

ORIGINAL ARTICLE OPEN ACCESS

Digital Health Research in Clinical Neurosciences

Explainable Artificial Intelligence to Predict Neurocognitive Disorder Progression in Multiple Sclerosis Using MRI and Clinical Data

Loredana Storelli¹ | Damiano Mistri¹ | Alice Mastropasqua^{1,2} | Marta Grosselle¹ | Paolo Preziosa^{1,2,3} | Lucrezia Rossi¹ | Massimo Filippi^{1,2,3,4,5} | Maria A. Rocca^{1,2,3} ¹Neuroimaging Research Unit, Division of Neuroscience, IRCCS San Raffaele Scientific Institute, Milan, Italy | ²Vita-Salute San Raffaele University, Milan, Italy | ³Neurology Unit, IRCCS San Raffaele Scientific Institute, Milan, Italy | ⁴Neurorehabilitation Unit, IRCCS San Raffaele Scientific Institute, Milan, Italy | ⁵Neurophysiology Service, IRCCS San Raffaele Scientific Institute, Milan, Italy**Correspondence:** Maria A. Rocca (rocca.mara@hsr.it)**Received:** 16 July 2025 | **Revised:** 12 November 2025 | **Accepted:** 3 March 2026**Keywords:** cognitive dysfunction | deep learning | explainable artificial intelligence | magnetic resonance imaging | multiple sclerosis

ABSTRACT

Background: Cognitive impairment is common in multiple sclerosis (MS), yet the application of diagnostic frameworks of Neurocognitive Disorders (NCDs) is limited. Additionally, the integration of multimodal data for predicting cognitive outcomes using artificial intelligence (AI) remains underexplored. This study aimed to characterize NCDs in MS and predict cognitive worsening using an explainable deep learning model trained on MRI and clinical data.

Methods: Two-hundred twenty-four MS patients and 115 healthy controls (HC) underwent 3.0T MRI and clinical assessment at baseline. MS patients also completed neuropsychological testing, including estimation of z-cognitive reserve, at baseline and after a median follow-up of 3.4 (interquartile range = [2.0; 6.1]) years. MS patients were classified as Mild or Major NCD according to the Diagnostic and Statistical Manual of Mental Disorders criteria at baseline, and as “stable” or “worsened” based on cognitive changes at follow-up. A deep learning model was trained on baseline T1-weighted MRI, demographic, clinical, and brain volumetric data to predict cognitive decline, with explainability methods used to interpret the model's decisions.

Results: At baseline, 4% of patients had Mild and 11% Major NCD. At follow-up, 12% showed cognitive decline. The deep learning model predicted follow-up cognitive status with 90% accuracy. Explainability models identified the most relevant predictors, in order of importance: cortical gray matter volume, age, thalamic and hippocampal volumes, T2 lesion volume, and z-cognitive reserve.

Conclusions: The proposed multimodal AI approach demonstrated robust performance and highlighted relevant brain regions associated with cognitive worsening, underscoring its potential for personalized cognitive assessment and monitoring in MS.

1 | Introduction

Cognition deficits affect a large proportion of patients with multiple sclerosis (MS), typically involving learning and memory, information processing speed, and executive functions [1]. Although a few studies have referred to cognitive deficits in MS patients as “subcortical dementia”, the term “dementia” has not permeated

the field of MS, largely because it is commonly associated with aging or used as a synonym for Alzheimer's disease [2]. One study [3] explored the prevalence of dementia in MS and found that 22% of patients met the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision (DSM-IV-TR) [4] criteria for its diagnosis. The concept of dementia has evolved over time to encompass numerous neurological and systemic disorders,

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.

© 2026 The Author(s). *European Journal of Neurology* published by John Wiley & Sons Ltd on behalf of European Academy of Neurology.

including frontotemporal degeneration, Lewy body disease, vascular disease, traumatic brain injury and HIV infection [5]. Moreover, diagnostic criteria have been developed to distinguish patients who exhibit cognitive dysfunction that does not significantly affect their daily functioning (i.e., mild cognitive impairment, MCI) from those who have lost functional ability [2]. The DSM-5 [6] has recently introduced the terms of Mild and Major Neurocognitive Disorders (NCDs), replacing the nosographic categories of MCI and dementia [2]. To date, only one study has investigated the prevalence of NCDs in MS, reporting that 21% of MS patients met the criteria for Major NCD and 67% for Mild NCD [7]. However, the clinical and MRI profiles of MS patients who meet the criteria for NCD have not yet been explored.

Artificial intelligence (AI) algorithms can capture complex patterns in data and, in recent years, the number of studies employing these algorithms on MRI-derived measures to classify cognitive impairment in MS has rapidly increased [8–10]. These studies have overall demonstrated that AI is a powerful tool for identifying MS patients with cognitive impairment based on T2 hyperintense white matter (WM) lesion volume (LV) [8, 10], cortical gray matter (GM) [10], thalamic [9, 10] and hippocampal volume [10], with an accuracy ranging from 69% to 91% [8–10]. Aside from machine learning techniques, only a previous study applied a deep learning approach directly to raw MRI data to predict cognitive worsening after 2 years of follow-up, potentially providing valuable insights to support clinicians in tailoring cognitive training interventions [11]. Despite promising results, this study classified MS patients based on a single cognitive test, the Symbol Digit Modalities Test (SDMT), and predominantly relied on a single data modality (T1-weighted and T2-weighted MRI scans) [11], without integrating other information (e.g., demographic and clinical data), thereby providing only a partial knowledge of the factors associated with cognitive deterioration. Additionally, given their complexity, these systems have become “black boxes”, raising concerns about the transparency of their decision-making processes [12]. This lack of transparency has posed a significant barrier to their adoption in healthcare. As a result, there has been increasing interest in explainable AI, which aims to develop methods that make models more interpretable and understandable [13–16].

Against this background, the aims of this study were twofold. First, we assessed the prevalence of Mild and Major NCDs in a relatively large cohort of MS patients using a comprehensive neuropsychological evaluation focused on daily functioning and characterized their demographic and clinical data, and MRI measures of brain tissue volumes. Second, we developed a multimodal deep learning model integrating T1-weighted MRI, demographic, clinical, and brain volumetric measures to predict future NCD development and identify key predictive factors using explainable AI.

2 | Methods

2.1 | Participants

Approval was obtained from the local institutional ethical standards committee on human experimentation (Protocol number 2015–33). All participants provided written informed consent according to the Declaration of Helsinki.

A cohort of 224 MS patients [17] and 115 healthy controls (HCs) was retrospectively selected from the data repository of Neuroimaging Research Unit, IRCCS San Raffaele Scientific Institute (Milan, Italy). Inclusion criteria were: (1) age ≥ 18 years, (2) Italian native speakers and (3) no systemic, psychiatric or neurologic disease (other than MS). In addition, MS patients were selected if they had undergone two clinical and neuropsychological evaluations separated by a minimum follow-up of 1 year, were relapse-free, steroid-free, and had been on a stable disease-modifying treatment (DMT) for at least 3 months prior to clinical, cognitive and MRI assessments.

2.2 | Clinical and Neuropsychological Assessment

At baseline, all MS patients underwent neurological examination with a rating of Expanded Disability Status Scale (EDSS) score [18], recording of current DMT, and definition of clinical phenotype (i.e., relapsing–remitting [RR] or progressive [P] MS). The Montgomery-Asberg Depression Rating Scale (MADRS) [19] was administered to assess depressive symptoms, the Modified Fatigue Impact Scale (MFIS) [20] was used to evaluate fatigue and the Multiple Sclerosis Quality of Life-54 (MSQOL-54) [21] was employed to assess health-related quality of life. Premorbid intelligence quotient (IQ) was estimated using the Italian version of the National Adult Reading Test [22]. For each patient, premorbid IQ and years of education were converted into z-score based on the distributions in the patient sample and then were averaged to obtain a cognitive reserve z-score [23].

