogeneity in the imaging protocols (Fortin et al., 2017). For these reasons, before pooling our multicenter MRI-derived volumes data, we harmonized them to minimize the "center-effect" on MRIs while preserving between-subject biological variability. In particular, we used the NeuroComBat package v. 0.1.dev0 (freely available at https://github.com/ncullen93/neuroCombat), an open-source and easy-to-use Python module that can be integrated into any existing processing pipelines (Fortin et al., 2018). For each MS subject, MRI-derived volumes are known to be influenced by demographic, clinical, and NP factors, such as age (Courchesne et al., 2000), sex (Goldstein, 2001), education (Arenaza-Urquijo et al., 2013), disease duration, EDSS score (Rusz et al., 2019), clinical phenotype, and SDMT. For this reason, these variables were included in the harmonization process as a source of intersubject biological variability. The harmonization process was performed after the training/validation and test set split (details in "Training, validation, and test" section) to avoid any potential data leakage.

Then, for both training/validation and test sets, an analysis of covariance (ANCOVA) was run to evaluate the existence of the "center effect" on MRI-derived volumes before and after the harmonization step, considering the effects of different demographic and clinical data (i.e., age, sex, education level, disease duration, EDSS score, clinical phenotype, and SDMT).

All subsequent analyses used harmonized MRI-derived volumes.

2.2.4 | Prediction of the cognitive performance using ML techniques

After MRI preprocessing, tissue segmentation, and MRI-derived volumes harmonization, we predicted the cognitive performance of MS

patients using advanced ML techniques (Figure 1). Indeed, we carried out both a classification and a regression task by also evaluating the potentials of several feature combinations fed in input, as detailed in the following.

Classification task: MS patients were subdivided into "IPS-preserved (IPS-p)" and "IPS-impaired (IPS-i)" subgroups, based on their SDMT Z-scores. 1 SD below the mean (i.e., SDMT Z scores ≤−1.0) was selected as the cutoff for IPS deficit (Buyukturkoglu et al., 2021). Thus, 285 MS patients were classified as IPS-p and 255 as IPS-i. In this task, we performed a prediction of the patient class label (IPS-p vs. IPS-i).

Regression task: we performed a direct prediction of the SDMT Z-score of each patient.

For both the classification and regression tasks, we trained, validate, and tested the eXtreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016). XGBoost is a scalable end-to-end tree boosting system that is widely used to achieve state-of-the-art performance on many recent ML challenges (Chen & Guestrin, 2016). One of its major benefits regards the great potential interpretability due to its recursive tree-based decision system.

To examine the contribution of nonimaging and imaging features to SDMT performance and to assess the most contributing features, we considered (1) a priori knowledge-based sets of features coupled with (2) a data-driven approach in which we automatically select the best combination of features without preconfigured sets. For the a priori sets of features, we built different combinations of demographic, clinical, and MRI-derived features, starting from a simple model including only demographic and clinical features and gradually increasing complexity to reach a comprehensive model that included all variables. These different combinations of features were inferred from the literature and clinical practice through highly qualified MS neurologists (details in Table 3). Briefly, we first considered a model combining demographic and the main clinical features in MS research, such as the disease duration and the EDSS score. Structural neuroimaging metrics, that is, T2-WM lesion volume (accounting for WM lesions extent) and brain volumes, were progressively introduced in the analyses to consider further the impact of different structural alterations on cognitive performance. Although controlled through the use of normative data, the potential residual effect of age, sex, and education on cognitive performance was accounted for by including these variables in each feature combination. For the data-driven approach, we applied an automated feature selection procedure through an XGBoost estimator as proposed recently by Yan et al. (2020). For each feature, the XGBoost algorithm estimates the importance gain, that is, the improvement in performance brought by each feature. Thus, we iteratively retrained a new XGBoost model using the top n features in the feature ranking obtained with the combination "All" (see Table 3) using $n = 1, 2, \ldots, 16$. We then observed the potential increase in performance by adding, one by one, the features with the top importance gain. The final selection of the best feature set (from a priori and data-driven approaches) was based on the highest performance and, in the case of equal performance, we preferred the feature set with the lowest number of features, following Occam's razor principle and reducing potential overfitting (Witten & Frank, 2016).

TABLE 3 Combination of features used for both the classification and regression task

Combination name	Features
Clinical	Age, sex, education, EDSS, disease duration
Whole brain	Age, sex, education, BV
GM + WM	Age, sex, education, cGMV, WMV, ThalV, AccuV, PutaV, CaudV, PallV, AmygV, HippV
GM + WM + cerebellum	Age, sex, education, cGMV, WMV, ThalV, AccuV, PutaV, CaudV, PallV, AmygV, HippV, CerebellumV
Whole brain + les	Age, sex, education, BV, T2LV
GM + WM + cerebellum + les	Age, sex, education, cGMV, WMV, ThalV, AccuV, PutaV, CaudV, PallV, AmygV, HippV, CerebellumV, T2LV
All	Age, sex, education, EDSS, disease duration, cGMV, WMV, ThalV, AccuV, PutaV, CaudV, PallV, AmygV, HippV, CerebellumV, T2LV

Abbreviations: AccuV, nucleus accumbens volume; AmygV, amygdala volume; BV, whole brain volume; CaudV, caudate nucleus volume; CerebellumV, cerebellum volume; cGMV, cortical grey matter volume; EDSS, Expanded Disability Status Scale; HippV, hippocampus volume; les, WM lesions; PallV, globus pallidus volume; PutaV, putamen volume; T2LV, lesions load; ThalV, thalamus volume; WMV, white matter volume.

Training, validation, and test

The XGBoost model has been trained, validated, and tested using the following approach: 80% of the entire data set (i.e., 432 randomly chosen patients) were considered as the training/validation set, and the remaining 20% (i.e., 108 patients) as the test set. On the training/validation data, each model has been trained and validated using a nested k-fold cross-validation (CV) strategy (stratified for the classification task) to estimate the unbiased generalization performance of the models along with performing, at the same time, data standardization, hyperparameters optimization, and feature selection (Varma & Simon, 2006). In detail, the inner loop was used for searching for the best data standardization approach and optimizing the estimator hyperparameters, and the outer loop for the feature selection. Specifically, the *Grid Search* parameters space was composed of different transformers for data standardization, that is, standard, robust scaling, and quantile transformation, and of a set of hyperparameters of the XGBoost estimator (see details in Supporting Table S1). The feature selection has been performed according to the performance in the outer loop of the nested CV (i.e., the validation set).

Moreover, since the selected features combination may vary depending on how the training/validation data are split in each fold of the nested CV, the latter has been repeated 10 times using random splits. A detailed diagram of the validation scheme has been reported in Supporting Figure S1. The average and standard deviation of the performance on the unseen test sets across all repetitions were computed to get the final scores. In particular, for the classification and regression tasks, the performance was quantified in terms of the area under the receiver operating characteristic curve (AUROC) and mean absolute error (MAE), respectively.

Experimental tests

The extraction of advanced neuroimaging features was carried out on a Dell PowerEdge T620 workstation equipped with two 8-core Intel Xeon E5–2640 v2, for a total of 32 CPU threads and 128 GB RAM, using the Oracle Grid Engine scheduler. For each subject, the processing time of a single-core CPU required approximately 20 and 15 min for the quantification of cerebral and cerebellar features, respectively.

The training, validation, and test of the pipelines were carried out using a custom-made code in Python language (v. 3.8.1) using the following modules: graphviz v.0.15, matplotlib v.3.3.4, numpy v.1.18.1, pandas v.1.0.2, pingouin v.0.3.5, scikit-learn v.0.22.2.post1 (Pedregosa et al., 2011), seaborn v.0.11.0, and xgboost v.1.2.1. In particular, we used *XGBClassifier* and *XGBRegressor* estimators for the classification and regression task, respectively. The total computation time for the training, validation, and test was about 5 days on a single core of a Linux workstation equipped with a 4-core (eight threads) Intel i7-7700K CPU and 64 GB RAM.

3 | RESULTS

3.1 | Data harmonization

Before data harmonization, ANCOVA results showed highly significant differences in MRI-derived volumes among different INNI centers (p-values $<10^{-3}$ for all structures except for NT2LV in the training/validation set and p-values $<10^{-2}$ for all structures except for NT2LV and NCaudmV in the test set). In particular, cortical GM, WM, and cerebellar volumes showed the most relevant differences, while the volumes of the whole brain and subcortical structures showed less pronounced differences (Table 4 and Figures 2 and 3). After data harmonization, volume differences among groups were either removed (ANCOVA test p-values >.05) or highly reduced (the partial η^2 coefficients relating to the group effects were reduced) in all structures for both the training/validation and test sets (Table 4 and Figures 2 and 3).

