applied sciences

MDPI

Review

A Systematic Review of AI Methods Across the MRI Analysis Pipeline for Multiple Sclerosis Progression Prediction

Umayal Venkatasamy [1], Alan Wang [1,2,3,4,5,6], William Schierding [3,7,8], Eryn Kwon [2,3,4,5], Helen V. Danesh-Meyer [3,7,8], Samantha Holdsworth [1,2,3] and Catherine Shi [3,9,*]

1 Faculty of Medical and Health Sciences, The University of Auckland, Auckland 1023, New Zealand; uven077@aucklanduni.ac.nz (U.V.); alan.wang@auckland.ac.nz (A.W.); s.holdsworth@auckland.ac.nz (S.H.)
2 Centre for Brain Research, The University of Auckland, Auckland 1142, New Zealand; e.kwon@auckland.ac.nz
3 Mātai Medical Research Institute, Tairāwhiti-Gisborne 4010, New Zealand; w.schierding@auckland.ac.nz (W.S.); h.daneshmeyer@auckland.ac.nz (H.V.D.-M.)
4 Medical Imaging Research Centre, The University of Auckland, Auckland 1010, New Zealand
5 Auckland Bioengineering Institute, The University of Auckland, Auckland 1142, New Zealand
6 Centre for Co-Created Ageing Research, The University of Auckland, Auckland 1010, New Zealand
7 Vision Research Foundation, 123 Remuera Road, Remuera, Auckland 1050, New Zealand
8 Department of Ophthalmology, The University of Auckland, Grafton, Auckland 1023, New Zealand
9 Department of Data Science and Artificial Intelligence, Auckland University of Technology, Auckland 1010, New Zealand
* Correspondence: catherine.shi@aut.ac.nz

Abstract

Magnetic resonance imaging (MRI)-based prediction of multiple sclerosis (MS) progression depends on how imaging data are prepared, represented, modelled, and evaluated. This systematic review synthesized artificial intelligence (AI) methods across the MRI-to-prediction pipeline. PubMed, Scopus, Web of Science, and Google Scholar were searched for studies published from 2010 to March 2026. Eligible studies included patients with MS or clinically isolated syndrome, used brain MRI directly or as the source of predictors, evaluated future progression-related outcomes, and reported quantitative predictive performance. Thirty-seven studies were included and synthesized narratively because of methodological and outcome heterogeneity. Prediction targets included CIS-to-MS/CDMS conversion, future MRI disease activity, PIRA/PIRMA, disability worsening or confirmed progression, future EDSS or disability status, and RRMS-to-SPMS conversion. Lesion morphology and location, radiomics, tissue atrophy, longitudinal imaging changes, and multimodal variables provided useful prognostic information. Classical machine learning (ML) remained effective for structured biomarkers, while deep learning (DL) enabled direct image-based modelling; neither approach was uniformly superior. PROBAST + AI assessment rated model Development as High concern in 35 of 37 studies and Unclear in two, Evaluation as High risk of bias in all 37 studies, and Applicability as Low concern in 32, Unclear in two, and High in three. Overall, MRI-based AI shows promising prognostic potential, while stronger data separation, calibration, transparent reporting, and independent validation are needed to support reproducible clinical translation.

Keywords: multiple sclerosis; magnetic resonance imaging; machine learning; deep learning; lesion segmentation; radiomics; progression prediction

Check for updates

Academic Editor: Miguel Alcaraz Baños

Received: 22 July 2026
Revised: 4 September 2026
Accepted: 6 September 2026
Published: 9 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Multiple sclerosis is a chronic immune-mediated disorder of the central nervous system characterized by demyelination and neuroaxonal injury, leading to heterogeneous neurological deficits and long-term disability [1,2]. Prognosis is difficult at the individual level because disease trajectories vary widely across phenotypes and over time, yet early identification of patients at higher risk of progression is important for treatment optimization and monitoring strategies [2]. MS is commonly described across clinical phenotypes including clinically isolated syndrome (CIS), relapsing remitting MS (RRMS), secondary progressive MS (SPMS), and primary progressive MS (PPMS), with RRMS representing most initial presentations and many patients transitioning to SPMS over the disease course [3,4]. Beyond physical disability, MS may also involve cognitive impairment, although the present review focuses primarily on MRI-based prediction of disease conversion, disability progression, relapse-independent progression, and future lesion activity [5].

MRI is central to MS diagnosis and longitudinal monitoring because it sensitively detects focal inflammatory lesions and supports evidence of dissemination in space and time [6]. Conventional sequences such as T2-W and FLAIR identify white matter (WM) hyperintensities that form the radiological foundation of MS assessment, and lesion burden and lesion distribution are associated with future outcomes [6,7]. However, conventional lesion measures correlate only moderately with clinical disability, often described as the clinico-radiological paradox, highlighting that lesion count/volume alone cannot fully explain progression and disability accumulation [8]. Recent MS imaging syntheses emphasize that MRI prognostication increasingly depends on extracting richer quantitative information from lesions and normal-appearing tissue, while also requiring careful standardization to support reliable interpretation across scanners and centers [9].

As MRI datasets become larger and more complex, AI methods are increasingly used to convert raw images into quantitative markers and predictions. Rather than being applied only at the end of analysis, AI can contribute at several stages of the MRI workflow. For example, deep learning has shown that lesion activity, such as enhancing lesions, can be predicted from non-contrast MRI in clinically relevant MS applications [10]. ML methods can operate on structured radiomic descriptors extracted from multi sequence MRI to classify MS lesion activity [11]. These developments motivate a review structure that follows the MRI analysis pipeline itself. Choices made early in the workflow, such as preprocessing, harmonization, segmentation and feature representation, directly influence what downstream predictive models can learn and how well they generalize.

A key practical challenge is that MRI derived biomarkers and AI models are highly sensitive to variability in acquisition and analysis, and multi-center studies face differences in scanners, vendors, protocols and longitudinal scanner changes that can introduce non biological variability and lead to inconsistent biomarker estimates and spurious associations with clinical outcomes if not addressed [12]. Consensus efforts in the MS imaging community therefore emphasize that harmonized acquisition and standardized post processing are essential prerequisites for using quantitative MRI biomarkers, and any AI models built on them, as clinically credible tools [12]. At the quantitative analysis stage, DL-based automated MS lesion segmentation offers a scalable alternative to manual delineation; however, segmentation accuracy and robustness to domain shift remain critical factors governing the reliability of derived quantitative imaging biomarkers [13]. Beyond lesion volume, richer feature representations such as lesion geometry, spatial distribution, and radiomics/texture features are designed to capture subtle, disease relevant microstructural patterns that are not discernible through routine MRI reading, and recent multicenter evidence indicates that FLAIR-based radiomics can contribute to predicting disability

worsening in MS, particularly when variability in image quality and spatial resolution is carefully addressed [14]. Together, these findings support the idea that the type of MRI-derived representation (e.g., voxel-level images, masks, engineered radiomics, longitudinal change maps) strongly influences the modelling strategy that is feasible and the prediction task that can be addressed.

Several reviews have already summarized the application of ML and DL methods in MS. These include reviews that organize studies according to broad clinical application domains, such as diagnosis, disease progression, and treatment response [15], reviews that categorize MRI-based prognostic modelling according to specific outcome measures, including conversion from CIS, Expanded Disability Status Scale (EDSS)-related disability, cognitive outcomes, motor impairment, and disease activity [16], and reviews that focus specifically on DL approaches for automated MS lesion segmentation [13].

While these reviews provide valuable overviews of the field, models are typically grouped according to task category or algorithmic approach, which can obscure how individual methodological choices relate to one another across the modelling process. In practice, successful clinical translation depends on the coherence of the entire methodological workflow, encompassing MRI acquisition and standardization, preprocessing and harmonization, biomarker extraction and representation, predictive modelling, and rigorous evaluation and validation. Explicit consideration of this end-to-end process is therefore essential for understanding the strengths, limitations, and translational potential of ML-based approaches in MS.

Accordingly, this review contributes a pipeline-oriented synthesis of AI methods across the MRI-to-prediction workflow for MS progression. We structure the literature according to the functional role of AI within the workflow: (i) MRI data preparation and preprocessing (including harmonization and quality control), (ii) imaging biomarker extraction and representation (lesion/tissue quantification, spatial and morphological characterization, radiomics, representation learning), and (iii) modelling strategies that map these MRI-derived representations to clinically meaningful progression endpoints.

This review examines progression-related prediction in four clinically relevant settings: conversion from CIS to MS or clinically definite MS (CDMS), transition from RRMS to SPMS, relapse-independent disability progression including PIRA-related outcomes, and future inflammatory activity indicated by new or enlarging lesions on follow-up MRI. These settings represent different stages and manifestations of disease evolution. They have important implications for prognosis, treatment planning, and long-term monitoring. Prediction of CIS conversion focuses on identifying individuals who are likely to develop clinically established MS at an early stage. Disability progression and PIRA-related outcomes reflect the accumulation of neurological impairment over time. New or enlarging lesions serve as MRI markers of ongoing inflammatory activity. The transition from RRMS to SPMS was also considered because it represents a clinically important milestone in disease progression. However, direct AI-based prediction of this transition was rarely investigated in the eligible studies, highlighting an important gap in the current literature. By comparing AI approaches across these progression settings, this review evaluates the potential of MRI-derived information to support risk stratification and to guide more individualized monitoring and management strategies. By linking MRI-derived feature representations with modelling strategies and prediction tasks, this pipeline-based framework helps identify which model types are commonly used for specific outcomes, where limitations in generalizability arise, and which stages of the workflow may constrain clinical translation.

2. Materials and Methods

2.1. Reporting Guideline and Literature Search

This systematic review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement [17]. The review was not prospectively registered, and no separate review protocol was prepared.

A structured literature search was conducted in PubMed, Scopus, Web of Science, and Google Scholar to identify studies applying AI, ML, or DL to brain MRI for the prediction of MS progression-related outcomes. The search covered publications from 2010 to March 2026. Search terms represented four main concepts: (i) MS and CIS; (ii) MRI and related imaging modalities; (iii) AI, ML, DL, neural networks, and radiomics; and (iv) progression-related outcomes, including conversion, disability progression, EDSS, progression independent of relapse activity (PIRA), disease activity, and new or enlarging lesions.

Search syntax was adapted to the requirements of each database. The PubMed strategy used Medical Subject Headings (MeSH) and title/abstract terms, while Scopus and Web of Science used database-specific field searching and Boolean operators. Google Scholar was used as an additional search source using the terms MS progression prediction + AI + DL + ML methods. The complete search strings, search limits, dates of searching, and numbers of records retrieved from each source are provided in Supplementary Table S1.

Previously published reviews on AI, MRI, and MS progression were consulted for background and contextual discussion only and were not used as an additional source for study identification or inclusion. All studies considered for inclusion were identified through the four documented database searches (PubMed, Scopus, Web of Science, and Google Scholar), and no records were added through citation searching, websites, organizations, or other review-based searching.

2.2. Eligibility Criteria

Studies were eligible if they: (i) included human participants with MS or CIS; (ii) performed patient-level prediction of a future MS-related outcome; (iii) used brain MRI directly as model input or as the source of predictors in an evaluated prediction model; (iv) applied AI, ML, or DL either for outcome prediction or for generating MRI-derived predictors used in an evaluated multivariable prediction model; (v) used MRI alone or combined with clinical, demographic, treatment, cerebrospinal fluid, electrophysiological, laboratory, or other variables; (vi) ensured the predictor-assessment period preceded the outcome-prediction period; and (vii) reported quantitative predictive performance measures, such as AUC/AUROC, C-index, accuracy, sensitivity, specificity, precision, F1 score, Brier score, MAE, or RMSE.

Prediction was not limited to baseline data. Longitudinal predictors were eligible if all predictor information was available before the start of the outcome-prediction period. Studies were excluded if predictor information was obtained during or after outcome assessment, or if the same follow-up MRI was used both as a model input and to define the outcome at that time.

Eligible outcomes included conversion from CIS to MS or CDMS; future physical disability status, worsening, or confirmed progression; PIRA, progression independent of relapse and MRI activity (PIRMA), or related relapse-independent progression outcomes; future patient-level radiological activity, including occurrence, count, or burden of new or enlarging T2 lesions or gadolinium-enhancing lesions; and transition from RRMS to SPMS.

Formal diagnostic or disability composite outcomes were eligible if they represented a prespecified progression endpoint or if all components fell within eligible outcome categories. Generic disease-activity composites combining heterogeneous clinical and radiological events were excluded when an ineligible component could independently

determine the outcome and eligible components were not separately extractable. Relapse alone, cognitive decline alone, and treatment response alone were ineligible unless a separately extractable eligible progression-related outcome was also reported.

Studies limited to cross-sectional diagnosis, disease-subtype classification, current-status classification, association analysis without prediction, or lesion detection/segmentation without future patient-level prediction were excluded. Voxel-, lesion-, or region-level prediction without a patient-level outcome was also excluded. MRI used only for participant selection, outcome definition, preprocessing, segmentation, or descriptive analysis was insufficient unless the resulting MRI-derived information was included in an eligible future-outcome prediction model.

Eligible publications were full-length original research articles published in English as peer-reviewed journal papers or full conference/workshop proceedings papers between January 2010 and March 2026. Reviews, meta-analyses, editorials, commentaries, protocols, letters lacking complete original-study information, case reports or case series, animal studies, non-original book chapters, and conference abstracts or posters without a full report were excluded. Full original research papers published as chapters in peer-reviewed conference proceedings volumes were eligible.

2.3. Study Selection

All retrieved records were imported into Rayyan, which was used for duplicate management and screening. A total of 1532 records were identified: 326 from PubMed, 665 from Scopus, 520 from Web of Science, and 21 from Google Scholar. After removal of 662 duplicate records, 870 records remained for title and abstract screening.

Title/abstract screening and full-text eligibility assessment were initially performed by the first author (U.V.) using the predefined eligibility criteria. No automated tool was used to make study-eligibility decisions. The final set of included studies and the full-text exclusion decisions were independently verified by a second author (A.W.). Any uncertainties identified during verification were discussed and resolved by consensus.

Following title and abstract screening, 805 records were excluded, and 65 reports were sought for full-text retrieval. Two reports could not be retrieved, leaving 63 reports for full-text eligibility assessment. Twenty-six reports were excluded after full-text assessment, resulting in 37 studies included in the systematic review. The study-selection process and full-text exclusion categories are presented in the PRISMA 2020 flow diagram Figure 1.

All studies included in the final review were represented in the documented database-search records and corresponding PRISMA flow diagram; no included studies were introduced from a separate or parallel review.

2.4. Data Extraction

Data were extracted using a predefined structured framework. For each eligible study, the following information was recorded: author and reference; cohort, study design, follow-up and DMT exposure; prediction outcome, definition and horizon; input data; MRI preprocessing and segmentation; feature extraction and quantification; data processing and validation; model or architecture; performance metrics; and key MRI predictors/contribution. MRI variables were distinguished as quantitative or image-derived features and categorical radiological variables, and whether MRI variables were retained in the final prediction model was recorded where reported. Information was extracted as reported in the original publications. Where information was unavailable, it was recorded as “not reported” or “unclear” rather than inferred. Initial data extraction was performed by the first author (U.V.). The principal data extracted for the final 37 included studies were independently verified by a second author (A.W.).