At baseline and follow-up, MS patients were administered the Brief Repeatable Battery of Neuropsychological Tests [24, 25] and the computerized version of the Wisconsin Card Sorting Test (WCST) [26] to assess: learning and memory (Selective Reminding Test and 10/36 Spatial Recall Test) [24, 25], complex attention (SDMT and Paced Auditory Serial Addition Test) [24, 25] and executive functions (Word List Generation [24, 25], WCST) [26]. Raw scores of cognitive tests were standardized into z-scores based on normative data [24–26]. Mild impairment in a cognitive domain was defined as two or more z-scores below -1.5 , while severe impairment required at least two z-scores below -2.0 [27]. An experienced neuropsychologist conducted a clinical interview to evaluate the impact of cognitive dysfunction on complex instrumental activities of daily living (i.e., managing medications and paying bills), and then determined whether each MS patient met the criteria for Mild NCD or Major NCD according to the DSM-5-TR criteria [5].

At follow-up, MS patients were classified into two groups based on changes in their diagnostic status: stable (i.e., patients who remained cognitively preserved at both timepoints, or maintained a diagnosis of Mild or Major NCD) and worsened (i.e., patients who were cognitively preserved at baseline but developed NCD at follow-up, or who progressed from Mild to Major NCD).

2.3 | MRI Acquisition

Brain scans were acquired from all subjects. The MRI protocol was performed using two 3.0T scanners (Scanner 1: Achieva,

Scanner 2: Ingenia; Philips Medical Systems, Eindhoven, The Netherlands). Detailed MRI acquisition parameters are provided in the [Supporting Information](#).

2.4 | Image Pre-Processing

T2-hyperintense WM LV, as well as normalized volumes of the brain (NBV), cortical GM (NCGMV), WM (NWMV), bilateral hippocampus and thalamus were calculated as described in [Supporting Information](#). T1-weighted images for each patient were then registered into the MNI 152 atlas space (<https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases>) using an affine transformation (FLIRT, FSL toolbox, version 5.0.5) to be used as input for the AI.

2.5 | Model Configuration and Input Data

As shown in Figure 1, the proposed deep learning algorithm for the prognosis of the cognitive status at follow-up was built on a supervised neural network with a “hybrid architecture”, consisting of two components: (i) a 3D convolutional neural network (3D-CNN) on T1-weighted images, and (ii) the integration of MRI-derived brain tissue volumes along with non-imaging data, such as demographic and clinical information, at the fully connected layer. T1-weighted images were selected as the sole imaging input due to their superior anatomical details and relevance to structural features, while LV from T2-weighted sequences was incorporated as tabular data. This choice reduced model complexity and improved interpretability. Clinical and demographic data presumably contain additional information that can contribute with MRI features to the algorithm decision. The demographic variables included in the model were age and sex, while the clinical information comprised EDSS, disease duration (years), follow-up

duration (years), treatment classification (no treatment/moderate efficacy/high efficacy), and a cognitive reserve z-score. According to prior literature [28], we found that incorporating MRI-derived volumetric measures could enhance the model's performance. Therefore, as tabular data, we also included baseline measures that exhibited slightly distinct value distributions between stable and worsened patients at follow-up; that is T2-hyperintense LV, as well as NWMV, NCGMV, and volumes of the thalamus and hippocampus.

To ensure robust model generalization, a stratified 5-fold cross-validation was employed. At each fold, 60% of the dataset is assigned for model training, 20% for model validation, and the remaining 20% for model testing.

2.6 | AI Neural Network Implementation

The method was implemented in Python 3.7 using PyTorch 1.0.0 and is based on a modified ResNet-10 architecture pre-trained on MedicalNet [29]. This 10-layer network includes $3 \times 3 \times 3$ convolutional kernels, batch normalization, residual skip connections, and two max-pooling layers (3×3 and 6×6 kernels) for feature reduction and spatial invariance. Extracted features were standardized via z-score normalization to improve training stability. Tabular data were standardized and concatenated with the flattened imaging features to form a combined input vector. This vector fed into a Multilayer Perceptron (MLP) classifier consisting of four fully connected layers with Tanh activations, enabling the model to handle both positive and negative values effectively. The MLP balanced the contribution of imaging and clinical features, unlike traditional image-only classifiers, ensuring integrated learning from both modalities. Additionally, Monte Carlo Dropout (MC-Dropout) was employed during both training and inference to estimate model uncertainty via multiple stochastic

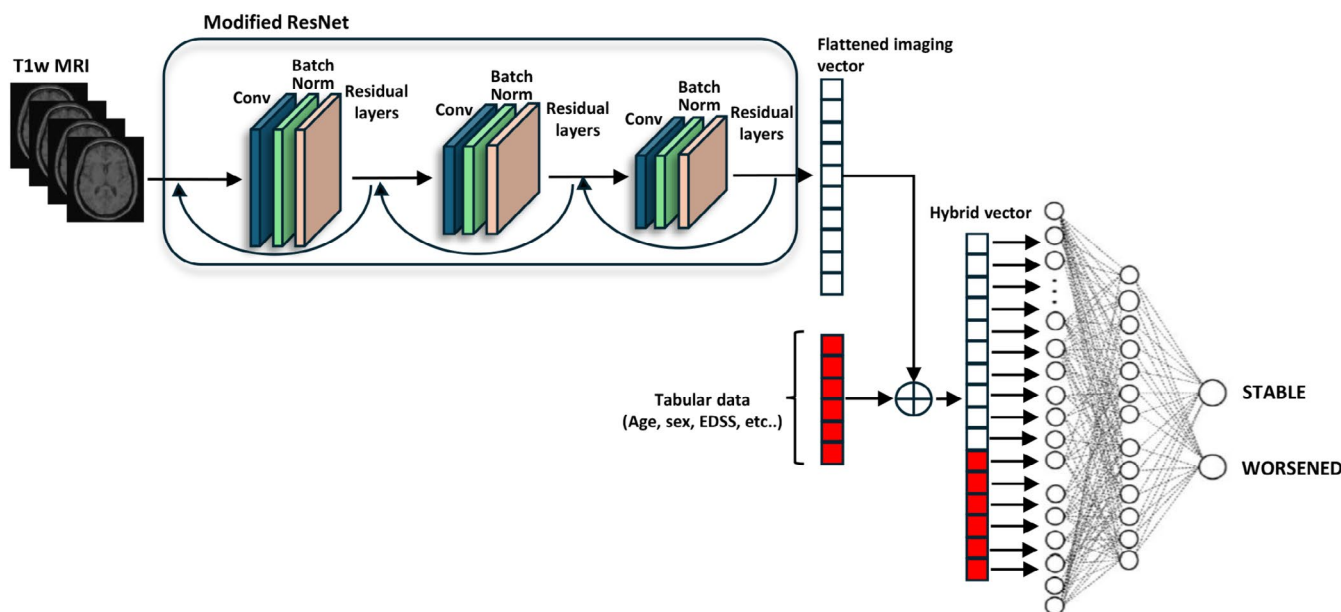


FIGURE 1 | A schematic representation of the proposed deep neural network architecture used to predict cognitively stable and cognitively worsened multiple sclerosis patients at follow-up. Batch Norm, batch normalization; Conv, convolution; EDSS, Expanded Disability Status Scale.

forward passes [30]. Implementation details are described in the [Supporting Information](#).

2.7 | Explainability Models

Explainability techniques were integrated into the AI predictive framework to interpret the neural network's decision-making process, with modality-specific approaches applied to imaging and tabular data, as shown in Figure 2. For imaging data, Gradient-weighted Class Activation Mapping (Grad-CAM) was used to identify anatomically relevant regions in T1-weighted MRI volumes [31]. After a forward pass and gradient-based backpropagation, class-specific weighted activations were computed and combined into Grad-CAM heatmaps. These were upsampled via trilinear interpolation and overlaid onto original MRI slices to generate full-volume visualizations. To identify class-level patterns, individual Grad-CAM maps were averaged within each diagnostic group (cognitively stable vs. worsened), and a voxel-wise differential attention map was computed. Anatomical localization of key regions was performed using the Automated Anatomical Labeling (AAL) atlas [32], with brain regions ranked according to the magnitude of importance and spatial extent (i.e., number of voxels).

