



Can Machine Learning Detect and Predict Outcomes in Multiple Sclerosis? A Review of Diagnostic and Prognostic Approaches

✉ Murat Emec

Istanbul Nisantasi University, School of Civil Aviation, Department of Aviation Electrical and Electronics, Istanbul, Turkiye

Abstract

Multiple sclerosis (MS) diagnosis relies on clinical presentation, magnetic resonance imaging (MRI), and the McDonald criteria, whereas long-term management requires anticipating disability progression and symptom burden. This narrative review evaluates current evidence on machine learning (ML) approaches for (i) detecting and classifying MS and (ii) predicting disability progression and symptom severity in patients with established MS. Evidence was synthesized from 25 sources published between 2012 and 2026 (23 peer-reviewed journal publications, one arXiv preprint, and one article published in a journal outside the major biomedical indexes), encompassing MS diagnosis [MRI, optical coherence tomography (OCT), fluid biomarkers, and differential diagnosis] and prognostic outcome prediction (disability progression, symptom-specific severity, and context-specific changes in disability). MRI remains the predominant data source for ML-based MS detection: deep learning models applied to structural MRI achieved accuracies of 70.2-98.8% in distinguishing MS from healthy controls, whereas conventional ML models achieved accuracies of 83.7-90.0%. A broader meta-review of more than 100 studies reported diagnostic accuracies ranging from 61% to 97%, while differentiation of MS from neuromyelitis optica spectrum disorder achieved 74-88% accuracy, depending on data integration. For prognostic applications, tree-based ensemble models generally outperformed simpler regression models: a 231-patient cohort achieved an R^2 of 0.97 for two-year disability score prediction, although this estimate warrants cautious interpretation because of the strong autocorrelation of baseline Expanded Disability Status Scale scores. A 662-woman cohort achieved 85-88% classification accuracy for postpartum changes in disability, whereas a 23-patient cohort achieved 92.65% accuracy in predicting restless legs syndrome severity, with sleep quality identified as the leading predictor. Overall, ML can effectively detect MS-associated patterns in MRI and is increasingly being applied to OCT and fluid biomarkers. It can also generate clinically meaningful prognostic predictions after MS diagnosis, although categorical outcome prediction appears to generalize better than continuous regression. Dataset size, single-center study designs, and limited external validation remain the principal barriers to clinical deployment across both domains. Multimodal data integration and prospective multicenter validation therefore represent important priorities for future research.

Keywords: Multiple sclerosis, machine learning, deep learning, diagnosis, prognosis, disease progression

Introduction

Multiple sclerosis (MS) is a chronic inflammatory and neurodegenerative disease of the central nervous system, and achieving an accurate diagnosis remains a persistent clinical challenge. The 2017 McDonald criteria, which integrate clinical findings with magnetic resonance imaging (MRI) evidence of lesion dissemination in space and time, have improved diagnostic sensitivity and facilitated earlier diagnosis;

however, their accurate application remains highly dependent on clinician expertise, and misdiagnosis continues to occur, particularly when MS must be differentiated from other white matter or inflammatory conditions (1).

Over the past decade, researchers have developed machine learning (ML) and deep learning (DL) methods to assist in this diagnostic process. Early applications focused primarily on automating the labor-intensive task of MS lesion segmentation

Address for Correspondence: Asst. Prof., Murat Emec, Istanbul Nisantasi University, Department of Aviation Electrical and Electronics, Istanbul, Turkiye

E-mail: murat.emc@gmail.com **ORCID-ID:** orcid.org/0000-0002-9407-1728

Received: 13.07.2026 **Accepted:** 17.09.2026 **Publication Date:** 15.10.2026

Cite this article as: Emec M. Can machine learning detect and predict outcomes in multiple sclerosis? A review of diagnostic and prognostic approaches. J Mult Scler Res. 2026;6:71-78



on MRI, which has traditionally been performed manually and is subject to substantial inter-rater variability (2,3). As computational methods have advanced, their applications have expanded to full diagnostic classification, including differentiation of MS from healthy controls and, more recently, from clinically similar disorders such as neuromyelitis optica spectrum disorder (NMOSD) (1,4). In parallel, researchers have increasingly explored non-MRI modalities, including optical coherence tomography (OCT) and blood and cerebrospinal fluid (CSF) biomarkers, as additional and less invasive substrates for ML-based detection (5,6).

Beyond diagnosis, ML has also been applied to patients with established MS to forecast disability accumulation and symptom burden, i.e., tasks that are prognostic rather than diagnostic. Nevertheless, these applications share the same underlying methodology and are frequently discussed alongside diagnostic applications in the literature (4,7). Disability, most commonly quantified using the Expanded Disability Status Scale (EDSS), is challenging to predict because it reflects the heterogeneous and non-linear accumulation of impairment across multiple neurological domains. Recent cohort studies have applied ensemble ML methods to predict future EDSS scores, disease-specific symptom severity, and changes in disability during clinically important periods, such as the postpartum period, often achieving higher classification accuracy than continuous regression approaches for disability scores.