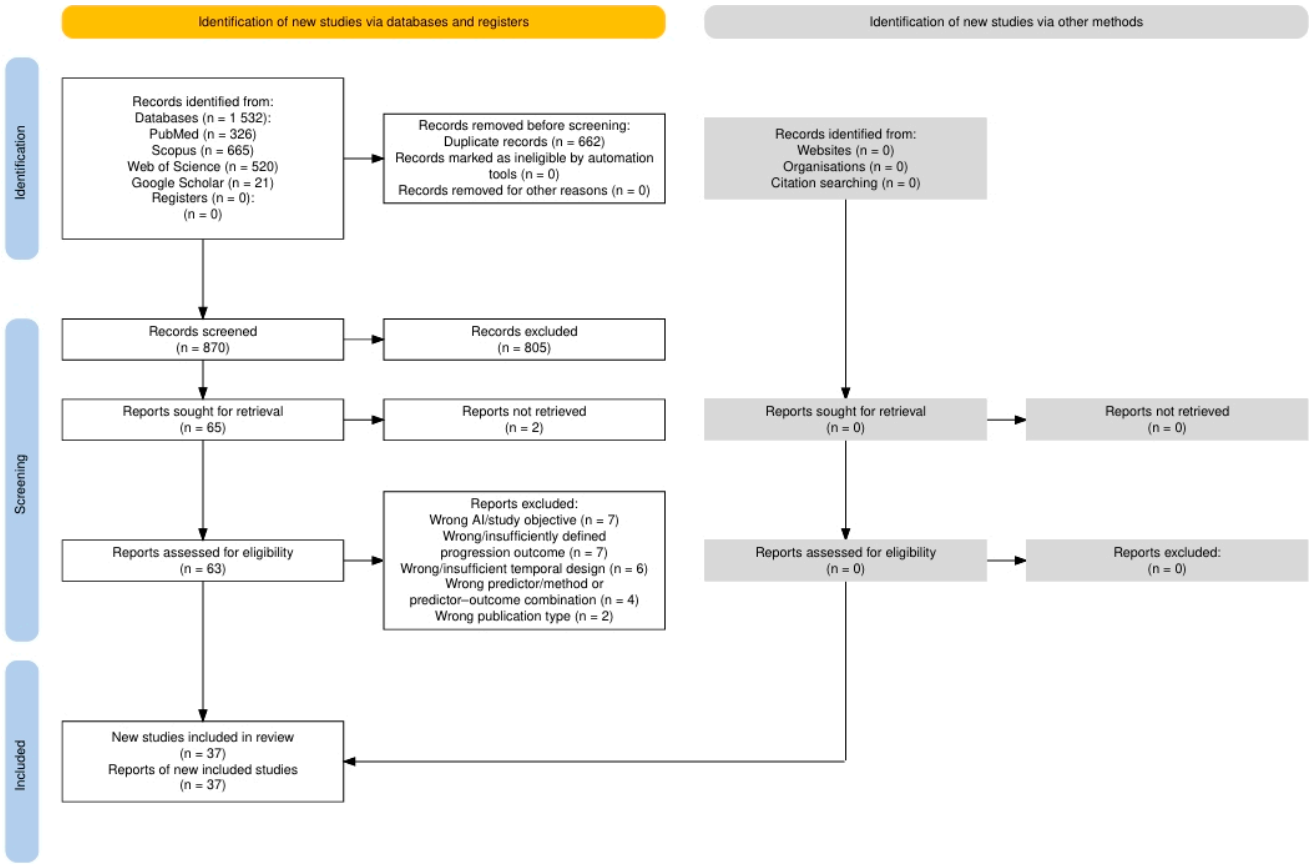


Figure 1. PRISMA 2020 flow diagram of database searching and study selection.

2.5. Methodological Quality, Risk of Bias, and Applicability Assessment

Methodological quality, risk of bias, and applicability were assessed using PROBAST + AI and its accompanying Explanation and Elaboration guidance [18]. For each study, relevant information on participants, predictors, outcome, sample size, missing data, class imbalance, feature selection, model development, validation, and performance reporting were extracted from the article and available Supplementary Material and assessed against the PROBAST + AI signaling questions.

PROBAST + AI assesses model Development quality (Dev) and model Evaluation risk of bias (Eval) separately across four domains: participants and data sources (D1), predictors (D2), outcome (D3), and analysis (D4). Development focuses on how appropriately the model was constructed, including sample size, predictor handling, missing data, class imbalance, and overfitting control. Evaluation focuses on how reliably performance was assessed, including evaluation sample size, data leakage, resampling procedures, and performance assessment. Applicability (App) was assessed separately for D1-D3 according to the match between the study and the review question.

Signaling questions were rated as Yes, Probably Yes, Probably No, No, No Information, or Not Applicable, and overall judgements as Low, Unclear, or High. Unreported information was not inferred. Materially different models, outcomes, prediction horizons, redevelopment procedures, or validation stages were assessed separately where required. Detailed PROBAST + AI assessments were conducted for the included studies and distinct model or evaluation stages, together with a study-level summary of overall judgements by MS progression-prediction outcome category.

2.6. Data Synthesis

Given the heterogeneity in MS phenotype, cohort characteristics, progression outcomes, MRI inputs, prediction horizons, feature representations, modelling approaches, and validation strategies, the findings were synthesized narratively and in structured tables rather than by meta-analysis. Studies were compared across the MRI analysis pipeline, including image preparation and preprocessing, biomarker extraction and representation, predictive modelling, and evaluation, with emphasis on the relationships among MRI-derived information, model design, progression outcome, and validation strategy. For consistency, AUC/AUROC values reported as percentages were converted to the 0–1 scale.

3. AI Methods for MRI Data Preparation and Preprocessing

MRI-based AI studies in MS increasingly rely on multicenter and longitudinal datasets, but scanner- and site-related differences introduce substantial non-biological variability. Variations in vendor, field strength, pulse sequence, software version, and acquisition protocol can reduce model robustness and obscure true disease-related signals when data are pooled without harmonization. For this reason, harmonization and quality-controlled preprocessing are now recognized as essential prerequisites for reliable quantitative MRI biomarkers in multicenter MS research [12].

3.1. Image Denoising and Artifact Correction

Brain MRI images are commonly affected by noise and motion artifacts that can reduce the quality of downstream analyses [19]. In MS MRI, typical artifacts include motion artifacts [20,21], susceptibility distortions [22], gradient nonlinearity [23], and radio-frequency inhomogeneities [24]. Each of these artifacts can significantly impact downstream analyses [19]. DL methods have shown strong performance in addressing both noise and motion artifacts, often outperforming classical filtering approaches by learning to separate signal from noise in a brain anatomy-aware manner [25].

For denoising, Kidoh et al. tested a DL-based reconstruction (dDLR) approach for brain MRI and reported statistically higher PSNR and SSIM than comparator convolutional neural networks (CNN) denoisers in experiments, along with improved qualitative assessments; in a volunteer study across multiple sequences (including FLAIR and MPRAGE), dDLR-processed lower-averaging acquisitions (NAQ2) achieved significantly better SSIM/PSNR than unprocessed NAQ2 and was qualitatively comparable to or better than higher-averaging acquisitions (NAQ5) for most criteria. This supports the idea that AI denoising can be used upstream to improve the quality of inputs that feed segmentation and feature extraction [26].

For motion artifacts, Pawar et al. evaluated a DL motion-correction method on clinical T1-W MPRAGE images identified as motion degraded, reporting reductions in ghosting/ringing/blurring and significant improvements in visual grading scores across multiple neuroanatomical regions; they also reported that a portion of repeated scans could be avoided, indicating operational relevance for clinical-scale datasets [27]. The proposed two stage multi loss MC Net combines novel hybrid L1 loss in the first stage to suppress global artifacts with a Total Variation (TV) term in the second stage to refine boundaries, delivering higher peak signal to noise ratio (PSNR)/structural similarity index measure (SSIM) than single stage L1 or L1 + TV models while preserving image context [28]. Finally, a multiscale fully convolutional network estimates motion artifacts from large amounts of simulated motion corrupted data, subtracts the artifact only estimate from the input, and produces corrected images with significantly reduced artifacts [29].

For ringing artifacts, Zhang et al. developed a CNN approach to reduce Gibbs-ringing artifacts; they reported that their method removed ringing without noticeable smoothing

and improved quantitative similarity to fully sampled references (e.g., root-mean-square error (RMSE)/PSNR/SSIM) relative to existing methods, illustrating how AI can correct a concrete reconstruction-related artifact before downstream analysis [30].

3.2. Bias Field Correction and Intensity Normalization

Bias field (intensity inhomogeneity) and inconsistent intensity scaling are long-recognized challenges in MRI because raw MR intensities are not standardized physical units, and they can vary across acquisitions even for the same tissue class [31]. These issues matter directly for ML/DL because many segmentation and feature extraction methods depend on intensity distributions and tissue contrast.

For bias field correction, the widely used N4ITK method is an extension/improvement of the classical normalization (N3) algorithm and introduces algorithmic modifications (including B-spline fitting/optimization changes) to improve bias field correction performance. N4 remains a common upstream step in contemporary MRI pipelines (including neuroimaging and MS contexts) because it reduces non-biological intensity nonuniformity that otherwise confounds tissue/lesion separation [32].

For intensity normalization, Shinohara et al. show that conventional MRI is measured in arbitrary units that are not comparable across subjects or visits, introduce formal "statistical principles of image normalization" (SPIN), and propose tissue-based "white stripe" methods that match the mean and variance of normal appearing white matter (NAWM) across scans. Across thousands of ADNI, AIBL, and MS scans, principled normalization greatly reduces between scan tissue intensity variability and prevents spurious grey matter (GM) loss from naive histogram matching, making robust normalization essential in large multi-site studies [31].

Critically for MS-specific longitudinal workflows, Roy et al. propose a 4D longitudinal intensity normalization method "in the presence of MS lesions," explicitly aiming to yield "stable and consistent longitudinal segmentations" and positioning the method as a preprocessing step usable with any segmentation method; they model intensity variability linked to lesion dynamics while keeping lesion information unaffected. This paper provides direct justification for putting intensity normalization explicitly in the preprocessing stage of MS MRI pipelines rather than treating it as an incidental detail [33].

Finally, more general tissue-based standardization approaches, as highlighted by De Nunzio et al., emphasize that the lack of fixed tissue-specific intensity affects post-processing tasks such as automatic segmentation and classification, reinforcing the automatic link between normalization choices and downstream ML/DL stability [34].

3.3. MRI Harmonization Across Scanners and Sites

A central preprocessing challenge in MS is harmonizing MRI acquisition, analysis, and storage across sites. A harmonization-focused MS review from the MAGNIMS Study Group explicitly motivates harmonization because cooperative, large-scale datasets are needed to develop diagnostic and prognostic biomarkers, yet differences in acquisition parameters, image analysis, and data storage across sites can introduce bias and represent a substantial constraint; the review surveys state-of-the-art harmonization developments and provides recommendations [12].

This problem is not theoretical: harmonization studies show that scanner effects can be large enough to hinder inference and even generate spurious findings. ComBat-based methods model additive and multiplicative scanner effects and, in both cross-sectional and longitudinal extensions, have been shown to remove unwanted site variability, increase power to detect biological effects, and improve type I error control compared with unharmonized data or simple scanner covariate adjustment [35,36]. In line with this, recent

MS progression work has extracted radiomics features from WML and NAWM using PyRadiomics, harmonized them with Longitudinal ComBat, and then applied recursive feature elimination before training ML models to predict EDSS worsening from multi-center FLAIR MRI, further illustrating that harmonization of derived MRI biomarkers is becoming standard practice prior to modelling in longitudinal multi scanner MS studies [14].

Deep learning can also be used for image-level or contrast-level harmonization. Dewey et al. propose DeepHarmony, a U-Net-based approach trained with a small overlap cohort scanned with two protocols; they report improved consistency of volume quantification across scanning protocols and show in an MS longitudinal dataset that atrophy calculations were substantially affected by protocol change, with those differences substantially reduced after DeepHarmony-based harmonization. This directly supports the pipeline claim that harmonization affects the reliability of derived quantitative biomarkers (atrophy/volumetrics) in MS [37].

3.4. Image Registration and Spatial Alignment

Registration and spatial alignment are essential in MS MRI pipelines because (i) multi-sequence analysis requires co-registration across contrasts, and (ii) longitudinal progression endpoints (new/enlarging lesions, atrophy trajectories, change maps) rely on consistent within-subject spatial correspondence across time [38].

Classical registration frameworks such as ANTs model spatial alignment as an optimization over a transformation (rigid, affine, elastic, or diffeomorphic) and a similarity metric (e.g., mean squared difference, cross correlation, mutual information), often within a multi resolution scheme. Symmetric Normalization (SyN) in ANTs estimates diffeomorphic mappings between images by integrating a regularized velocity field in a symmetric fashion, and underpins groupwise template building (SyGN) and downstream tasks such as atlas-based labelling and cortical thickness analysis [39]. In many recent MS DL prognostic studies, however, registration is implemented using conventional affine normalization to standard templates (e.g., FLIRT to MNI 152) primarily to ensure comparable anatomical locations and voxel wise correspondence across patients and scanners, rather than as a methodological focus in itself [40].

Balakrishnan et al. introduce VoxelMorph, a CNN-based framework for deformable registration, reporting that the unsupervised model's accuracy is comparable to state-of-the-art methods while operating orders of magnitude faster, and that auxiliary segmentation information can further improve registration performance [41]. This complements the more traditional affine normalization used in MS progression-prediction pipelines, where PD/T2, T1 and lesion masks are skull-stripped, intensity-normalized and then affinely registered to a common or MNI152 space before being passed to CNNs for disease-worsening or conversion prediction [42].

An MS-specific longitudinal lesion-change pipeline illustrates the role of alignment and related preprocessing. In a fully automatic method for segmenting changes in FLAIR-hyperintense WML on serial MS MRI, the authors describe preprocessing steps including intrasubject registration and bias field correction/intensity scaling prior to voxel-level analysis of intensity changes in lesions and NAWM-based statistical testing. This provides direct evidence that longitudinal MS endpoints require explicit alignment as a preprocessing stage [38].

3.5. AI-Assisted Quality Control

Quality control (QC) is an important preliminary step in MRI analysis because low-quality scans can bias subsequent processing and analysis. Esteban et al. addressed this problem with MRI Quality Control tool (MRIQC), which extracts image quality metrics,

generates visual reports, and applies a binary accept/exclude classifier to support more objective and scalable QC. However, although MRIQC achieved about 76% accuracy on unseen-site data, its performance remained affected by site effects, showing that generalization across acquisition centers is still challenging [43]. Samani et al. extended this direction in diffusion MRI through QC-Automator, a CNN-based transfer-learning framework that detects specific artifacts such as motion, multiband interleaving, ghosting, susceptibility, herringbone, and chemical shift [44]. Together, these studies show that AI-assisted QC can reduce reliance on manual inspection, but robust cross-site and cross-protocol performance remains a key requirement for reliable downstream neuroimaging analysis.

3.6. Domain Adaptation and Harmonization Networks for Cross-Site Variability

Cross-site variability is a major challenge in MS lesion segmentation because differences across scanners, sites, and imaging protocols can reduce model generalization to new domains. Ackaouy et al. addressed this through Seg-JDOT, an unsupervised domain adaptation framework based on optimal transport that aligns source and target representations without requiring manual target-domain annotations. They also showed that domain shift persists even after image standardization, suggesting that preprocessing alone cannot fully remove cross-site bias [45]. Consistent with this, Zeng et al. noted that current MS lesion adaptation methods mainly rely on feature adaptation and suggested that future work should combine both image-level and feature-level adaptation [13]. Therefore, domain adaptation and harmonization-oriented networks are likely to be important for building MS imaging models that remain reliable across centers.

4. AI Methods for Imaging Biomarker Extraction

Imaging biomarker extraction transforms pre-processed MRI into quantitative features for prediction. This includes automated lesion/tissue segmentation to compute lesion volumes and brain atrophy; characterization of lesion morphology and spatial distribution; high-dimensional radiomic/texture descriptors; and temporal change maps from longitudinal scans. Classical pipelines (e.g., FreeSurfer/SPM tissue segmentation, LST/BIANCA lesion segmentation) and DL models (U-Net, nnU-Net) have been used. Each method has limitations (e.g., segmentation errors for small/unclear lesions, variability across scanners). Radiomics features capture sub-resolution patterns but require standardized inputs. Learned representations (deep features) promise richer biomarkers but also demand large data or pretraining. In all cases, consistent extraction under domain shift is a key challenge.

4.1. Lesion Detection and Segmentation

Automated lesion segmentation tools differ in both their input requirements and their underlying modelling strategies, and several automatic MRI-based methods have been developed to address these challenges. A comparison of commonly used segmentation methods, their required inputs, preprocessing steps, and extracted outputs is provided in Table 1. MSmetrix is a fully automated unsupervised method that uses 3D T1-W and FLAIR MRI in a probabilistic framework to detect WM lesions as an outlier to normal brain while simultaneously segmenting GM, WM, and CSF, and it does not require training data [46]. Lesion-TOADS is a fully automatic atlas-based method that uses multichannel MRI together with topological and statistical atlases to delineate MS lesions while segmenting the brain into its major structures [47]. BIANCA is a fully automated supervised method based on the k-nearest neighbor (k-NN) algorithm; it is described as multimodal and allows incorporation of spatial information and local spatial intensity averaging [48]. BIANCA-MS was developed from the original BIANCA specifically for MS lesion segmentation and introduced two key modifications: a harmonized setting tested across different MRI proto-

cols and a cleaning step intended to improve consistency between automated and manual segmentations [49]. The authors further reported that BIANCA-MS outperformed other available tools and showed comparable performance across different scanning protocols, modality sets, and image resolutions, supporting its use under heterogeneous imaging conditions [49].

