European Journal of Neurology

european journal of neurology
the official journal of the european academy of neurology

ORIGINAL ARTICLE OPEN ACCESS

Machine Learning Analysis Applied to Prediction of Early Progression Independent of Relapse Activity in Multiple Sclerosis Patients

Valentina Poretto[1] | Walter Endrizzi[2,3] | Matteo Betti[4] | Stefano Bovo[2] | Angelo Bellinvia[4] | Flavio Ragni[2] | Caterina Lapucci[5] | Monica Moroni[2] | Sabrina Marangoni[1] | Emilio Portaccio[4] | Chiara Longo[1] | Lorenzo Gios[2] | Marco Chierici[2] | Giuseppe Jurman[2,6] | Bruno Giometto[1,7] | Matilde Inglese[5,8] | Venet Osmani[2] | Manuela Marenco[5] | Antonio Uccelli[5] | Maria Pia Amato[4,9]

[1]Neurology Unit, Azienda Provinciale per i Servizi Sanitari (APSS), Trento, Italy | [2]Data Science for Health, Fondazione Bruno Kessler, Trento, Italy | [3]Department of Cellular, Computational and Integrative Biology, University of Trento, Trento, Italy | [4]Department of NEUROFARBA, University of Florence, Florence, Italy | [5]IRCCS Ospedale Policlinico San Martino, Genoa, Italy | [6]Department of Biomedical Sciences, Humanitas University, Milan, Italy | [7]Centro Interdipartimentale di Scienze Mediche (CISMed), Facoltà di Medicina e Chirurgia, Università di Trento, Trento, Italy | [8]Department of Neurology, Rehabilitation, Ophthalmology, Genetics, Maternal and Child Health (DINOGMI), University of Genoa, Genoa, Italy | [9]IRCCS Don Carlo Gnocchi Foundation, Florence, Italy

Correspondence: Valentina Poretto (valentina.poretto@gmail.com)

Received: 26 March 2025 | **Revised:** 21 July 2025 | **Accepted:** 29 September 2025

Funding: This work was supported by the Italian Ministry of Health, grant number NET-2018-12366666 NeuroArtP3. Moreover this work was partially funded under the National Plan for Complementary Investments to the NRRP, project "D34H—Digital Driven Diagnostics, prognostics and therapeutics for sustainable Health care" (project code: PNC0000001), Spoke 2: "Multilayer platform to support the generation of the Patients' Digital Twin", CUP: B53C22006170001, funded by the Italian Ministry of University and Research. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the article.

Keywords: machine learning | multiple sclerosis | personalized | prediction | prognosis | progression independent of relapse activity

ABSTRACT

Background: Predicting prognosis in people with multiple sclerosis (pwMS) at early disease stages still remains an unmet need. Machine learning (ML) strategies demonstrated good reliability when applied for prediction in medicine. This study aimed at developing a predictive algorithm comparing different ML approaches, by using routine demographic, clinical and radiological data from a large multicentric cohort of newly diagnosed pwMS.

Methods: Demographic, clinical, radiological and biochemical data were retrospectively collected at three Italian MS centers at baseline and four timepoints thereafter (6, 12, 24, and 36 months). Data from the first evaluation and subsequent 2-year follow-up were analyzed, comparing different ML models (Random Forest, Extra Trees, XGBoost, Logistic Regression and Support Vector Classifier) to predict progression independent of relapse activity (PIRA) at year 3. To understand how features impacted the selected model's output, a ML explainability analysis was performed on the whole cohort and on specific subsets of patients, those aged under 45 and those NEDA-3 at the 2-year follow-up.

Results: Data from 719 pwMS (age 34.6 ± 11.2 years); female sex 501 (70%) were analyzed. Ninety-two pwMS (13%) developed PIRA at year 3. Random Forest achieved the highest score, with a test set area under the ROC curve (AUC) of 0.75 ± 0.06. Features with the highest predictive impact were Expanded Disability Status Scale at 24 months, age at symptom onset and disease duration at baseline.

Conclusion: Our results showed the feasibility of applying ML techniques to predict short-term PIRA in newly diagnosed pwMS by using routine clinical practice data, paving the way for tailored and personalized approaches.

1 | Introduction

Multiple Sclerosis (MS) is a neuroinflammatory and degenerative disease characterized by a heterogeneous and unpredictable course [1], which can lead to significant disability in young adults. Besides the amount of worsening associated with relapse activity (RAW), there is increasing evidence in the literature that most disability accrual is independent of relapse activity, so-called progression independent of relapse activity (PIRA) [2, 3]. This "silent" progression is not exclusive to the progressive course of MS and its underlying pathological mechanisms begin even in the earliest phases of the disease [3–6], remaining often unnoticed by physicians during the initial phase of the disease [7].

To date, predictors of PIRA at MS onset have not yet been outlined and defining the prognosis of MS patients at early disease stages remains a challenging and unresolved issue [8]. Therefore, the identification of reliable prognostic markers of MS progression detectable since MS onset is an urgent need in order to improve proper stratification of each patient's risk of PIRA and to develop personalized and effective therapeutic strategies [9].

In the last years, machine learning (ML) strategies have demonstrated good reliability when applied for prediction in medicine [10], due to their capability of deriving complex and hidden patterns in high-dimensional data, overcoming the limits of traditional statistical techniques, which can typically handle only a few input variables and are often based on strict assumptions [11]. Previous studies used various ML approaches to predict progression in patients with MS (pwMS) analyzing clinical [12, 13], radiological [14, 15], and genetic and -omics biomarkers as well [16], obtaining variable results. Despite these efforts predicting disability progression with high accuracy still remains challenging [12, 17, 18], especially in a purely clinical setting.

In this framework, the Italian project "Artificial intelligence of imaging and clinical neurological data for predictive, preventive and personalized (P3) medicine" (NeuroArtP3, grant number NET-2018-12366666), a multi-site project co-funded by the Italian Ministry of Health, aims at identifying a personalized approach to patients with neurological diseases through the use of artificial intelligence (AI). Specifically, this study aims at developing a predictive algorithm comparing different ML approaches, by using routine demographic, clinical and radiological data collected from clinical practice on a large multicentric cohort of newly diagnosed MS patients, to find early predictors of disability progression.

2 | Material and Methods

2.1 | Dataset Description

Data from 722 patients was collected from three Italian MS centers (Careggi University Hospital, SOD Riabilitazione Neurologica, Florence; IRCCS Ospedale Policlinico San Martino Genova; Azienda Provinciale per i Servizi Sanitari, Trento).