Despite the growing body of research across both domains, the translation of ML tools into routine clinical practice remains limited. Reported performance varies considerably among studies, and this heterogeneity is frequently attributed to differences in dataset size, imaging or clinical feature protocols,

and the selection of comparison or outcome groups (1,7). The present review synthesizes current evidence on ML-based detection and classification of MS together with ML-based prediction of disability progression and symptom severity to clarify (i) which data modalities and model families have demonstrated the strongest performance in each domain, (ii) how performance differs across tasks of increasing clinical difficulty, and (iii) which methodological barriers must be addressed before these tools can be considered for routine clinical deployment.

This is a narrative, rather than a systematic, review. Sources were identified through purposive searches of PubMed and Google Scholar (last search: July 2026) using combinations of the terms “multiple sclerosis,” “machine learning,” “deep learning,” “diagnosis,” “prognosis,” and “disability progression,” supplemented by manual screening of the reference lists of recently identified reviews. We retained 25 sources published between 2012 and 2026 based on their direct relevance to ML-based MS diagnosis or disability/symptom prognosis, methodological clarity, and, for primary studies, availability of quantitative performance metrics. This purposive approach does not guarantee exhaustive coverage and reflects the authors’ judgment regarding representative and influential work rather than a systematic sample.

Chronological Evolution of the Field

Table 1 summarizes 25 key sources in chronological order, illustrating how the field has evolved from early, narrowly focused lesion-segmentation tools to broader diagnostic frameworks and, more recently, to prognostic and disability-prediction applications.

Table 1. Overview of key studies and reviews on machine learning applications for MS diagnosis and prognosis (2012-2026)

Year	Author(s)	Data modality	Method	Focus	Contribution/milestone
2022	Filippi et al. (1)	MRI	Narrative review	Diagnostic work-up	Summarized the evolution of the McDonald criteria and AI-detectable markers
2018	Danelakis et al. (2)	MRI	Segmentation survey	Lesion segmentation	Cataloged the growth of available methods and highlighted the lack of standardization
2012	Lladó et al. (3)	MRI	Segmentation review	Lesion segmentation	Early comprehensive survey of automated segmentation methods
2021	Moazami et al. (4)	MRI	SVM, RF, CNN	Diagnosis, staging	Proposed a four-category taxonomy for ML applications in MS imaging
2022	Toosy and Eshaghi (5)	OCT	ML review	Diagnosis (OCT)	Identified OCT-based ML as a promising but less mature diagnostic modality
2024	Arrambide et al. (6)	Blood/CSF	Narrative review	Fluid biomarkers	Highlighted the underexplored potential of AI applied to fluid biomarkers
2024	Pilehvari et al. (7)	MRI + clinical	Meta-review (103 studies)	Diagnosis, risk	Quantified performance ranges across ML/DL studies
2012	Lladó et al. (8)	Serial MRI	Change detection	Lesion segmentation	Demonstrated the feasibility of automated new-lesion detection using longitudinal scans
2012	Jacob et al. (9)	Clinical/serology	Diagnostic criteria	Differential diagnosis	Defined the NMOSD concept, which was later used as an ML comparator class