The Lesion Segmentation Tool (LST) for SPM provides two lesion-segmentation algorithms: the lesion growth algorithm (LGA), which requires T1 plus FLAIR, and the lesion prediction algorithm (LPA), which requires FLAIR only; the toolbox also includes a longitudinal pipeline [50]. More recent DL methods include LST-AI, which is presented as an advanced DL-based extension of LST built as an ensemble of three 3D U-Nets [51]. Another CNN-based method is nicMSlesions, which is based on a cascade of two 3D patch-wise CNN, where the first network is designed to be more sensitive to candidate lesion voxels and the second to reduce false positives from the first stage [52]. Automated MS lesion-segmentation methods therefore encompass probabilistic, atlas-based, k-NN-based, and CNN-based approaches. However, robustness across heterogeneous imaging protocols remains an important requirement for broader clinical application.

4.2. Tissue Segmentation and Structural Quantification

Structural quantification in progression-focused MS studies has been performed using automated morphometric tools; for example, regional cortical and deep GM atrophy in CIS was assessed using FreeSurfer to derive progression-relevant tissue measures [53]. Beyond lesion masks, tissue segmentation and structural quantification provide volumetric markers of diffuse neurodegeneration that complement focal lesion-based assessment in MS progression research [54]. In CIS, selective GM atrophy has shown prognostic value, with atrophy of the superior frontal gyrus, thalamus, and cerebellum identified as independent predictors of conversion to MS [53]. Early longitudinal CIS work further showed that global brain volume loss and GM volume loss occur within the first year, and that brain volume loss predicted conversion to MS [55]. In RRMS, tissue-compartment measures also retain prognostic relevance, as the GM:NAWM ratio at diagnosis has been reported to predict both disability progression and conversion to SPMS [56]. At a regional level, deep GM degeneration, particularly thalamic atrophy, appears especially informative, with multicenter evidence showing that deep GM volume loss drives disability worsening, while both baseline thalamic atrophy and more recent AI-measured thalamic atrophy have been associated with later disability progression [57–59]. Together, tissue segmentation and structural quantification enable objective assessment of global and regional neurodegeneration, particularly GM and deep-GM damage, and provide complementary information beyond conventional lesion-based measures.

4.3. Spatial and Morphological Lesion Characterization

Several studies have shown that automated structural characterization of lesions provides prognostic information beyond lesion count alone. In CIS, shape-based features extracted from automated lesion masks were used to quantify lesion morphology, and important predictors included mean lesion volume, minimum sphericity, and minimum surface-volume-ratio, with converters showing larger and less round lesions than non-converters [60]. Beyond simple geometric descriptors, lesion geometry has also been incorporated into Support Vector Machine (SVM)-based prediction models for conversion from CIS to CDMS, supporting the value of structural lesion characterization in early prognostic modelling [61]. Spatial lesion distribution has also shown predictive relevance. In a recent explainable ML study, periventricular and infratentorial lesion locations were identified among the significant imaging variables associated with CIS-to-CDMS conver-

sion [62]. These findings indicate that lesion geometry and spatial distribution provide prognostic information beyond conventional measures of lesion burden alone. Representative FLAIR MRI and lesion-mask visualizations illustrating lesion segmentation and spatial characterization are shown in Figure 2. The examples demonstrate how visible white-matter lesions can be represented as segmentation masks and further characterized according to their anatomical distribution, including periventricular, cortex-contacting, and infratentorial locations.

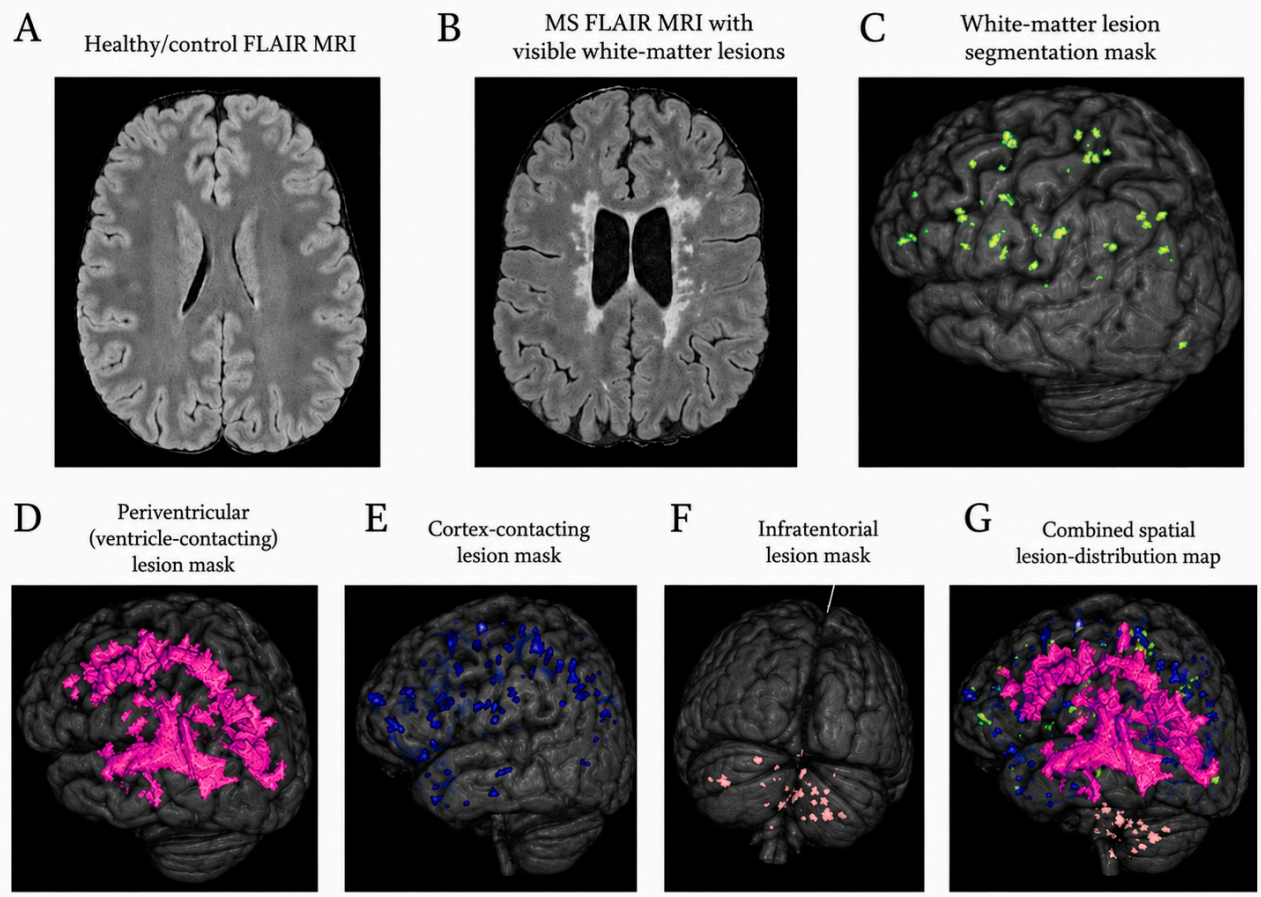


Figure 2. Representative FLAIR MRI and lesion-mask visualizations illustrating MS lesion segmentation and spatial characterization. (**A**) Healthy/control FLAIR MRI; (**B**) MS FLAIR MRI showing visible white-matter lesions; (**C**) green indicates the white-matter lesion segmentation mask; (**D**) magenta indicates the periventricular (ventricle-contacting) lesion mask; (**E**) blue indicates the cortex-contacting lesion mask; (**F**) light pink indicates the infratentorial lesion mask; and (**G**) shows the combined spatial lesion-distribution map. These examples illustrate how conventional MRI appearance can be converted into lesion masks and spatially characterized lesion representations for subsequent quantitative imaging analysis and progression modelling. The representative MRI data were obtained from the access-controlled "MRI dataset for susceptibility-based radiomic feature extraction in healthy controls and patients with multiple sclerosis" hosted on Zenodo [63]. The lesion-mask visualizations were generated by the authors from the MRI data.

4.4. Radiomics and Feature Engineering

Radiomics and feature engineering aim to extract large numbers of quantitative features from medical images to characterize image phenotypes beyond visual inspection. In the radiomics workflow described by Lambin et al., the process begins with high-quality standardized imaging, followed by definition of the region of interest through automated segmentation or expert delineation, after which quantitative imaging features are extracted from the defined region. These features include descriptors of shape and texture [64]. In ML-

based MS progression prediction, radiomics and feature engineering transform MRI data into explicit handcrafted predictors, whereas DL approaches derive feature representations implicitly within the network architecture.

The extracted features are generally grouped into several major classes. First-order features describe the distribution of voxel intensities within the segmented region. Shape features quantify geometric properties such as size, compactness, elongation, and surface characteristics, and are derived from the segmentation mask rather than from grey-level values. Texture features describe spatial relationships and heterogeneity patterns using matrices such as the grey-level co-occurrence matrix (GLCM), grey-level run length matrix (GLRLM), grey-level size zone matrix (GLSZM), neighboring grey tone difference matrix (NGTDM), and grey-level dependence matrix (GLDM) [65].

Radiomics studies often generate very large numbers of features, making feature engineering necessary to improve robustness and reduce redundancy. This process commonly includes reproducibility filtering, correlation reduction, and selection of the most informative features before model building. To promote consistency across studies, the Image Biomarker Standardization Initiative (IBSI) established standardized definitions, benchmark values, and reporting guidance for radiomic features and radiomics workflows [66].

Reporting of radiomics standardization was heterogeneous across the included studies. Khan et al. reported the use of PyRadiomics 2.20 and stated that image discretization was aligned with IBSI recommendations [14]. In contrast, other studies described reproducibility or standardization procedures without explicitly reporting IBSI compliance. Peng et al. and Shi et al. assessed feature reproducibility using intraclass correlation coefficients [67,68], while Elhaie et al. reported the PyRadiomics extension in 3D Slicer together with standardized resampling and fixed-bin-width discretization [69]. However, explicit verification of extracted features against IBSI reference values was not consistently reported. This variation limits direct comparison of radiomic features across studies and reinforces the need for more complete reporting of feature-extraction settings and standardization procedures.

Radiomic and engineered features therefore provide quantitative representations of image intensity, morphology, and heterogeneity beyond conventional lesion burden; however, their reproducibility and cross-study comparability depend on standardized extraction and transparent reporting.

4.5. Representation Learning from MRI

In DL approaches, baseline MRI scans are used directly as model inputs, and prognostic imaging biomarkers are computationally derived as latent feature representations within the network instead of being manually specified as handcrafted features. Tousignant et al. used multimodal MRI volumes in an end-to-end 3D CNN framework for future disability progression prediction [70], while Storelli et al. showed that baseline T1-W and T2-W MRI can support prediction of clinical and cognitive worsening without explicit handcrafted feature extraction [40]. Taloni et al. further showed that ResNet50-based analysis of 3D T1-W MRI identified informative frontal GM regions [71], and Coll et al. extended this paradigm to survival modelling of PIRA using EfficientNet-derived features from routine MRI at the first demyelinating event [72]. Accordingly, these approaches derive progression-relevant imaging representations directly from MRI rather than relying on predefined handcrafted biomarkers.

Table 1. Comparison of automatic MRI lesion segmentation algorithms for multiple sclerosis, including required inputs, preprocessing workflows, and extracted lesion-related outputs.

Method	Required Inputs	Preprocessing	Final Data Extraction
Lesion-TOADS [47]	MPRAGE images T2-FLAIR images	Skull and extracranial tissues are removed prior to segmentation using the MPRAGE scan.	Outputs: volumes of T2-FLAIR -hyperintense WM lesions, sulcal CSF, ventricular CSF, cortical GM, total cerebral WM (including lesions), cerebellar GM, cerebellar WM, putamen, caudate, thalamus, and brainstem
MSmetrix [46]	3D T1-w images 3D T2-FLAIR images	The input T2-FLAIR image is rigidly co-registered with the input T1-w image. The T1-W input image is skull striped. The probabilistic anatomical priors for GM, WM and CSF are transferred to the skull-stripped T1-W space.	WM GM CSF WM lesions
LST—Lesion Growth Algorithm (LGA) [50]	T1-W images T2-FLAIR images	Segments the T1-W images into three main tissue classes: CSF, GM, and WM. Combines this information with co-registered T2-FLAIR intensities to calculate lesion belief maps. Applies an initial threshold to generate a binary lesion map. Grows the lesion map along voxels that appear hyperintense on T2-FLAIR.	T2-FLAIR image (Bias corrected) Lesion probability map .mat file
LST—Lesion Prediction Algorithm (LPA) [5,50,60]	T2-FLAIR images	Requires T2-FLAIR image only. Optional reference image for co registration prior to segmentation.	Lesion probability estimates for each voxel
BIANCA [48]	T2-FLAIR images (mandatory) T1-W image (optional but recommended) Manually segmented WMH masks (training subjects) Brain mask Spatial location features (optional: MNI coordinates)	FLIRT registration between modalities, Linear registration to MNI for spatial coordinates, Intensity normalization, Patch options, Training point selection (critical), Exclusion mask (CSF-based).	Voxel-wise WMH probability map User-defined probability threshold applied Final WMH mask is thresholded at 0.9
BIANCA-MS [49]	Primary Reference Image: FLAIR or PD. Supporting Modalities: T1-W, T2-W, and PD (if available). Training Data: Manually segmented lesion masks required for retraining.	Visual inspection for artifacts, All modalities co-registered to the reference sequence using FLIRT, Brain extraction from T1-W images using BET, Registered to main modality for tissue-specific masks, GM and cerebellar exclusion masks are applied to reduce false positives, Pure WM, GM, and CSF masks obtained using FAST, with lesion clusters removed.	Binary lesion mask of WM lesions in the brain; Lesion volume calculated from the output mask

Table 1. *Cont.*

Method	Required Inputs	Preprocessing	Final Data Extraction
LST-AI [51]	T1-W image, FLAIR image	T1-W and FLAIR images are registered to MNI ICBM152 space. Images are resampled to a consistent $1 \times 1 \times 1$ mm^3 resolution and orientation. Skull stripping using HD-BET, Centre-cropping to $192 \times 192 \times 192$ voxels and intensity-normalized to [0, 1]	Three individual lesion probability maps from the ensemble, Averaged probability map, Binary lesion mask obtained by thresholding the averaged map at 0.5, then inverse-transformed to the original FLAIR space.
nicMSlesions (cascaded 3D CNN) [52]	FLAIR images; T1-W images. Additional T2-W or PD images can be incorporated if available. Manual lesion masks are required for training/retraining.	Images are skull stripped using BET and intensity corrected/normalized using N3. Modalities are co-registered into the same image space. Intensity normalization is applied and multi-sequence 3D image patches are extracted as separate input channels.	Voxel-wise WM lesion probability map (from the second CNN), Binary lesion segmentation mask after thresholding.

3D, three-dimensional; AI, artificial intelligence; BET, Brain Extraction Tool; BIANCA, Brain Intensity AbNormality Classification Algorithm; BIANCA-MS, BIANCA for multiple sclerosis; CNN, convolutional neural network; CSF, cerebrospinal fluid; FAST, FMRIB's Automated Segmentation Tool; FLAIR, fluid-attenuated inversion recovery; FLIRT, FMRIB's Linear Image Registration Tool; GM, grey matter; HD-BET, High-Definition Brain Extraction Tool; ICBM, International Consortium for Brain Mapping; LGA, lesion growth algorithm; LPA, lesion prediction algorithm; LST, Lesion Segmentation Tool; LST-AI, Lesion Segmentation Tool–Artificial Intelligence; MNI, Montreal Neurological Institute; MPRAGE, magnetization-prepared rapid gradient-echo; MRI, magnetic resonance imaging; MS, multiple sclerosis; N3, nonparametric nonuniform intensity normalization; PD, proton density; T1-W, T1-weighted; T2-W, T2-weighted; WM, white matter; WMH, white matter hyperintensity.