As part of the standard clinical routine practice, pwMS were enrolled using the following inclusion criteria: (1) diagnosis according to 2010 McDonald criteria [19], independently of clinical phenotype; (2) age 18–65 years; (3) availability of at least one clinical evaluation with Expanded Disability Status Scale (EDSS) score; and (4) one brain magnetic resonance imaging (MRI) performed every 12 months for the first 3 years of disease.

Clinical assessment including Expanded Disability Status Score (EDSS) evaluation was performed at baseline (T0), 6 months (T1), 12 months (T2), 24 months (T3), and 36 months (T4) post-baseline. Baseline (T0) was defined as the first visit at MS center. The study protocol was approved by Local Ethics Committees of Firenze, Genova and Trento. All subjects underwent a 1.5 or 3T MRI examination including at least the following pulse sequences: 3D T1-weighted (3D-T1) and fuid-attenuated inversion recovery (FLAIR). Patients underwent MRI evaluation at baseline and at least at three following timepoints (T2, T3, T4), where data regarding lesion number, localization (periventricular, cortical/iuxtacortical, infratentorial, optic nerve and spinal), new lesions compared to baseline and contrast-enhancement were collected. All images were assessed by consensus of two experienced observers, who were blinded to patient identity.

Pseudoanonymized data were collected through a dedicated REDCap database (*research electronic data capture [REDCap]*) [20] in order to guarantee homogeneous and standardized data management among different centers.

The outcome variable was defined as PIRA at T4, which is the presence of disease progression measured through EDSS variation between time points T3 and T4 (increase of >1.5 if EDSS at T3 was 0.0, of >1.0 if EDSS at T3 was between 1.0 and 5.0, of >0.5 if EDSS at T3 was ≥ 5.5) in the absence of relapses in the time-interval between 3 months before and 1 month after T4 visit, as described by Muller et al. [21]. Importantly, T4 EDSS measurement was used solely to define the outcome variable and was not included among the predictive features in any of our models, which relied exclusively on assessments conducted at T0–T3.

2.2 | Data Preprocessing

Data preprocessing followed a two-step process: an attribute-engineering process and a filtering one. In the attribute-engineering process, the following novel attributes were generated from raw features. Age at symptom onset and "Treatment Lag" were computed as the difference, in years, between symptom onset date and birth date and between the first treatment date and symptom onset, respectively. "Treatment lag" was further converted into a categorical variable according to the lag duration (less than 1 year, between 1 and 3 years, more than 3 years, no treatments received). Alongside the no treatment category, disease-modifying treatments (DMTs) present in the raw dataset were grouped into two categories: low-efficacy treatments (interferons, glatiramer acetate, fumarates,

teriflunomide) and high-efficacy treatments (cladribine, cyclophosphamide, mitoxantrone, fingolimod and other immunomodulators, natalizumab, alemtuzumab, ocrelizumab and rituximab). Relapse attributes were converted into binary attributes describing the presence of at least one relapse in the considered time interval.

The filtering process allowed for the reduction of the total number of features through the removal of features with more than 20% of missing values, features with zero variance, imbalanced binary features (with less than 10% of values in the minority class), non-binary features with heavily imbalanced distributions and other non-clinically relevant features (for a full list of features see Table S1).

Missing values in the dataset were imputed using the median for numerical attributes and the most frequent value for categorical attributes including nominal, ordinal and binary types. Numerical features were then scaled to ensure range consistency. In contrast, nominal features were transformed using one-hot encoding, creating binary indicators for each category.

2.3 | Classification

Binary classification of patients as PIRA was performed using different models: Random Forest (RF), ExtraTrees Classifier (ETC), Extreme Gradient Boosting (XGB), Logistic Regression (LR) and Support Vector Classifier (SVC).

The classification process was carried out through a Randomized Nested Grid Search Cross Validation (RNGCV) strategy (Figure 1).

This method takes the entire dataset as input and consists of two nested loops: an outer loop for dataset splitting and testing, and an inner loop for parameter optimization and cross-validation.

In the external loop the dataset is randomly split into a training (80%) and test (20%) set preserving the class distribution.

The training set is used in the inner loop to find the best model configuration through a randomized grid search over a predefined parameter space exploring 20 parameter combinations. For each configuration, a repeated threefold cross-validation is performed and the model obtaining the best-averaged Matthews Correlation Coefficient (MCC) is selected as the best model, refitted to the entire training set and returned to the outer loop, where its performance (MCC, AUC-ROC, F1-Score and features importance) is computed on the held-out test set.

The full procedure was repeated for 30 iterations and the reported results are the average and standard deviations across the different iterations.

2.4 | Importance Analysis

Together with performance metrics, the RNGCV framework computed the permutation importance of each predictor on the best model to return an estimate of its contribution to classification.

Permutation importance is a technique that quantitatively assesses the importance of each feature by measuring changes in model performance when the values of that feature are randomly shuffled [22].

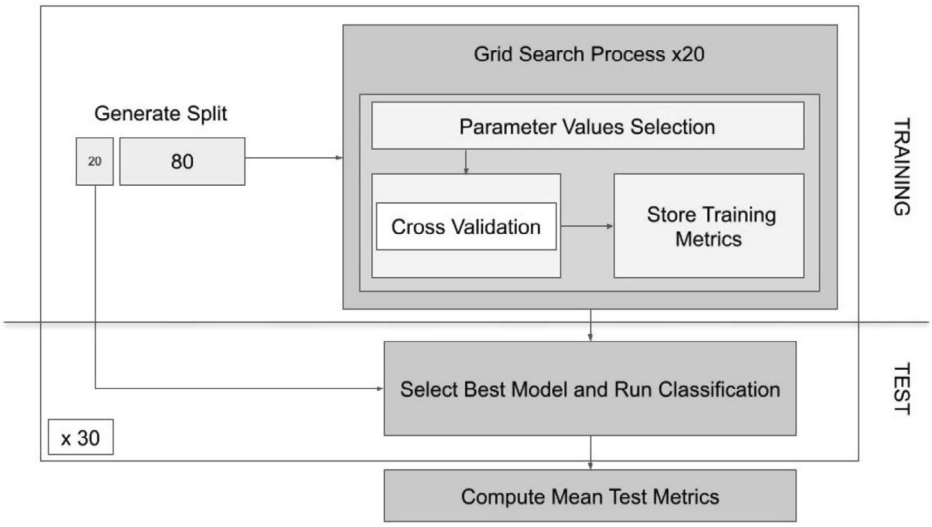

FIGURE 1 | Schematic representation of the random nested grid search cross validation (RNGCV) strategy. This strategy consists of two nested loops: An outer loop for dataset splitting and testing, and an inner loop for parameter optimization and cross-validation. The external loop consists of thirty independent iterations to evaluate model robustness. In each iteration, the input data is randomly split between train and test sets using an 80/20% proportion and a randomized grid search explores 20 different parameter combinations. For each configuration, a repeated stratified threefold cross-validation is used. At the end of each outer loop iteration, the optimized model (selected using the inner loop cross-validation and based on MCC) is evaluated on the 20% held-out test set. Classification metrics (MCC, AUC-ROC, and F1-Score) and feature importances are stored at each iteration, generating 30 unique results. Finally, their mean and standard deviation are computed.