Table 1. Continued					
Year	Author(s)	Data modality	Method	Focus	Contribution/milestone
2021	Shoeibi et al. (10)	MRI	CNN/DL	Automated detection	Synthesized DL architectures used for MS detection
2021	Andropova et al. (11)	MRI	LDA, ANN, SVM	Computer-aided diagnosis	Reviewed CAD components designed to support the radiologist workflow
2021	Petzold et al. (12)	OCT	Consensus + AI	Imaging quality control	Extended the OSCAR-IB criteria for AI-based OCT analysis
2022	Aslam et al. (13)	MRI + clinical	SVM, RF, CNN	Diagnosis	Identified leading approaches from 2011-2022 and highlighted translational barriers
2022	Bonacchi et al. (14)	Multimodal	Narrative review	Clinical practice	Examined the real-world translation of AI-based diagnostic tools
2023	Shah et al. (15)	MRI	Bibliometric review	Lesion detection	Characterized methodological and citation trends over the preceding decade
2023	Siriratnam et al. (16)	Clinical/imaging	Narrative review	Differential diagnosis	Updated advances in NMOSD relevant to MS-mimic tasks
2024	Collorone et al. (17)	MRI	Narrative review	Key challenges	Highlighted heterogeneity, generalizability, and validation gaps
2024	Amin et al. (18)	Multimodal	Narrative review	Overview	Provided a broad synthesis spanning MS diagnosis and management
2025	Szekely-Kohn et al. (19)	MRI	Narrative review	MRI interpretation	Identified U-Net as the dominant architecture and highlighted emerging lesion subtypes (PRL, CVS)
2025	Kuntal et al. (20)	MRI	CNN	Detection	Described a technical CNN-based pipeline for MS detection
2025	Thabet et al. (21)	MRI	ML/DL comparison	Classification, segmentation	Provided a recent comparison of classification and segmentation methods
2025	Abdul-Salam et al. (22)	OCT/retina	DL	Diagnosis (retina)	Established retinal imaging as a potential source of non-invasive biomarkers
2025	Kaya et al. (23)*	Clinical, questionnaire	Decision Tree, RF, XGBoost	Symptom severity	Developed the first ML model for RLS severity in an MS cohort
2026	Ozdogar et al. (24)*	Clinical, demographic, and cognitive	XGBoost, RF, Linear Reg	Disability progression	Reported high-accuracy EDSS prediction without imaging ($R^2=0.97$)
2026	Emec et al. (25)*	Clinical, reproductive, obstetric	RF, XGBoost, Elastic Net, SVC	Postpartum disability	Developed the first model using reproductive history for MS prognosis

Asterisk () denotes primary cohort studies; the remaining twenty-two entries are literature reviews or methodological surveys. ANN: Artificial neural network, CAD: Computer-aided diagnosis, CNN: Convolutional neural network, CSF: Cerebrospinal fluid, CVS: Central vein sign, EDSS: Expanded Disability Status Scale, DL: Deep learning, LDA: Linear discriminant analysis, ML: Machine learning, MRI: Magnetic resonance imaging, MS: Multiple sclerosis, NMOSD: Neuromyelitis optica spectrum disorder, OCT: Optical coherence tomography, PRL: Paramagnetic rim lesion, RF: Random forest, RLS: Restless legs syndrome, SVC: Support vector classifier, SVM: Support vector machine, AI: Artificial intelligence

As listed in Table 1, the earliest work (2012) focused almost exclusively on automated lesion segmentation using MRI (3,8,9), reflecting the practical need to reduce the burden associated with manual lesion annotation. Between 2018 and 2021, the field shifted toward DL architectures capable of full diagnostic classification rather than segmentation alone, alongside early consensus efforts to standardize non-MRI modalities, such as OCT (2,10,12). From 2022 onward, the literature has increasingly emphasized translational challenges, including generalizability, differential diagnosis, and multimodal data integration, rather than purely technical improvements in model performance (1,13,14,17). The most recent sources (2024-2026) reflect a further expansion in scope along two directions: toward fluid biomarkers and retinal imaging on the diagnostic side (6,19,22)

and toward ensemble ML models for predicting disability progression, postpartum disability change, and symptom-specific severity on the prognostic side (23-25). This latter shift suggests that, once a diagnosis is established, the same model families originally developed for classification are increasingly being repurposed to address a clinically distinct question: how will this patient's disease evolve? It should be noted that 22 of the 25 sources summarized in Table 1 are literature reviews or methodological surveys rather than primary studies. In particular, the performance ranges reported in Table 2 are derived from two review articles (1,7) rather than recalculated from the primary studies they cite and should therefore be interpreted as review-level summaries rather than independently verified findings from the primary studies.

Table 2. Summary of prognostic and disability/symptom-severity prediction studies included in this review

Study	n	Outcome predicted	Best model/performance	Top predictors
Kaya et al. (23)	23	Restless legs syndrome severity	XGBoost: MAE: 1.94, accuracy: 92.65%	Sleep quality (PSQI)
Ozdogar et al. (24)	231	EDSS at 2-year follow-up	XGBoost (CorBoost): R ² =0.97; MAE: 0.24	Baseline EDSS, disease duration, age
Emec et al. (25)	662 women/909 pregnancies	Postpartum disability change (Δ EDSS)	SVC (classification): accuracy: 88.5%; F1=0.87	Postpartum relapse, disease duration, age at pregnancy

The 88.5% figure is the accuracy of the single best-performing classification model (SVC) reported by Emec et al. (25); the 85-88% range cited in Section 4.3 spans all classification models evaluated in that study. The cohort comprised 662 women, with 909 pregnancies in total. As some women contributed more than one pregnancy, these pregnancy-level observations are not fully independent, which should be considered when interpreting the reported accuracy

EDSS: Expanded Disability Status Scale, MAE: Mean absolute error, PSQI: Pittsburgh Sleep Quality Index, SVC: Support vector classifier