For tabular data, Permutation Feature Importance (PFI) was employed [33]. The method involved randomly permuting individual features across 5000 iterations and measuring the corresponding decrease in model performance (R^2 score). The

averaged importance scores from all iterations were used to rank the features by their contribution to the model's predictions, ensuring robustness and interpretability.

2.8 | Statistical Analysis

Differences of demographic, clinical and neuropsychological variables between MS patients and HC, as well as MS patients according to the presence of cognitive impairment were assessed using appropriate statistical tests, including the Chi-square, Fisher's exact, two-sample t , and Mann-Whitney U tests. Between-group comparisons of brain volumetric measures were analyzed using age-, sex-, and scanner-adjusted linear models. Binary logistic regression evaluated associations between cognitive stability at follow-up and key tabular predictors identified via PFI.

Evaluation metrics included accuracy, F1-score, precision, recall, and the area under the curve (AUC). To address limitations of accuracy in imbalanced datasets, the F1-score (harmonic mean of precision and recall) was employed, balancing false positives and false negatives. Precision measures the proportion of true positives among positive predictions ($TP/(TP + FP)$), while recall quantifies the proportion of actual positives correctly identified ($TP/(TP + FN)$). The AUC assessed the trade-off between true positive rate (recall) and false positive rate (1-specificity) across thresholds. All statistical analyses were conducted using R version 4.2.2.

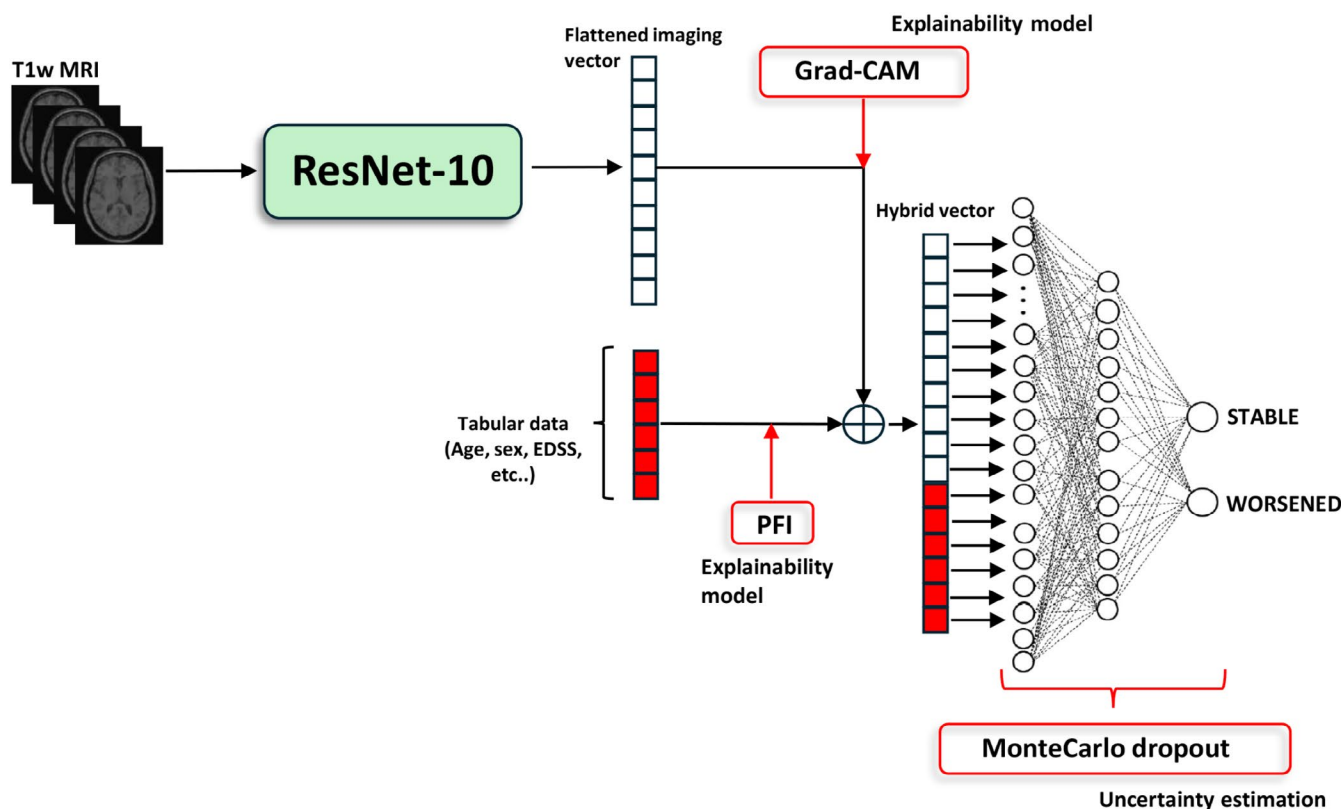


FIGURE 2 | A schematic representation of the integration of explainability models from imaging (Grad-CAM) and tabular data (Permutation Feature Importance), along with the uncertainty estimation model (MonteCarlo dropout), into the proposed deep neural network architecture. EDSS, Expanded Disability Status Scale; PFI, Permutation Feature Importance.

TABLE 1 | Main demographic, clinical, and MRI characteristics of healthy controls and MS patients grouped according to the presence of NCDs.

	HC (n = 115)	MS (n = 224)	MS vs. HC p	CP (n = 190)	Mild NCD (n = 10)	Major NCD (n = 24)	CP vs. mild NCD vs. major NCD p
Age, median (IQR), y	42.1 (30.1; 53.7)	40.8 (32.5–46.4)	0.42	39.4 (31.2; 45.9)	41.5 (39.2; 53.4)	43.8 (37.6; 54.5)	0.049 ^a
Female, N (%)	64 (56)	130 (58)	0.76	108 (57)	6 (60)	16 (67)	≥ 0.39
Education, median (IQR), y	—	13.0 (13.0; 6.0)	—	13.0 (13.0; 16.0)	13.0 (10.0; 13.0)	11.0 (8.0; 13.0)	0.008 ^a
FU duration, median (IQR), y	—	3.4 (2.0; 6.1)	—	3.5 (2.0; 6.4)	2.0 (1.3; 3.0)	3.2 (1.7; 5.5)	0.028 ^b
Baseline disease duration, median (IQR), y	—	9.3 (3.9; 16.4)	—	9.0 (3.7; 5.3)	7.7 (1.3; 20.0)	16.0 (9.4; 22.7)	0.010 ^a
Baseline EDSS, median (IQR)	—	2.0 (1.5; 4.0)	—	2.0 (1.5; 3.0)	4.0 (2.0; 5.0)	4.0 (3.0; 6.0)	< 0.001 ^a
FU EDSS, median (IQR)	—	2.0 (1.0; 4.5)	—	1.5 (1.0; 3.5)	5.5 (2.0; 6.0)	4.5 (3.5; 6.0)	0.035 ^b
MS phenotype: RRMS/PMS, N (%)	—	173 (77)/51 (23)	—	158 (83)/32 (17)	5 (50)/5 (50)	10 (42)/14 (58)	≤ 0.021 ^{a,b}
MS phenotype change: change, N (%)	—	RRMS to PMS, 16 (7)	—	RRMS to PMS, 15 (8)	RRMS to PMS, 0 (0)	RRMS to PMS, 1 (4)	≥ 0.70
Baseline DMTs ^d : No treatment/moderate efficacy/high efficacy, N (%)	—	36 (16)/114 (51)/74 (33)	—	30 (16)/97 (51)/63 (33)	3 (30)/3 (30)/4 (40)	3 (13)/14 (58)/7 (29)	≥ 0.28
DMT ^d change: yes/no, N (%)	—	93 (41)/131 (59)	—	80 (42)/110 (58)	3 (30)/7 (70)	10 (42)/14 (58)	≥ 0.53
ARR at FU, mean (range)	—	0.1 (0.0; 0.8)	—	0.1 (0.0; 0.8)	0.0 (0.0; 0.0)	0.1 (0.0; 0.3)	≥ 0.69
FU CP/Mild NCD/Major NCD, N (%)	—	167 (74)/17 (8)/40 (18)	—	167 (88)/11 (6)/12 (6)	0 (0)/6 (60)/4 (40)	0 (0)/0 (0)/24 (100)	< 0.001 ^{a,b,c}
z cognitive reserve, median (IQR)	—	0.0 (−0.4; 0.5)	—	0.1 (−0.3; 0.5)	−0.2 (−1.2; 0.3)	−0.8 (−1.7; 0.3)	≤ 0.001 ^b
MADRS, median (IQR)	—	8.0 (4.0; 15.0)	—	8.0 (4.0; 15.0)	7.0 (4.0; 18.0)	10.0 (6.0; 15.0)	0.99