3.2 | Classification task

For the prediction of the cognitive class (IPS-p vs. IPS-i), the AUROC scores in the validation set are reported in Table 5 and represented in Figure 4. All the models showed good performance (average AUROC in the range 0.71–0.74). In particular, the best performance, that is, AUROC of 0.74 (0.01) [mean (standard deviation, SD)], was achieved by the following features' combinations: *Whole brain + les* (i.e., age, sex, education, brain volume, T2LV) and *Auto 4*

TABLE 4 ANCOVA test p-values and partial η^2 coefficients relating to the group, before and after the harmonization step, are reported for both the training/validation and test sets. After data harmonization, volume differences among groups were either removed (p-values >.05) or highly reduced (reduced partial η^2 coefficients relating to the group effects) in all structures

	Before harmonization				After harmonization			
	Training/validation		Test		Training/validation		Test	
Volume	p-Value	Partial η^2	p-Value	Partial η^2	p-Value	Partial η^2	p-Value	Partial η^2
NBV	2E-5	0.06	9E-4	0.17	0.02	0.03	.47	0.04
NWMV	2E-106	0.69	1E-26	0.73	0.34	0.01	.38	0.04
cpGMV	6E-93	0.64	2E-25	0.71	0.00	0.04	.26	0.05
NthalmV	3E-9	0.10	6E-4	0.18	0.04	0.02	.83	0.02
NhippmV	7E-19	0.19	4E-5	0.23	0.12	0.02	.93	0.01
NamygmV	6E-14	0.15	3E-3	0.15	0.73	0.00	.96	0.01
NaccumV	2E-34	0.32	3E-9	0.37	0.25	0.01	.99	0.00
NcaudmV	3E-4	0.05	0.15	0.07	0.09	0.02	.83	0.01
NpallmV	2E-16	0.17	2E-05	0.24	0.53	0.01	.82	0.02
NputamV	2E-24	0.24	2E-7	0.31	0.34	0.01	.92	0.01
Ncerebellum_mV	3E-54	0.46	4E-15	0.53	0.97	0.00	.98	0.00
NT2LV	0.76	0.004	0.62	0.03	0.61	0.01	.82	0.02

Abbreviations: NAccumV, normalized mean accumbens volume; NAmygmV, normalized mean amygdala volume; NBV, normalized whole brain volume; NCaudmV, normalized mean caudate volume; NCerebellm_mV, normalized mean cerebellum volume; NcGMV, normalized cortical gray matter volume; NHippmV, normalized mean hippocampus volume; NWMV, normalized white matter volume; NPallmV, normalized mean pallidus volume; NPutamV, normalized mean putament volume; NT2LV, normalized lesion volume; NThalmV, normalized mean thalamus volume.

FIGURE 2 Box plot of the volumes of different structures among groups in the training/validation set. The horizontal line inside each box represents the median value of the plotted data. The box shows the first and third quartiles, while the whiskers extend to show the rest of the distribution except for points that are determined to be outliers. Features' acronyms are described in Table 4

(i.e., thalamus, cortical GM, hippocampus volumes, and T2LV) (Figure 4a). With the same highest performance but fewer predictors, the latter combination was considered the best set of predictors for the cognitive class. This set showed an increase in AUROC of 2.78% (Figure 4b) compared to the classification performance obtained using the XGBoost model trained using only the most important

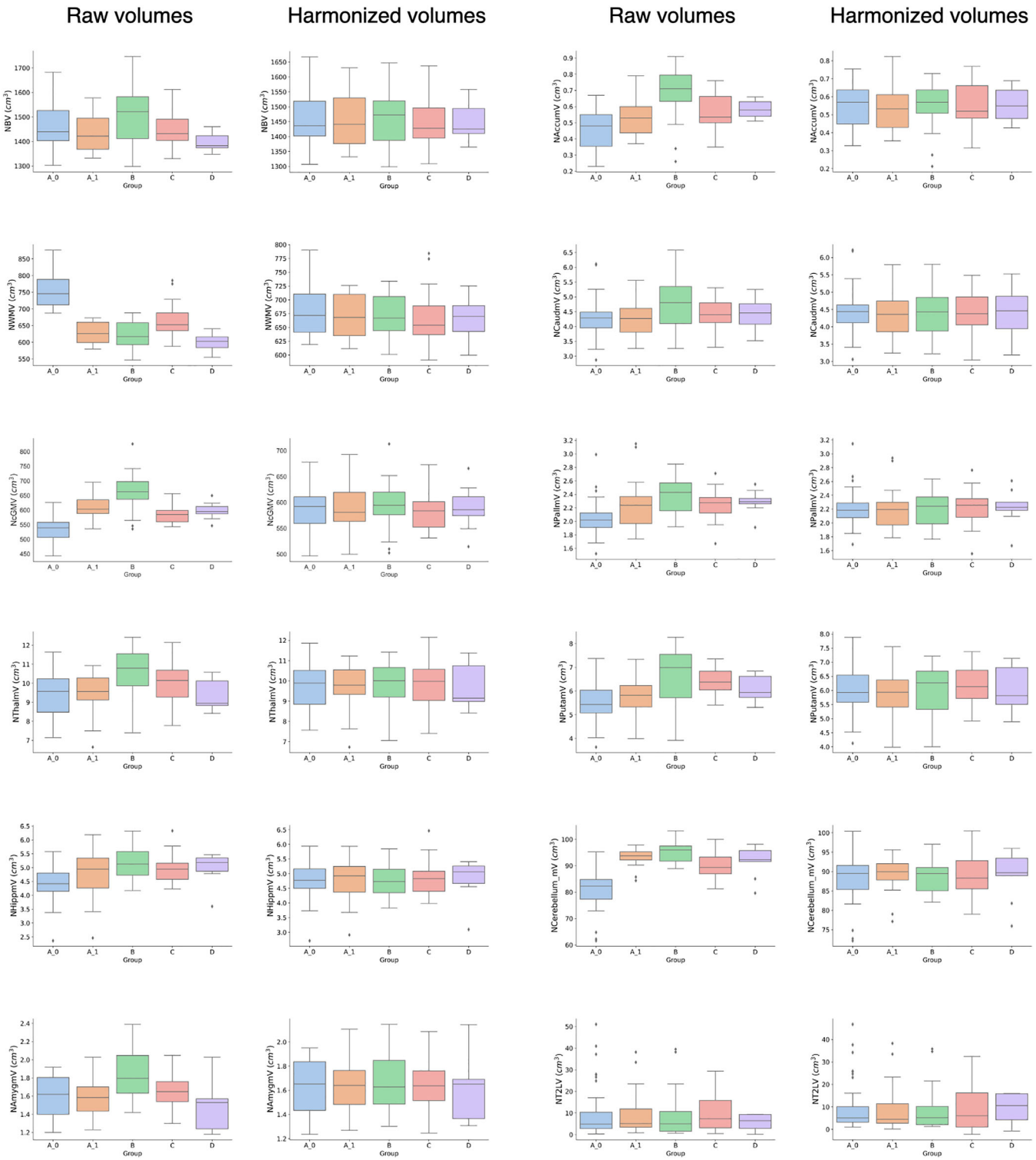

FIGURE 3 Box plot of the volumes of different structures among groups in the test set. The horizontal line inside each box represents the median value of the plotted data. The box shows the first and third quartiles, while the whiskers extend to show the rest of the distribution except for points that are determined to be outliers. Features' acronyms are described in Table 4

predictor, that is, the thalamus volume. Then, the final XGBoost classifier trained with the thalamus, cortical GM, hippocampus, and lesions volumes was tested on the unseen test set data obtaining an AUROC of 0.69 (0.03) [mean (SD)].

3.3 | Regression task

To predict the SDMT z-scores, the average MAE values in the validation set are reported in Table 5 and Figure 5. The best performance

TABLE 5 Performances in the validation set. Mean values (standard deviation) of 10 repetitions are reported. For the classification task (IPS-p vs. IPS-i), we computed AUROC values, and for the regression task (SDMT Z-score prediction), we showed the MAE values. The combination of features automatically selected are graphically reported in Figures 4a and 5b for the classification and regression task, respectively

Feature combination	Classification AUROC	Regression MAE
Clinical	0.71 (0.01)	1.05 (0.01)
Whole brain	0.73 (0.01)	0.99 (0.01)
GM + WM	0.72 (0.01)	0.97 (0.01)
GM + WM + cerebellum	0.72 (0.01)	0.97 (0.00)
Whole brain + les	0.74 (0.01)	0.97 (0.01)
GM + WM + cerebellum + les	0.73 (0.02)	0.96 (0.01)
All	0.72 (0.02)	0.96 (0.01)
Auto 1	0.72 (0.01)	1.06 (0.01)
Auto 2	0.73 (0.01)	1.02 (0.01)
Auto 3	0.72 (0.01)	0.99 (0.01)
Auto 4	0.74 (0.01)	0.98 (0.01)
Auto 5	0.73 (0.01)	0.97 (0.01)
Auto 6	0.73 (0.01)	0.98 (0.01)
Auto 7	0.73 (0.01)	0.95 (0.01)
Auto 8	0.73 (0.01)	0.95 (0.01)
Auto 9	0.73 (0.01)	0.96 (0.01)
Auto 10	0.73 (0.01)	0.96 (0.01)
Auto 11	0.73 (0.01)	0.96 (0.01)
Auto 12	0.73 (0.01)	0.96 (0.01)
Auto 13	0.73 (0.01)	0.96 (0.01)
Auto 14	0.73 (0.02)	0.96 (0.01)
Auto 15	0.73 (0.02)	0.96 (0.01)

Abbreviations: AUROC, area under the receiver operating characteristic curve; Auto, the combination of features automatically selected; GM, gray matter; les, WM lesions; MAE, mean absolute error; SDMT, Symbol Digit Modalities Test; WM, white matter.

(MAE = 0.95 (0.01) [mean (SD)]) has been achieved by the XGBoost regressor trained with cortical GM and thalamus volumes, EDSS, nucleus accumbens, lesions, putamen volumes, and age. This feature combination, chosen automatically during the training phase, showed a decrease in MAE score of 10.38%, compared with a regression performed using an XGBoost model trained using only the best predictor, that is, the cortical GM volume. Finally, this model has been tested on the unseen test data, obtaining an MAE score of 1.02 (0.01) [mean (SD)].