ML Approaches for MS Diagnosis

MRI-based Classification: Conventional ML versus DL

MRI remains the principal data source for ML-based MS detection (1,4,10). Two broad methodological families predominate in this literature. DL approaches [most commonly convolutional neural networks (CNNs) and U-Net variants] operate directly on raw T2-weighted, fluid-attenuated inversion recovery, or susceptibility-weighted images and have achieved accuracies ranging from 70.2% to 98.8% in distinguishing MS from healthy controls (1). In contrast, conventional ML approaches [including support vector machines, random forests (RFs), and logistic regression] typically rely on engineered features derived from quantitative, diffusion-weighted, or resting-state functional MRI and have reported somewhat narrower accuracy ranges of 83.7-90.0% (1). Notably, the higher peak performance observed with DL models does not necessarily indicate consistent superiority, as performance in both methodological families depends heavily on dataset composition, imaging modality, and the definition of the diagnostic task (13,21).

The diagnostic reference standard itself is also evolving. The 2017 McDonald criteria, against which nearly all models discussed in this review were trained and validated, are themselves undergoing revision: proposed updates incorporate the central vein sign and paramagnetic rim lesions as imaging markers of MS-specific pathology, kappa free light chains as a fluid biomarker, and the optic nerve as an additional topography for demonstrating dissemination in space. Several of these candidate markers are precisely the targets of the lesion-subtype segmentation tools described in Section 3.2. Models trained and benchmarked under the current criteria may therefore require retraining or revalidation as the reference standard they were developed to approximate continues to evolve.

Automated Lesion Segmentation

Lesion segmentation is the most methodologically mature ML application in MS, with a research history extending back to at least the early 2010s (3,8). Manual lesion assessment is time-consuming and subject to inter-rater variability, which prompted the early adoption of automated approaches (2). U-Net-type CNN architectures now predominate in this subfield,

and current tools target not only conventional T2 lesions but also gadolinium-enhancing lesions, central vein sign-positive lesions, paramagnetic rim lesions, and cortical and spinal cord lesions (19). Despite this progress, no segmentation method has achieved universal acceptance in clinical practice, and expert expectations regarding reliability continue to exceed the performance delivered by most current algorithms (2,15). It is important to distinguish regulatory clearance from demonstrated clinical utility: several MS lesion quantification tools have obtained Conformite Europeenne marking or U.S. Food and Drug Administration (FDA) clearance, but such clearance primarily certifies technical performance and safety rather than demonstrating benefits to patient outcomes or constituting formal endorsement by clinical practice guidelines. Cleared tools also remain sensitive to scanner hardware, acquisition protocols, and field strength and typically require post-deployment performance monitoring when used outside the conditions under which they were validated.

Differential Diagnosis: MS versus Clinical Mimics

A clinically important extension of ML-based classification is the differentiation of MS from disorders with overlapping clinical presentations, most notably NMOSD (9,16). RF models using gray-matter measures alone have achieved approximately 74% accuracy in distinguishing MS from NMOSD. In contrast, multiparametric models incorporating MRI with clinical and neuropsychological data have achieved approximately 88% accuracy (1). This gap illustrates a recurring theme: models trained only to distinguish MS from healthy controls may perform considerably worse when applied to realistic differential-diagnosis comparisons, such as vascular white-matter lesions or other inflammatory disorders (1).

Beyond MRI: OCT, Retinal Imaging, and Fluid Biomarkers

The field is expanding beyond structural MRI. OCT, which quantifies retinal nerve fiber and ganglion cell layer thickness, has shown promising preliminary results for distinguishing MS from healthy controls. However, sample sizes remain small, and standardization of quality control is still evolving (5,12). Consensus efforts, such as the artificial intelligence (AI)-extended OSCAR-IB criteria (OCT Standardization of Image Acquisition and Analysis

in MS) criteria, aim to address this standardization gap (12). DL applied to retinal imaging is increasingly being positioned as a non-invasive, low-cost complement to neurological assessment (22). Separately, ML applied to blood and CSF biomarkers remains comparatively underexplored relative to imaging-based approaches, despite its potential to provide an additional minimally invasive diagnostic substrate (6).

Clinical Integration and Multimodal Approaches

Several reviews argue that multimodal integration, i.e., combining MRI with clinical, demographic, OCT, or fluid biomarker data, is the most credible path toward improved diagnostic utility, as no single marker captures the full heterogeneity of MS pathology (6,17,18). At the same time, reviews assessing real-world clinical translation caution that most current AI tools remain assistive rather than autonomous, requiring human oversight and further validation before routine deployment (14,19). This assistive positioning is consistent with the regulatory status of most tools in this space: regulatory clearance indicates that a tool performs a specified technical function safely and reproducibly, rather than demonstrating that it improves diagnostic accuracy or patient outcomes in clinical practice. Guideline endorsement of autonomous decision-making remains absent across the tools discussed in this review.