5. AI Methods for Modelling MRI Features and Predicting Clinical Outcomes

AI methods for MS progression prediction use either explicit MRI-derived biomarkers or MRI directly to model future clinical outcomes. MRI-based ML reviews consistently describe this literature as using imaging variables such as lesion measures and tissue-volume markers, often together with demographic and clinical covariates, for tasks including CIS-to-MS conversion, EDSS-related disability prediction, and other progression outcomes [16,62].

5.1. Classical Machine Learning Approaches

Classical ML is well suited to settings in which MRI information is represented as structured input variables rather than raw images. In this setting, the model operates on predefined biomarkers such as lesion geometry, lesion location, lesion burden, and GM-based measures, usually combined with baseline demographic or clinical data [61]. In CIS-to-CDMS prediction, SVMs and ensemble tree-based models are among the most frequently applied approaches. Bendfeldt et al. [61] used SVMs with lesion geometry metrics derived from Minkowski functionals together with global GM volume information and concluded that both lesion geometry and GM-based information can aid prediction of conversion to clinically definite MS [61].

More recently, Rasouli et al. [62] developed an explainable eXtreme Gradient Boosting (XGBoost) model using demographic, clinical, and imaging variables and identified age, gender, schooling, motor symptoms, infratentorial lesions, periventricular lesions, oligoclonal bands in CSF, and lesion/symptom type as the most influential predictors of CIS-to-CDMS conversion [62]. A separate 2024 comparative CIS study benchmarked CatBoost, XGBoost, LightGBM, random forest, SVM, and logistic regression on the same public

CIS cohort, showing that boosted-tree and other feature-based classifiers remain competitive when heterogeneous structured inputs are used [73]. Earlier work by Zhao et al. [74] showed that SVMs incorporating short-term clinical and brain MRI data outperformed logistic regression for predicting MS disease course in that study [74].

Classical ML therefore provides a suitable framework for modelling pre-extracted and interpretable MRI biomarkers, particularly when these features are integrated with clinical or demographic covariates. SVM, XGBoost, and related ensemble methods are among the commonly applied approaches in progression-focused studies.

5.2. Deep Learning Models for End-to-End Prediction

DL models for end-to-end prediction use MRI directly as model input, allowing prognostic representations to be learned automatically within the network rather than being manually predefined as handcrafted biomarkers. Tousignant et al. [70] introduced the first automatic end-to-end DL framework for predicting future disability progression from baseline multimodal brain MRI in MS. Their model used a 3D CNN with parallel convolutional pathways to learn multi-scale MRI patterns directly from the images, and prediction improved when lesion label masks were included as supplementary inputs. The study also showed that uncertainty estimates derived from Monte Carlo dropout were useful for identifying less reliable predictions and scans that might require closer review [70].

Subsequent studies extended this strategy to other progression-related outcomes. Storelli et al. [40] developed a CNN-based model using baseline T1-W and T2-W MRI to predict clinical worsening, cognitive deterioration, or both at 2 years in a multicenter MS cohort [40]. Taloni et al. [71] likewise used baseline 3D T1-W MRI for disability progression prediction and found that the most informative predictive regions were located mainly in the frontal GM, showing that DL can derive anatomically meaningful prognostic patterns directly from structural MRI [71].

A recent application of end-to-end DL is the prediction of PIRA following a first demyelinating event. Coll et al. [72] trained an EfficientNet-based DL survival model on routine baseline T1-W and T2-FLAIR MRI to estimate the cumulative probability of progression independent of relapse activity (PIRA). Their study further showed that the DL-derived PIRA risk score improved a conventional age-adjusted Cox model, indicating that end-to-end MRI features added prognostic value beyond classical covariates. Interpretability analysis also highlighted the frontoparietal cortex as an important region in the learned prediction [72].

Evidence from broader systematic evaluations remains more conservative. A 2025 systematic review and meta-analysis concluded that MRI-driven AI models predict MS disability progression with only moderate accuracy, and that DL has shown only marginal, non-significant improvement over strong classical baselines, while external validation and calibration remain limited across studies [75].

End-to-end DL provides a direct framework for MRI-based prognosis in MS; however, reported performance gains over carefully engineered feature-based models remain modest. The literature remains dominated by established CNN architecture, indicating a need for further evaluation of alternative DL designs and their robustness across independent datasets.

5.3. Multimodal and Hybrid Models

Multimodal and hybrid models combine MRI-derived quantitative features and/or categorical radiological variables with clinical, demographic, biochemical, or other variables to improve MS progression prediction, rather than relying on imaging alone [76,77].

In progression-focused studies, multimodal models commonly integrate routine MRI lesion descriptors, including lesion location, lesion type, new lesion activity, and contrast enhancement, with demographic and clinical variables and, in some studies, CSF biomarkers. Rasouli et al. [62] used XGBoost on demographic, clinical, and imaging data, and identified periventricular lesions, infratentorial lesions, and CSF oligoclonal bands among the key contributors to CIS-to-CDMS prediction [62].

Poretto et al. [76] similarly integrated demographic, clinical, biochemical, and routine MRI-derived radiological variables, including lesion number, lesion localization, new lesions, and contrast enhancement, for early PIRA prediction [76]. A related comparative CIS study expanded the multimodal feature space by combining MRI indicators, CSF measures, and clinical and environmental-history variables [78].

Yoo et al. [77] employed a modelling strategy that combined deep-learned lesion-pattern representations with user-defined MRI and clinical features for CIS-to-CDMS prediction. The final model integrated CNN-derived latent imaging features with explicitly engineered radiological and clinical variables [77].

Multimodal modelling enables MRI-derived information to be evaluated alongside patient-level clinical and demographic characteristics. According to the reviewed studies, these models commonly combined lesion descriptors with clinical, demographic, and selected biological variables. Some approaches further integrated deep-learned imaging representations with explicitly engineered radiological and clinical predictors. This extends multimodal modelling beyond conventional structured feature combinations and may capture complementary information from both learned and handcrafted features.

5.4. Longitudinal and Temporal Modelling

Longitudinal and temporal modelling aims to capture disease evolution across time rather than relying only on a single baseline scan. In the studies selected here, temporal information is introduced mainly in two ways: (i) through direct comparison of baseline and follow-up MRI to detect new or evolving lesion activity, and (ii) through forecasting future lesion outcomes from baseline imaging and later follow-up labels [79,80].

A direct paired-image comparison strategy is exemplified by Salem et al. (2022) [80]. Their method uses baseline and follow-up 3D T2-FLAIR scans together, first aligning them and then learning a deformation field that non-linearly registers the baseline image to the follow-up space. The network then uses the baseline image, follow-up image, and learned deformation field jointly to segment new lesions. The model therefore incorporates both spatial and appearance changes between scans rather than relying solely on voxel-wise intensity differences, supporting the detection of newly developed lesions [80].

Durso-Finley et al. (2022) [79] used baseline multi-sequence MRI to predict future new and enlarging T2 lesion counts on follow-up MRI across multiple treatments, thereby framing lesion activity prediction as a forecasting problem from baseline imaging [79]. A related but simpler temporal design appears in Afzal et al. (2018) [81], who used baseline and 1-year follow-up T2-W MRI in a LeNet-style CNN for CIS-to-MS prediction. The study used two time-point MRI inputs but did not describe a separate temporal module such as recurrence or explicit subtraction. Thus, temporal information was incorporated mainly through the inclusion of both scans as model input [81].

In Yılmaz Acar et al. (2022) [82], temporal modelling was approached as future disease-activity prediction using 3D-DWT features extracted from MRI. The study supports a feature-engineered forecasting strategy, rather than a direct baseline–follow-up change-detection framework [82]. Hechenberger et al. (2022) [5] is more appropriately considered a longitudinal prognostic study than a temporal AI model. The study related information

processing speed, together with MRI measures such as lesion load and brain volumes, to later disability progression [5].

Across the selected studies, temporal information was represented primarily through serial MRI inputs or follow-up-defined lesion outcomes. These approaches demonstrate the potential of longitudinal information to support prediction of lesion activity and disease evolution beyond static baseline characterization.

6. Evaluation Strategies and Performance Metrics

The studies summarized in Table 2 evaluated predictive performance using a wide range of metrics, reflecting differences in prediction targets and model outputs. Across the table, the most common measures were accuracy, sensitivity, specificity, and AUC, especially for classification tasks such as CIS to CDMS conversion, lesion activity classification, and general progression prediction. This suggests that, despite differences in the clinical prediction tasks, most studies evaluated model performance using classification-based metrics.

A smaller subset of studies used metrics that were more closely matched to non-classification targets. Kırkbir et al. evaluated 5-year EDSS prediction with RMSE, which is more appropriate for continuous outcome prediction than accuracy or AUC [83]. Elliott et al. used hazard ratios, confidence intervals, and p values within Cox regression and related statistical models [84], showing that time to event or progression risk studies require different evaluation criteria from standard classification frameworks.

Table 2 also shows that evaluation metrics change when the target was lesion activity or lesion detection rather than patient level progression status. Yılmaz Acar et al. reported F1 score, precision, and recall for future disease activity prediction [82], while Salem et al. used F1, Dice, Positive Predictive Value per Lesion (PPVL), and sensitivity for new lesion detection [80]. Durso Finley et al. reported average precision for treatment specific prediction of future new/enlarging T2 lesion burden [79]. These examples show that when the task shifts from patient level outcome prediction to lesion level forecasting or segmentation style detection, overlap based and precision–recall type measures become more informative than accuracy alone.

In terms of validation design, cross-validation (CV) was the most common internal evaluation strategy, but its implementation varied widely across studies. Table 2 includes seven-fold CV, three-fold CV with bootstrapping, repeated stratified 10-fold CV, 10-fold CV, leave one out CV, 5-fold CV, and nested CV, depending on the study design and sample size. This variation makes direct numerical comparison across studies difficult, because reported performance depends not only on the model but also on the way training, validation, and testing were partitioned.

A more distinctive observation is that external validation remained relatively uncommon. Storelli et al. used an independent test set from a different scanner [40], and Coll et al. performed external validation on an independent cohort [72], which are particularly valuable strategies in MRI-based MS prediction because scanner and protocol differences can otherwise lead to over optimistic internal results. Compared with the larger number of studies relying only on internal cross-validation, these examples highlight that true generalizability testing is still limited.

Table 2 further shows that dataset imbalance was handled in several different ways. Some studies used class weighted loss, balanced minibatch sampling, or explicit class imbalance handling during modelling, while others addressed imbalance earlier by removing severely imbalanced variables or by using augmentation strategies such as rotations, reflections, flipping, scaling, and patch resampling. This suggests that imbalance was widely recognized as an important methodological issue, but no single correction strategy dominated across the reviewed studies.

Benchmarking was usually performed within individual studies by comparing several candidate models or feature sets on the same dataset. Table 2 includes comparisons among RF, XGBoost, decision tree, logistic regression, SVM, k NN, naive Bayes, Extra Trees, hist gradient boosting, and DL architectures, depending on the task. In other studies, benchmarking was done by comparing shape-based, intensity-based, and combined feature models [60], or by contrasting wavelet-based and non-wavelet approaches [82]. This indicates that benchmarking was mainly used to identify the strongest model within a study rather than to establish a universal benchmark across the field.

Reproducibility was addressed most clearly in radiomics-based studies. Radiomics-based studies applied ICC-based reproducibility filtering to retain only stable radiomic features before subsequent feature selection [67,68]. Elhaie et al. reduced redundancy through collinearity removal, RFECV, and an additional L1 regularization check [69]. These studies show that, when very large numbers of engineered imaging features are extracted, stability filtering and feature selection procedures become central to reproducible evaluation.

Overall, Table 2 shows that model assessment in MS prediction studies is highly heterogeneous, spanning classification, regression, survival, lesion detection, and treatment response metrics, with cross-validation as the dominant validation strategy and external validation still comparatively rare. It also shows that imbalance handling, within study benchmarking, and feature stability procedures are essential for interpreting reported performance, particularly when models rely on heterogeneous MRI derived biomarkers and high dimensional feature spaces.

Table 2. Summary of the reviewed studies describing AI models, MRI features/variables, and MS progression prediction.

Author & Reference	Cohort, Study Design, Follow-Up and DMT Exposure	Prediction Outcome, Definition and Horizon	Input Data	MRI Preprocessing and Segmentation	Feature Extraction and Quantification	Data Processing and Validation	Model/ Architecture	Performance Metrics	Key MRI Predictors/Contribution	Methodological Considerations
CIS-to-MS/CDMS conversion										
Tayyab et al., (2023)—Prediction of New Disease Activity [85]	Secondary analysis of a multicentre phase III trial involving 142 adults with CIS from 12 Canadian centres; 72 received minocycline and 70 placebo; follow-up up to 24 months.	New disease activity within 24 months, defined as CIS-to-MS conversion according to the 2005 McDonald criteria; 81 converted, 29 remained stable and 32 had uncertain outcomes.	Baseline PDw, T2w and T1w MRI used to derive volumetric measures; final model inputs were eight DGM volumes, T2 lesion volume and treatment assignment.	Bias correction, skull stripping, lesion filling, registration and intensity normalization; semiautomatic lesion segmentation and ANTs multi-atlas deep-grey-matter segmentation.	T2 lesion volume and bilateral thalamus, caudate, putamen and globus-pallidus volumes.	Three approaches to uncertain labels were compared using nested stratified seven-fold cross-validation; the predictor set had been selected previously using the same dataset; no external validation.	Random forest excluding uncertain cases, random forest treating uncertain outcomes as stable, and probabilistic random forest modelling label uncertainty.	AUC: 0.69 for RF-exclude, 0.71 for RF-naïve and 0.76 for probabilistic RF; corresponding F1 scores were 82.6%, 76.8% and 86.6%.	T2 lesion burden and thalamic volumes were among the most informative imaging features; modelling uncertain labels with PRF gave the best performance.	Secondary analysis of a relatively small multicentre trial dataset. Nested 7-fold CV was used, although the predictors had been selected previously from the same dataset. Thirty-two patients had uncertain outcomes, and independent validation was not performed.
Aksoy et al., (2025)—Prediction of CIS-to-MS Conversion [86]	Secondary analysis of an open prospective Mexican CIS cohort; 177 of 273 patients entered complete-case analysis, with 141 used for training and 36 for testing. Follow-up duration and DMT exposure were not reported in the model paper.	Binary prediction of CIS-to-CDMS conversion; 97/177 converted. The diagnostic definition and prediction horizon were not reported in the model paper.	Thirty demographic, medical-history, clinical, CSF, evoked-potential, symptom and binary MRI lesion-location variables.	Not reported; the model used pre-existing binary indicators for periventricular, cortical, infratentorial and spinal-cord lesions rather than raw MRI.	No direct image-feature extraction; MRI locations, oligoclonal bands and evoked potentials were encoded as binary variables, with symptom categories one-hot encoded.	Complete-case analysis, stratified 80/20 train–test split and five-fold cross-validation; several models and ablations were compared on the same small holdout set; no external validation.	Two-stage logistic regression: demographic baseline-risk model followed by a clinical/symptom model incorporating the Stage 1 probability; SHAP used for explanation.	Holdout AUC 0.909, accuracy 0.806, precision 0.842, recall 0.800 and F1 0.821; five-fold cross-validation AUC 0.838 ± 0.095.	Periventricular MRI lesions were the strongest imaging predictor; infratentorial lesions also increased predicted conversion risk. MRI inputs were binary lesion-location indicators rather than quantitative image features.	Complete-case analysis reduced the cohort from 273 to 177 patients, leaving 36 for holdout testing. The CDMS definition and prediction horizon were not clearly reported. Test AUC was higher than the 5-fold CV estimate, and there was no external multicentre validation.