2.5 | Stratification Strategies

To gain deeper insights into the provided dataset, the RNGCV framework was also applied to specific clusters of patients who met clinical criteria relevant in the context of MS. A first cluster was composed of all patients 45 years or younger [23], while a second cluster was composed of all patients with NEDA-3 (No Evidence of Disease Activity) [24, 25] at T3, defined as the simultaneous absence of relapses, disability worsening and MRI activity.

For each identified cluster, the raw dataset describing the subpopulation was processed as described earlier and the RNGCV strategy was applied to the preprocessed subset of data.

3 | Results

3.1 | Data Preprocessing

Data collection resulted in a total of 722 patients, three of whom were excluded during the data preprocessing (due to "first treatment" date missing). The final dataset consisted of 719 patients with 28 attributes (16 at baseline, 4 at T1, 4 at T2 and 4 at T3).

Table 1 presents descriptive statistics of some informative variables at baseline (for a full description see Table S1).

3.2 | Prediction Results

The target variable, that is, PIRA at T4, was positive in 92 patients (13%), while the remaining 627 patients (87%) showed no progression.

The RNGCV strategy returned a robust estimation of the performance of different ML models in predicting disease progression at T4. Among all tested classifiers (Table 2), Random Forest achieved the best predictive performance with AUC-ROC of 0.75 ± 0.06, MCC of 0.28 ± 0.1 and F1-Score of 0.36 ± 0.08 (mean ± SD) (Figure 2).

Besides returning the model's performances, the RNGCV strategy identified the following attributes as consistently informative for the prediction of disease progression: EDSS at T3, age at symptom onset, lag between symptom onset and first evaluation, EDSS at T1 (Figure 2).

3.3 | Stratification Strategies

The predictive ability of the same ML models was computed on two clusters of patients, defined according to age and NEDA-3.

The cluster of patients aged 45 years or less consisted of 583 patients, of whom 59 (10%) showed disease progression at T4. The

TABLE 1 | Demographic and clinical characteristics at baseline.

Feature name		Counts/statistics
F, *n* (%)		501 (70%)
Disease duration (years), mean ± SD		1.8 ± 2.8 25th percentile: 0.18, 75th percentile: 2.03
Age at onset (years), mean ± SD		34.57 ± 11.23
Baseline EDSS, median (min–max)		1.5 (0.0–7.5)
Baseline MS Phenotype, *n* (%)	CIS	102 (14.1%)
	RRMS	575 (79.6%)
	PMS-A	36 (5%)
	PMS-NA	9 (1.3%)
Disease onset–to-treatment time span *n* (%)	Treated within 1 year	319 (44.36%)
	Treated within 3 years	174 (24.20%)
	Treated after 3 years	132 (18.35%)
	No treatment	94 (13.07%)
Onset topography *n* (%)	Motor Supratentorial	218 (30.31%)
	Sensory Supratentorial	108 (15.02%)
	Infratentorial	156 (21.69%)
	Visual	178 (24.75%)
	Spinal	119 (16.55%)

Abbreviations: CIS, clinically isolated syndrome; EDSS, Expanded Disability Status Scale; PMS-A, active progressive multiple sclerosis; PMS-NA, non-active progressive multiple sclerosis; RRMS, relapsing–remitting multiple sclerosis.

preprocessing steps applied to this specific subset resulted in the selection of 31 attributes (16 at baseline, 5 at T1, 5 at T2, 5 at T3). For the attributes description and descriptive statistics, see Tables S2 and S4.

Among all tested classifiers (Table S4), Random Forest achieved the best predictive performance with AUC-ROC of 0.77 ± 0.05, MCC of 0.26 ± 0.11 and F1-Score of 0.34 ± 0.1 (mean ± SD) (Figure 3). The most informative attributes for disease progression in the considered subgroup included: EDSS at T3, age at symptoms onset, lag between symptoms onset and first evaluation, EDSS at T1 (Figure 3).

TABLE 2 | Classification results of the randomized nested grid search cross validation (RNGCV) analysis.

Classifier	MCC	F1-Score	AUC-ROC
ETC	0.20 ± 0.12	0.29 ± 0.11	0.69 ± 0.07
LR	0.11 ± 0.08	0.24 ± 0.05	0.61 ± 0.07
RF	0.28 ± 0.1	0.36 ± 0.08	0.75 ± 0.06
SVC	0.17 ± 0.08	0.27 ± 0.07	0.59 ± 0.06
XGB	0.27 ± 0.09	0.35 ± 0.06	0.74 ± 0.06

Note: The performance of five Machine learning (ML) models—that is, ETC, extra trees classifier; LR, logistic regression, RF, random forest, SVC, support vector classifier, XGB, extreme gradient boosting—was assessed using Matthew Correlation Coefficient (MCC), F1-Score and Area Under the Curve Receiving Operating Characteristic (AUC-ROC).
The reported results represent the mean ± standard deviation of each classification metric computed over the 30 iterations of the RNGCV analysis. Results for the best-performing model are highlighted in green.

NEDA-3, that is, no evidence of disease activity [24, 25] is a well-established indicator in MS treatment and an outcome measure in clinical and research settings. NEDA-3 patients in the considered time interval have no relapses, no increase in EDSS and neither new nor active (enhancing or new lesions) lesions on their MRI scans.

For this analysis, only patients fulfilling the definition of NEDA-3 (i.e., no relapses, no MRI activity and EDSS worsening) at the T3 time interval were selected. The preprocessing steps and the RNGCV strategy described in the main text were applied to the resulting subset of data to predict PIRA at T4.