4 | DISCUSSION

In this study, we applied ML techniques to predict a proxy (i.e., the SDMT score) of the cognitive status of MS patients. We performed both a classification task (IPS-p vs. IPS-i MS patients) and a regression task (SDMT score prediction) combining the information obtained from demographic, clinical, and MRI-derived volumes data of 540 MS patients belonging to the large, multicenter, INNI repository. An XGBoost estimator was trained, validated, and tested using a combined hold-out/CV scheme (80% of subjects in the training/validation sets and 20% in the test set). In the training/validation set, the model was trained and validated using a nested CV strategy (stratified for the classification task) to perform hyperparameters optimization and feature selection. Moreover, since the decisions may vary depending on how the training/validation data are split in each fold of the nested CV, the nested CV procedure was repeated 10 times using different random splits. Our results showed that all the features' combinations showed a good performance. For the classification task, the XGBoost classifier trained with thalamus, cortical GM, hippocampus, and lesions volumes, achieved an AUROC score of 0.74 (0.01) in the validation set, and an AUROC score of 0.69 (0.03) on the (unseen) test set data. On the other side, in the regression task, the best performance was achieved by the XGBoost regressor trained with cortical GM and thalamus volumes, EDSS, nucleus accumbens, lesions, putamen volumes, and age, obtaining an MAE equal to 0.95 (0.01) in the validation set, and an MAE = 1.02 (0.01) on the (unseen) test set.

Our findings confirm that the diffuse damage to the structural brain architecture subtended to MS pathology may predict consequences on the cognitive status of MS patients (Meijer et al., 2018), which cannot be sufficiently explained using clinical data alone. Beyond the model showing the best performance, we were interested in unveiling the smallest feature set (i.e., that with fewer features) best predicting the SDMT score and less prone to overfitting. For example, for the classification task, we observed two models with the same best performance (i.e., AUROC = 0.74 (0.01)), and we selected the feature combination with fewer features (i.e., thalamus, cortical GM, hippocampus volumes, and T2LV). Basically, we showed that, in the classification task, the thalamus, cortical GM, hippocampus volumes, and T2 lesion volume are a dense representation of all MRI-related and clinical/demographic features. We feel that this is an important result, in line with previous studies (Benedict et al., 2013; Bergsland et al., 2016; Bisecco et al., 2015; Bisecco et al., 2018; Burggraaff et al., 2020). Indeed, cortical atrophy, in particular localized area of the prefrontal, parietal, and temporal cortex, is known to be a critical substrate for CI (Amato et al., 2004; Benedict, Carone, & Bakshi, 2004; Benedict, Weinstock-Guttman, et al., 2004; Nocentini et al., 2014; Zivadinov et al., 2001). The thalamus, with its extensive afferent and efferent connections with the midbrain and the cerebral cortex, serves as a crucial "cognitive hub" and, thus, its degeneration is likely to contribute to IPS dysfunction (Minagar et al., 2013) and consequently to a global cognitive dysfunction. At the same time, it is well known that the thalamic volume highly correlates with the whole-brain volume in MS populations with a relatively high disease duration—like in our cohort (mean 10.8 years, SD 8.7 years) (Eshaghi et al., 2018). This may explain, for example, why combinations with the volume of specific/localized brain regions or the whole-brain may be equally valuable. Besides the thalamus, another relevant "cognitive structure," such as the hippocampus, was found to contribute to cognitive dysfunction in MS patients. The hippocampus is a predilected site for demyelinated

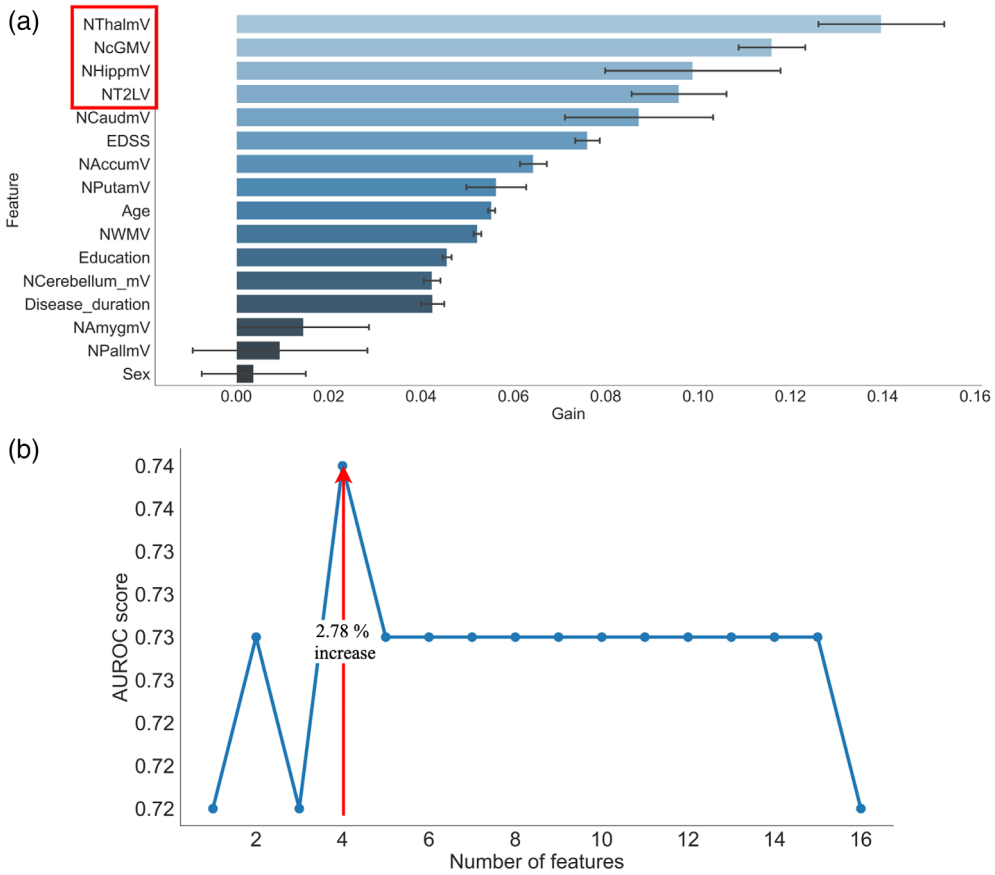

FIGURE 4 Classification task using the XGBoost estimator for the automated features selection. (a) Feature ranking and (b) area under the ROC curve (AUROC) as a function of the number of features. In both panels, average values (using 10 repetitions of the fivefold nested stratified cross-validation [CV]) in the validation set are reported. In Panel (a), the black lines with caps indicate the standard deviation, and the features in the red rectangle are those that together get the best AUROC in the validation set. Features' acronyms are described in Table 3

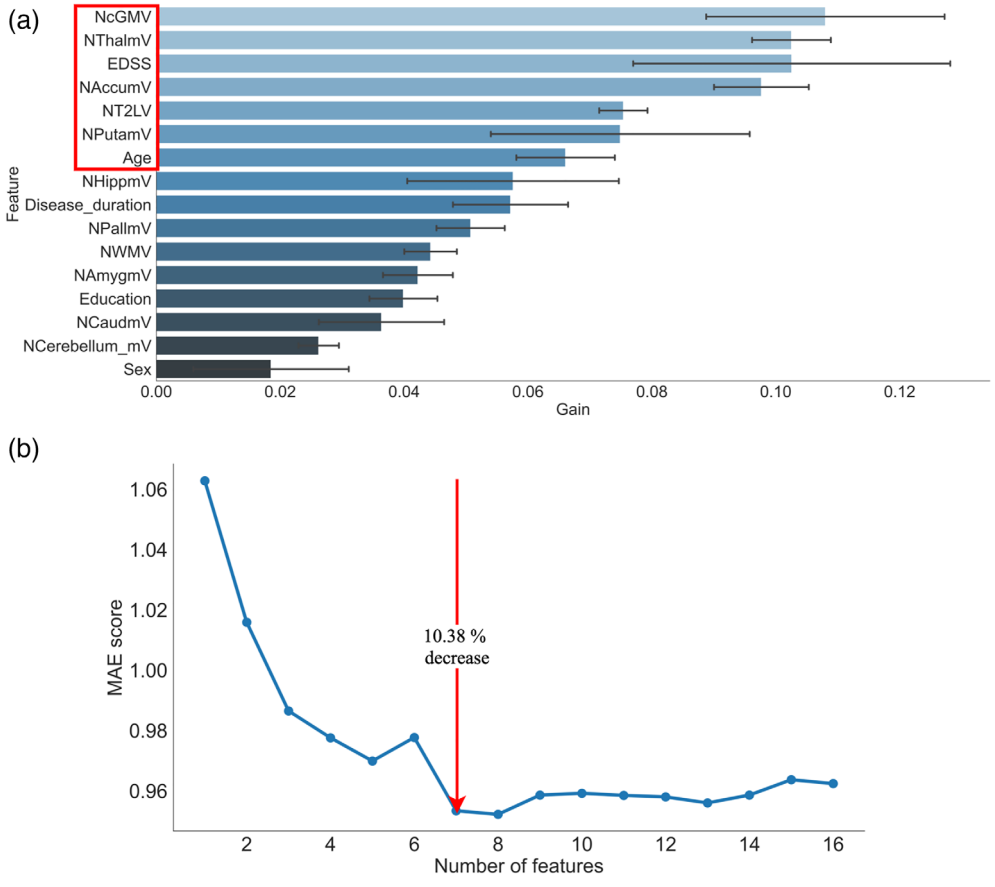

FIGURE 5 Regression task using the XGBoost estimator for the automated features selection. (a) Feature ranking and (b) mean absolute error (MAE) as a function of the number of features. In both panels, average values (using 10 repetitions of the fivefold nested cross-validation [CV]) in the validation set are reported. In panel (a), the black lines with caps indicate the standard deviation, and the features in the red rectangle are those that together get the best MAE in the validation set. Features' acronyms are described in Table 3

lesions (Benedict et al., 2020; Geurts et al., 2007), directly involved in learning and memory functions. The cortical–thalamic–hippocampal disruption affects cognitive performance in MS with mild to minimal CI (Kern et al., 2015).