Table 2. *Cont.*

Author & Reference	Cohort, Study Design, Follow-Up and DMT Exposure	Prediction Outcome, Definition and Horizon	Input Data	MRI Preprocessing and Segmentation	Feature Extraction and Quantification	Data Processing and Validation	Model/ Architecture	Performance Metrics	Key MRI Predictors/Contribution	Methodological Considerations
Wottschel et al., (2019)—Prediction of Conversion from CIS to CDMS [87]	Retrospective multicentre cohort of 400 CIS patients from six European MAGNIMS centres; baseline MRI obtained within 14 weeks of CIS onset (reported SD 7 weeks) and one-year follow-up; 91 patients converted to CDMS. Treatment information was unavailable.	Second clinical relapse and CIS-to-CDMS conversion within one year.	Baseline T1-W MRI, PD/T2- or FLAIR-derived lesion masks, age, sex, EDSS and CIS presentation type.	N4 bias correction, lesion filling, registration, GIF brain parcellation/segmentation and DiReCT cortical-thickness estimation.	Regional lesion load, grey-matter probability, cortical thickness, brain volumes and clinical/demographic features.	Bootstrap-averaged RFE with class undersampling and 2-, 5-, 10-fold and leave-one-out CV. The paper does not clearly state that feature selection and stopping were confined to the training folds.	Linear SVM with recursive feature elimination and bootstrap weight averaging.	All-centre accuracy: 64.8% with 2-fold CV, 68.5% with 5-fold CV and 70.8% with leave-one-out CV; smaller single-centre estimates were higher but more susceptible to optimism.	White-matter lesion load was a major predictor, together with regional GM probability, cortical thickness and brain volumes; CIS type was the only selected non-imaging feature.	The multicentre cohort included 91 converters among 400 patients and 213–214 candidate features. Accuracy varied considerably with centre size and CV method, with smaller single-centre cohorts giving higher estimates. RFE and its stopping step were not clearly described as training-fold-specific, and no independent test cohort was used.
Yoo et al., (2016)—Predicting Future Disease Activity [77]	Cohort of 140 patients with early symptoms of MS who did not meet the 2005 McDonald criteria at baseline; 80 met the conversion criteria and 60 did not within two years. Recruitment details and DMT exposure were not reported.	Conversion to meet the 2005 McDonald criteria within two years.	Baseline T2- and proton-density-weighted MRI; the CNN used Euclidean-distance representations of white-matter lesion masks rather than raw MRI intensities.	Skull stripping, intensity normalisation, semi-automated multimodal lesion segmentation, affine MNI152 registration and downsampling to $128 \times 128 \times 30$ voxels.	Euclidean distance transform increased lesion-mask information density; spatial lesion-pattern features were learned automatically by the CNN. Total lesion burden was used as a comparator.	Seven-fold CV with 120 training and 20 test patients per fold; training data were augmented fourfold and class weights were adjusted within each fold. Four CNN configurations were compared.	Nine-layer 3D CNN with three convolution/max-pooling blocks, two fully connected layers and logistic output; a convolutional deep-belief network provided unsupervised pretraining.	CNN with distance transform and pretraining: accuracy 72.9%, sensitivity 78.6%, specificity 65.1% and AUC 71.8%. Lesion-burden logistic regression: accuracy 65.0% and AUC 67.6%.	Spatial lesion-distribution patterns provided more prognostic information than total lesion burden, although the learned features were not anatomically interpretable.	The study included 140 patients and compared four CNN configurations within the same 7-fold CV. Early stopping was based on cross-validation, but a separate validation set within each training fold was not clearly described. Performance was not tested in an independent cohort.

Table 2. *Cont.*

Author & Reference	Cohort, Study Design, Follow-Up and DMT Exposure	Prediction Outcome, Definition and Horizon	Input Data	MRI Preprocessing and Segmentation	Feature Extraction and Quantification	Data Processing and Validation	Model/ Architecture	Performance Metrics	Key MRI Predictors/Contribution	Methodological Considerations
Wottschel et al., (2015)—Prediction of Conversion to MS [88]	Retrospective subgroup of a CIS cohort recruited during 1995–2004. Seventy-four patients were assessed at one year and 70 at three years; none received DMT.	Conversion to CDMS after a second demyelinating attack lasting more than 24 h and occurring at least four weeks after the first attack. Conversion occurred in 22/74 patients by one year and 31/70 by three years.	Baseline PD- and T2-W brain MRI, eight lesion features, and age, sex, CIS presentation type and EDSS.	Baseline 1.5T dual-echo MRI; semi-automated lesion marking by an experienced neurologist, followed by diffeomorphic registration to MNI152 space.	Lesion count, load, PD/T2 intensity, spatial-distribution measures and lesion-size profile, giving 12 candidate predictors with the clinical variables.	Majority-class undersampling was repeated 100 times, producing balanced sets of 44 patients at one year and 62 at three years; each was assessed with LOOCV. Forward feature selection and polynomial kernels (degrees 1–5) were compared within the same analysis.	Support-vector machine with polynomial kernels of degrees 1–5.	Across 100 balanced resamples, 1-year accuracy was 71.4% (sensitivity 77%, specificity 66%); 3-year accuracy was 68.0% (sensitivity 60%, specificity 76%).	Lesion load was the only imaging feature in the best 1-year model. At three years, lesion count, PD intensity and two lesion-location/distance measures were selected.	The cohort was small, with 22 converters at one year and 31 at three years. Repeated undersampling produced balanced subsets for LOOCV, while feature and kernel selection used the same internal evaluation framework. The authors noted that LOOCV may overestimate absolute accuracy, and no independent cohort was tested.
Yoo et al., (2017)—Predicting MS Conversion from CIS [89]	Multicentre cohort of 140 patients with CIS recruited from 12 clinics; 80 converted and 60 did not under the 2005 McDonald criteria within two years. Nine non-converters had only 12-month follow-up; DMT exposure was not reported.	Conversion to MS under the 2005 McDonald criteria within two years, based on new T2 lesions, enhancing lesions and/or clinical relapse.	Euclidean-distance-transformed baseline lesion masks plus lesion volume, brain parenchymal fraction, diffusely abnormal white matter, sex, CIS location, EDSS and mono-/multifocal presentation.	Standardised multisite PD/T2 MRI; skull stripping, intensity normalisation, affine MNI152 registration, semi-automated lesion segmentation with expert seeding and quality review, followed by downsampling to 128 × 128 × 30.	A Euclidean distance transform increased lesion-mask information density, and the CNN learned latent spatial lesion patterns. Eleven standardised clinical and MRI measurements were fused with the learned features.	Sevenfold CV used 120 training and 20 test patients per fold, with training-set augmentation and class weighting. Replication/scaling parameters for feature fusion were selected using the same CV rather than a fully nested procedure.	Nine-layer 3D CNN with three convolution/max-pooling blocks, two fully connected layers and a logistic-regression (softmax) output; convolutional DBN pretraining was used.	Combined CNN: accuracy 75.0%, sensitivity 78.7%, specificity 70.4% and AUC 74.6%. Lesion-pattern-only CNN achieved 72.9% accuracy; random forest with user-defined predictors achieved 67.9%.	Learned spatial lesion patterns outperformed lesion volume alone. Adding clinical and MRI measurements modestly increased accuracy from 72.9% to 75.0%; EDT and pretraining were important for optimisation.	Same 140-patient cohort as Yoo et al. (2016) [77]; nine patients classified as non-converters had follow-up only to month 12. Fusion parameters were tuned using the same 7-fold CV, and the authors acknowledged possible bias because fully nested CV was not used. No independent cohort was tested.

Table 2. *Cont.*

Author & Reference	Cohort, Study Design, Follow-Up and DMT Exposure	Prediction Outcome, Definition and Horizon	Input Data	MRI Preprocessing and Segmentation	Feature Extraction and Quantification	Data Processing and Validation	Model/ Architecture	Performance Metrics	Key MRI Predictors/Contribution	Methodological Considerations
Haouam and Benmalek (2024)—Early Prediction of MS Progression [78]	Retrospective analysis of a Mexican mestizo CIS dataset from NINN, Mexico City, later described as open-source. The dataset contained 273 records, although reported class counts totalled 271 (125 CDMS, 146 non-CDMS). Recruitment dates, follow-up and DMT exposure were not reported.	Binary classification of CIS conversion to CDMS versus non-CDMS. CDMS criteria, outcome ascertainment, follow-up duration and prediction horizon were not reported.	Demographic, medical-history, symptom, oligoclonal-band, evoked-potential and categorical MRI lesion-location variables; initial and final EDSS were removed because of extensive missingness.	Not reported. MRI acquisition, sequences, scanners, image processing, segmentation, quality control and reader reliability were unspecified.	No raw-image, volumetric or radiomic extraction was performed; MRI findings were pre-existing categorical variables. Feature selection was mentioned but its method, selected features and timing were not reported.	Stratified training/test split with 82 patients in the test set. Grid-search CV was used for tuning, but the CV design and feature-selection procedure were not described. The final algorithm was chosen according to test-set performance.	Random Forest, XGBoost, Decision Tree and Logistic Regression; Random Forest was selected as the best-performing model.	Random Forest: AUC 0.93, accuracy 0.87, CDMS precision 0.83 and recall 0.86. Logistic Regression achieved AUC/accuracy 0.90/0.83, Decision Tree 0.86/0.82, and XGBoost 0.90/0.79.	Periventricular MRI was the highest-ranked predictor, followed by age and infratentorial MRI.	Reporting was limited: CDMS criteria and follow-up were not given, and MRI assessment and feature-selection methods were not described. The dataset contained 273 records, but the reported class counts totalled 271. The final algorithm was chosen using the same 82-patient test set, and no independent validation cohort was used.
Bendfeldt et al., (2019)—Prediction of CIS to CDMS Conversion [61]	Post-hoc analysis of the 98-centre BENEFIT phase III trial (468 randomized patients). Baseline MRI was available for 364 patients, although the main geometry and cortical-GM analyses used selected subsets of 160 and 167 patients, respectively. Patients received interferon β-1b or placebo for two years or until CDMS.	Baseline prediction of conversion from CIS to CDMS within two years. CDMS was diagnosed using modified Poser criteria and confirmed by a central committee.	Baseline T1 post-contrast and T2 MRI, lesion masks, lesion count/load, lesion geometry, GM measures, age, sex and EDSS. The cortical-GM analysis used voxel wise GM maps.	Centralised 1.5-T MRI acquisition, assessment and quality control. Trained radiologists marked lesions and technicians quantified volumes. Geometry analysis used registered lesion masks; cortical-GM processing included FAST segmentation, lesion filling, FLIRT, SPM8/VBM8 and DARTEL.	Minkowski functionals measured lesion volume, surface area, mean breadth and Euler–Poincaré characteristic. Features were summarised globally and across 13 ROIs; GM ratios and voxel wise cortical-GM values were also evaluated.	Non-linear SVMs were evaluated with 10-fold CV; parameters were optimized by grid search and significance assessed with 1000 label permutations. The paper does not clearly state whether tuning and feature-reduction steps were repeated within each training fold. Cortical-GM analysis used LOOCV and 500 bootstraps.	RBF-kernel SVMs combined clinical, lesion, GM and geometry features; a separate linear SVM used voxel wise cortical-GM maps. The best non-linear model, M9, used 15 features.	Best non-linear model: 70.4% balanced accuracy in the IFN-β group; best significant placebo model: 67.6%. Cortical-GM analysis reported 71.2% accuracy, 64.0% sensitivity and 78.3% specificity in placebo patients, versus 48% accuracy in treated patients.	Lesion geometry added information beyond conventional lesion count/load and was relatively more important in the IFN-β group. Cortical-GM patterns predicted conversion in placebo-treated patients but not in the treated group.	Although the source trial was large and multicentre, the principal classifiers used selected MRI subsets of only 160–167 patients. Several feature combinations were examined within internal CV, and the paper does not clearly describe training-fold-specific tuning or feature reduction. Performance was moderate and was not tested in an independent cohort.

Table 2. *Cont.*

Author & Reference	Cohort, Study Design, Follow-Up and DMT Exposure	Prediction Outcome, Definition and Horizon	Input Data	MRI Preprocessing and Segmentation	Feature Extraction and Quantification	Data Processing and Validation	Model/ Architecture	Performance Metrics	Key MRI Predictors/Contribution	Methodological Considerations
Rasouli et al., (2024)—Prediction of CIS to MS Conversion [62]	Secondary analysis of a public prospective cohort of 273 Mexican mestizo patients newly diagnosed with CIS during 2006–2010. Patients were assessed before treatment and followed for ten years; subsequent DMT exposure was not reported.	Ten-year prediction of conversion from CIS to CDMS based on final diagnosis using 2010 McDonald criteria. Overall, 125/273 patients converted.	Baseline demographic, clinical, CSF and categorical MRI variables, including age, sex, schooling, symptoms, oligoclonal bands and periventricular, cortical, infratentorial and spinal-cord lesions.	Not reported. MRI was represented by categorical lesion-location findings; acquisition, segmentation, reader assessment, quality control and reliability were unspecified.	A count of positive MRI lesion-location categories and a count of symptom types were engineered. SHAP ranked predictor contributions, and repeated feature-addition experiments identified a nine-feature model.	Variables with $\geq$50% missingness were removed and remaining missing values were imputed with MICE. A 25% hold-out test set was reserved for final evaluation. Hyperparameters were tuned with 1000 training-set trials, and feature selection used stratified 10-fold CV repeated five times. Training-only fitting of MICE was not stated.	XGBoost classifier with SHAP-based global and individual explanations. The final model used nine predictors; tuned hyperparameter values were not reported.	Repeated CV: AUC 0.918, accuracy 0.836, sensitivity 0.839 and specificity 0.834. Hold-out test: AUC 0.858, accuracy 0.783, sensitivity 0.750 and specificity 0.811. Confidence intervals were not reported.	Periventricular MRI was the strongest predictor, followed by OCB; infratentorial MRI was also influential, whereas spinal-cord MRI contributed little.	Same 273-patient Mexican CIS cohort used in other included studies. A 25% hold-out set was reserved for final testing, and model/feature selection was performed on the training data. However, MICE was described before the train–test split, and training-only imputation was not stated. Hold-out performance was lower than internal CV, and no external cohort was tested.
Zhang et al., (2019)—CIS to MS Prediction [60]	Prospective single-centre cohort of 84 CIS patients followed for at least three years. Three patients had received treatment before baseline MRI; the remaining 81 were untreated at baseline. Treatment exposure during follow-up was not reported.	Three-year conversion from CIS to MS under 2010 McDonald criteria. Overall, 66/84 converted: 33 clinically and 33 radiologically; 18 remained non-converters.	Baseline 3D FLAIR and 3D T1 brain MRI from one 3-T scanner. No clinical or demographic predictors were included.	FLAIR resolution $1.03 \times 1.03 \times 1.5$ mm^3 and T1 resolution 1-mm isotropic. Primary masks used computer-assisted manual segmentation by an experienced neuroradiologist; automated LST masks provided a sensitivity analysis. Reader blinding and reliability were not reported.	Lesion volume, intensity histogram measures, surface area, sphericity and surface-to-volume ratio. Patient-level features were the minimum, maximum, mean and SD across lesions, plus total lesion volume and lesion count.	Oblique-RF hyperparameters were selected using out-of-bag error, and performance was evaluated with one random 3-fold CV. The paper does not state whether hyperparameter selection was repeated within each training fold. Feature importance was estimated from 100 bootstrap iterations.	Oblique Random Forest classifiers using intensity, shape or combined features. The best model was shape-based with 300 trees and mtry = 3; baseline dissemination in space was the comparator.	Manual-segmentation shape RF: accuracy 0.85, sensitivity 0.94, specificity 0.50, balanced accuracy 0.72 and DOR 15.50. Automated-LST shape RF: accuracy 0.82 and balanced accuracy 0.64. DIS comparator: accuracy 0.79 and balanced accuracy 0.58; McNemar p = 0.03.	Mean lesion volume, minimum sphericity and minimum surface-to-volume ratio were most influential. Converters had larger, less spherical lesions; lesion count had low importance, and intensity features did not improve prediction.	Small, imbalanced cohort (66 converters, 18 non-converters); the 85% accuracy corresponded to 50% specificity and 72% balanced accuracy. Only one 3-fold CV was used, training-fold-specific tuning was not described, and there was no independent validation.