ML for Disability Progression and Symptom Prediction

Once a diagnosis of MS is established, a distinct but methodologically related question arises: how will the patient's disability and symptom burden evolve? Three recent cohort studies illustrate this prognostic application of ML, summarized chronologically in Table 2.

EDSS and Disability Progression Prediction

Ozdogar et al. (24) developed an XGBoost-based model (CorBoost) to predict 2-year EDSS scores in 231 people with MS, using 126 baseline demographic, clinical, physical, and cognitive features that were reduced to 42 through correlation-based feature selection. The model achieved a coefficient of determination (R^2) of 0.97 and a mean absolute error (MAE) of 0.24 EDSS points, outperforming the RF and linear regression comparators. Baseline EDSS, disease duration, and age were the most influential predictors. Notably, the model achieved this performance using only routinely collected clinical data, without MRI or fluid biomarker input. This very high R^2 should nonetheless be interpreted with caution because EDSS is strongly autocorrelated over a 2-year interval and baseline EDSS was the leading predictor; thus, much of the explained variance may reflect a patient's baseline status rather than genuine predictive insight into disease evolution. The reported MAE of 0.24 EDSS points is notably below the scale's own inter-rater measurement error, further suggesting that the model's practical incremental value over baseline disability alone may be more modest than the headline R^2 implies.

Symptom-specific Severity Prediction

Kaya et al. (23) applied DecisionTree, RF, and XGBoost models to predict restless legs syndrome (RLS) severity in a small cohort of 23 people with MS, using sleep, mobility, and quality-of-life questionnaire data. XGBoost performed best (MAE: 1.94; accuracy: 92.65%), with the Pittsburgh Sleep Quality Index emerging as the dominant predictor across all three algorithms, consistent with conventional correlation analysis. Reporting both a continuous error metric (MAE) and a categorical accuracy measure indicates that the authors evaluated continuous as well as categorical formulations of RLS severity; readers are referred to the primary source for the exact sex composition of the 23-person cohort. This study illustrates that ML can extend beyond global disability metrics to model specific, clinically burdensome MS symptoms, although the very small sample size limits generalizability.

Context-specific Disability Prediction: The Postpartum Period

Emec et al. (25) modeled postpartum disability change (Δ EDSS) in a large real-world cohort of 662 women with MS contributing 909 pregnancies, comparing regression models (RF, XGBoost, Elastic Net, Support Vector Regression) with classification models predicting improved, stable, or worsened disability categories. Classification models substantially outperformed regression models (test accuracy, 85-88%; F1: 0.84-0.87; versus test $R^2=0.31-0.39$, for regression), a pattern the authors attributed to the discretized, threshold-like distribution of clinically meaningful EDSS change. Postpartum relapse was the strongest predictor of disability worsening, followed by disease duration and age at pregnancy, with obstetric variables such as delivery type and breastfeeding contributing secondary predictive value.

It should be noted that references 23, 24, and 25 share substantial overlap in authorship and appear to draw on a largely common patient population and research group, so they are better interpreted as a coherent program of work from one center rather than as three independent replications. Independent multicenter prognostic studies are still needed before the pattern described below can be generalized with confidence. Across all three studies, tree-based ensemble methods (RF, XGBoost) consistently matched or outperformed simpler linear models, and categorical outcome framing tended to generalize better than continuous regression, mirroring a pattern also noted in the diagnostic literature, where task definition strongly influences reported performance (13,21).

Diagnostic and Prognostic Performance Summary

Table 2 summarizes the three prognostic and disability-prediction studies described above, whereas Table 3 consolidates diagnostic performance measures by task and data modality.

Table 4 summarizes qualitative risk-of-bias judgments for the prognostic studies; similarly, the diagnostic performance

measures in Table 3 vary in the completeness of their reporting. Table 5 makes this limitation explicit by summarizing, for each row of Table 3, the validation strategy, aggregate sample size, and class balance reported in the underlying source. When a source did not report one of these elements, we state this explicitly rather than leaving the field blank, because the absence of this information is itself relevant to interpreting the reported accuracy measures.

Challenges and Limitations

Several methodological limitations recur across both domains. Most studies rely on small, single-center datasets, thereby limiting generalizability across scanners, acquisition protocols, and patient populations; this limitation is particularly apparent in symptom-prediction studies involving cohorts as small as 23 patients (1,10,23). Diagnostic models are frequently validated

against healthy controls rather than clinically representative comparator groups, whereas prognostic regression models often exhibit substantial overfitting between training and test datasets (1,16,25). DL as well as ensemble tree-based models remain relatively opaque, which can hinder clinician confidence, interpretability, and regulatory evaluation (10,19). Finally, prospective multicenter validation, which is a prerequisite for broader clinical deployment, remains uncommon, with most available evidence derived from retrospective, single-center study designs (1,17,24,25).