As regards WM T2LV, although it was retained in our models among the main predictors of the cognitive status in MS patients, it could not fully explain the severity of CI in MS patients, once again confirming the concept of the "clinic-radiological" paradox observed in MS (Barkhof, 2002). An increasing number of studies, on the other hand, have shown that focal MS-related WM damage is the tip of the iceberg, representing just the visible inflammatory processes (Meijer et al., 2018; Miller et al., 2003; Moll et al., 2011; Rao et al., 2014), while the most extensive damage is represented by the widespread, microscopic, involvement of normal-appearing WM as well as cortical and deep GM (Popescu et al., 2015; Rocca et al., 2017).

4.1 | Methodological considerations on the ML approach

We explored the predictive abilities of a wide set of demographic, clinical, and neuroimaging features with an ML approach, in line with the goals of evaluating cognitive performance on an individual basis. This approach differs from conventional regression analysis applied to the entire data set in which the possibility of overfitting may not be negligible. In particular, we used the hold-out method to split the entire data set into a training/validation set (80%), and test set (20%), and, in the training/validation set only, a fivefold nested CV scheme was applied to perform, simultaneously, hyperparameters optimization and feature selection. The combination of the hold-out method along with the nested CV in the training/validation set allowed us to take all the decisions on the training/validation data only and to evaluate the performance of the final model on unseen data, thus preventing any form of peeking (Diciotti et al., 2013).

Ideally, any ML model should be evaluated on samples that were not used to train or fine-tune (e.g., through hyperparameter optimization) the model so that they provide an unbiased assessment of the generalization error, or in other words, a "*sense of model effectiveness*" (Kuhn & Johnson, 2013). However, and unfortunately, many studies in the literature do not use a truly test set with samples unseen during the training and hyperparameter optimization (Bendfeldt et al., 2019; Wottschel et al., 2015; Wottschel et al., 2019; Zhang et al., 2019; Zurita et al., 2018), leading to a risk of overfitting and overly optimistic results. The lack of data never used during the "decisional" phase (hyperparameters optimization, and feature selection) does not allow an unbiased evaluation of the ability of these advanced algorithms to learn from data and generalize. To the best of our knowledge, only Buyukturkoglu et al. (2021) applied a nested CV scheme in order to perform hyperparameters optimization in the inner loop, but the feature selection was carried out in the outer CV, thus making their results noncompletely reliable.

Comparing, in the classification task, the AUROC = 0.74 in the validation set with the AUROC = 0.69 in the test set, we found a drop of 0.05. This drop value has been considered "modest" in a recent systematic review comparing the performance of deep learning algorithms on the internal and external data sets (Yu et al., 2022). It is well known from the ML theory that a drop in the performance in the test set may be present and can be due to several factors, including (i) a very different sample distribution in the training/validation and test sets, (ii) possible overfitting in the model selection procedure, and (iii) small size of the data set and specifically, of the validation set (Müller & Guido, 2016). In our study, (i) we split the training/validation and test set from the same sample population, and we did not notice different distributions of the features, for example, due to the random sampling; (ii) we tried not to make the XGBoost hyperparameter optimization too complex because when the space of models searched over becomes richer, the probability of incurring overfitting is increased; and (iii) we adopted a fivefold nested CV in the training/validation set, because it offers a favorable bias-variance trade-off (Hastie et al., 2013; Lemm et al., 2011) and is also adequate for model selection (Breiman & Spector, 1992). Although we observe this residual effect, we highlight that our study applied a rigorous split of training/validation and test sets for the first time in predicting a cognitive score in an MS population using a valuable multicenter data set.

4.2 | Methodological considerations on the multicenter data set and the need for MRI data harmonization

Multicenter studies confer many distinct advantages, including larger sample sizes and allowing to find more generalizable findings, sharing resources among collaborative sites, and promoting networking (Cheng et al., 2017; Localio et al., 2001). Well-executed multicenter studies are more likely to improve performance and/or have a positive impact on research and clinical outcomes (Cheng et al., 2017; Huggett et al., 2011; O'Sullivan et al., 2010; Payne et al., 2011; Schwartz et al., 2016). In recent years, multicenter neuroimaging studies in the field of MS have rapidly increased (Chitnis et al., 2013; Hagens et al., 2018; Preziosa et al., 2016; Storelli et al., 2019). Even in multicenter MRI studies with consistent scanner field strength, systematic differences in scanner manufacturers and acquisition parameters can lead to severe biases in volumetric analyses (Shinohara et al., 2017), particularly when subtle differences in tissue volume are being searched for along with association with cognitive functions. These nonbiological confounders typically have a priori unpredictable effects, and several statistical approaches attempted to handle this source of variability (Fortin et al., 2017). To this aim, in this multicenter study, we harmonized the MRI-derived volumes by using NeuroComBat (Fortin et al., 2017), a technique formerly proposed for genetic data (Johnson et al., 2007). Recently, the same approach has been successfully applied to diffusion tensor imaging data (Fortin et al., 2017), cortical thickness measurements (Fortin et al., 2018; Radua et al., 2020), and subcortical volumes (Pomponio et al., 2020; Radua et al., 2020). Moreover, among the advantages of this harmonization technique, the possibility of applying it directly to MRI-derived volumes, regardless of how the images were acquired (different scanners and different acquisition protocols), is the most important.

Indeed, the INNI repository currently collects retrospective 3T MRI data from four core centers where different scanners and acquisition protocols were used for specific research purposes.

4.3 | Limitations and future developments

This study presents some limitations. First, we predicted the cognitive performance in a sample of MS patients with any phenotype of the disease. It is not yet clear, indeed, whether different MS phenotypes have overlapping pathophysiological substrates of CI, although similar NP profiles have been described in all MS courses (Benedict et al., 2020; De Sonneville et al., 2002; Huijbregts et al., 2006). Unfortunately, we were not able to perform sub-group analyses due to the paucity of some phenotypes—that is, primary progressive MS, clinically isolated syndrome, and benign MS—and a different distribution within the participating centers. Future ML studies should investigate whether different MS phenotypes have different structural brain MRI predictors of CI.

Second, the INNI repository currently contains MRI data acquired with imaging protocols set by each center independently. A recent study concluded that "The use of standardized protocols yielded up to a five-fold reduction in required sample sizes to detect disease-related neuroanatomical changes, and is particularly beneficial for detecting subtle effects" (George et al., 2020). For these reasons, and according to the INNI main future goals (Filippi et al., 2017), standardized acquisition protocols of advanced structural and functional MRI data set will be advocated.

Finally, in this study, we evaluated the relationship between the cognitive status, measured through the SDMT score, and the volumetric data extracted from anatomical T1w and T2w scans. Future research should investigate predictors of cognitive performance using other single/combination of NP tests as well as other MRI metrics, such as diffusion-weighted imaging- and, especially, functional MRI-derived metrics.

5 | CONCLUSION

Our ML approach using a comprehensive set of brain structural measures extracted from a large multicenter 3T-MRI data set showed a good performance in predicting CI in MS. This novel approach confirmed how the involvement of some cognitive hubs of the brain, such as the thalamus and the hippocampus, are more relevant than focal WM damage (i.e., T2LV) in the prediction of cognitive performance in MS.

ACKNOWLEDGMENT

INNI Network.

FUNDING INFORMATION

The Italian Neuroimaging Network Initiative (INNI) (https://database.inni-ms.org) is a multicenter multimodal repository, financially supported by a research grant from the Fondazione Italiana Sclerosi Multipla (FISM2018/S/3), and financed or cofinanced with the "5 per mille" public funding.