Table 2. *Cont.*

Author & Reference	Cohort, Study Design, Follow-Up and DMT Exposure	Prediction Outcome, Definition and Horizon	Input Data	MRI Preprocessing and Segmentation	Feature Extraction and Quantification	Data Processing and Validation	Model/ Architecture	Performance Metrics	Key MRI Predictors/Contribution	Methodological Considerations
Future lesion or MRI disease activity										
Yılmaz Acar et al., (2022)—Prediction of Future MS Disease Activity [82]	Secondary analysis of the MICCAI 2021 MSSEG-2 dataset containing 40 patients with MS and MRI at two time points, 1–3 years apart; 29 developed new lesions and 11 did not. Clinical characteristics and DMT exposure were not reported.	New lesions on follow-up MRI 1–3 years after baseline. This was future radiological activity, not disability progression or PIRA.	Baseline whole-brain 3D FLAIR MRI only.	Z-score normalization, resizing to 192 × 192 × 192 voxels, and ANIMA skull stripping; no lesion segmentation in the prediction pipeline.	Two-level 3D-DWT was applied using 10 wavelet families. The second-level LLL approximation sub-band was converted to feature vectors, and the 500 highest-ranked coefficients were retained using ANOVA.	Stratified patient-level 5-fold CV was used. The published workflow appears to perform ANOVA feature selection before CV; training-fold-only feature selection was not described. Different feature-set sizes, 10 wavelets and six classifiers were compared using the same dataset.	SVM, random forest, k-nearest neighbours, decision tree, naïve Bayes and logistic regression.	SVM with SYM4, SYM8 or Haar features achieved F1, precision and recall of 95%. The positive class was absence of new lesions, so recall measured identification of radiologically stable patients.	Second-level LLL wavelet features from whole-brain baseline FLAIR supported prediction of later lesion activity; no specific lesion or anatomical predictor was identified.	Small cohort (n = 40; 11 stable). ANOVA feature selection appears to precede 5-fold CV, and multiple feature-set sizes, wavelets and classifiers were compared on the same data. No independent test cohort was used.
Mohammadi Sepahvand et al., (2019)—Prediction of Future Disease Activity [90]	Proprietary multicentre, multicancer clinical-trial cohort of 1068 RRMS patients followed for two years; 937 had sufficient clinical outcome data for prediction. The characteristics of excluded patients and treatment allocation/exposure were not reported.	MRI activity within two years, defined as at least one new/enlarging T2 lesion or gadolinium-enhancing lesion at year 1 or 2. Approximately 75% were active.	Baseline T1w, T2w, FLAIR and PDw MRI. In lesion-informed models, PDw was replaced by either semi-manual expert or automatically inferred T2 lesion labels.	Brain extraction, N3 bias correction, Nyul intensity normalisation and MNI registration. Semi-manual lesion masks were algorithmically generated and expert-corrected; automated masks used a modified 3D U-Net.	Features were learned from complete 3D MRI volumes; lesion masks provided an additional channel directing the CNN toward baseline lesions.	Stratified 80:10:10 training-validation-test split of 937 patients, with inactive patients oversampled during training; final test n = 95. The segmentation network used scans from all time points, but patient-wise separation of longitudinal scans across segmentation splits was not described.	Five-block 3D CNN for activity prediction and a modified 3D U-Net for lesion segmentation.	MRI only: AUC 0.47; accuracy 26.31%, precision 50.45%, sensitivity 8.45%, specificity 79.16%. MRI + inferred T2 labels: AUC 0.74; accuracy 58.95%, precision 88.09%, sensitivity 54.12%, specificity 79.16%. MRI + expert T2 labels: AUC 0.88; accuracy 80.21%, precision 91.82%, sensitivity 80.11%, specificity 79.16%. AUC values were reported in the ROC figures.	Lesion masks substantially improved prediction over MRI alone. Automated masks performed less well because the segmentation network often missed small lesions, which represented about 40% of lesions.	Of 1068 trial participants, 937 had sufficient outcome data for prediction, but the excluded patients and treatment exposure were not characterised. Prediction performance came from one internal 80:10:10 split. The segmentation network used scans from all time points, but patient-wise separation of longitudinal scans was not described. No independent cohort was tested.

Table 2. *Cont.*

Author & Reference	Cohort, Study Design, Follow-Up and DMT Exposure	Prediction Outcome, Definition and Horizon	Input Data	MRI Preprocessing and Segmentation	Feature Extraction and Quantification	Data Processing and Validation	Model/ Architecture	Performance Metrics	Key MRI Predictors/Contribution	Methodological Considerations
Durso-Finley et al., (2022)—Prediction of Future Lesion Activity [79]	Pooled analysis of 1817 patients with RRMS from four multicentre randomized trials: BRAVO, OPERA 1/2 and DEFINE. Five groups were evaluated: placebo, no-, lower-, moderate- and high-efficacy treatment. MRI was obtained before treatment and at years 1 and 2.	Baseline prediction of NE-T2 lesion count occurring between years 1 and 2 under each treatment. MEDA-T2 was defined as fewer than three lesions. Predicted treatment-specific counts were used to estimate CATE relative to placebo.	Baseline T1, gadolinium-enhanced T1, T2, FLAIR and PD MRI with expert Gad and T2 lesion masks. Imaging-clinical models also used age, sex and EDSS.	Denoising, intensity correction and normalisation; rigid multimodal registration to T1 and stereotaxic space; 1-mm resampling. Network inputs were standardised, resampled to 2-mm isotropic resolution and cropped to 72 × 76 × 52 voxels. Lesion-annotation methods were not described.	A ResNet learned latent features directly from multimodal MRI and lesion masks. Pooled imaging features were concatenated with normalised clinical variables; no handcrafted radiomics or anatomical importance analysis was reported.	Four-by-four nested cross-validation was used. Four inner-fold models formed an ensemble, and performance was calculated from aggregated outer-fold test predictions. No independent external validation or calibration analysis was reported.	TARNET-based model with a shared three-block ResNet encoder and five treatment-specific MLP heads. Regression predicted log-transformed lesion counts; classification predicted MEDA-T2. The regression model with clinical features was used for CATE estimation.	Preferred model: AP 0.79–0.996 and ROC-AUC 0.488–0.836 across treatment groups; log-count MSE 0.043–1.062. In the HE arm, 99.6% had MEDA-T2, giving AP 0.996 but ROC-AUC 0.488.	The multimodal MRI-clinical model identified groups with different treatment-related lesion suppression, but no specific anatomical imaging biomarker was reported.	Large cohort with 4 × 4 nested CV. The HE arm was highly imbalanced (99.6% MEDA-T2), giving AP 0.996 but ROC-AUC 0.488. Individual treatment effects could not be directly validated; no external cohort was used.
PIRA/PIRMA prediction										
Marastoni et al., (2025)—PIRA Prediction [91]	Prospective single-centre cohort of 80 treatment-naïve newly diagnosed RRMS patients; 5-year follow-up; all subsequently received dimethyl fumarate or teriflunomide.	Five-year PIRA, defined as relapse-independent EDSS worsening confirmed ≥6 months later; PIRMA additionally required no new/enlarging or gadolinium-enhancing MRI lesions within ±180 days of the event.	Baseline 3T brain MRI and 68 CSF inflammatory mediators; age and sex were included in subsequent regression/ROC analyses. OCB and CSF/serum albumin ratio were also recorded.	Automated white-matter lesion segmentation, lesion filling, FreeSurfer/MALF volumetric analysis, and manual cortical-lesion assessment.	White-matter and cortical lesion burden, cortical thickness, regional grey-matter volumes, and CSF inflammatory biomarker concentrations.	Random forest feature selection and classification using an 80/20 train–test split, followed by Cox regression and ROC analysis.	Random forest	PIRA: RF test accuracy 0.69 (CSF) and 0.75 (MRI); combined CSF-MRI AUC 0.87 (95% CI 0.78–0.96). PIRMA: RF test accuracy 0.56 for CSF and 0.50 for MRI.	Higher cortical lesion number, lower thalamic volume and lower middle-frontal-gyrus thickness independently predicted PIRA.	Small cohort (80 patients; 23 PIRA events) with many CSF/MRI predictors. RF used one 80/20 split, but held-out evaluation of the subsequent Cox/ROC models was not described; AUC 0.87 is therefore not clearly an independent test estimate. No external validation.

Table 2. *Cont.*

Author & Reference	Cohort, Study Design, Follow-Up and DMT Exposure	Prediction Outcome, Definition and Horizon	Input Data	MRI Preprocessing and Segmentation	Feature Extraction and Quantification	Data Processing and Validation	Model/ Architecture	Performance Metrics	Key MRI Predictors/Contribution	Methodological Considerations
Ianniello et al., (2026)—Predicting short-term PIRA [92]	Single-centre retrospective study of 186 patients with relapsing-course MS; 81 with longitudinal MRI entered the 36-month analysis. Baseline treatment comprised no DMT (10%), high-efficacy DMT (33%) or low/moderate-efficacy DMT (57%).	Confirmed PIRA at 24 and 36 months; progression was confirmed after 24 weeks, sustained for ≥12 months and temporally independent of relapses.	Clinical and treatment variables plus conventional brain/cervical-cord MRI measures, lesion burden, atrophy and regional ChaCo structural-disconnection scores. The 36-month model also included MRI changes through Month 24.	LST-AI lesion segmentation with visual correction, lesion filling, FSL/SIENA brain volumetry, manual cervical-cord area measurement and NeMo-based ChaCo analysis.	Brain and cervical-cord lesion burden, brain volume change, cervical-cord area/change and regional structural-disconnection scores.	Feature selection, SMOTE and Naïve Bayes classification using stratified 80/20 train–test splits and 10-fold cross-validation; no external validation.	Feature-selected Naïve Bayes classifiers for 24-month and 36-month PIRA.	24 months: AUC 0.73, accuracy 84.6%, PIRA precision 0.35 and recall 0.41. 36 months: AUC 0.83, accuracy 83.1%, PIRA precision 0.56 and recall 0.69.	Baseline cervical-cord area and regional structural disconnection contributed to 24-month prediction. The 36-month model additionally selected baseline brain volume, brain-volume loss, new cervical-cord lesions and cingulate/brainstem disconnection.	Few PIRA events (23/186 at 24 months; 16/81 at 36 months). The 24-month accuracy was mainly driven by stable patients (PIRA recall 0.41). The 36-month model used a selected longitudinal subset and month-24 MRI changes. No external validation was performed.
Cagol et al., (2026)—Prediction of Clinical Outcomes [93]	Prospective study of two independent Basel cohorts: Cohort 1, 120 patients with MS, median follow-up 3.5 years; Cohort 2, 279 patients, median follow-up 7.3 years. Most participants received DMTs.	Time to PIRA and PIRMA over variable follow-up, defined using roving-reference, relapse-independent EDSS worsening confirmed after ≥6 months. PIRA occurred in 28/120 and 79/279 patients.	Baseline clinical/demographic variables, brain and cervical-cord MRI biomarkers, and serum sNfL and sGFAP; Cohort 1 additionally included advanced quantitative MRI.	Deep-learning white-matter lesion segmentation with manual refinement, SAMSEG T1-lesion segmentation, FreeSurfer morphometry, expert lesion assessment and Spinal Cord Toolbox measurements.	Brain and spinal-cord volumes, cortical thickness, lesion burden, regional qMRI metrics and serum sNfL/sGFAP z-scores.	Boruta feature selection followed by RF survival/classification models evaluated with 10-fold CV; the main report does not state whether Boruta selection was repeated within each training fold. Analogous biomarker-selection analyses were performed independently in Cohort 2. A four-variable ridge-Cox score developed in Cohort 2 was subsequently evaluated in Cohort 1.	Boruta/random-forest survival and classification models; ridge-penalised Cox model for the parsimonious risk score.	Time-to-PIRA C-index: 0.67 in Cohort 1 and 0.70 in Cohort 2. Four-variable risk score: development C-index 0.64; evaluation in Cohort 1 C-index 0.67, HR 3.60 (95% CI 1.57–8.25).	Cervical spinal-cord atrophy, cortical thinning and deep-grey-matter/brain volume loss were the most reproducible MRI predictors of future PIRA across cohorts.	Two independent prospective cohorts supported reproducibility, although they differed in follow-up and available biomarkers. Boruta selection was not clearly nested within CV. The four-variable risk score developed in Cohort 2 was directly evaluated in Cohort 1.

Table 2. *Cont.*

Author & Reference	Cohort, Study Design, Follow-Up and DMT Exposure	Prediction Outcome, Definition and Horizon	Input Data	MRI Preprocessing and Segmentation	Feature Extraction and Quantification	Data Processing and Validation	Model/ Architecture	Performance Metrics	Key MRI Predictors/Contribution	Methodological Considerations
L. Coll et al., (2025)—Predicting PIRA [72]	Retrospective analysis of 259 prospectively followed patients after a first demyelinating event; 58 (22%) developed PIRA. Median total follow-up was 7.7 years. DMT was not included in the DL model. External validation included 32 MS PATHS patients followed for a median 5.45 years.	Time to first PIRA, defined in the original cohort as 6-month confirmed relapse-independent EDSS worsening. External validation used a surrogate of ≥1-point PDDS worsening without self-reported relapse.	Baseline T1-W and T2-FLAIR brain MRI only, acquired near the first demyelinating event. Clinical, demographic, lesion-count and treatment variables were not DL inputs.	Bias correction, skull stripping, MNI152 registration, T2-FLAIR/T1 co-registration and min–max intensity normalisation.	EfficientNet processed the central 40 axial slices from each sequence; median slice-level predictions produced patient survival estimates. Deep SHAP generated relevance maps.	Five-fold CV with separate training, validation and test folds. Five fitted fold models were applied unchanged to the external cohort without retraining or recalibration, and their predictions were averaged.	ImageNet-pretrained EfficientNet-B0 adapted for discrete-time survival prediction across 11 annual intervals.	Time-dependent C-index 0.72 and IBS 0.10 internally. Threshold accuracy was 78.4% internally and 71.9% externally, but sensitivity was only 25.9% and 25.0%, respectively. Adding DL risk increased Harrell's C from 0.62 to 0.74 internally.	Deep SHAP highlighted frontal and parietal cortex, lateral ventricles and periventricular/frontoparietal white matter as influential regions.	External validation was performed without retraining, but the cohort was very small (32 patients, 4 events) and used a PDDS/self-reported-relapse surrogate rather than the original EDSS-based PIRA definition. Accuracy was driven mainly by specificity, with sensitivity only about 25% in both cohorts.
Poretto et al., (2025)—Early PIRA [76]	Retrospective three-centre Italian study of 719 patients after three exclusions from 722 records. Clinical assessments occurred at baseline and Months 6, 12, 24 and 36. DMT initiation was within one year for 319 patients, within three years for 174, after three years for 132, and absent for 94; treatments were grouped as low or high efficacy.	Non-confirmed PIRA at Month 36, defined by EDSS worsening from Month 24 to Month 36 using Month-24 EDSS-dependent thresholds, without relapse from 3 months before to 1 month after the outcome visit. Predictors through Month 24 provided a 12-month prediction interval.	Routine demographic, clinical, treatment, relapse and conventional MRI variables collected through Month 24. The main processed dataset contained 28 predictors: 16 baseline and four at each follow-up time point.	Brain MRI used 1.5-T or 3-T scanners with at least 3D T1 and FLAIR. Two experienced observers, blinded to patient identity, assessed lesion number, location, new lesions and enhancement. No automated segmentation, volumetry, radiomics, harmonisation or reliability analysis was reported.	Age at onset, treatment lag, DMT efficacy and interval-specific relapse indicators were engineered. Sparse, invariant, imbalanced or clinically irrelevant variables were removed; remaining data were imputed, scaled and one-hot encoded. Training-specific fitting of preprocessing was not clearly reported.	Thirty repeated outer 80:20 stratified train–test splits; within each training set, 20 hyperparameter configurations were compared using repeated stratified 3-fold CV and MCC. Performance was averaged across the 30 outer held-out tests.	Random Forest, Extra Trees, XGBoost, logistic regression and SVC; Random Forest performed best. Permutation importance was calculated across outer iterations.	Random Forest: AUC 0.75 ± 0.06, MCC 0.28 ± 0.10, F1 0.36 ± 0.08. Age ≤ 45 subgroup: AUC 0.77 ± 0.05; NEDA-3 subgroup: AUC 0.80 ± 0.09.	Month-24 EDSS, age at onset, disease duration/onset-to-first-assessment lag and Month-6 EDSS were the leading predictors. Conventional MRI contributed little; baseline periventricular lesion burden was the main MRI feature.	Large multicentre cohort with repeated nested internal validation, but only 92/719 PIRA events and the outcome was not confirmed at a later visit. Training-only fitting of preprocessing was not stated, and no independent cohort was tested.