To move this observation from assertion toward demonstration, Table 4 provides a brief, Prediction Model Risk Of Bias Assessment Tool (PROBAST)-informed qualitative appraisal of the three primary prognostic studies discussed in Section 4. This appraisal is based solely on information reported in the published articles

Table 3. Reported diagnostic performance of ML approaches for MS, by task and data modality

Diagnostic task	Data/model type	Accuracy (%)	Sensitivity (%)	Specificity (%)	Source
MS versus healthy controls	DL (CNN, U-Net) on structural MRI	70.2-98.8	—	—	Filippi et al. (1)
MS versus healthy controls	Conventional ML (SVM, RF, logistic regression) on quantitative MRI/DWI/rs-fMRI	83.7-90.0	—	—	Filippi et al. (1)
Overall ML/DL diagnostic performance	Mixed ML/DL, meta-review of 103 studies	61-97	60-98	60-98	Pilehvari et al. (7)
Classification studies specifically	Mixed ML/DL, same meta-review subset	71-97.9	78-97.3	65-99.3	Pilehvari et al. (7)
MS versus NMOSD (gray matter only)	RF	74	—	—	Filippi et al. (1)
MS versus NMOSD (multiparametric)	MRI + clinical + neuropsychological model	88	—	—	Filippi et al. (1)

“—” indicates that sensitivity or specificity was not reported separately from overall accuracy in the source study

MS: Multiple sclerosis, CNN: Convolutional neural network, DL: Deep learning, DWI: Diffusion-weighted imaging, ML: Machine learning, MRI: Magnetic resonance imaging, NMOSD: Neuromyelitis optica spectrum disorder, RF: Random forest, rs-fMRI: Resting-state functional magnetic resonance imaging, SVM: Support vector machine

Table 4. Brief PROBAST-informed qualitative risk-of-bias appraisal of the three primary prognostic studies discussed in Section 4

Study	Participants	Predictors	Outcome	Analysis	Overall judgment
Kaya et al. (23)	High concern (n=23; single-center convenience sample)	Low concern (clearly defined questionnaire and clinical predictors)	Low concern (RLS severity clearly defined)	High concern (no external validation; high model-to-sample-size ratio)	High risk of bias
Ozdogar et al. (24)	Unclear concern (single-center; recruitment not fully detailed)	Low concern (routinely collected clinical, demographic, and cognitive features)	Low concern (EDSS is a standardized, well-defined outcome)	High concern (no external validation; correlation-based feature selection from 126 to 42 features in n=231; strong baseline-outcome autocorrelation, as discussed in Section 4.1)	High risk of bias
Emec et al. (25)	Unclear concern (larger cohort, but repeated pregnancies per woman were not treated as clustered observations)	Low concern (reproductive, obstetric, and clinical predictors clearly defined)	Low concern (categorical Δ EDSS outcome clearly defined)	Unclear concern (no external validation reported; non-independence of repeated pregnancies)	Unclear risk of bias

Judgments reflect information available in the published reports and the summaries in Section 4 of this review; they were not derived from a full protocol-level PROBAST assessment

EDSS: Expanded Disability Status Scale, PROBAST: Prediction Model Risk of Bias Assessment Tool, RLS: Restless legs syndrome

Table 5. Validation strategies, sample sizes, and class balance underlying the diagnostic performance measures reported in Table 3

Diagnostic task	Data/model type	Validation strategy	Sample size	Class balance
MS versus healthy controls	DL (CNN, U-Net) on structural MRI	Not reported in source review; varies by underlying study	Not aggregated across underlying studies	Not reported
MS versus healthy controls	Conventional ML (SVM, RF, logistic regression) on quantitative MRI/DWI/rs-fMRI	Not reported in source review; varies by underlying study	Not aggregated across underlying studies	Not reported
Overall ML/DL diagnostic performance	Mixed ML/DL, meta-review of 103 studies	Mixed across the 103 studies reviewed; predominantly internal cross-validation or hold-out	Aggregated across 103 studies; not stated as a single figure	Not reported
Classification studies specifically	Mixed ML/DL, same meta-review subset	Mixed across studies reviewed; predominantly internal cross-validation or hold-out	Aggregated across studies; not stated as a single figure	Not reported
MS versus NMOSD (gray matter only)	RF	Not reported in source review	Not reported in source review	Not reported
MS versus NMOSD (multiparametric)	MRI + clinical + neuropsychological model	Not reported in source review	Not reported in source review	Not reported