CONFLICT OF INTEREST

The authors declare the following conflicts of interest. Chiara Marzi: None. Alessandro d'Ambrosio: None. Stefano Diciotti: None. Alvino Bisecco received speaker's honoraria and/or compensation for consulting service and/or speaking activities from Biogen, Roche, Merck, Celgene and Genzyme. Manuela Altieri: None. Massimo Filippi is Editor-in-Chief of the Journal of Neurology and Associate Editor of Human Brain Mapping; received compensation for consulting services and/or speaking activities from Almirall, Alexion, Bayer, Biogen Idec, Celgene, Eli Lilly, Genzyme, Merck-Serono, Novartis, Roche, Sanofi, Takeda, and Teva Pharmaceutical Industries; and receives research support from Biogen Idec, Merck-Serono, Novartis, Roche, Teva Pharmaceutical Industries, Italian Ministry of Health, Fondazione Italiana Sclerosi Multipla, and ARiSLA (Fondazione Italiana di Ricerca per la SLA). Maria Assunta Rocca received speakers' honoraria from Bayer, Biogen Idec, Bristol Myers Squibb, Celgene, Genzyme, Merck Serono, Novartis, Roche and Teva, and receives research support from the MS Society of Canada and Fondazione Italiana Sclerosi Multipla. Loredana Storelli: None. Patrizia Pantano has received funding for travel from Novartis, Genzyme, and Bracco and a speaking honorarium from Biogen. She received research support from Italian Ministry of Foreign Affairs and Fondazione Italiana Sclerosi Multipla. Silvia Tommasin: None. Rosa Cortese: None. Nicola De Stefano has received honoraria from Biogen-Idec, Bristol Myers Squibb, Celgene, Genzyme, Immunic, Merck Serono, Novartis, Roche and Teva for consulting services, speaking, and travel support. He serves on advisory boards for Merck, Novartis, Biogen-Idec, Roche, and Genzyme, Immunic and he has received research grant support from the Italian MS Society. Gioacchino Tedeschi is speaker, consulting fees and research support from Biogen, Genzyme, Merck Serono, Mylan, Novartis, Roche, Teva, Allergan, Abbvie and Lundbeck. Research support from Fondazione Italiana Sclerosis Multipla. Antonio Gallo received speaker and consulting fees from Biogen, Genzyme, Merck Serono, Mylan, Novartis, Roche, and Teva, and receives research support from Fondazione Italiana Sclerosi Multipla.

DATA AVAILABILITY STATEMENT

Authors elect to not share data.

PATIENT CONSENT STATEMENT

Informed consent was obtained from all individual participants included in the study.

ORCID

Chiara Marzi https://orcid.org/0000-0002-1791-3573
Stefano Diciotti https://orcid.org/0000-0001-8778-7819
Alvino Bisecco https://orcid.org/0000-0002-7202-4445
Manuela Altieri https://orcid.org/0000-0003-0483-4478
Massimo Filippi https://orcid.org/0000-0002-5485-0479
Maria Assunta Rocca https://orcid.org/0000-0003-2358-4320
Loredana Storelli https://orcid.org/0000-0002-4979-613X

Patrizia Pantano https://orcid.org/0000-0001-9659-8294
Silvia Tommasin https://orcid.org/0000-0001-9088-7968
Rosa Cortese https://orcid.org/0000-0002-9803-7914
Nicola De Stefano https://orcid.org/0000-0003-4930-7639
Gioacchino Tedeschi https://orcid.org/0000-0002-3321-1125
Antonio Gallo https://orcid.org/0000-0002-2203-6237

REFERENCES

Amato, M. P., Bartolozzi, M. L., Zipoli, V., Portaccio, E., Mortilla, M., Guidi, L., Siracusa, G., Sorbi, S., Federico, A., & De Stefano, N. (2004). Neocortical volume decrease in relapsing-remitting MS patients with mild cognitive impairment. *Neurology*, *63*, 89–93. https://doi.org/10.1212/01.wnl.0000129544.79539.d5

Amato, M. P., Portaccio, E., Goretti, B., Zipoli, V., Ricchiuti, L., De Caro, M. F., Patti, F., Vecchio, R., Sorbi, S., & Trojano, M. (2006). The Rao's brief repeatable battery and Stroop test: Normative values with age, education and gender corrections in an Italian population. *Multiple Sclerosis*, *12*, 787–793. https://doi.org/10.1177/1352458506070933

Arenaza-Urquijo, E. M., Landeau, B., La Joie, R., Mevel, K., Mézenge, F., Perrotin, A., Desgranges, B., Bartrés-Faz, D., Eustache, F., & Chételat, G. (2013). Relationships between years of education and gray matter volume, metabolism and functional connectivity in healthy elders. *NeuroImage*, *83*, 450–457. https://doi.org/10.1016/j.neuroimage.2013.06.053

Barkhof, F. (2002). The clinico-radiological paradox in multiple sclerosis revisited. *Current Opinion in Neurology*, *15*, 239–245.

Battaglini, M., Jenkinson, M., & De Stefano, N. (2012). Evaluating and reducing the impact of white matter lesions on brain volume measurements. *Human Brain Mapping*, *33*, 2062–2071. https://doi.org/10.1002/hbm.21344

Battaglini, M., Jenkinson, M., De Stefano, N., & for the Alzheimer's Disease Neuroimaging Initiative. (2018). SIENA-XL for improving the assessment of gray and white matter volume changes on brain MRI: SIENA-XL for brain atrophy. *Human Brain Mapping*, *39*, 1063–1077. https://doi.org/10.1002/hbm.23828

Bendfeldt, K., Taschler, B., Gaetano, L., Madoerin, P., Kuster, P., Mueller-Lenke, N., Amann, M., Vrenken, H., Wottschel, V., Barkhof, F., Borgwardt, S., Klöppel, S., Wicklein, E.-M., Kappos, L., Edan, G., Freedman, M. S., Montalbán, X., Hartung, H.-P., Pohl, C., … Nichols, T. E. (2019). MRI-based prediction of conversion from clinically isolated syndrome to clinically definite multiple sclerosis using SVM and lesion geometry. *Brain Imaging and Behavior*, *13*, 1361–1374. https://doi.org/10.1007/s11682-018-9942-9

Benedict, R. H., Hulst, H. E., Bergsland, N., Schoonheim, M. M., Dwyer, M. G., Weinstock-Guttman, B., Geurts, J. J., & Zivadinov, R. (2013). Clinical significance of atrophy and white matter mean diffusivity within the thalamus of multiple sclerosis patients. *Multiple Sclerosis*, *19*, 1478–1484. https://doi.org/10.1177/1352458513478675

Benedict, R. H. B., Amato, M. P., DeLuca, J., & Geurts, J. J. G. (2020). Cognitive impairment in multiple sclerosis: Clinical management, MRI, and therapeutic avenues. *Lancet Neurology*, *19*, 860–871. https://doi.org/10.1016/S1474-4422(20)30277-5

Benedict, R. H. B., Carone, D. A., & Bakshi, R. (2004). Correlating brain atrophy with cognitive dysfunction, mood disturbances, and personality disorder in multiple sclerosis. *Journal of Neuroimaging*, *14*, 36S–45S. https://doi.org/10.1177/1051228404266267

Benedict, R. H. B., Fischer, J. S., Archibald, C. J., Arnett, P. A., Beatty, W. W., Bobholz, J., Chelune, G. J., Fisk, J. D., Langdon, D. W., Caruso, L., Foley, F., LaRocca, N. G., Vowels, L., Weinstein, A., DeLuca, J., Rao, S. M., & Munschauer, F. (2002). Minimal neuropsychological assessment of MS patients: A consensus approach. *The Clinical Neuropsychologist*, *16*, 381–397. https://doi.org/10.1076/clin.16.3.381.13859

Benedict, R. H. B., Ramasamy, D., Munschauer, F., Weinstock-Guttman, B., & Zivadinov, R. (2009). Memory impairment in multiple sclerosis: Correlation with deep grey matter and mesial temporal atrophy. *Journal of Neurology, Neurosurgery, and Psychiatry*, *80*, 201–206. https://doi.org/10.1136/jnnp.2008.148403

Benedict, R. H. B., Weinstock-Guttman, B., Fishman, I., Sharma, J., Tjoa, C. W., & Bakshi, R. (2004). Prediction of neuropsychological impairment in multiple sclerosis: Comparison of conventional magnetic resonance imaging measures of atrophy and lesion burden. *Archives of Neurology*, *61*, 226–230. https://doi.org/10.1001/archneur.61.2.226

Bergsland, N., Zivadinov, R., Dwyer, M. G., Weinstock-Guttman, B., & Benedict, R. H. (2016). Localized atrophy of the thalamus and slowed cognitive processing speed in MS patients. *Multiple Sclerosis*, *22*, 1327–1336. https://doi.org/10.1177/1352458515616204

Bisecco, A., Rocca, M. A., Pagani, E., Mancini, L., Enzinger, C., Gallo, A., Vrenken, H., Stromillo, M. L., Copetti, M., Thomas, D. L., Fazekas, F., Tedeschi, G., Barkhof, F., Stefano, N. D., Filippi, M., & MAGNIMS Network. (2015). Connectivity-based parcellation of the thalamus in multiple sclerosis and its implications for cognitive impairment: A multicenter study. *Human Brain Mapping*, *36*, 2809–2825. https://doi.org/10.1002/hbm.22809

Bisecco, A., Stamenova, S., Caiazzo, G., d'Ambrosio, A., Sacco, R., Docimo, R., Esposito, S., Cirillo, M., Esposito, F., Bonavita, S., Tedeschi, G., & Gallo, A. (2018). Attention and processing speed performance in multiple sclerosis is mostly related to thalamic volume. *Brain Imaging and Behavior*, *12*, 20–28. https://doi.org/10.1007/s11682-016-9667-6

Breiman, L., & Spector, P. (1992). Submodel selection and evaluation in regression. The X-random case. *International Statistical Review/Revue Internationale de Statistique*, *60*, 291. https://doi.org/10.2307/1403680

Burggraaff, J., Liu, Y., Prieto, J. C., Simoes, J., de Sitter, A., Ruggieri, S., Brouwer, I., Lissenberg-Witte, B. I., Rocca, M. A., Valsasina, P., Ropele, S., Gasperini, C., Gallo, A., Pareto, D., Sastre-Garriga, J., Enzinger, C., Filippi, M., De Stefano, N., Ciccarelli, O., … MAGNIMS Study Group. (2020). Manual and automated tissue segmentation confirm the impact of thalamus atrophy on cognition in multiple sclerosis: A multicenter study. *NeuroImage: Clinical*, *29*, 102549. https://doi.org/10.1016/j.nicl.2020.102549