Patients with no evidence of disease activity (NEDA-3) at T3 were 311, of whom 35 (11%) had PIRA at T4. The application of the aforementioned preprocessing steps to this subset of patients led to a preprocessed dataset that consisted of 27 predictors (see Table S3). Results of all tested classifiers are reported in Table S5 while the AUC-ROC curve and permutation importances of the best performing classifier (RF) are depicted in Figure 4.

The best performing model across iterations was Random Forest, with an average AUC of 0.8 ± 0.09. The most informative features highlighted by the explainability analysis were EDSS score at T3, time interval between symptom onset and first evaluation and brainstem involvement at disease onset.

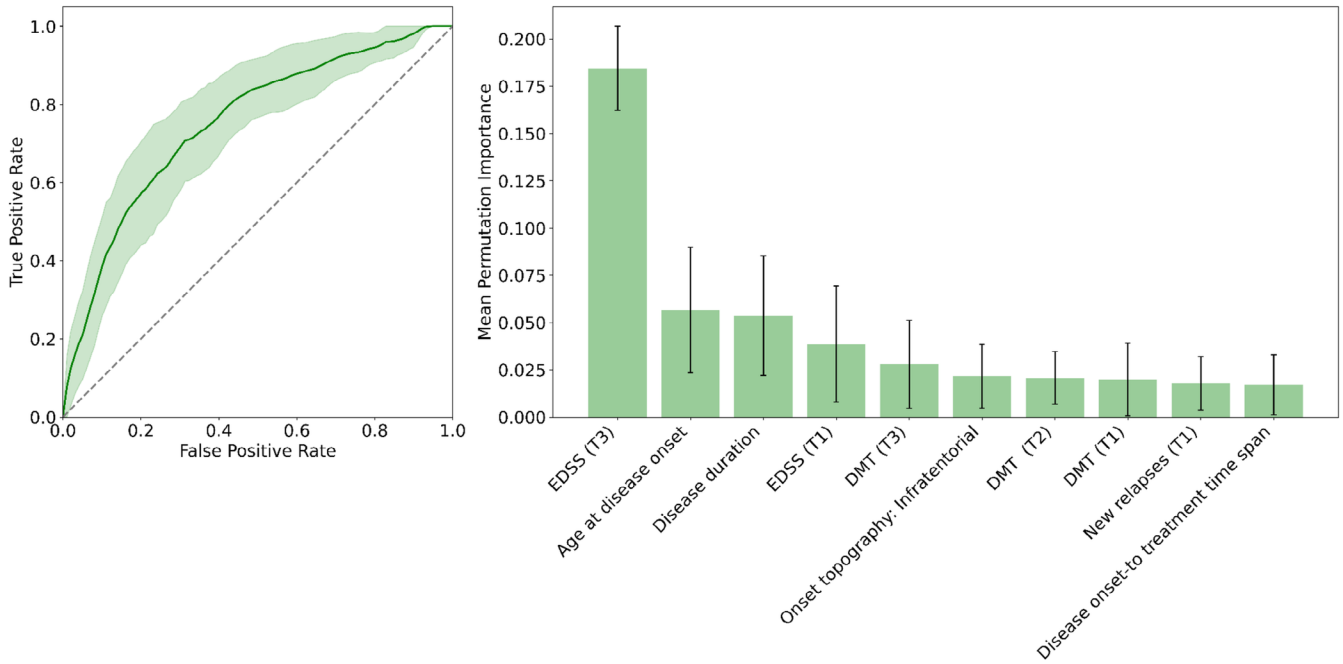

FIGURE 2 | Classification results for the random forest (RF) classifier on the main analysis. The Left Panel illustrates the AUC-ROC curve for the RF classifier. The solid green line represents the mean AUC across the 30 iterations of the Randomized Nested Grid Cross-Validation (RNGCV) analysis, while the surrounding shaded area depicts the standard deviation (SD). The dashed line represents a random guess baseline (AUC = 0.5). The horizontal axis corresponds to the False Positive Rate (FPR), and the vertical axis corresponds to the True Positive Rate (TPR). The right panel exhibits the most predictive features of the RF classifier. Vertical bars indicate the average permutation importance values, while the error bars represent the standard deviation computed across the 30 iterations of the RNGCV analysis.