Entries reflect what is reported in the citing review articles (1,7) rather than the primary studies they summarize (see Section 2); "not reported" is itself an informative finding regarding the granularity of the currently available evidence

MS: Multiple sclerosis, CNN: Convolutional neural network, DL: Deep learning, DWI: Diffusion-weighted imaging, ML: Machine learning, MRI: Magnetic resonance imaging, NMOSD: Neuromyelitis optica spectrum disorder, RF: Random forest, rs-fMRI: Resting-state functional magnetic resonance imaging, SVM: Support vector machine

or summarized above and should not be considered a substitute for a formal PROBAST or TRIPOD+AI audit conducted with access to the full study protocols; rather, it is intended to illustrate, without exhaustively documenting, the methodological domains most likely to influence generalizability.

Future Directions

Priorities for future work include: (i) developing multimodal models that integrate MRI with OCT, fluid biomarkers, and clinical history to improve upon single-modality systems (6,18); (ii) establishing open, standardized, multicenter datasets and implementing federated learning to address generalizability limitations across diagnostic as well as prognostic applications (19); (iii) placing greater emphasis on model interpretability and prospective validation in clinically realistic differential-diagnosis and longitudinal-prognosis settings (1,19); and (iv) externally validating prognostic models across healthcare systems while incorporating imaging or fluid biomarkers alongside the clinical variables evaluated to date (24,25). Given that DL as well as ensemble tree-based models remain largely opaque (Section 6), explainable artificial intelligence frameworks, such as SHAP and LIME, are increasingly being proposed as potential approaches to bridging the gap between model outputs and clinical trust. By attributing a model's output to specific input features on a per-patient basis, these methods could enable clinicians to scrutinize individual predictions against established MS pathophysiology rather than treating model outputs as opaque scores, potentially supporting regulatory evaluation and guideline acceptance of these tools.

Conclusion

ML can detect MS primarily by identifying lesions and disease-associated patterns on MRI, with an expanding evidence base for applications involving OCT and fluid biomarkers. Once a diagnosis is established, the same broad model families (particularly tree-based ensembles such as RF and XGBoost) can generate clinically meaningful predictions of disability progression, context-specific disability change, and symptom-specific severity, often achieving higher reported performance when outcomes are framed categorically rather than continuously. Across both diagnostic and prognostic applications, the strongest and most consistent evidence supports computer-assisted analysis rather than autonomous clinical decision-making; performance remains sensitive to dataset size and quality, single-center study designs, and the clinical realism of validation settings. Headline accuracy measures should also be interpreted cautiously when model performance may be driven substantially by strongly autocorrelated baseline variables (Section 4.1) rather than by a novel predictive signal. Continued progress will depend less on isolated algorithmic improvements and more on multimodal data integration, standardized open datasets, and rigorous prospective, multicenter validation across the diagnostic as well as prognostic dimensions of MS care.

Footnotes

Financial Disclosure: The author declared that this study received no financial support.