Buyukturkoglu, K., Zeng, D., Bharadwaj, S., Tozlu, C., Mormina, E., Igwe, K. C., Lee, S., Habeck, C., Brickman, A. M., Riley, C. S., De Jager, P. L., Sumowski, J. F., & Leavitt, V. M. (2021). Classifying multiple sclerosis patients on the basis of SDMT performance using machine learning. *Multiple Sclerosis*, *27*, 107–116. https://doi.org/10.1177/1352458520958362

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, *15*, 233–234. https://doi.org/10.1038/nmeth.4642

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Presented at the KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Cheng, A., Kessler, D., Mackinnon, R., Chang, T. P., Nadkarni, V. M., Hunt, E. A., Duval-Arnould, J., Lin, Y., Pusic, M., & Auerbach, M. (2017). Conducting multicenter research in healthcare simulation: Lessons learned from the INSPIRE network. *Advances in Simulation*, *2*, 6. https://doi.org/10.1186/s41077-017-0039-0

Chiaravalloti, N. D., & DeLuca, J. (2008). Cognitive impairment in multiple sclerosis. *Lancet Neurology*, *7*, 1139–1151. https://doi.org/10.1016/S1474-4422(08)70259-X

Chitnis, T., Guttmann, C. R., Zaitsev, A., Musallam, A., Weinstock-Guttman, B., Yeh, A., Rodriguez, M., Ness, J., Gorman, M. P., Healy, B. C., Kuntz, N., Chabas, D., Strober, J. B., Waubant, E.,

Krupp, L., Pelletier, D., Erickson, B., Bergsland, N., Zivadinov, R., & U.S. Network of Pediatric MS Centers of Excellence. (2013). Quantitative MRI analysis in children with multiple sclerosis: A multicenter feasibility pilot study. *BMC Neurology*, *13*, 173. https://doi.org/10.1186/1471-2377-13-173

Christodoulou, C., Krupp, L. B., Liang, Z., Huang, W., Melville, P., Roque, C., Scherl, W. F., Morgan, T., MacAllister, W. S., Li, L., Tudorica, L. A., Li, X., Roche, P., & Peyster, R. (2003). Cognitive performance and MR markers of cerebral injury in cognitively impaired MS patients. *Neurology*, *60*, 1793–1798. https://doi.org/10.1212/01.wnl.0000072264.75989.b8

Courchesne, E., Chisum, H. J., Townsend, J., Cowles, A., Covington, J., Egaas, B., Harwood, M., Hinds, S., & Press, G. A. (2000). Normal brain development and aging: Quantitative analysis at in vivo MR imaging in healthy volunteers. *Radiology*, *216*, 672–682. https://doi.org/10.1148/radiology.216.3.r00au37672

De Sonneville, L. M. J., Boringa, J. B., Reuling, I. E. W., Lazeron, R. H. C., Adèr, H. J., & Polman, C. H. (2002). Information processing characteristics in subtypes of multiple sclerosis. *Neuropsychologia*, *40*, 1751–1765. https://doi.org/10.1016/s0028-3932(02)00041-6

Diciotti, S., Ciulli, S., Mascalchi, M., Giannelli, M., & Toschi, N. (2013). The "peeking" effect in supervised feature selection on diffusion tensor imaging data. *AJNR. American Journal of Neuroradiology*, *34*, E107. https://doi.org/10.3174/ajnr.A3685

Dobson, R., & Giovannoni, G. (2019). Multiple sclerosis—A review. *European Journal of Neurology*, *26*, 27–40. https://doi.org/10.1111/ene.13819

Dolan, R. J. (2008). Neuroimaging of cognition: Past, present, and future. *Neuron*, *60*, 496–502. https://doi.org/10.1016/j.neuron.2008.10.038

Eshaghi, A., Marinescu, R. V., Young, A. L., Firth, N. C., Prados, F., Jorge Cardoso, M., Tur, C., De Angelis, F., Cawley, N., Brownlee, W. J., De Stefano, N., Laura Stromillo, M., Battaglini, M., Ruggieri, S., Gasperini, C., Filippi, M., Rocca, M. A., Rovira, A., Sastre-Garriga, J., … Ciccarelli, O. (2018). Progression of regional grey matter atrophy in multiple sclerosis. *Brain*, *141*, 1665–1677. https://doi.org/10.1093/brain/awy088

Filippi, M., Bar-Or, A., Piehl, F., Preziosa, P., Solari, A., Vukusic, S., & Rocca, M. A. (2018). Multiple sclerosis. *Nature Reviews Disease Primers*, *4*, 43. https://doi.org/10.1038/s41572-018-0041-4

Filippi, M., Preziosa, P., Langdon, D., Lassmann, H., Friedemann, P., Rovira, À., Schoonheim, M. M., Solari, A., Stankoff, B., & Rocca, M. A. (2020). Identifying progression in multiple sclerosis: New perspectives. *Annals of Neurology*, *88*, 438–452. https://doi.org/10.1002/ana.25808

Filippi, M., Tedeschi, G., Pantano, P., De Stefano, N., Zaratin, P., Rocca, M. A., & for the INNI Network. (2017). The Italian Neuroimaging Network Initiative (INNI): Enabling the use of advanced MRI techniques in patients with MS. *Neurological Sciences*, *38*, 1029–1038. https://doi.org/10.1007/s10072-017-2903-z

Foong, J., Rozewicz, L., Chong, W. K., Thompson, A. J., Miller, D. H., & Ron, M. A. (2000). A comparison of neuropsychological deficits in primary and secondary progressive multiple sclerosis. *Journal of Neurology*, *247*, 97–101. https://doi.org/10.1007/pl00007804

Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., Adams, P., Cooper, C., Fava, M., McGrath, P. J., McInnis, M., Phillips, M. L., Trivedi, M. H., Weissman, M. M., & Shinohara, R. T. (2018). Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, *167*, 104–120. https://doi.org/10.1016/j.neuroimage.2017.11.024

Fortin, J.-P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., Schultz, R. T., Verma, R., & Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, *161*, 149–170. https://doi.org/10.1016/j.neuroimage.2017.08.047

George, A., Kuzniecky, R., Rusinek, H., Pardoe, H. R., & for the Human Epilepsy Project Investigators. (2020). Standardized brain MRI acquisition protocols improve statistical power in multicenter quantitative morphometry studies. *Journal of Neuroimaging*, *30*, 126–133. https://doi.org/10.1111/jon.12673

Geurts, J. J. G., Bö, L., Roosendaal, S. D., Hazes, T., Daniëls, R., Barkhof, F., Witter, M. P., Huitinga, I., & van der Valk, P. (2007). Extensive hippocampal demyelination in multiple sclerosis. *Journal of Neuropathology and Experimental Neurology*, *66*, 819–827. https://doi.org/10.1097/nen.0b013e3181461f54

Goldstein, J. M. (2001). Normal sexual dimorphism of the adult human brain assessed by In vivo magnetic resonance imaging. *Cerebral Cortex*, *11*, 490–497. https://doi.org/10.1093/cercor/11.6.490

Hagens, M. H. J., Burggraaff, J., Kilsdonk, I. D., de Vos, M. L., Cawley, N., Sbardella, E., Andelova, M., Amann, M., Lieb, J. M., Pantano, P., Lissenberg-Witte, B. I., Killestein, J., Oreja-Guevara, C., Ciccarelli, O., Gasperini, C., Lukas, C., Wattjes, M. P., Barkhof, F., & MAGNIMS Study Group. (2018). Three-Tesla MRI does not improve the diagnosis of multiple sclerosis: A multicenter study. *Neurology*, *91*, e249–e257. https://doi.org/10.1212/WNL.0000000000005825

Hastie, T., Tibshirani, R., & Friedman, J. (2013). *The elements of statistical learning, springer series in statistics*. Springer. https://doi.org/10.1007/978-0-387-84858-7

Houtchens, M. K., Benedict, R. H. B., Killiany, R., Sharma, J., Jaisani, Z., Singh, B., Weinstock-Guttman, B., Guttmann, C. R. G., & Bakshi, R. (2007). Thalamic atrophy and cognition in multiple sclerosis. *Neurology*, *69*, 1213–1223. https://doi.org/10.1212/01.wnl.0000276992.17011.b5

Huggett, K. N., Gusic, M. E., Greenberg, R., & Ketterer, J. M. (2011). Twelve tips for conducting collaborative research in medical education. *Medical Teacher*, *33*, 713–718. https://doi.org/10.3109/0142159X.2010.547956

Huijbregts, S. C. J., Kalkers, N. F., de Sonneville, L. M. J., de Groot, V., & Polman, C. H. (2006). Cognitive impairment and decline in different MS subtypes. *Journal of the Neurological Sciences*, *245*, 187–194. https://doi.org/10.1016/j.jns.2005.07.018

Ion-Mărgineanu, A., Kocevar, G., Stamile, C., Sima, D. M., Durand-Dubief, F., Van Huffel, S., & Sappey-Marinier, D. (2017). Machine learning approach for classifying multiple sclerosis courses by combining clinical data with lesion loads and magnetic resonance metabolic features. *Frontiers in Neuroscience*, *11*, 398. https://doi.org/10.3389/fnins.2017.00398