(Continues)

TABLE 1 | (Continued)

	HC (n = 115)	MS (n = 224)	MS vs. HC p	CP (n = 190)	Mild NCD (n = 10)	Major NCD (n = 24)	CP vs. mild NCD vs. major NCD p
MFIS, median (IQR)	—	30.0 (18.2; 44.0)	—	29.0 (16.0; 43.0)	23.0 (17.0; 37.0)	44.0 (31.0; 54.0)	0.015^a
MSQOL-54 PHCS, median (IQR)	—	70.3 (59.1; 85.2)	—	77.4 (62.1; 86.1)	74.0 (46.7; 83.7)	60.6 (51.2; 72.2)	0.037^a
MSQOL-54 MHCS, median (IQR)	—	76.8 (60.5; 86.6)	—	77.0 (60.6; 86.8)	85.2 (62.0; 92.2)	72.6 (59.0; 80.0)	0.41
Subjects scanned with: Scanner 1/Scanner 2, N (%)	66 (57)/49 (43)	150 (67)/74 (33)	0.11	124 (65)/66 (35)	6 (60)/4 (40)	20 (83)/4 (17)	≥ 0.10
Baseline T2 LV, median (IQR), mL	0.0 (0.0; 0.2)	4.2 (1.5; 9.6)	<0.001	3.6 (1.3; 8.1)	12.5 (7.8; 15.9)	10.2 (4.3; 20.6)	≤ 0.046^{a,b}
Baseline NBV, median (IQR), mL	1582 (1555; 1620)	1544 (1494; 1600)	<0.001	1558 (1508; 1605)	1523 (1456; 1568)	1476 (1421; 1534)	<0.001^a
Baseline NCGMV, median (IQR), mL	636 (598; 667)	615 (534; 651)	<0.001	621 (576; 658)	606 (537; 643)	551 (520; 585)	<0.001^a
Baseline NWMV, median (IQR), mL	783 (694; 818)	756 (696; 799)	<0.001	757 (696; 801)	750 (667; 782)	755 (729; 796)	0.008^a
Baseline normalized Thal V, median (IQR), mL	22.4 (21.5; 23.2)	21.0 (19.2; 22.2)	<0.001	21.3 (20.0; 22.5)	18.7 (17.6; 19.5)	18.4 (16.8; 19.9)	<0.003^{a,b}
Baseline normalized Hipp V, median (IQR), mL	10.6 (10.1; 11.2)	10.2 (9.4; 11.0)	<0.001	10.3 (9.5; 11.1)	9.6 (9.0; 10.2)	8.8 (7.3; 10.2)	<0.036^{a,b,c}

Note: Comparisons performed by Chi-square test (sex, MS phenotype, MS phenotype change, treatment at baseline, treatment change, cognitive status and scanner), Mann-Whitney U test (age, education, FU duration, disease duration, EDSS, ARR, z cognitive reserve, MADRS, MFIS and MSQOL-54 composite scores) and age-, sex-, and scanner-adjusted linear models (MRI variables). Bold p-values indicate a statistically significant result. Abbreviations: ARR, annualized relapse rate; CP, cognitively preserved patients; DMT, disease-modifying treatment; EDSS, Expanded Disability Status Scale; FU, follow-up; HC, healthy controls; Hipp, hippocampus; IQR, interquartile range; LV, lesion volume; MADRS, Montgomery-Åsberg Depression Rating Scale; MFIS, Modified Fatigue Impact Scale; MHCS, Mental Health Composite Score; MS, multiple sclerosis; MSQOL-54, Multiple Sclerosis Quality of Life-54; N, number; NBV, normalized brain volume; NCD, neurocognitive disorder; NCGMV, normalized cortical gray matter volume; NWMV, normalized white matter volume; PHCS, Physical Health Composite Score; PMS, progressive multiple sclerosis; RRMS, relapsing-remitting multiple sclerosis; Thal, thalamus; V, volume; z, z-score.

^aCP vs. Major NCD.

^bCP vs. Mild NCD.

^cMild vs. Major NCD.

^dClassification of DMTs: moderate efficacy = interferon beta, glatiramer acetate, teriflunomide, dimethyl fumarate; high efficacy = fingolimod, natalizumab, ocrelizumab.

TABLE 2 | Mean and standard deviation (SD) of the evaluation metrics obtained for each fold of the cross-validation.

Mean (SD)	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Accuracy	0.923 (0.039)	0.894 (0.090)	0.861 (0.097)	0.888 (0.067)	0.947 (0.049)
F1 score	0.830 (0.103)	0.731 (0.075)	0.898 (0.174)	0.770 (0.201)	0.796 (0.197)
Precision	0.969 (0.087)	0.949 (0.078)	0.946 (0.091)	0.983 (0.053)	0.958 (0.100)
Recall	0.839 (0.120)	0.769 (0.239)	0.808 (0.134)	0.872 (0.241)	0.807 (0.229)