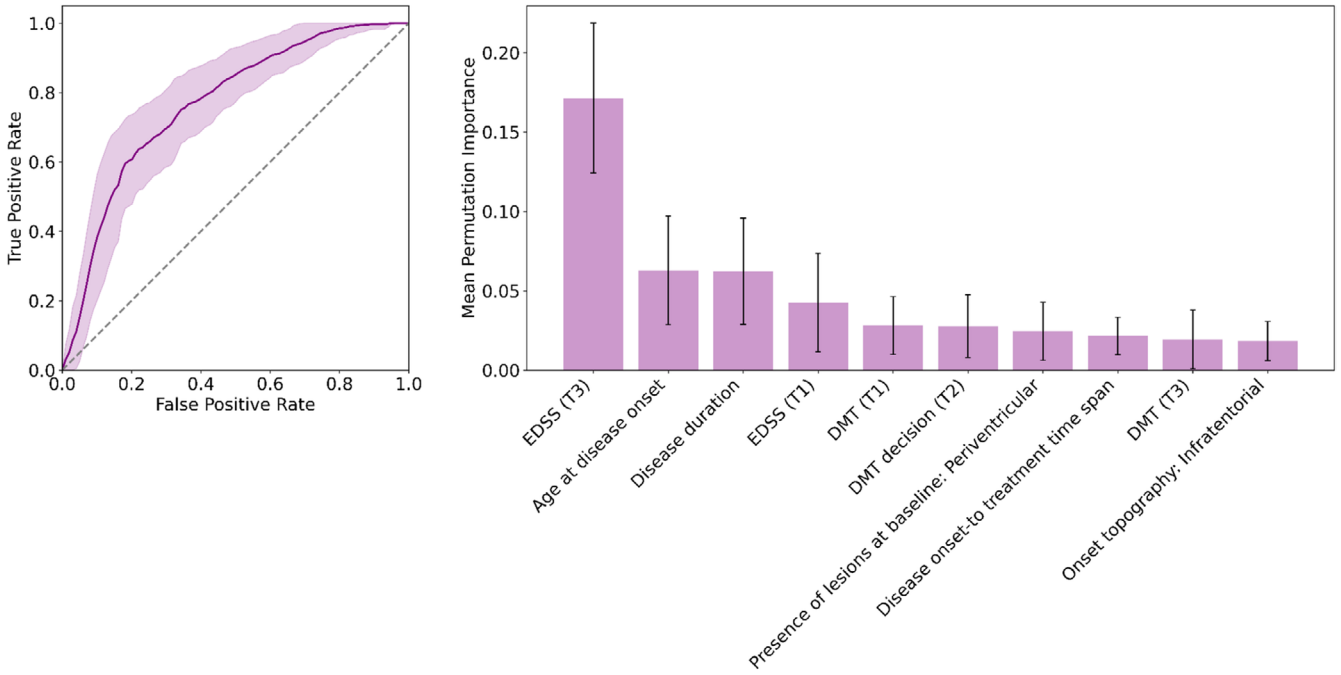

FIGURE 3 | Classification results for the random forest (RF) classifier on subset of patients aged 45 years or less. The Left Panel illustrates the AUC-ROC curve for the RF classifier. The solid purple line represents the mean AUC across the 30 iterations of the Randomized Nested Grid Cross-Validation (RNGCV) analysis, while the surrounding shaded area depicts the standard deviation (SD). The dashed line represents a random guess baseline (AUC = 0.5). The horizontal axis corresponds to the False Positive Rate (FPR), and the vertical axis corresponds to the True Positive Rate (TPR). The right panel exhibits the most predictive features of the RF classifier. Vertical bars indicate the average permutation importance values, while the error bars represent the standard deviation computed across the 30 iterations of the RNGCV analysis.

4 | Discussion

Our study supports the feasibility of applying ML techniques in predicting PIRA in pwMS in clinical routine, with a good level of accuracy. Importantly, although MS progression is inherently stochastic, our results proved that relevant historical demographic, clinical and radiological data, collected as part of routine clinical care, can lead to high discrimination performance and good calibration.

During two years of follow-up, about 13% of patients in our cohort experienced PIRA, in agreement with data reported in the literature [26, 27]. Among the different variables analyzed in this study, demographic and clinical data emerged as the most important predictors of PIRA in our model, namely EDSS at T3, age at onset and disease duration. On the other hand, ML analysis did not show a significant impact of MRI in predicting PIRA in our cohort of pwMS, with just a limited contribution of the amount of periventricular white matter (WM) damage at baseline MRI. These findings are in line with previous studies (typically prospective studies in which MRI data are included in all records) which indicate that clinical data have discriminative values for prognosis, even when used together with the results of MRI images [28, 29].

Furthermore, recent studies based on ML approaches suggest that MRI features collected in clinical practice can add relatively marginal predictivity to ML models compared to clinical data [12, 15, 16, 30]. This finding could be explained by the clinico-radiological paradox of MS [31] and seems particularly true for conventional MRI data, which seem to fail to capture the diffuse effect of the disease beyond focal lesions. Conversely, the use of post-processing MRI data [32, 33] such as T2 total lesion volume [34] gray matter evaluation [35], thalamic volume [36] seems to perform significantly better than clinical data [14, 15]. The lack of data regarding other more recent MRI biomarkers, such as paramagnetic rim lesions, gray matter damage (both cortical and subcortical) or spinal cord atrophy is another possible explanation [37]. Besides, recent studies suggest that even additional features, such as "omics" techniques and genetic analysis, may provide limited value over clinical and MRI data for predicting disability progression [16].

Compared to previous works using ML analysis in pwMS to predict disease progression [13, 14], our cohort was composed of patients at a very early stage of the disease. We thus considered those predictors that are available at disease diagnosis and during the first two years of follow-up, developing a prediction model particularly suited for routine clinical practice, useful to support tailoring and personalization of DMTs choice during the optimal window of treatment opportunity.

Our model confirms the well-known role of age, EDSS at disease onset as PIRA predictors [26, 27, 38]. It also consolidates the contribution of disease duration and time lag between onset and treatment start in determining disability progression. These last parameters should be interpreted as an indirect indicator of accurate patient stratification and timely initiation of proper treatment, strengthening previous evidence [39–42] regarding the paramount importance of early treatment initiation. This is also reinforced by the prognostic