Table 2. *Cont.*

Author & Reference	Cohort, Study Design, Follow-Up and DMT Exposure	Prediction Outcome, Definition and Horizon	Input Data	MRI Preprocessing and Segmentation	Feature Extraction and Quantification	Data Processing and Validation	Model/ Architecture	Performance Metrics	Key MRI Predictors/Contribution	Methodological Considerations
Disability worsening or confirmed progression										
Aerts et al., (2026)—Prediction of Disability Worsening in MS [94]	Retrospective single-centre cohort of 127 patients contributing 424, sometimes overlapping, prediction episodes; 22 episodes (5.2%) showed disability worsening. Follow-up was 2–3 years. Most patients (76.4%) received moderate/high-efficacy DMTs, but treatment was not modelled.	Unconfirmed EDSS worsening 2–3 years after baseline, defined using baseline EDSS-dependent thresholds; 22/424 episodes worsened.	Baseline FLAIR MRI radiomics, motor evoked-potential time series, EDSS, age and gender.	FLAIR denoising and N4 correction; super-resolution reconstruction for lower-resolution protocols; SAMSEG anatomical segmentation, LST lesion segmentation, intracranial-volume normalization and adaptive-histogram intensity normalization.	Anatomical and lesion volumes; shape, first-order and texture radiomics from WML, NAWM, GM, CSF and thalamus; and conventional/high-dimensional motor-EPTS features.	Twenty repeated patient-grouped stratified 3-fold CV runs; all episodes from one patient remained in the same fold. Fold-specific preprocessing included scaling, Boruta, ADASYN, optional Platt calibration and threshold selection. Hyperparameters had been selected beforehand using a preliminary full-dataset CV.	Logistic regression, random forest, balanced random forest and LightGBM; optional Platt calibration and SHAP interpretation.	Best model by Brier score: combined MRI + EPTS LGBM, Brier score 0.062. Exact AUROC/AP values for this specific winning configuration were not numerically reported in the main article.	MRI radiomics dominated the best model, particularly WML/NAWM texture and NAWM shape features. Motor-EPTS features provided complementary information, and combined models gave the best overall balance of performance.	Only 22 worsening episodes among 424 episodes from 127 patients. Patient-grouped repeated CV prevented within-patient leakage, but hyperparameters were selected before the main CV using the full dataset. Disability worsening was unconfirmed, and there was no external validation.
Wang et al., (2025)—Prediction of Disability Progression [95]	Retrospective two-centre RRMS study (n = 191): Centre A provided 111 training and 47 internal-validation patients; Centre B provided 33 external-validation patients. Follow-up was 2 years. DMT information was collected, although treatment status was inconsistently described.	Two-year disability progression based on EDSS increase and CW based on SDMT decline; neither outcome was subsequently confirmed or required to be relapse independent.	Baseline 3T T1-W, FLAIR and DIR MRI plus demographic, clinical, treatment and disease-involvement variables.	Bias correction, spatial normalisation, automated lesion segmentation with manual correction, lesion filling, CAT12 cortical analysis and choroid-plexus segmentation.	355 regional cortical morphometric features plus 3111 lesion-based shape, first-order and texture radiomic features from 3D T1W, FLAIR and DIR.	Group testing, correlation filtering and cross-validated LASSO feature selection in the training cohort; internal and second-centre validation. CovBat harmonisation incorporated both centres.	Logistic-regression nomograms combining clinical predictors, cortical morphometric scores and radiomics scores; SHAP used for interpretation.	Disability-progression AUC: 0.950 internally and 0.904 externally. Cognitive-worsening AUC: 0.916 internally and 0.889 externally. Calibration curves were also reported.	Cortical atrophy and reduced cortical complexity, together with heterogeneous lesion radiomic features, were associated with future DP and CW.	Small external cohort (n = 33; 13 disability progression, 12 Cognitive worsening). Outcomes were based on single unconfirmed EDSS/SDMT assessments. CovBat used imaging features from both centres, so external validation was not fully independent of preprocessing.

Table 2. *Cont.*

Author & Reference	Cohort, Study Design, Follow-Up and DMT Exposure	Prediction Outcome, Definition and Horizon	Input Data	MRI Preprocessing and Segmentation	Feature Extraction and Quantification	Data Processing and Validation	Model/ Architecture	Performance Metrics	Key MRI Predictors/Contribution	Methodological Considerations
Shen et al., (2025)—Prediction of Clinical Progression [96]	Post hoc analysis of 350 interferon β-1a-treated OPERA-I participants over 96 weeks. A separate 340-patient OPERA-II cohort was used to evaluate transferability of the selected DBM regions.	Time to 24-week confirmed composite disability progression (EDSS, T25FW or 9HPT) or 24-week-sustained ≥20% T25FW worsening. CCDP24 occurred in 81/350 patients; neither outcome required relapse independence.	Baseline T1W brain MRI plus age, sex, years since disease onset, weight, total brain and T2-lesion volumes, EDSS, T25FW and 9HPT.	Resampling, N4 bias correction, denoising and ANTs registration to a healthy-control template; Jacobian maps quantified local volume differences.	Median and SD of regional Jacobian values from 51 atlas regions produced 102 DBM features; variance filtering retained 56. LOCO-MP identified six leading regional DBM features.	Random-survival-forest models used six repeated five-fold cross-validation; CNNs used 30 random 80:20 splits. Feature selection preceded cross-validation. OPERA II repeated model training using the transferred regions.	Random survival forest, penalized Cox models and full-brain or region-based 3D CNNs.	Top DBM Random survival forest models achieved approximately C-index 0.67 for CCDP24 and up to about 0.73 for S25FW. Region-CNN-derived features achieved median C-index 0.61 after RSF retraining.	The most stable DBM predictors involved the third and lateral ventricles, precuneus, cerebellar white matter, lingual and parahippocampal regions. Sparse regional DBM features outperformed the full DBM set.	LOCO-MP feature selection used the full OPERA-I cohort before CV. OPERA II supported transferability of the selected regions, but models were retrained rather than externally tested unchanged. Event information was limited, especially for S25FW.
Bejarano et al., (2011)—Short-Term MS Prognosis Prediction [97]	Prospective single-centre cohort of 51 patients with CIS or MS followed quarterly for two years; 28 received DMT at baseline. Independent validation included 96 patients from a second centre.	Two-year final EDSS category, six-month-confirmed disability progression, and relapse-free status. Fifteen patients progressed and 21 remained relapse-free; progression was not required to be relapse-independent.	Baseline disability and functional measures, MS phenotype, demographics, DMT use, MRI lesion/brain volumes and motor-evoked-potential measures.	Baseline 1.5T MRI; manual T1/T2 lesion delineation (ICC 0.892) and SPM2-based GM/WM volumetry using an MS-optimised VBM procedure. MRI analysis was blinded to MEP results.	T1, T2 and enhancing-lesion burden plus total grey- and white-matter volumes. Classifier-specific wrapper selection identified informative predictor subsets.	Wrapper-based predictor selection followed by 10-fold CV repeated 10 times. Training-fold-specific feature selection was not stated, and a separate inner validation set for neural-network early stopping was not described. The selected EDSS neural network was then evaluated in an independent 96-patient cohort.	Naïve Bayes, boosted logistic regression, decision-tree methods and multilayer-perceptron neural networks.	Internal NNet: EDSS accuracy 80%, AUC 0.76; confirmed progression accuracy 75%, AUC 0.74; relapse-free accuracy 67%, AUC 0.65. External EDSS-model accuracy was 81%.	GM atrophy was the MRI measure most related to future disability, but EDSS and MEP measures were more influential. The externally tested neural network did not include MRI variables.	Small development cohort (n = 51; 15 progression events). Wrapper selection was not clearly nested within CV, and the early-stopping procedure did not describe a separate inner validation set. External testing covered only the EDSS/MEP model and reported accuracy only.

Table 2. *Cont.*

Author & Reference	Cohort, Study Design, Follow-Up and DMT Exposure	Prediction Outcome, Definition and Horizon	Input Data	MRI Preprocessing and Segmentation	Feature Extraction and Quantification	Data Processing and Validation	Model/ Architecture	Performance Metrics	Key MRI Predictors/Contribution	Methodological Considerations
Storelli et al., (2022)—Predicting Disease Progression [40]	Multicentre INNI cohort of 373 patients-299 RRMS and 74 progressive MS-with follow-up at 2 ± 0.5 years. At baseline, 14% were untreated, 50% received first-line DMT and 36% received second-line DMT.	Two-year EDSS worsening and ≥4-point SDMT worsening were modelled separately. A combined EDSS + SDMT model was also reported, but its exact binary outcome rule was not clearly stated. Outcomes were not subsequently confirmed or required to be relapse independent.	Baseline whole-brain T1-W and T2-W MRI only; clinical, treatment and conventional MRI measurements were not CNN inputs.	INNI quality control, brain extraction, T1/T2 alignment and affine registration to MNI152 space. Lesions were manually segmented and centrally checked but were not direct inputs.	Imaging features were learned automatically from complete T1 and T2 volumes. Layer-wise relevance propagation provided qualitative saliency maps.	Development used 325 patients from A1/B/C/D, divided 70:30 into training and validation subsets. Final testing used 48 patients from scanner A2, a different scanner/protocol at the same centre as A1; no independent-centre validation was performed.	Dual-branch 3D CNN with three convolutional blocks per MRI sequence, followed by feature merging and binary classification.	Test accuracy was 83.3% for EDSS worsening, 67.7% for SDMT worsening and 85.7% for the combined model; reported sensitivities were 57.1%, 60.0% and 75.0%, respectively. Test-set AUCs were not reported.	LRP maps mainly highlighted periventricular regions, WM lesions and the corpus callosum; these were qualitative saliency findings rather than quantified biomarkers.	Scanner-held-out test set was small (n = 48) and came from the same centre. Worsening events were few and outcomes were unconfirmed. Reported sensitivities do not fully reconcile with the stated event counts.
Zhao et al., (2020)—Predicting MS Disease Course [98]	Retrospective two-centre cohort of 181 patients with MS (136 RRMS, 45 progressive); mean follow-up 3.94 years, with 62/181 progressing. Treatment changes and relapses during follow-up were not modelled.	EDSS worsening at five years, defined as an increase of ≥1.5 points from baseline. Predictors were collected during the first two years; worsening was neither confirmed nor required to be relapse independent.	Demographic, clinical and MRI-derived longitudinal variables. The complete CLIMB model contained 198 derived features; external models used variables shared with EPIC.	MRI variables included lesion volume and change, new lesions, enhancing-lesion count and brain parenchymal fraction. Acquisition, segmentation, harmonisation and reproducibility procedures were not reported.	Lagged changes were generated from repeated clinical and MRI measures; categorical variables were one-hot encoded. EPIC preprocessing produced 105 features.	Nested stratified 10-fold CV was used in CLIMB, with hyperparameters and minority-class weights tuned in the inner CV. Several decision thresholds were evaluated to examine sensitivity-specificity trade-offs. For external validation, models were rebuilt in CLIMB using annual observations and variables shared with EPIC, then applied directly to EPIC.	SVM, logistic regression, random forest, XGBoost, LightGBM and a stacked logistic meta-learner.	Meta-learner AUC: 0.79 using all CLIMB features, 0.78 using the reduced common-feature CLIMB model and 0.83 in EPIC. At threshold 0.50, the EPIC meta-learner achieved sensitivity 0.78, specificity 0.76 and accuracy 0.77.	EDSS/EDSS change and pyramidal function were the most consistent predictors. MRI contributed through lesion activity and BPF in CLIMB, while GM and ventricular-CSF volumes were prominent in EPIC.	Strong design with nested CV and geographical EPIC validation, although only the reduced common-feature model could be transferred. Prediction relied heavily on EDSS changes during the first two years; treatment was not modelled and the five-year EDSS outcome was unconfirmed.

Table 2. *Cont.*

Author & Reference	Cohort, Study Design, Follow-Up and DMT Exposure	Prediction Outcome, Definition and Horizon	Input Data	MRI Preprocessing and Segmentation	Feature Extraction and Quantification	Data Processing and Validation	Model/ Architecture	Performance Metrics	Key MRI Predictors/Contribution	Methodological Considerations
Taloni et al., (2022)—Disability Progression Prediction [71]	Retrospective two-centre cohort of 181 patients with MS (136 RRMS, 45 progressive); mean follow-up 3.94 years, with 62/181 progressing. Treatment changes and relapses during follow-up were not modelled.	Three-month-confirmed EDSS progression assessed 2–6 years after baseline using baseline-dependent EDSS thresholds. Progression was not required to be relapse independent.	Baseline 3D T1-W brain MRI only; no clinical, treatment, lesion-mask or demographic predictors were used.	Bias correction, brain extraction, linear MNI registration and intensity normalisation. Volumes were divided into axial, sagittal and coronal JPEG slices resized to 224 × 224 pixels.	ResNet50 learned features from individual 2D slices grouped into four anatomical slabs per orientation. Informative regions were estimated by intersecting qualifying slices.	One hundred overlapping stratified 90:10 splits were used; each split contained 163 training patients and 18 validation patients, including only six progressors in validation. A total of 1200 slab-specific classifiers were generated, with slice-level results compared against shuffled-label null models.	ImageNet-pretrained ResNet50 was first adapted to brain MRI using 1833 independent scans and then fine-tuned for progression classification. Separate 2D classifiers were trained for four slabs in each coronal, sagittal and axial orientation.	Among selected slice–classifier combinations passing the null-model criteria, AUCs were 0.72–0.93 coronally, 0.81–0.90 sagittally and 0.69–0.86 axially. Most slice-level AUC distributions were centred near 0.5, and no integrated patient-level performance was reported.	Selected discriminative slices converged mainly in frontal grey-matter regions; this localisation was exploratory and was not independently validated.	Pilot study with overlapping 90:10 splits and only six progressors per validation set. High AUCs came from selected slice classifiers rather than a single patient-level model. No independent test cohort was used.
Zhao et al., (2017)—Predicting MS Disease Course [74]	Prospective CLIMB cohort of 1693 patients with a five-year visit, ≥3 clinical assessments and smoking history. Analyses used 1352 patients at baseline, 1331/1236 for 12-/24-month clinical models and 574 for MRI models. Treatment variables were excluded.	Five-year EDSS worsening, defined as an increase ≥ 1.5 points from baseline. Predictors were collected at baseline or through 12 or 24 months; worsening was not formally confirmed or required to be relapse independent.	Demographic and clinical variables, EDSS and functional-system measures, plus brain parenchymal fraction and whole-brain T2-lesion volume. Longitudinal models included changes from baseline.	Annual MRI was processed using a previously described semi-automated, template-driven segmentation method. Acquisition, scanner and harmonisation details were not reported in this paper.	Whole-brain T2-lesion volume and brain parenchymal fraction; no regional, radiomic or raw-image features.	Missing longitudinal values were interpolated/extrapolated within each patient's time series. Bagging with majority-class undersampling produced 10 balanced classifiers. Performance used 10-fold CV; multiple cost ratios were compared, with no independent validation. Restriction of imputation to the 1-/2-year predictor windows was not stated.	Logistic regression and linear/nonlinear SVMs. Final analyses emphasised linear SVM with bagging, under sampling and false-negative cost ratios of 1–3.	Baseline SVM accuracy increased from 0.64 to 0.70 with MRI. At 12 months, MRI-SVM sensitivity/specificity/accuracy were 0.71/0.68/0.69; at 24 months, 0.65/0.74/0.71. Increasing false-negative cost raised 12-month sensitivity to 0.86 but reduced specificity to 0.47–0.49.	BPF ranked highly for non-worsening, whereas T2 lesion volume ranked among predictors of worsening. MRI improved one-year prediction; extending observation to two years gave little additional overall benefit.	Only 574 patients entered the MRI models. Results came from internal 10-fold CV, with several cost settings compared on the same data. Imputation used longitudinal patient records, but restriction to the 1-/2-year predictor windows was not stated. Treatment was not modelled and there was no external validation.