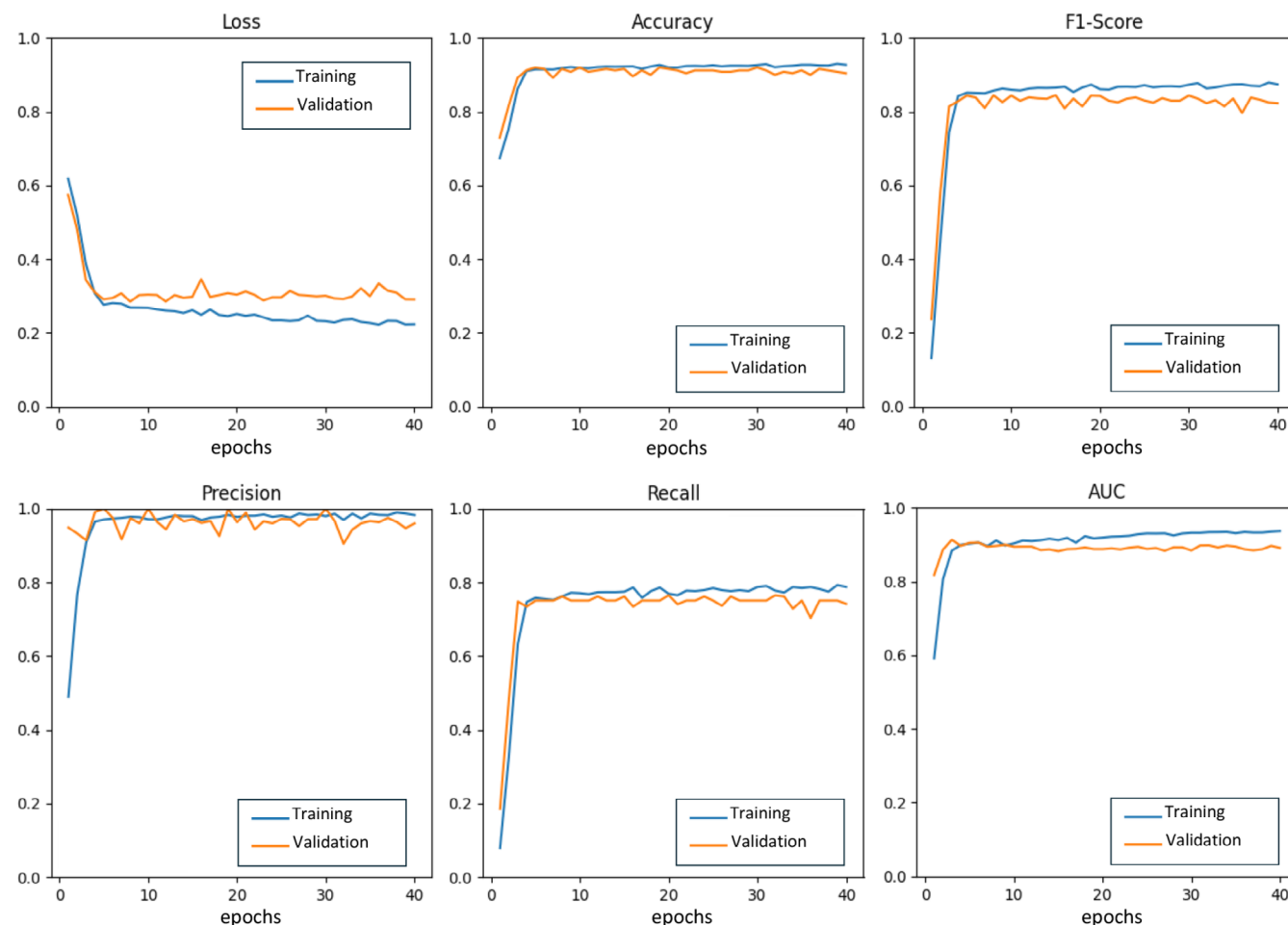


FIGURE 3 | The mean evaluation metrics (Loss, Accuracy, F1-score, Precision, Recall, and AUC) obtained from training and 5-fold cross-validation phases are plotted across the epochs.

3 | Results

3.1 | Demographic, Clinical and MRI Characteristics

The main baseline features of MS patients and HC are summarized in Table 1.

At baseline, 10 MS patients (4%) met the criteria for Mild NCD and 24 (11%) for Major NCD. Compared to patients with preserved cognition, those with Major NCD were older ($p=0.049$), had lower years of education ($p=0.010$), lower cognitive reserve ($p=0.002$), longer disease duration ($p=0.010$), higher EDSS

($p<0.001$), higher prevalence of PMS ($p<0.001$), higher MFIS scores ($p=0.015$) and reported a lower quality of life as measured by the physical health composite score of the MSQOL-54 ($p=0.037$). Patients with Mild NCD showed a higher prevalence of PMS compared to cognitively preserved ones ($p=0.021$). Compared to cognitively preserved patients, those with Mild and Major NCD had higher T2 LV ($p=0.007$ and $p=0.046$, respectively), lower thalamic ($p=0.003$ and $p<0.001$, respectively) and hippocampal volume ($p=0.036$ and $p<0.001$, respectively). Finally, patients with Major NCD had lower NBV ($p<0.001$), NCGMV ($p<0.001$), NWMV ($p=0.008$) compared to patients with preserved cognition and lower hippocampal volume compared to patients with Mild NCD ($p=0.033$) (Table 1).

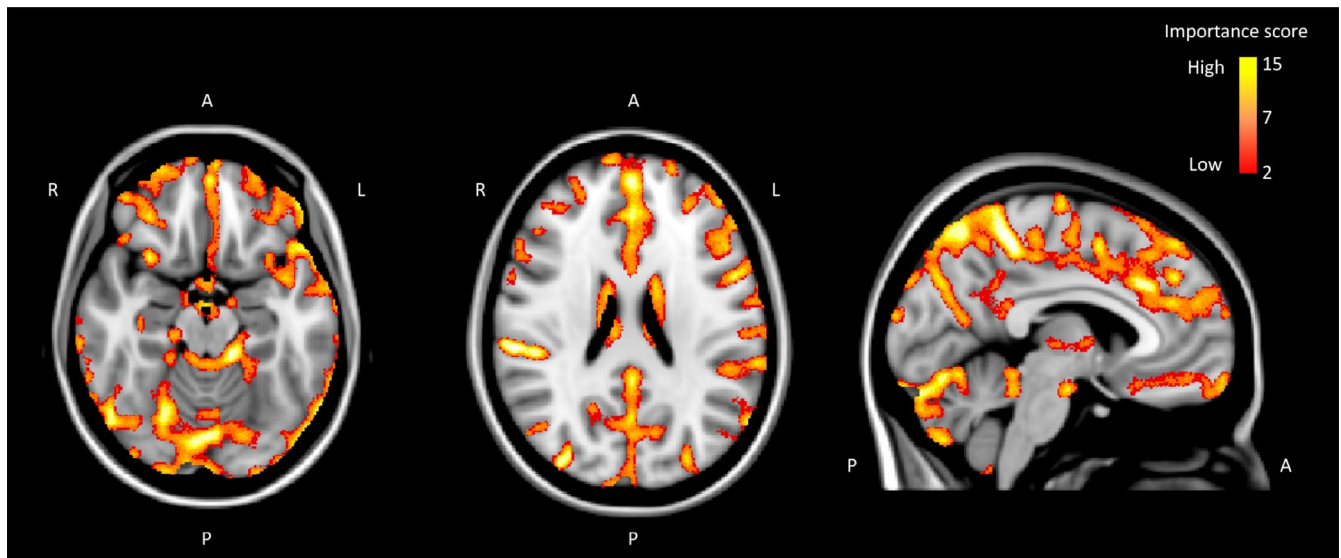


FIGURE 4 | The color-coded differential activation heatmap, obtained using the Grad-CAM method, is overlaid on the MNI-152T1-weighted atlas. Regions with greater importance in the stable group compared to the worsened group are presented in two axial slices and one sagittal slice. The importance score values represent normalized activation intensities obtained via Grad-CAM (without physical units) reflecting the relative weight of each voxel in the classification. A, anterior; L, left; P, posterior; R, right.

After a median follow-up of 3.4 (interquartile range=[2.0; 6.1]) years, 27 out of 224 MS patients (12%) experienced cognitive worsening. Of the 190 cognitively preserved patients at baseline, 11 (6%) developed Mild NCD and 12 (6%) developed Major NCD, while 4 out of 10 (40%) patients with Mild NCD at baseline progressed to Major NCD. In the entire MS sample, the mean annualized relapse rate at follow-up was 0.1 (range: 0.0–0.8); 16 patients (7%) with RRMS developed PMS, and 93 patients (41%) changed DMTs.

3.2 | Model Performance

During the training phase, the algorithm was halted after 40 epochs, based on the early stopping criterion that monitored the validation loss (0.05 after 40 epochs). The optimized hybrid classification model achieved a mean training accuracy of 91%, with an F1-score of 83%. During the training phase, the model also attained a mean precision of 96% and a recall of 84%, demonstrating a strong balance between the two.

The validation of the model showed a slight drop in the metrics, suggesting a good generalization of the model with no overfitting. The mean validation accuracy was 90%, with a mean F1-score of 81%. During the validation phase, the model achieved a mean precision of 95% and a mean recall score of 82%, indicating a strong ability to correctly identify positive instances while capturing a substantial proportion of relevant cases. Table 2 lists the validation performance scores obtained with the proposed system in terms of mean and standard deviation across the cross-validated folds.

As shown in Figure 3, the system achieved an average AUC of 0.91 (± 0.06) and 0.89 (± 0.09) to predict cognitive classes during training and validation phases, respectively. Figure 3 summarizes all the evaluation metrics and ROC curves obtained by the model for both training and validation phases.

MC-Dropout results revealed a mean uncertainty of 3.42% ($\pm 0.70\%$) in predictions across the entire dataset. When separating correctly classified from incorrectly classified patients, a slightly lower uncertainty was observed for the correctly classified (mean = $3.07\% \pm 0.53\%$) compared to the incorrectly classified (mean = $3.77\% \pm 0.87\%$) patients ($p = 0.08$).

Because patients with major NCD at baseline may still experience clinical worsening without transitioning to a more severe cognitive category at follow-up, we performed a sensitivity analysis excluding those patients classified as having major NCD at baseline from the whole dataset, to verify the robustness of the model. The results of this analysis are presented in the [Supporting Information](#).

3.3 | Explainability models

Figure 4 presents the mean attention heatmap generated using the Grad-CAM method, overlaid on the MNI-152T1-weighted atlas. The heatmap highlights regions with greater importance for the model's decision in the stable group compared to the worsened group.

Following the application of the AAL atlas to the Grad-CAM maps, we identified the brain regions most relevant to the algorithm's decision-making process, ranked by cluster size. These regions are listed in Table 3 in descending order of importance.