References

- Filippi M, Preziosa P, Arnold DL, Barkhof F, Harrison DM, Maggi P, Mainero C, Montalban X, Sechi E, Weinshenker BG, Rocca MA. Present and future of the diagnostic work-up of multiple sclerosis: the imaging perspective. *J Neurol*. 2023;270:1286-1299.
- Danelakis A, Theoharis T, Verganelakis DA. Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging. *Comput Med Imaging Graph*. 2018;70:83-100.
- Lladó X, Oliver A, Cabezas M, Freixenet J, Vilanova JC, Quiles A, Valls L, Ramió-Torrentá L, Rovira À. Segmentation of multiple sclerosis lesions in brain MRI: a review of automated approaches. *Inf Sci*. 2012;186:164-185.
- Moazami F, Lefevre-Utile A, Papaloukas C, Soumelis V. Machine learning approaches in study of multiple sclerosis disease through magnetic resonance images. *Front Immunol*. 2021;12:700582.
- Toosy AT, Eshaghi A. Machine learning utility for optical coherence tomography in multiple sclerosis: is the future now? *Neurology*. 2022;99:453-454.
- Arrambide G, Comabella M, Tur C. Big data and artificial intelligence applied to blood and CSF fluid biomarkers in multiple sclerosis. *Front Immunol*. 2024;15:1459502.
- Pilehvari S, Morgan Y, Peng W. An analytical review on the use of artificial intelligence and machine learning in diagnosis, prediction, and risk factor analysis of multiple sclerosis. *Mult Scler Relat Disord*. 2024;89:105761.
- Lladó X, Ganiler O, Oliver A, Martí R, Freixenet J, Valls L, Vilanova JC, Ramió-Torrentá L, Rovira À. Automated detection of multiple sclerosis lesions in serial brain MRI. *Neuroradiology*. 2012;54:787-807.
- Jacob A, McKeon A, Nakashima I, Sato DK, Elson L, Fujihara K, de Seze J. Current concept of neuromyelitis optica (NMO) and NMO spectrum disorders. *J Neurol Neurosurg Psychiatry*. 2013;84:922-930.
- Shoeibi A, Khodatars M, Jafari M, Moridian P, Rezaei M, Alizadehsani R, Khozeimeh F, Gorriz JM, Heras J, Panahiazar M, Nahavandi S, Acharya UR. Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: a review. *Comput Biol Med*. 2021;136:104697.
- Andropova P, Cheremisin D, Meldo A. A computer-aided diagnosis system in the diagnosis of multiple sclerosis. In: *Innovations in Computational Intelligence and Computer Vision: Proceedings of ICICV 2021. Advances in Intelligent Systems and Computing*, vol. 1424. Springer, Singapore; 2022.
- Petzold A, Albrecht P, Balcer L, Bekkers E, Brandt AU, Calabresi PA, Deborah OG, Graves JS, Green A, Keane PA, Nij Bijvank JA, Sander JW, Paul F, Saidha S, Villoslada P, Wagner SK, Yeh EA; IMSVISUAL, ERN-EYE Consortium. Artificial intelligence extension of the OSCAR-IB criteria. *Ann Clin Transl Neurol*. 2021;8:1528-1542.
- Aslam N, Khan IU, Bashamakh A, Alghool FA, Aboulmour M, Alsuwayan NM, Alturaif RK, Brahimi S, Aljameel SS, Al Ghamdi K. Multiple sclerosis diagnosis using machine learning and deep learning: challenges and opportunities. *Sensors (Basel)*. 2022;22:7856.
- Bonacchi R, Filippi M, Rocca MA. Role of artificial intelligence in MS clinical practice. *Neuroimage Clin*. 2022;35:103065.
- Shah A, Al-Shaibani MS, Ahmad M, Bunyan R. A bibliography of multiple sclerosis lesion detection methods using brain MRIs. 2023;arXiv preprint arXiv:2302.09516.
- Siriratnam P, Huda S, Butzkueven H, van der Walt A, Jokubaitis V, Monif M. A comprehensive review of the advances in neuromyelitis optica spectrum disorder. *Autoimmun Rev*. 2023;22:103465.
- Collorone S, Coll L, Lorenzi M, Lladó X, Sastre-Garriga J, Tintoré M, Montalban X, Rovira À, Pareto D, Tur C. Artificial intelligence applied to MRI data to tackle key challenges in multiple sclerosis. *Mult Scler*. 2024;30:767-784.
- Amin M, Martínez-Heras E, Ontaneda D, Prados Carrasco F. Artificial intelligence and multiple sclerosis. *Curr Neurol Neurosci Rep*. 2024;24:233-243.
- Szekely-Kohn AC, Castellani M, Espino DM, Baronti L, Ahmed Z, Manifold WKG, Douglas M. Machine learning for refining interpretation of magnetic resonance imaging scans in the management of multiple sclerosis: a narrative review. *R Soc Open Sci*. 2025;12:241052.
- Kuntal RS, Naik PK, Gupta R, Patil V, Polasi S, S V, Kavin BP. Multiple sclerosis detection with convolutional neural networks. *Int J Environ Sci*. 2025;11:954-969.
- Thabet RM, Shedeed HA, Al-Berry MN, Khattab D. Multiple sclerosis classification and segmentation in neuroimaging MRI using different machine and deep learning techniques: a review. *Artif Intell Rev*. 2025;58:255.
- Abdul-Salam SE, Florescu RM, Sfiredel V, Alexandru DO, Șerbănescu MS, Rotaru-Zăvăleanu AD. Deep learning and the retina: a new frontier in multiple sclerosis diagnosis. *Curr Health Sci J*. 2025;51:26-36.
- Kaya E, Emec M, Özdoğan AT, Zengin ES, Karakaş H, Daştan S, Baba C, Özcanhan MH, Özakbaş S. Development of restless legs syndrome severity prediction models for people with multiple sclerosis using machine learning. *Turk J Neurol*. 2025;31:440-449.
- Ozdogar AT, Emec M, Kaya E, Zengin ES, Ozcanhan MH, Ozakbas S. Development of a machine learning model to predict the expanded disability status scale in multiple sclerosis patients. *Mult Scler Relat Disord*. 2026;107:106937.
- Emec M, Alizada S, Simsek Y, Samadzade U, Ozcanhan MH, Ozakbas S. Integrating reproductive and clinical variables to predict postpartum disability outcomes in multiple sclerosis using machine learning. *Mult Scler Relat Disord*. 2026;109:107107.