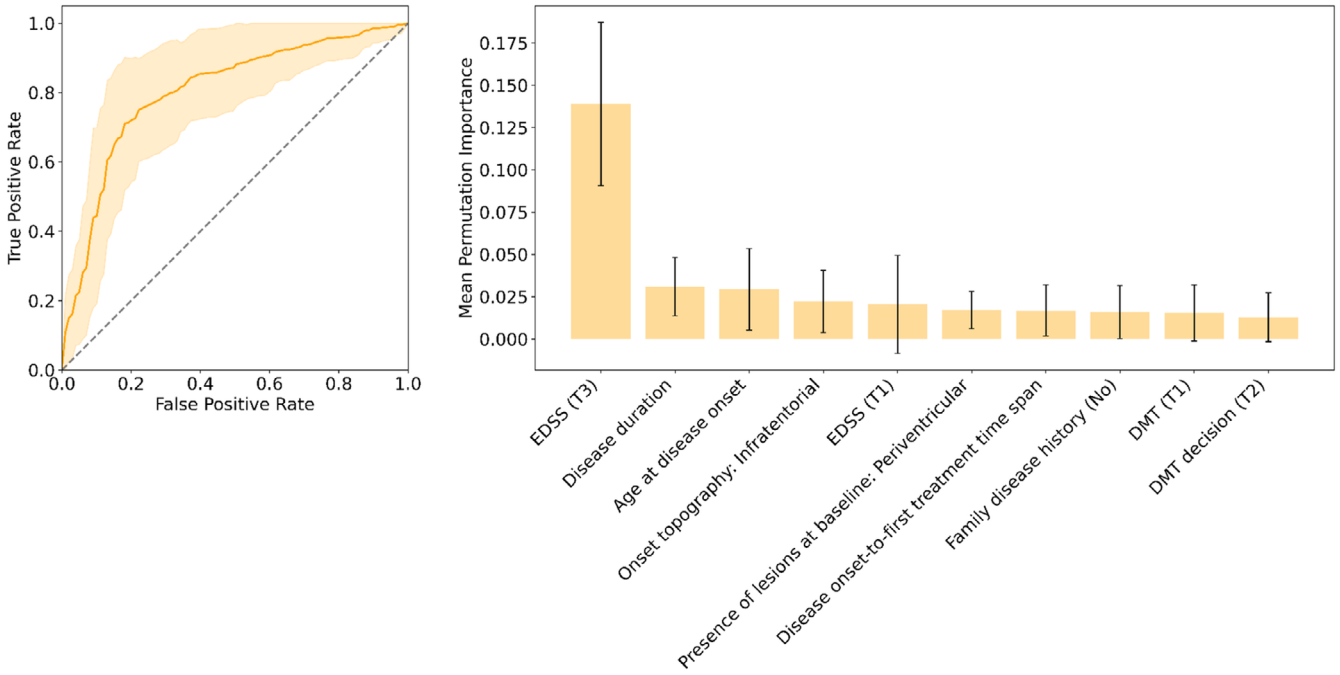

FIGURE 4 | Classification results for the random forest (RF) classifier on subset of patients with NEDA positive at T3. The Left Panel illustrates the AUC-ROC curve for the RF classifier. The solid orange line represents the mean AUC across the 30 iterations of the Randomized Nested Grid Cross-Validation (RNGCV) analysis, while the surrounding shaded area depicts the standard deviation (SD). The dashed line represents a random guess baseline (AUC = 0.5). The horizontal axis corresponds to the False Positive Rate (FPR), and the vertical axis corresponds to the True Positive Rate (TPR). The right panel exhibits the most predictive features of the RF classifier. Vertical bars indicate the average permutation importance values, while the error bars represent the standard deviation computed across the 30 iterations of the RNGCV analysis.

role of DMT, which also appeared to be a predictor of PIRA, although with a weaker performance compared to age at onset, disease duration and EDSS.

To further deepen the reliability of our results, ML techniques were applied to two different patients' subgroups, specifically patients younger than 45 years and patients with NEDA-3 after a 2-year follow-up. Both sub-analyses showed good accuracy, comparable to that observed in the overall cohort and confirmed the predictive role of EDSS, age at symptoms onset and disease duration. However, they differ in terms of secondary predictors and underline different nuances based on the subgroup considered. In patients younger than 45 years, the role of DMTs and the time lag between disease onset and treatment start gained more relevance in assessing individual risk of PIRA, corroborating the increasing data supporting the prompt initiation of treatment [42, 43].

In the subset of NEDA-3 patients, clinical presentation, namely infratentorial symptoms at onset, appeared to have a greater role as an early predictor of PIRA. Even though infratentorial involvement in MS is a well-known risk factor for poor prognosis [44], this parameter, when applied to patient with NEDA-3 in the very first years of disease, and therefore with apparent disease stability, might contribute to early identify those at higher risk of disability progression and therefore drive modification of the therapeutic strategy.

When interpreting the results of our study, some limitations have to be taken into account. Firstly, our model was not validated on an external cohort due to the lack of publicly available dataset with a comparable set of features derived from routine clinical practice. However, to ensure the stability of the results, an intra-study validation of ML models was performed by means of a RNGCV strategy. By training and testing each model on a different combination of input data, this cross-validation strategy ensured that both the reported performances and the set of informative variables were robust independently on the specific composition of the training set. Secondly, in the RNGCV strategy it was preferred to limit the feature engineering to favor a more objective data driven approach. A different formulation of some variables or different data processing techniques (such as different approaches for handling missing values) could have brought out aspects of interest even if this clashes with the problem constituted by the class distribution of each categorical feature. Thirdly, due to the retrospective nature of this study, less granularity of radiological, compared to clinical data, could have affected their predictive potential.

Furthermore, our outcome was a non-confirmed PIRA, due to lack of a confirmation timepoint. By definition, PIRA, did not consider radiological evaluation; besides disability accrual was evaluated only through EDSS while instruments such as the Multiple Sclerosis Functional Composite (MSFC) [45] could achieve higher sensitivity. It has to be noted, however, that previous studies using other clinical endpoints such as 9-hole peg test, Timed 25-Foot Walk and Single Digit Modalities Test modifications did not demonstrate a clear superiority in the prediction performance [15]. Finally, although predicting MS progression over a three-year follow-up is clinically relevant, we fully acknowledge that extending this duration, incorporating additional clinically attainable variables (e.g., non-conventional MRI

biomarkers or omics-based immunological profiling), as well as leveraging multimodal data integration across clinical, imaging, and molecular domains, will be essential to enhance our models' predictive power and therefore their clinical relevance.

5 | Conclusions

Despite the above limitations, our results support the feasibility of applying ML techniques to determine the prediction of PIRA in early stage pwMS, using only data generally available in clinical practice. These results would allow subsequent tailoring of treatment and definitely a better long-term prognosis.

Disability prediction can be further enhanced when considering specific subgroups of patients with homogeneous features, such as those with younger age or NEDA-3 score.

Acknowledgments

The authors would like to thank all the patients involved in the NeuroArtP3 project, as well as the colleagues from the different institutions and departments contributing to this initiative and the Italian Ministry of Health, for financing NET-2018-12366666.

Conflicts of Interest

The authors declare no conflicts of interest.

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. C. Confavreux and S. Vukusic, "The Clinical Course of Multiple Sclerosis," *Handbook of Clinical Neurology* 122 (2014): 343–369, https://doi.org/10.1016/B978-0-444-52001-2.00014-5.

2. L. Kappos, J. S. Wolinsky, G. Giovannoni, et al., "Contribution of Relapse-Independent Progression vs Relapse-Associated Worsening to Overall Confirmed Disability Accumulation in Typical Relapsing Multiple Sclerosis in a Pooled Analysis of 2 Randomized Clinical Trials," *JAMA Neurology* 77, no. 9 (2020): 1132–1140, https://doi.org/10.1001/jamaneurol.2020.1568.

3. F. D. Lublin, D. A. Häring, H. Ganjgahi, et al., "How Patients With Multiple Sclerosis Acquire Disability," *Brain* 145, no. 9 (2022): 3147–3161, https://doi.org/10.1093/brain/awac016.

4. A. Scalfari, A. Traboulsee, J. Oh, et al., "Smouldering-Associated Worsening in Multiple Sclerosis: An International Consensus Statement on Definition, Biology, Clinical Implications, and Future Directions," *Annals of Neurology* 96, no. 5 (2024): 826–845, https://doi.org/10.1002/ana.27034.