The application of the PFI explainability method to tabular data yielded a ranked list of the most important features arranged in order of decreasing importance. As shown in Figure 5, the six tabular features most influential in the algorithm's decision-making included NCGMV, age, thalamic volume, hippocampal volume, T2 LV, and cognitive reserve.

TABLE 3 | Brain regions most strongly contributing to the model's identification of cognitively stable MS patients.

Bilateral anatomical region (AAL)	Activation	Cluster extent
Middle frontal gyrus	6.0	278
Superior frontal gyrus, medial	5.9	1078
Middle occipital gyrus	5.3	28
Postcentral gyrus	5.1	28
Middle temporal gyrus	5.1	62
Hippocampus	4.8	67
Inferior frontal gyrus, pars orbitalis	4.7	13
Temporal pole: superior temporal gyrus	4.7	12
SupraMarginal gyrus	4.6	30
Superior frontal gyrus, pars orbitalis	4.6	50
Calcarine fissure and surrounding cortex	4.4	16
Superior frontal gyrus, medial orbital	4.3	12
Inferior frontal gyrus, triangular part	4.3	8
Lobule IX of cerebellar hemisphere	4.2	16
Temporal pole: middle temporal gyrus	4.2	5
Angular gyrus	4.1	14
Anterior cingulate & paracingulate gyri	4.1	23
Insula	4.1	8
Superior frontal gyrus, dorsolateral	4.1	13
Lobule IV, V of vermis	4.0	12
Inferior temporal gyrus	4.0	10
Superior parietal gyrus	4.0	6
Parahippocampal gyrus	4.0	3
Inferior occipital gyrus	4.0	2
Precentral gyrus	4.0	6

Abbreviation: AAL, automated anatomical labeling.

Once the PFI explainability method identified the most relevant tabular predictors, a binary logistic regression analysis was employed to explore the direction of associations with cognitive worsening. Although tabular data did not reach statistical significance in any of the binary logistic regression models, β values were used to interpret the direction of the correlations. Cognitive

worsening was associated with older age ($\beta=0.24$), male sex ($\beta=0.14$), higher EDSS scores ($\beta=0.20$), and lower cognitive reserve ($\beta=-0.0000001$). Patients treated with high-efficacy DMTs ($\beta=0.02$) and moderate-efficacy DMTs ($\beta=0.10$) were more likely to develop NCD at follow-up compared to untreated patients. Furthermore, cognitive worsening was associated with lower baseline NWMV ($\beta=-0.004$), NCGMV ($\beta=-0.004$), thalamic ($\beta=-0.008$), and hippocampal volume ($\beta=-0.19$). Finally, a higher T2 LV ($\beta=0.20$) was observed in the worsened patients' group compared to the stable patients' group.

4 | Discussion

This longitudinal study investigated demographic, clinical, and MRI features associated with NCDs in a large and comprehensively characterized cohort of MS patients, with a focus on identifying the potential predictors of NCD development and worsening using AI. The integration of multiple data modalities and explainability techniques provides valuable insights that may contribute to a better understanding of cognitive impairment in MS and the potential for clinical translation of AI in this domain.

Previous studies often dichotomized cognitive function in MS (preserved vs. impaired), oversimplifying cognitive symptoms. A recent investigation [27] validated specific impairment thresholds to distinguish distinct cognitive phenotypes in MS, characterized by different levels of impairment. However, the DSM-5-TR [5] criteria applied in this study require not only assessing the degree of cognitive dysfunction but also its impact on complex instrumental activities, allowing for a more valid description of patients' real-world functioning [34].

At baseline, the overall prevalence of NCDs among MS patients was 15% (4% Mild NCD and 11% Major NCD), which increased to 25% (8% Mild NCD and 17% Major NCD) after a median follow-up duration of 3.4 years. This result is consistent with the Major NCD frequency (21%–22%) reported by previous investigations in MS [3, 7], while the higher prevalence of Mild NCD (67%) found by Hancock and colleagues [7] likely reflects distinct neuropsychological approaches, as the diagnosis was based on evidence of a “modest” cognitive decline without specific threshold scores. In line with our expectations, patients with Mild and Major NCD exhibited a more severe clinical and MRI profile compared to cognitively preserved MS patients. Of note, the only significant difference between Mild and Major NCD in MS patients was reduced hippocampal volume in the latter, suggesting that hippocampal atrophy reflects the trajectory of cognitive decline and may serve as a key biomarker in MS [35].

The hybrid classification model implemented showed strong performance in predicting NCD progression, with early stopping effectively preventing overfitting. Validation metrics were stable, demonstrating good generalizability and well-balanced precision and recall. Additionally, AUC values over 0.89 in both training and validation confirmed the model's high ability to accurately classify cognitive status at follow-up. These results highlight the promising role of AI in advancing predictive accuracy and personalized prognosis of cognitive impairment in MS.

MC-Dropout analysis showed low overall prediction uncertainty (~3.4%), with slightly higher (not significant) uncertainty in misclassified cases, indicating the model's confidence aligns with its accuracy. This uncertainty measure is important for clinical use, helping to stratify risk and flag cases needing expert review. Consistent with our previous work [11], T1-weighted MRI likely represents the main contributor to model performance, as structural brain features are highly informative for predicting cognitive decline. The inclusion of clinical and demographic variables further enhanced predictive accuracy, indicating that while both modalities offer complementary information, MRI data remain the primary drivers of the observed performance. This finding is particularly relevant for future implementation, as it suggests that robust performance may still be achievable in settings where only imaging data are available, while the integration of tabular data could further refine predictions when accessible.

Another key aspect of this study, particularly regarding prospective clinical implementation, is the comprehensive use of explainability methods to elucidate the contribution of all features employed by the algorithm in reaching its final decision. The Grad-CAM attention maps localized key brain regions implicated in the classification, predominantly involving areas known for their role in cognition, thus biologically validating the model's focus. The region importance ranking derived from the AAL atlas corroborates established neuroanatomical substrates affected in MS cognitive impairment. Notably, with 10 out of 25 areas identified by the Grad-CAM attention maps, the frontal lobe emerged as the most relevant predictor of cognitive worsening. Due to its involvement in working memory, processing speed, attention and executive function, structural damage in this key cognitive structure has been consistently linked to

cognitive impairment in MS patients [1, 36]. Beside the frontal lobe, the Grad-CAM attention maps identified six areas in the temporal lobe, four in the parietal lobe, three in the occipital lobe and two in the cerebellum. These findings align with previous results suggesting that, in MS patients, cortical and cerebellar atrophy are strongly associated with cognitive dysfunction [1, 10, 35–37]. Specifically, structural damage in the temporal lobe has been linked to episodic memory deficits [35, 36], parietal lobe involvement to impairment in attention and visuospatial processing [38, 39], structural damage in the occipital lobe to visual processing dysfunction [36, 40], and cerebellar involvement to impairments in working memory, processing speed, and motor function [37, 41].

Concurrently, PFI analysis of tabular clinical and imaging features identified NCGMV, age, thalamic and hippocampal volumes, T2 LV, and cognitive reserve as the primary contributors to the model's decision-making process. These results highlight the synergistic impact of neurodegeneration, lesion burden, and demographic/clinical factors on cognitive trajectories in these patients [1, 36]. MRI-derived features therefore appear to be the main drivers of model performance, while clinical and demographic data provide complementary contextual information that enhances robustness and interpretability. This interplay underscores the interconnected nature of multimodal data and highlights the importance of considering cross-modal effects when interpreting model explainability results. Future studies should aim to disentangle these interactions and examine how variations in one modality might affect model behavior and identified regions of importance. Such investigations could deepen our understanding of multimodal explainability and support the development of more generalizable predictive frameworks across different clinical settings.