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL. *NeuroImage*, *62*, 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015

Johnen, A., Landmeyer, N. C., Bürkner, P.-C., Wiendl, H., Meuth, S. G., & Holling, H. (2017). Distinct cognitive impairments in different disease courses of multiple sclerosis—A systematic review and meta-analysis. *Neuroscience and Biobehavioral Reviews*, *83*, 568–578. https://doi.org/10.1016/j.neubiorev.2017.09.005

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*, 118–127. https://doi.org/10.1093/biostatistics/kxj037

Kalb, R., Beier, M., Benedict, R. H., Charvet, L., Costello, K., Feinstein, A., Gingold, J., Goverover, Y., Halper, J., Harris, C., Kostich, L., Krupp, L., Lathi, E., LaRocca, N., Thrower, B., & DeLuca, J. (2018). Recommendations for cognitive screening and management in multiple sclerosis care. *Multiple Sclerosis*, *24*, 1665–1680. https://doi.org/10.1177/1352458518803785

Kern, K. C., Gold, S. M., Lee, B., Montag, M., Horsfall, J., O'Connor, M.-F., & Sicotte, N. L. (2015). Thalamic-hippocampal-prefrontal disruption in relapsing-remitting multiple sclerosis. *NeuroImage: Clinical*, *8*, 440–447. https://doi.org/10.1016/j.nicl.2014.12.015

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Kurtzke, J. F. (1983). Rating neurologic impairment in multiple sclerosis: An Expanded Disability Status Scale (EDSS). *Neurology*, *33*, 1444–1452. https://doi.org/10.1212/WNL.33.11.1444

Langdon, D., Amato, M., Boringa, J., Brochet, B., Foley, F., Fredrikson, S., Hämäläinen, P., Hartung, H.-P., Krupp, L., Penner, I., Reder, A., &

Benedict, R. (2012). Recommendations for a brief international cognitive assessment for multiple sclerosis (BICAMS). *Multiple Sclerosis*, *18*, 891–898. https://doi.org/10.1177/1352458511431076

Lemm, S., Blankertz, B., Dickhaus, T., & Müller, K.-R. (2011). Introduction to machine learning for brain imaging. *NeuroImage*, *56*, 387–399. https://doi.org/10.1016/j.neuroimage.2010.11.004

Localio, A. R., Berlin, J. A., Ten Have, T. R., & Kimmel, S. E. (2001). Adjustments for center in multicenter studies: An overview. *Annals of Internal Medicine*, *135*, 112–123. https://doi.org/10.7326/0003-4819-135-2-200107170-00012

Mato-Abad, V., Labiano-Fontcuberta, A., Rodríguez-Yáñez, S., García-Vázquez, R., Munteanu, C. R., Andrade-Garda, J., Domingo-Santos, A., Galán Sánchez-Seco, V., Aladro, Y., Martínez-Ginés, M. L., Ayuso, L., & Benito-León, J. (2019). Classification of radiologically isolated syndrome and clinically isolated syndrome with machine-learning techniques. *European Journal of Neurology*, *26*, 1000–1005. https://doi.org/10.1111/ene.13923

Meijer, K. A., van Geest, Q., Eijlers, A. J. C., Geurts, J. J. G., Schoonheim, M. M., & Hulst, H. E. (2018). Is impaired information processing speed a matter of structural or functional damage in MS? *NeuroImage: Clinical*, *20*, 844–850. https://doi.org/10.1016/j.nicl.2018.09.021

Miller, D. H., Thompson, A. J., & Filippi, M. (2003). Magnetic resonance studies of abnormalities in the normal appearing white matter and grey matter in multiple sclerosis. *Journal of Neurology*, *250*, 1407–1419. https://doi.org/10.1007/s00415-003-0243-9

Minagar, A., Barnett, M. H., Benedict, R. H. B., Pelletier, D., Pirko, I., Sahraian, M. A., Frohman, E., & Zivadinov, R. (2013). The thalamus and multiple sclerosis: Modern views on pathologic, imaging, and clinical aspects. *Neurology*, *80*, 210–219. https://doi.org/10.1212/WNL.0b013e31827b910b

Moll, N. M., Rietsch, A. M., Thomas, S., Ransohoff, A. J., Lee, J.-C., Fox, R., Chang, A., Ransohoff, R. M., & Fisher, E. (2011). Multiple sclerosis normal-appearing white matter: Pathology-imaging correlations. *Annals of Neurology*, *70*, 764–773. https://doi.org/10.1002/ana.22521

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: A guide for data scientists* (1st ed.). O'Reilly Media.

Neeb, H., & Schenk, J. (2019). Multivariate prediction of multiple sclerosis using robust quantitative MR-based image metrics. *Zeitschrift für Medizinische Physik*, *29*, 262–271. https://doi.org/10.1016/j.zemedi.2018.10.004

Nocentini, U., Bozzali, M., Spanò, B., Cercignani, M., Serra, L., Basile, B., Mannu, R., Caltagirone, C., & De Luca, J. (2014). Exploration of the relationships between regional grey matter atrophy and cognition in multiple sclerosis. *Brain Imaging and Behavior*, *8*, 378–386. https://doi.org/10.1007/s11682-012-9170-7

O'Sullivan, P. S., Stoddard, H. A., & Kalishman, S. (2010). Collaborative research in medical education: A discussion of theory and practice. *Medical Education*, *44*, 1175–1184. https://doi.org/10.1111/j.1365-2923.2010.03768.x

Parmenter, B. A., Weinstock-Guttman, B., Garg, N., Munschauer, F., & Benedict, R. H. (2007). Screening for cognitive impairment in multiple sclerosis using the symbol digit modalities test. *Multiple Sclerosis*, *13*, 52–57. https://doi.org/10.1177/1352458506070750

Patenaude, B., Smith, S. M., Kennedy, D. N., & Jenkinson, M. (2011). A Bayesian model of shape and appearance for subcortical brain segmentation. *NeuroImage*, *56*, 907–922. https://doi.org/10.1016/j.neuroimage.2011.02.046

Paulus, M. P., Kuplicki, R., & Yeh, H.-W. (2019). Machine learning and brain imaging: Opportunities and challenges. *Trends in Neurosciences*, *42*, 659–661. https://doi.org/10.1016/j.tins.2019.07.007

Payne, S., Seymour, J., Molassiotis, A., Froggatt, K., Grande, G., Lloyd-Williams, M., Foster, C., Wilson, R., Rolls, L., Todd, C., & Addington-Hall, J. (2011). Benefits and challenges of collaborative research: Lessons from supportive and palliative care. *BMJ Supportive & Palliative Care*, *1*, 5–11. https://doi.org/10.1136/bmjspcare-2011-000018

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pomponio, R., Erus, G., Habes, M., Doshi, J., Srinivasan, D., Mamourian, E., Bashyam, V., Nasrallah, I. M., Satterthwaite, T. D., Fan, Y., Launer, L. J., Masters, C. L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S. C., Fripp, J., Koutsouleris, N., Wolf, D. H., … Davatzikos, C. (2020). Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, *208*, 116450. https://doi.org/10.1016/j.neuroimage.2019.116450

Popescu, V., Klaver, R., Voorn, P., Galis-de Graaf, Y., Knol, D. L., Twisk, J. W. R., Versteeg, A., Schenk, G. J., Van der Valk, P., Barkhof, F., De Vries, H. E., Vrenken, H., & Geurts, J. J. G. (2015). What drives MRI-measured cortical atrophy in multiple sclerosis? *Multiple Sclerosis*, *21*, 1280–1290. https://doi.org/10.1177/1352458514562440

Preziosa, P., Rocca, M. A., Pagani, E., Stromillo, M. L., Enzinger, C., Gallo, A., Hulst, H. E., Atzori, M., Pareto, D., Riccitelli, G. C., Copetti, M., De Stefano, N., Fazekas, F., Bisecco, A., Barkhof, F., Yousry, T. A., Arévalo, M. J., Filippi, M., & MAGNIMS Study Group. (2016). Structural MRI correlates of cognitive impairment in patients with multiple sclerosis: A multicenter study. *Human Brain Mapping*, *37*, 1627–1644. https://doi.org/10.1002/hbm.23125

Radua, J., Vieta, E., Shinohara, R., Kochunov, P., Quidé, Y., Green, M. J., Weickert, C. S., Weickert, T., Bruggemann, J., Kircher, T., Nenadić, I., Cairns, M. J., Seal, M., Schall, U., Henskens, F., Fullerton, J. M., Mowry, B., Pantelis, C., Lenroot, R., … Pineda-Zapata, J. (2020). Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. *NeuroImage*, *218*, 116956. https://doi.org/10.1016/j.neuroimage.2020.116956

Rao, S. M. (1991). *A manual for the brief, repeatable battery of neuropsychological tests in multiple sclerosis*. National Multiple Sclerosis Society.