5. E. Portaccio, A. Bellinvia, M. Fonderico, et al., "Progression Is Independent of Relapse Activity in Early Multiple Sclerosis: A Real-Life Cohort Study," *Brain* 145, no. 8 (2022): 2796–2805, https://doi.org/10.1093/brain/awac111.

6. M. Calabrese, P. Preziosa, A. Scalfari, et al., "Determinants and Biomarkers of Progression Independent of Relapses in Multiple Sclerosis," *Annals of Neurology* 96, no. 1 (2024): 1–20, https://doi.org/10.1002/ana.26913.

7. University of California, San Francisco MS-EPIC Team, B. A. C. Cree, J. A. Hollenbach, et al., "Silent Progression in Disease Activity-Free Relapsing Multiple Sclerosis," *Annals of Neurology* 85, no. 5 (2019): 653–666, https://doi.org/10.1002/ana.25463.

8. T. Kalincik, G. Cutter, T. Spelman, et al., "Defining Reliable Disability Outcomes in Multiple Sclerosis," *Brain* 138, no. Pt 11 (2015): 3287–3298, https://doi.org/10.1093/brain/awv258.

9. E. Monreal, J. I. Fernández-Velasco, R. Álvarez-Lafuente, et al., "Serum Biomarkers at Disease Onset for Personalized Therapy in Multiple Sclerosis," *Brain* 147, no. 12 (2024): 4084–4093, https://doi.org/10.1093/brain/awae260.

10. J. H. Chen and S. M. Asch, "Machine Learning and Prediction in Medicine – Beyond the Peak of Inflated Expectations," *New England Journal of Medicine* 376, no. 26 (2017): 2507–2509, https://doi.org/10.1056/NEJMp1702071.

11. D. Bzdok, N. Altman, and M. Krzywinski, "Statistics Versus Machine Learning," *Nature Methods* 15, no. 4 (2018): 233–234, https://doi.org/10.1038/nmeth.4642.

12. Y. Zhao, B. C. Healy, D. Rotstein, et al., "Exploration of Machine Learning Techniques in Predicting Multiple Sclerosis Disease Course," *PLoS One* 12, no. 4 (2017): e0174866, https://doi.org/10.1371/journal.pone.0174866.

13. E. De Brouwer, T. Becker, L. Werthen-Brabants, et al., "Machine-Learning-Based Prediction of Disability Progression in Multiple Sclerosis: An Observational, International, Multi-Center Study," *PLOS Digit Health* 3, no. 7 (2024): e0000533, https://doi.org/10.1371/journal.pdig.0000533.

14. S. Tommasin, S. Cocozza, A. Taloni, et al., "Machine Learning Classifier to Identify Clinical and Radiological Features Relevant to Disability Progression in Multiple Sclerosis," *Journal of Neurology* 268, no. 12 (2021): 4834–4845, https://doi.org/10.1007/s00415-021-10605-7.

15. S. Noteboom, M. Seiler, C. Chien, et al., "Evaluation of Machine Learning-Based Classification of Clinical Impairment and Prediction of Clinical Worsening in Multiple Sclerosis," *Journal of Neurology* 271, no. 8 (2024): 5577–5589, https://doi.org/10.1007/s00415-024-12507-w.

16. M. Andorra, A. Freire, I. Zubizarreta, et al., "Predicting Disease Severity in Multiple Sclerosis Using Multimodal Data and Machine Learning," *Journal of Neurology* 271, no. 3 (2024): 1133–1149, https://doi.org/10.1007/s00415-023-12132-z.

17. M. T. Law, A. L. Traboulsee, D. K. Li, et al., "Machine Learning in Secondary Progressive Multiple Sclerosis: An Improved Predictive Model for Short-Term Disability Progression," *Multiple Sclerosis Journal – Experimental, Translational and Clinical* 5, no. 4 (2019): 2055217319885983, https://doi.org/10.1177/2055217319885983.

18. M. F. Pinto, H. Oliveira, S. Batista, et al., "Prediction of Disease Progression and Outcomes in Multiple Sclerosis With Machine Learning," *Scientific Reports* 10, no. 1 (2020): 21038, https://doi.org/10.1038/s41598-020-78212-6.

19. C. H. Polman, S. C. Reingold, B. Banwell, et al., "Diagnostic Criteria for Multiple Sclerosis: 2010 Revisions to the McDonald Criteria," *Annals of Neurology* 69, no. 2 (2011): 292–302, https://doi.org/10.1002/ana.22366.

20. E. F. Patridge and T. P. Bardyn, "Research Electronic Data Capture (REDCap)," *Journal of the Medical Library Association* 106, no. 1 (2018): 142–144, https://doi.org/10.5195/jmla.2018.319.

21. J. Müller, A. Cagol, J. Lorscheider, et al., "Harmonizing Definitions for Progression Independent of Relapse Activity in Multiple Sclerosis: A Systematic Review," *JAMA Neurology* 80, no. 11 (2023): 1232–1245, https://doi.org/10.1001/jamaneurol.2023.3331.

22. F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-Learn: Machine Learning in Python," *Journal of Machine Learning Research* 12 (2011): 2825–2830.

23. B. L. Vollmer, A. B. Wolf, S. Sillau, J. R. Corboy, and E. Alvarez, "Evolution of Disease Modifying Therapy Benefits and Risks: An

Argument for De-Escalation as a Treatment Paradigm for Patients With Multiple Sclerosis," *Frontiers in Neurology* 12 (2022): 12, https://doi.org/10.3389/fneur.2021.799138.

24. C. Gasperini, L. Prosperini, M. Tintoré, et al., "Unraveling Treatment Response in Multiple Sclerosis:A Clinical and MRI Challenge," *Neurology* 92, no. 4 (2019): 180–192, https://doi.org/10.1212/WNL.0000000000006810.

25. L. Prosperini, S. Ruggieri, S. Haggiag, C. Tortorella, C. Pozzilli, and C. Gasperini, "Prognostic Accuracy of NEDA-3 in Long-Term Outcomes of Multiple Sclerosis," *Neurology Neuroimmunology & Neuroinflammation* 8, no. 6 (2021): e1059, https://doi.org/10.1212/NXI.0000000000001059.

26. C. Tur, P. Carbonell-Mirabent, Á. Cobo-Calvo, et al., "Association of Early Progression Independent of Relapse Activity With Long-Term Disability After a First Demyelinating Event in Multiple Sclerosis," *JAMA Neurology* 80, no. 2 (2023): 151–160, https://doi.org/10.1001/jamaneurol.2022.4655.