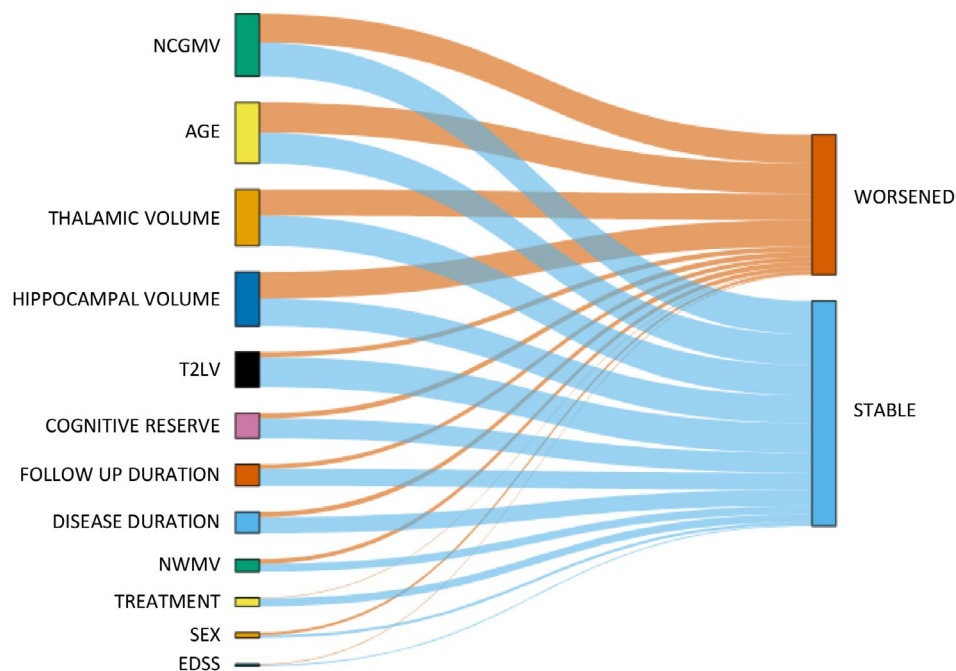


FIGURE 5 | Flow diagram illustrating the most important tabular features identified by the PFI method for classifying cognitively stable and worsened patients is shown. The plot displays the results for the entire dataset. The width of each link represents the strength of the association between the variable and the outcome. EDSS, Expanded Disability Status Scale; LV, lesion volume; NCGMV, normalized cortical gray matter volume; NWMV, normalized white matter volume.

Although logistic regression did not yield statistically significant associations, likely due to limited sample size or multicollinearity, the directional β coefficients are consistent with existing literature. Cognitive worsening correlated positively with older age, male sex, higher disability, and lesion burden, while negatively associated with cognitive reserve and brain volumes. Interestingly, patients treated with DMTs of varying efficacy showed a trend towards increased risk of cognitive decline, which might reflect confounding by indication or the presence of more aggressive disease phenotypes necessitating earlier or more effective treatments.

This study has several limitations. First, the distribution of cognitively stable vs worsened patients was unbalanced, with only 12% showing cognitive decline over the follow-up period, although this rate aligns with previous research when accounting for follow-up duration [42].

The smaller proportion of patients with Mild or Major NCD at baseline may have reduced statistical power for group comparisons, potentially obscuring subtle differences. Additionally, the cognitive reserve index excluded leisure activities due to missing data, using education as a proxy [43]. Anxiety was not assessed, though depression was measured and major psychiatric conditions excluded. Although we implemented several strategies to reduce the risk of overfitting and ensure the robustness of our results, including a combination of data augmentation and regularization techniques, we acknowledge that external validation on an independent cohort and further studies are needed to confirm the model's generalizability and investigate causal links between imaging biomarkers and cognitive decline.

To conclude, the multimodal AI implemented in this study demonstrated high accuracy and reliability, with explainability analyses confirming the involvement of brain regions linked to cognitive impairment. These results support the potential for AI-based tools in personalized cognitive assessment and monitoring in MS.

Author Contributions

Loredana Storelli: data curation, formal analysis, methodology, software, writing – original draft preparation. **Damiano Mistri:** investigation, formal analysis, writing – original draft preparation. **Alice Mastropasqua:** investigation, formal analysis, writing – original draft preparation. **Marta Grosselle:** investigation, writing – review and editing. **Paolo Preziosa:** validation, writing – review and editing. **Lucrezia Rossi:** validation, writing – review and editing. **Massimo Filippi:** conceptualization, data curation, funding acquisition, resources, supervision, writing – review and editing. **Maria A. Rocca:** conceptualization, formal analysis, supervision, writing – review and editing.

Acknowledgements

The authors have nothing to report.

Funding

This study was supported by FISM—Fondazione Italiana Sclerosi Multipla—cod. 2023/S/1 and financed or co-financed with the ‘5 per mille’ public funding.

Conflicts of Interest

L. Storelli, D. Mistri, M. Grosselle, A. Mastropasqua, and L. Rossi have nothing to disclose. P. Preziosa received speaker honoraria from Roche, Biogen, Novartis, Merck, Bristol Myers Squibb, Genzyme, Horizon and Sanofi, he has received research support from Italian Ministry of Health and Fondazione Italiana Sclerosi Multipla. M. Filippi is Editor-in-Chief of the *Journal of Neurology*, Associate Editor of *Human Brain Mapping*, *Neurological Sciences*, and *Radiology*; received compensation for consulting services from Almirall, Biogen, Bristol-Myers Squibb, Eli Lilly, Merck, Novartis, Roche, Sanofi; speaking activities from Amgen, Bayer, Biogen, Bristol-Myers Squibb, Celgene, Chiesi Italia SpA, Eisai, Eli Lilly, Fujirebio, Genzyme, Janssen, Merck, Neopharmed Gentili, Neuraxpharm, Novartis, Novo Nordisk, Roche, Sanofi, Takeda; participation in Advisory Boards for Alexion, Biogen, Bristol-Myers Squibb, Eli Lilly, GE Healthcare Ltd, Merck, Neuraxpharm, Novartis, Roche, Sandoz, Sanofi, Takeda; scientific direction of educational events for Biogen, Merck, Roche, Celgene, Bristol-Myers Squibb, Lilly, Novartis, Sanofi-Genzyme; he receives research support from Biogen Idec, Merck-Serono, Novartis, Roche, the Italian Ministry of Health, the Italian Ministry of University and Research, and Fondazione Italiana Sclerosi Multipla. M.A. Rocca received consulting fees from Biogen, Bristol Myers Squibb, Roche; and speaker honoraria from Alexion, Biogen, Bristol Myers Squibb, Celgene, Horizon Therapeutics Italy, Merck Serono SpA, Mitsubishi-Tanabe Pharma, Neuraxpharm, Novartis, Roche, Sandoz, and Sanofi. She receives research support from the MS Society of Canada, the Italian Ministry of Health, the Italian Ministry of University and Research, and Fondazione Italiana Sclerosi Multipla. She is Associate Editor for *Multiple Sclerosis and Related Disorders*; and Associate Co-Editor for Europe and Africa for *Multiple Sclerosis Journal*.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

1. M. A. Rocca, M. P. Amato, N. De Stefano, et al., “Clinical and Imaging Assessment of Cognitive Dysfunction in Multiple Sclerosis,” *Lancet Neurology* 14 (2015): 302–317.
2. H. J. Westervelt, “Dementia in Multiple Sclerosis: Why Is It Rarely Discussed?,” *Archives of Clinical Neuropsychology* 30 (2015): 174–177.
3. R. H. Benedict and J. H. Bobholz, “Multiple Sclerosis,” *Seminars in Neurology* 27 (2007): 78–85.
4. American Psychiatric A, “Diagnostic and Statistical Manual of Mental Disorders,” (2000).
5. American Psychiatric A, “Diagnostic and Statistical Manual of Mental Disorders,” (2022).
6. American Psychiatric A, “Diagnostic and Statistical Manual of Mental Disorders,” (2013).
7. L. M. Hancock, B. Hermann, M. M. Schoonheim, S. J. Hetzel, B. Brochet, and J. DeLuca, “Comparing Diagnostic Criteria for the Diagnosis of Neurocognitive Disorders in Multiple Sclerosis,” *Multiple Sclerosis and Related Disorders* 58 (2022): 103479.
8. T. Brummer, M. Muthuraman, F. Steffen, et al., “Improved Prediction of Early Cognitive Impairment in Multiple Sclerosis Combining Blood and Imaging Biomarkers,” *Brain Communications* 4 (2022): fca153.
9. K. Buyukturkoglu, D. Zeng, S. Bharadwaj, et al., “Classifying Multiple Sclerosis Patients on the Basis of SDMT Performance Using Machine Learning,” *Multiple Sclerosis* 27 (2021): 107–116.