Rao, S. M., Glatt, S., Hammeke, T. A., McQuillen, M. P., Khatri, B. O., Rhodes, A. M., & Pollard, S. (1985). Chronic progressive multiple sclerosis. Relationship between cerebral ventricular size and neuropsychological impairment. *Archives of Neurology*, *42*, 678–682. https://doi.org/10.1001/archneur.1985.04060070068018

Rao, S. M., Leo, G. J., Haughton, V. M., St Aubin-Faubert, P., & Bernardin, L. (1989). Correlation of magnetic resonance imaging with neuropsychological testing in multiple sclerosis. *Neurology*, *39*, 161–166. https://doi.org/10.1212/wnl.39.2.161

Rao, S. M., Martin, A. L., Huelin, R., Wissinger, E., Khankhel, Z., Kim, E., & Fahrbach, K. (2014). Correlations between MRI and information processing speed in MS: A meta-analysis. *Multiple Sclerosis International*, *2014*, 1–9. https://doi.org/10.1155/2014/975803

Rocca, M. A., Amato, M. P., De Stefano, N., Enzinger, C., Geurts, J. J., Penner, I.-K., Rovira, A., Sumowski, J. F., Valsasina, P., & Filippi, M. (2015). Clinical and imaging assessment of cognitive dysfunction in multiple sclerosis. *The Lancet Neurology*, *14*, 302–317. https://doi.org/10.1016/S1474-4422(14)70250-9

Rocca, M. A., Battaglini, M., Benedict, R. H. B., De Stefano, N., Geurts, J. J. G., Henry, R. G., Horsfield, M. A., Jenkinson, M., Pagani, E., & Filippi, M. (2017). Brain MRI atrophy quantification in MS: From methods to clinical application. *Neurology*, *88*, 403–413. https://doi.org/10.1212/WNL.0000000000003542

Ruano, L., Portaccio, E., Goretti, B., Niccolai, C., Severo, M., Patti, F., Cilia, S., Gallo, P., Grossi, P., Ghezzi, A., Roscio, M., Mattioli, F., Stampatori, C., Trojano, M., Viterbo, R. G., & Amato, M. P. (2017). Age and disability drive cognitive impairment in multiple sclerosis across disease subtypes. *Multiple Sclerosis*, *23*, 1258–1267. https://doi.org/10.1177/1352458516674367

Rusz, J., Vaneckova, M., Benova, B., Tykalova, T., Novotny, M., Ruzickova, H., Uher, T., Andelova, M., Novotna, K., Friedova, L., Motyl, J., Kucerova, K., Krasensky, J., & Horakova, D. (2019). Brain

volumetric correlates of dysarthria in multiple sclerosis. *Brain and Language*, *194*, 58–64. https://doi.org/10.1016/j.bandl.2019.04.009

Sanfilipo, M. P., Benedict, R. H. B., Weinstock-Guttman, B., & Bakshi, R. (2006). Gray and white matter brain atrophy and neuropsychological impairment in multiple sclerosis. *Neurology*, *66*, 685–692. https://doi.org/10.1212/01.wnl.0000201238.93586.d9

Schwartz, A., Young, R., Hicks, P. J., & Learn, A. (2016). Medical education practice-based research networks: Facilitating collaborative research. *Medical Teacher*, *38*, 64–74. https://doi.org/10.3109/0142159X.2014.970991

Shinohara, R. T., Oh, J., Nair, G., Calabresi, P. A., Davatzikos, C., Doshi, J., Henry, R. G., Kim, G., Linn, K. A., Papinutto, N., Pelletier, D., Pham, D. L., Reich, D. S., Rooney, W., Roy, S., Stern, W., Tummala, S., Yousuf, F., Zhu, A., … NAIMS Cooperative. (2017). Volumetric analysis from a harmonized multisite brain MRI study of a single subject with multiple sclerosis. *AJNR. American Journal of Neuroradiology*, *38*, 1501–1509. https://doi.org/10.3174/ajnr.A5254

Sicotte, N. L., Kern, K. C., Giesser, B. S., Arshanapalli, A., Schultz, A., Montag, M., Wang, H., & Bookheimer, S. Y. (2008). Regional hippocampal atrophy in multiple sclerosis. *Brain*, *131*, 1134–1141. https://doi.org/10.1093/brain/awn030

Smith, A. (1982). *Symbol digit modalities test: Manual*. Western Psychological Services.

Stankiewicz, J. M., Glanz, B. I., Healy, B. C., Arora, A., Neema, M., Benedict, R. H. B., Guss, Z. D., Tauhid, S., Buckle, G. J., Houtchens, M. K., Khoury, S. J., Weiner, H. L., Guttmann, C. R. G., & Bakshi, R. (2011). Brain MRI lesion load at 1.5T and 3T versus clinical status in multiple sclerosis. *Journal of Neuroimaging*, *21*, e50–e56. https://doi.org/10.1111/j.1552-6569.2009.00449.x

Storelli, L., Rocca, M. A., Pantano, P., Pagani, E., De Stefano, N., Tedeschi, G., Zaratin, P., Filippi, M., & for the INNI Network. (2019). MRI quality control for the Italian Neuroimaging Network Initiative: Moving towards big data in multiple sclerosis. *Journal of Neurology*, *266*, 2848–2858. https://doi.org/10.1007/s00415-019-09509-4

Tommasin, S., Cocozza, S., Taloni, A., Giannì, C., Petsas, N., Pontillo, G., Petracca, M., Ruggieri, S., De Giglio, L., Pozzilli, C., Brunetti, A., & Pantano, P. (2021). Machine learning classifier to identify clinical and radiological features relevant to disability progression in multiple sclerosis. *Journal of Neurology*, *268*, 4834–4845. https://doi.org/10.1007/s00415-021-10605-7

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging*, *29*, 1310–1320. https://doi.org/10.1109/TMI.2010.2046908

Van Schependom, J., D'Hooghe, M. B., Cleynhens, K., D'Hooge, M., Haelewyck, M. C., De Keyser, J., & Nagels, G. (2014). The symbol digit modalities test as sentinel test for cognitive impairment in multiple sclerosis. *European Journal of Neurology*, *21*, 1219–e72. https://doi.org/10.1111/ene.12463

Van Schependom, J., & Nagels, G. (2017). Targeting cognitive impairment in multiple sclerosis-the road toward an imaging-based biomarker. *Frontiers in Neuroscience*, *11*, 380. https://doi.org/10.3389/fnins.2017.00380

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*, 91. https://doi.org/10.1186/1471-2105-7-91

Witten, I. H., & Frank, E. (2016). *Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann Publishers.

Wottschel, V., Alexander, D. C., Kwok, P. P., Chard, D. T., Stromillo, M. L., De Stefano, N., Thompson, A. J., Miller, D. H., & Ciccarelli, O. (2015). Predicting outcome in clinically isolated syndrome using machine learning. *NeuroImage: Clinical*, *7*, 281–287. https://doi.org/10.1016/j.nicl.2014.11.021

Wottschel, V., Chard, D. T., Enzinger, C., Filippi, M., Frederiksen, J. L., Gasperini, C., Giorgio, A., Rocca, M. A., Rovira, A., De Stefano, N., Tintoré, M., Alexander, D. C., Barkhof, F., & Ciccarelli, O. (2019). SVM recursive feature elimination analyses of structural brain MRI predicts near-term relapses in patients with clinically isolated syndromes suggestive of multiple sclerosis. *NeuroImage: Clinical*, *24*, 102011. https://doi.org/10.1016/j.nicl.2019.102011

Yan, L., Zhang, H.-T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jing, L., Zhang, M., Huang, X., Xiao, Y., Cao, H., Chen, Y., Ren, T., Wang, F., Xiao, Y., Huang, S., Tan, X., … Yuan, Y. (2020). An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence*, *2*, 283–288. https://doi.org/10.1038/s42256-020-0180-7

Yu, A. C., Mohajer, B., & Eng, J. (2022). External validation of deep learning algorithms for radiologic diagnosis: A systematic review. *Radiology: Artificial Intelligence*, *4*, e210064. https://doi.org/10.1148/ryai.210064

Zhang, H., Alberts, E., Pongratz, V., Mühlau, M., Zimmer, C., Wiestler, B., & Eichinger, P. (2019). Predicting conversion from clinically isolated syndrome to multiple sclerosis—An imaging-based machine learning approach. *NeuroImage: Clinical*, *21*, 101593. https://doi.org/10.1016/j.nicl.2018.11.003

Zivadinov, R., Sepcic, J., Nasuelli, D., De Masi, R., Bragadin, L. M., Tommasi, M. A., Zambito-Marsala, S., Moretti, R., Bratina, A., Ukmar, M., Pozzi-Mucelli, R. S., Grop, A., Cazzato, G., & Zorzon, M. (2001). A longitudinal study of brain atrophy and cognitive disturbances in the early phase of relapsing-remitting multiple sclerosis. *Journal of Neurology, Neurosurgery, and Psychiatry*, *70*, 773–780. https://doi.org/10.1136/jnnp.70.6.773

Zurita, M., Montalba, C., Labbé, T., Cruz, J. P., Dalboni da Rocha, J., Tejos, C., Ciampi, E., Cárcamo, C., Sitaram, R., & Uribe, S. (2018). Characterization of relapsing-remitting multiple sclerosis patients using support vector machine classifications of functional and diffusion MRI data. *NeuroImage: Clinical*, *20*, 724–730. https://doi.org/10.1016/j.nicl.2018.09.002

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Marzi, C., d'Ambrosio, A., Diciotti, S., Bisecco, A., Altieri, M., Filippi, M., Rocca, M. A., Storelli, L., Pantano, P., Tommasin, S., Cortese, R., De Stefano, N., Tedeschi, G., Gallo, A., & the INNI Network (2023). Prediction of the information processing speed performance in multiple sclerosis using a machine learning approach in a large multicenter magnetic resonance imaging data set. *Human Brain Mapping*, *44*(1), 186–202. https://doi.org/10.1002/hbm.26106