27. E. Portaccio, M. Betti, E. De Meo, et al., "Progression Independent of Relapse Activity in Relapsing Multiple Sclerosis: Impact and Relationship With Secondary Progression," *Journal of Neurology* 271, no. 8 (2024): 5074–5082, https://doi.org/10.1007/s00415-024-12448-4.

28. V. Wottschel, D. C. Alexander, P. P. Kwok, et al., "Predicting Outcome in Clinically Isolated Syndrome Using Machine Learning," *NeuroImage: Clinical* 7 (2014): 281–287, https://doi.org/10.1016/j.nicl.2014.11.021.

29. Y. Yoo, L. Y. W. Tang, D. K. B. Li, et al., "Deep Learning of Brain Lesion Patterns and User-Defined Clinical and MRI Features for Predicting Conversion to Multiple Sclerosis From Clinically Isolated Syndrome," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 7 (2017): 250–259.

30. R. Seccia, D. Gammelli, F. Dominici, et al., "Considering Patient Clinical History Impacts Performance of Machine Learning Models in Predicting Course of Multiple Sclerosis," *PLoS One* 15, no. 3 (2020): e0230219, https://doi.org/10.1371/journal.pone.0230219.

31. F. Barkhof, "The Clinico-Radiological Paradox in Multiple Sclerosis Revisited," *Current Opinion in Neurology* 15, no. 3 (2002): 239–245, https://doi.org/10.1097/00019052-200206000-00003.

32. R. Bakshi, B. C. Healy, S. L. Dupuy, et al., "Brain MRI Predicts Worsening Multiple Sclerosis Disability Over 5 Years in the SUMMIT Study," *Journal of Neuroimaging* 30, no. 2 (2020): 212–218, https://doi.org/10.1111/jon.1268.

33. A. Cagol, S. Schaedelin, M. Barakovic, et al., "Association of Brain Atrophy With Disease Progression Independent of Relapse Activity in Patients With Relapsing Multiple Sclerosis," *JAMA Neurology* 79, no. 7 (2022): 682–692, https://doi.org/10.1001/jamaneurol.2022.1025.

34. C. Elliott, S. Belachew, J. S. Wolinsky, et al., "Chronic White Matter Lesion Activity Predicts Clinical Progression in Primary Progressive Multiple Sclerosis," *Brain* 142 (2019): 2787–2799.

35. A. Eshaghi, F. Prados, W. J. Brownlee, et al., "Deep Gray Matter Volume Loss Drives Disability Worsening in Multiple Sclerosis," *Annals of Neurology* 83, no. 2 (2018): 210–222, https://doi.org/10.1002/ana.25145.

36. M. Dwyer, C. Lyman, H. Ferrari, et al., "DeepGRAI (Deep Gray Rating via Artificial Intelligence): Fast, Feasible, and Clinically Relevant Thalamic Atrophy Measurement on Clinical Quality T2-FLAIR MRI in Multiple Sclerosis," *NeuroImage: Clinical* 30 (2021): 102652, https://doi.org/10.1016/j.nicl.2021.102652.

37. H. Yousef, B. Malagurski Tortei, and F. Castiglione, "Predicting Multiple Sclerosis Disease Progression and Outcomes With Machine Learning and MRI-Based Biomarkers: A Review," *Journal of Neurology* 271, no. 10 (2024): 6543–6572.

38. L. Lorefice, O. E. Ferraro, G. Fenu, et al., "Late-Onset Multiple Sclerosis: Disability Trajectories in Relapsing-Remitting Patients of the Italian MS Registry," *Journal of Neurology* 271, no. 4 (2024): 1630–1637, https://doi.org/10.1007/s00415-023-12152-9.

39. G. Giovannoni, H. Butzkueven, S. Dhib-Jalbut, et al., "Brain Health: Time Matters in Multiple Sclerosis," *Multiple Sclerosis and Related Disorders* 9, no. Suppl 1 (2016): S5–S48, https://doi.org/10.1016/j.msard.2016.07.003.

40. T. A. Chalmer, L. M. Baggesen, M. Nørgaard, et al., "Early Versus Later Treatment Start in Multiple Sclerosis: A Register-Based Cohort Study," *European Journal of Neurology* 25, no. 10 (2018): 1262-e110, https://doi.org/10.1111/ene.13692.

41. A. He, B. Merkel, J. W. L. Brown, et al., "Timing of High-Efficacy Therapy for Multiple Sclerosis: A Retrospective Observational Cohort Study," *Lancet Neurology* 19, no. 4 (2020): 307–316, https://doi.org/10.1016/S1474-4422(20)30067-3.

42. A. Cobo-Calvo, C. Tur, S. Otero-Romero, et al., "Association of Very Early Treatment Initiation With the Risk of Long-Term Disability in Patients With a First Demyelinating Event," *Neurology* 101, no. 13 (2023): e1280–e1292, https://doi.org/10.1212/WNL.0000000000207664.

43. K. Selmaj, B. A. C. Cree, M. Barnett, A. Thompson, and H. P. Hartung, "Multiple Sclerosis: Time for Early Treatment With High-Efficacy Drugs," *Journal of Neurology* 271, no. 1 (2024): 105–115, https://doi.org/10.1007/s00415-023-11969-8.

44. M. Tintore, A. Rovira, G. Arrambide, et al., "Brainstem Lesions in Clinically Isolated Syndromes," *Neurology* 75, no. 21 (2010): 1933–1938, https://doi.org/10.1212/WNL.0b013e3181feb26f.

45. M. L. Martínez-Ginés, A. Esquivel, Y. H. Hernández, L. A. Alvarez-Sala, and J. Benito-León, "Investigating the Relationship Between Multiple Sclerosis Disability and Driving Performance: A Comparative Study of the Multiple Sclerosis Functional Composite and Expanded Disability Status Scale," *Clinical Neurology and Neurosurgery* 244 (2024): 108431, https://doi.org/10.1016/j.clineuro.2024.108431.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Table S1:** Descriptive statistics of remaining variables of the main dataset. **Table S2:** Descriptive statistics of patients aged 45 years or less. **Table S3:** Descriptive statistics of patients with NEDA condition at T3. **Table S4:** Classification results of the randomized nested grid search cross validation (RNGCV) analysis on subset of patients aged 45 years or less. **Table S5:** Classification results of the randomized nested grid search cross validation (RNGCV) analysis on subset of patients with NEDA-3 positive at T3.