10. C. Marzi, A. d'Ambrosio, S. Diciotti, et al., "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 (2023): 186–202.
11. L. Storelli, M. Azzimonti, M. Gueye, et al., "A Deep Learning Approach to Predicting Disease Progression in Multiple Sclerosis Using Magnetic Resonance Imaging," *Investigative Radiology* 57 (2022): 423–432.
12. P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy* 23: 18 (2020).
13. F. J. Dongil-Moreno, M. Ortiz, A. Pueyo, et al., "Diagnosis of Multiple Sclerosis Using Optical Coherence Tomography Supported by Explainable Artificial Intelligence," *Eye (London, England)* 38 (2024): 1502–1508.
14. M. Hernandez, U. Ramon-Julvez, E. Vilades, B. Cordon, E. Mayor-domo, and E. Garcia-Martin, "Explainable Artificial Intelligence Toward Usable and Trustworthy Computer-Aided Diagnosis of Multiple Sclerosis From Optical Coherence Tomography," *PLoS One* 18 (2023): e0289495.
15. S. Rasouli, M. S. Dakkali, R. Azarbad, et al., "Predicting the Conversion From Clinically Isolated Syndrome to Multiple Sclerosis: An Explainable Machine Learning Approach," *Multiple Sclerosis and Related Disorders* 86 (2024): 105614.
16. M. A. Yamin, P. Valsasina, J. Tessadori, et al., "Discovering Functional Connectivity Features Characterizing Multiple Sclerosis Phenotypes Using Explainable Artificial Intelligence," *Human Brain Mapping* 44 (2023): 2294–2306.
17. A. J. Thompson, B. L. Banwell, F. Barkhof, et al., "Diagnosis of Multiple Sclerosis: 2017 Revisions of the McDonald Criteria," *Lancet Neurol* 17 (2018): 162–173.
18. J. F. Kurtzke, "Rating Neurologic Impairment in Multiple Sclerosis: An Expanded Disability Status Scale (EDSS)," *Neurology* 33 (1983): 1444–1452.
19. C. J. Hawley, T. M. Gale, T. Sivakumaran, and Hertfordshire Neuroscience Research g, "Defining Remission by Cut Off Score on the MADRS: Selecting the Optimal Value," *Journal of Affective Disorders* 72 (2002): 177–184.
20. O. Marchesi, C. Vizzino, A. Meani, et al., "Fatigue in Multiple Sclerosis Patients With Different Clinical Phenotypes: A Clinical and Magnetic Resonance Imaging Study," *European Journal of Neurology* 27 (2020): 2549–2560.
21. A. Solari, G. Filippini, L. Mendozzi, et al., "Validation of Italian Multiple Sclerosis Quality of Life 54 Questionnaire," *Journal of Neurology, Neurosurgery, and Psychiatry* 67 (1999): 158–162.
22. L. S. G. Colombo and C. Brivio, "Stima Del Quoziente Intellettivo Tramite L'applicazione Del TIB (Test Breve di Intelligenza)," *Giornale Italiano di Psicologia* 3 (2002): 613–638.
23. M. P. Amato, L. Razzolini, B. Goretti, et al., "Cognitive Reserve and Cortical Atrophy in Multiple Sclerosis: A Longitudinal Study," *Neurology* 80 (2013): 1728–1733.
24. B. Goretti, F. Patti, S. Cilia, et al., "The Rao's Brief Repeatable Battery Version B: Normative Values With Age, Education and Gender Corrections in an Italian Population," *Neurological Sciences* 35 (2014): 79–82.
25. N. Tedone, C. Vizzino, A. Meani, et al., "The Brief Repeatable Battery of Neuropsychological Tests (BRB-N) Version a: Update of Italian Normative Data From the Italian Neuroimaging Network Initiative (INNI)," *Journal of Neurology* 271 (2024): 1813–1823.
26. R. K. Heaton, G. J. Chelune, J. L. Talley, G. G. Kay, and G. Curtiss, *Wisconsin Card Sorting Test Manual: Revised and Expanded* (Psychological Assessment Resources Inc, 1993).
27. D. Mistri, N. Tedone, D. Biondi, et al., "Cognitive Phenotypes in Multiple Sclerosis: Mapping the Spectrum of Impairment," *Journal of Neurology* 271 (2024): 1571–1583.
28. S. Candemir, X. V. Nguyen, L. M. Prevedello, et al., "Predicting Rate of Cognitive Decline at Baseline Using a Deep Neural Network With Multidata Analysis," *Journal of Medical Imaging (Bellingham)* 7 (2020): 044501.
29. S. Chen, K. Ma, Y. Zheng, et al., "Med3D: Transfer Learning for 3D Medical Image Analysis. Computer Vision and Pattern Recognition," *arXiv* (2019): arXiv:1904.00625.
30. T. Zeevi, R. Venkataraman, L. H. Staib, and J. A. Onofrey, "Monte-Carlo Frequency Dropout for Predictive Uncertainty Estimation in Deep Learning," in *Ieee International Symposium on Biomedical Imaging* (Isbi, 2024).
31. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization," in *Ieee International Conference on Computer Vision (ICCV, 2017)*, 618–626.
32. E. T. Rolls, C. C. Huang, C. P. Lin, J. Feng, and M. Joliot, "Automated Anatomical Labelling Atlas 3," *NeuroImage* 206 (2020): 116189.
33. L. Breiman, "Random Forests," *Machine Learning* 45 (2001): 5–32.
34. P. S. Sachdev, D. Blacker, D. G. Blazer, et al., "Classifying Neurocognitive Disorders: The DSM-5 Approach," *Nature Reviews. Neurology* 10 (2014): 634–642.
35. M. A. Rocca, F. Barkhof, J. De Luca, et al., "The Hippocampus in Multiple Sclerosis," *Lancet Neurology* 17 (2018): 918–926.
36. P. Preziosa, M. A. Rocca, E. Pagani, et al., "Structural MRI Correlates of Cognitive Impairment in Patients With Multiple Sclerosis: A Multicenter Study," *Human Brain Mapping* 37 (2016): 1627–1644.
37. A. D'Ambrosio, E. Pagani, G. C. Riccitelli, et al., "Cerebellar Contribution to Motor and Cognitive Performance in Multiple Sclerosis: An MRI Sub-Regional Volumetric Analysis," *Multiple Sclerosis (Houndmills)* 23 (2017): 1194–1203.
38. I. Gabilondo, O. Rilo, N. Ojeda, et al., "The Influence of Posterior Visual Pathway Damage on Visual Information Processing Speed in Multiple Sclerosis," *Multiple Sclerosis* 23 (2017): 1276–1288.
39. K. A. Meijer, A. J. C. Eijlers, L. Douw, et al., "Increased Connectivity of Hub Networks and Cognitive Impairment in Multiple Sclerosis," *Neurology* 88 (2017): 2107–2114.
40. M. Pitteri, I. B. Galazzo, L. Brusini, et al., "Microstructural MRI Correlates of Cognitive Impairment in Multiple Sclerosis: The Role of Deep Gray Matter," *Diagnostics (Basel)* 11 (2021): 11.
41. M. Grothe, M. Lotze, S. Langner, and A. Dressel, "Impairments in Walking Ability, Dexterity, and Cognitive Function in Multiple Sclerosis Are Associated With Different Regional Cerebellar Gray Matter Loss," *Cerebellum* 16 (2017): 945–950.
42. M. Huiskamp, A. J. C. Eijlers, T. A. A. Broeders, et al., "Longitudinal Network Changes and Conversion to Cognitive Impairment in Multiple Sclerosis," *Neurology* 97 (2021): e794–e802.
43. Y. Chen, C. Lv, X. Li, et al., "The Positive Impacts of Early-Life Education on Cognition, Leisure Activity, and Brain Structure in Healthy Aging," *Aging (Albany NY)* 11 (2019): 4923–4942.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** Supporting Information.