Table 2. *Cont.*

Author & Reference	Cohort, Study Design, Follow-Up and DMT Exposure	Prediction Outcome, Definition and Horizon	Input Data	MRI Preprocessing and Segmentation	Feature Extraction and Quantification	Data Processing and Validation	Model/ Architecture	Performance Metrics	Key MRI Predictors/Contribution	Methodological Considerations
Khan et al., (2025)—Predicting Disability Worsening [14]	Retrospective two-centre study of 247 patients contributing 814 MRI sessions. Baseline analysis included all patients; six-month delta analysis included 166. DMT exposure was not reported or modelled.	EDSS worsening around two years using baseline-dependent thresholds. Baseline and outcome EDSS values were selected within broad temporal windows, and worsening was not subsequently confirmed.	T2-FLAIR MRI, age, sex and baseline EDSS. Longitudinal models used proportional imaging-feature changes over approximately six months.	Denoising, N4 correction, super-resolution of low-resolution images, intensity matching, SAMSEG tissue segmentation and LST lesion segmentation.	Two hundred WML and NAWM radiomic features, 39 regional volumes, lesion count and lesion volume. Volumes were normalised by intracranial volume.	Subject-grouped 60:20:20 train-validation-test split. Longitudinal ComBat, feature reduction and RFECV were training-derived; Optuna tuning/model selection used development data, with the test set reserved for final evaluation. Each MRI session was treated as a separate observation, while all sessions from one patient remained in the same partition.	Elastic-net logistic regression, balanced random forest and LightGBM, evaluated using clinical, baseline, six-month delta and combined feature approaches.	Best baseline model: PR-AUC 0.20, ROC-AUC 0.64, sensitivity 0.50 and specificity 0.81. Best longitudinal model: PR-AUC 0.11, ROC-AUC 0.69, sensitivity 1.00 and specificity 0.45. Combined models performed poorly.	Baseline WML and NAWM texture, together with left lateral-ventricle volume, were the strongest imaging contributors. Six-month lesion-texture changes showed limited predictive value.	Subject-grouped splitting and training-only processing reduced leakage, but worsening was rare and patients could contribute multiple MRI sessions. Centre-held-out models generalized poorly, so both centres were pooled for the final analysis. The EDSS outcome used broad, unconfirmed follow-up windows.
Andorra et al., (2024)—Predicting Disease Severity [99]	Prospective multicentre Sys4MS cohort of 322 MS patients and 98 controls from Barcelona, Genoa, Berlin and Oslo, followed for approximately two years. At baseline, 70% received DMDs; 22 started treatment and 17 escalated therapy during follow-up. A separate Barcelona cohort was used for model redevelopment.	Baseline prediction of two-year outcomes, principally six-month-confirmed disability accumulation on EDSS, 9HPT, T25WT, SDMT and SL25. Additional outcomes included NEDA, EDSS change and treatment initiation or escalation.	Approximately 100 clinical, demographic, disability, DMD, MRI, OCT, genetic, cytomic and phosphoproteomic variables. MRI predictors included lesion activity/load and normalized brain, grey-matter, white-matter and thalamic volumes.	Standardised 3-T T1-MPRAGE, T2-FLAIR and post-gadolinium T1 MRI. FLAIR was registered to T1 for manual lesion segmentation; lesion-filled T1 images were processed with FSL SIENAX. Lesion and volume analyses were performed centrally by the same operator.	Quantitative MRI measures included T2-lesion volume, enhancing and new/enlarging lesions, and normalized whole-brain, grey-matter, white-matter and thalamic volumes. Correlated variables were filtered, PCA was explored and Random Forest impurity importance ranked features.	Class-frequency weighting addressed imbalance. Missing values were partly handled by RF imputation with 10-fold CV using all available data. Final Sys4MS model-performance validation was not clearly separated from model development. The Barcelona analysis retrained the models because variable definitions differed between cohorts.	Random Forest classifiers using stepwise clinical-only and clinical-plus-imaging, genetic or omics feature sets. Hyperparameters and feature importance were optimized during training; SVMs did not improve performance.	Sys4MS C/I AUCs: EDSS-CDA 0.57, 9HPT-CDA 0.65, T25WT-CDA 0.50, SDMT-CDA 0.62, SL25-CDA 0.80 and NEDA 0.80. Clinical-only EDSS-CDA AUC was 0.62. The retrained Barcelona C/I EDSS-CDA model achieved AUC 0.75.	Baseline disability measures were the strongest predictors. Brain volume and T2 lesion load contributed to several models, but MRI provided little incremental value for EDSS-CDA.	Prospective multicentre cohort, but event numbers were small and missing-value imputation used all available data. Final Sys4MS validation was not clearly separated from model development. Barcelona provided independent replication, but models were retrained rather than externally tested unchanged.

Table 2. *Cont.*

Author & Reference	Cohort, Study Design, Follow-Up and DMT Exposure	Prediction Outcome, Definition and Horizon	Input Data	MRI Preprocessing and Segmentation	Feature Extraction and Quantification	Data Processing and Validation	Model/ Architecture	Performance Metrics	Key MRI Predictors/Contribution	Methodological Considerations
Pellegrini et al., (2020)—Prediction of Disability Progression [100]	Individual-patient-data analysis of 1582 RRMS patients from placebo arms of ADVANCE, DEFINE, CONFIRM and AFFIRM. Follow-up was up to two years; ADVANCE placebo participants switched to pegylated interferon β-1a after one year. Prior MS treatment was reported in 34.5%.	Time to 24-week-confirmed disability progression over up to two years. The primary endpoint combined EDSS progression with worsening on functional measures (T25FW, 9HPT, PASAT and VFT); alternative EDSS/MSFC-based definitions were also examined.	Baseline demographic, disease-history, EDSS, MSFC-4, SF-36 and structured MRI variables, including enhancing-lesion count, T1/T2 lesion volumes, brain-volume Z-score and BPF.	Models used precomputed MRI measures rather than raw images. MRI acquisition, segmentation and quality-control procedures were not reported in this paper. MRI was available for only 44% of DEFINE and 48% of CONFIRM participants.	No radiomic or deep-image features were used. MRI inputs were quantitative lesion and brain-volume measures; predictors were ranked and sequentially combined under each modelling approach.	MCMC multiple imputation produced 10 datasets; study identifier accounted for trial source. Bootstrap optimism-corrected survival C-indices and calibration slopes were used for internal validation. The main article does not state that the complete imputation/model-selection process was repeated within each bootstrap sample. No external validation was performed.	LASSO, ridge and elastic-net Cox models; Cox-based SVM; and unconditional and conditional survival random forests.	All modelling approaches showed limited discrimination (C-index $\leq$ 0.65). For the primary endpoint, the three-best-predictor model had C = 0.59 (95% CI 0.57–0.61), compared with C = 0.65 (0.61–0.69) for baseline MSFC-4 alone; interactions did not improve performance.	MRI measures were not among the strongest predictors of the primary outcome. PASAT, SF-36 physical score and VFT ranked highest, although predictor rankings varied across modelling methods.	Large pooled RCT sample with 24-week-confirmed progression and bootstrap internal validation. However, the placebo-arm populations were highly selected, and MRI was available for only 44–48% in DEFINE/CONFIRM, requiring substantial imputation. External validation was not pursued because discrimination was poor.
Tousignant et al., (2019)—Prediction of Disease Progression [70]	Placebo-arm completers from two multicentre/multiscanner RRMS trials: 465 patients contributing 1083 MRI prediction instances; treated patients and non-completers were excluded.	EDSS progression within one year after the prediction MRI, defined by a baseline-dependent EDSS increase of 0.5–1.5 points sustained for $\geq$12 weeks.	T1 pre- and post-contrast, T2W, PDW and FLAIR MRI; a second model additionally used expert-corrected T2 and Gd-enhancing lesion masks. No clinical or demographic predictors were used.	Site contrast was standardized using dummy scans. MRI underwent brain extraction, N3 bias correction, Nyul intensity normalization and MNI registration. T2 masks were automatically segmented and expert-corrected; gadolinium-enhancing masks were determined by two-expert consensus.	An end-to-end 3D CNN learned features directly from multimodal MRI. Parallel convolutional pathways captured imaging patterns at different spatial scales; no handcrafted radiomics or anatomical attribution analysis was performed.	Minority-class oversampling was applied during training. Four-fold CV was reported together with 75/15/10 train-validation-test proportions, but the relationship between these procedures was not explained. All scans from one patient remained in the same fold.	Inception-inspired 3D CNN with three four-path convolutional blocks and two fully connected layers. The approximately 14-million-parameter model used ReLU, dropout and sigmoid output; a VGG-like 3D CNN was the comparator.	MRI-only CNN: AUC 0.660 $\pm$ 0.055; MRI plus lesion masks: 0.701 $\pm$ 0.027; VGG-like comparator: 0.615 $\pm$ 0.053.	Adding T2 and Gd-enhancing lesion masks improved discrimination and reduced between-fold variability, but the learned imaging features were not anatomically interpreted.	465 patients contributed 1083 repeated scan instances, with relatively few progression-labelled scans for a ~14M-parameter CNN. Patient-level fold separation was maintained, but the 75/15/10 split–4-fold CV relationship was unclear. No independent external validation was performed.

Table 2. *Cont.*

Author & Reference	Cohort, Study Design, Follow-Up and DMT Exposure	Prediction Outcome, Definition and Horizon	Input Data	MRI Preprocessing and Segmentation	Feature Extraction and Quantification	Data Processing and Validation	Model/ Architecture	Performance Metrics	Key MRI Predictors/Contribution	Methodological Considerations
Future EDSS, Disability Status, or Severity Trajectory										
Roca et al., (2020)—Prediction of Clinical Disability [101]	Multicentre OFSEP challenge study. After 19 image-quality/FOV exclusions, 856 patients were used for training and 96 for validation; a separate 475-patient challenge set with withheld outcomes was used for final testing. Treatment information was unavailable.	Absolute EDSS score approximately two years after baseline MRI. Baseline EDSS was unavailable; therefore, EDSS change, confirmed progression and PIRA were not evaluated.	Baseline FLAIR, age and sex. CNN models learned features from FLAIR/WMH maps; classical models used ventricular volume, whole-brain lesion load and tract-specific lesion burden across 60 WM regions.	N4 bias correction, ANTs registration to a customised FLAIR template, intensity normalisation and resizing. A pretrained and retrained CNN generated white-matter lesion maps.	CNNs learned image features directly. Classical models used age, sex, acquisition type, ventricular volume, whole-brain lesion load and lesion burden in 60 white-matter tracts.	Stratified 90:10 training-validation split. Validation-set errors determined EDSS-range-specific ensemble weights. Final evaluation used a separate 475-patient challenge set whose EDSS outcomes were withheld from investigators; no external cohort outside OFSEP was tested.	Ensemble of two 3D CNN regressors, random-forest regressors and a UMAP-based predictor, combined using validation-derived weighted averaging.	Aggregated model MSE 2.21 on validation and 3.00 on challenge testing, corresponding to approximate RMSEs of 1.49 and 1.73 EDSS points. The reported "1.7-point error" should not be interpreted as MAE.	Tract-specific lesion burden, particularly in the posterior corona radiata, together with age and lateral-ventricle volume, contributed most strongly to prediction. Imaging-based models outperformed age alone.	Large multicentre dataset with a separate outcome-blinded challenge test. However, the target was absolute two-year EDSS rather than progression because baseline EDSS was unavailable. Development used one 90:10 split, and prediction was poorer at the EDSS extremes; treatment information was unavailable.
Campanioni et al., (2024)—Prediction of MS Progression [102]	β1 and β2 longitudinal EDSS trajectory descriptors, plus categorical EDSS prediction at 1, 2, 5 and 10 years. EDSS(0) was cross-sectional and should not be treated as future prediction.	β1 and β2 longitudinal EDSS trajectory descriptors, plus categorical EDSS prediction at 1, 2, 5 and 10 years. EDSS(0) was cross-sectional and should not be treated as future prediction.	Human-entered structured findings from baseline MRI plus sex and age at MS onset; XGBoost did not analyse MRI voxels directly.	Detailed MRI acquisition, preprocessing, harmonisation, segmentation and reader-reliability methods were not reported for the structured MRI findings.	Standardised ordinal and binary MRI findings were used; no raw-image, voxel or radiomic features were extracted.	Bayesian Hyperopt-tuned XGBoost with an 80/20 fitting-evaluation split and stratified five-fold CV. The relationship between the holdout split, CV and hyperparameter tuning was not clearly described. No external validation.	Linear and XGBoost regression for β1/β2; multinomial logistic regression and XGBoost classification for EDSS(t). Hyper optimised XGBoost performed best.	β1/β2 MAE: 0.11/5.18. β1-derived course classification: AUC 0.86, sensitivity 0.84, accuracy 0.88. Future EDSS(t) AUC/accuracy: 1 y 0.91/0.78; 2 y 0.86/0.75; 5 y 0.88/0.79; 10 y 0.89/0.73.	Age at onset dominated β1/β2 prediction, whereas the baseline MRI finding of ≥9 brain lesions was the strongest EDSS(t) classifier feature.	Later-horizon analyses had fewer patients (400 at 1 year; 279 at 10 years) and excluded rare EDSS classes. The relationship between the 80/20 split, five-fold CV and Hyperopt tuning was unclear. No external validation was performed.

Table 2. *Cont.*

Author & Reference	Cohort, Study Design, Follow-Up and DMT Exposure	Prediction Outcome, Definition and Horizon	Input Data	MRI Preprocessing and Segmentation	Feature Extraction and Quantification	Data Processing and Validation	Model/ Architecture	Performance Metrics	Key MRI Predictors/Contribution	Methodological Considerations
Kırkbir et al., (2025)—Predicting Disease Progression [83]	Retrospective two-centre MSBase cohort of 1000 RRMS patients selected from 3033 records, all with ≥5 years' follow-up. All were reported to receive first-line treatment; 89.8% started treatment within two years, but treatment changes were not modelled.	Regression prediction of continuous fifth-year EDSS from baseline and earlier follow-up variables. The timing of the second predictor visit is inconsistently reported as 12 versus 24 months. The target was future EDSS status, not confirmed progression or PIRA.	Demographic, relapse/treatment, KFS and EDSS variables plus radiologist-entered categorical MRI lesion descriptors. Raw MRI images were not used.	MRI lesion findings were recorded as categorical text variables; acquisition, segmentation, harmonisation, quality control and reader reliability were not reported.	Final correlation-based selection retained 12 age/KFS/EDSS variables; no MRI predictor remained in Model 3.	Five variables with >40% missingness were removed; CART was selected over MICE, RF and KNN for imputation, followed by standardisation. A random 70:30 development-test split was reported, but the validation fraction was inconsistent. Training-only preprocessing and outcome-based feature selection were not stated.	Three LSTM regression models: initial all-variable model, GridSearch-optimised model and feature-selected GridSearch model. The final model used two hidden layers of 10 and 50 units, Adam, ReLU, dropout 0.1, batch size 64 and 100 epochs. Sequence construction was insufficiently described.	Model 1 RMSE 1.46; GridSearch Model 2 RMSE 1.41; feature-selected Model 3 RMSE 1.33. The MRI-containing models were Models 1–2; Model 3 contained no MRI predictor.	MRI lesion descriptors showed some univariate associations with fifth-year EDSS, but none was retained in the final model; prior EDSS and pyramidal functional-system scores were stronger predictors.	Only 1000/3033 records were included. Training-only preprocessing and feature selection were not stated, and the validation split was inconsistently reported. All three models were compared on the same test set; the final RMSE 1.33 model contained no MRI predictor.
Al-Iedani et al. (2024)—Prediction of Cognitive Decline [103]	Prospective Australian RRMS cohort (n = 43) followed for 5 years. DMTs were dimethyl fumarate (n = 9), fingolimod (n = 21) and injectables (n = 13); 10 patients relapsed during follow-up.	Five-year changes in tARCS, SDMT, cognitive domains and EDSS. For this progression review, ΔEDSS is the relevant disability outcome; binary cognitive decline was defined post hoc as ΔtARCS < 0.	Seventy baseline multimodal MR measures from structural MRI/FLAIR, DTI and MRS plus six clinical predictors.	SIENAX/FSL/SPM12 and lesion-growth segmentation for structural/lesion measures; MRtrix/FSL/ANTs for DTI; LCModel for MRS.	Brain, GM, WM, CSF and lesion volumes; DTI metrics (FA, MD, RD, AD); regional MRS metabolite measures. Best model incorporated multimodal imaging features and disease duration.	GLMnet/caret framework. Ridge, Lasso, and Elastic Net evaluated using RMSE-based model selection. No external validation; internal validation procedure incompletely reported.	Penalised regression (Ridge, Lasso, Elastic Net). Lasso performed best for 5-year tARCS prediction.	ΔEDSS R^2 = 0.25. Secondary cognitive models: ΔtARCS R^2 = 0.54 and ΔSDMT R^2 = 0.39; post-hoc binary ΔtARCS decline AUC = 0.92.	For cognitive change, multimodal MR features outperformed conventional volumetrics alone. MRS metabolites, FA measures, CSF/lesion measures and disease duration contributed to the tARCS model; exact selected predictors for ΔEDSS were not reported in the main paper.	Only 43 patients were modelled from 76 candidate predictors. Regularisation was used, but protected internal validation was not clearly reported and there was no external validation. The AUC = 0.92 relates to an arbitrary cognitive-decline threshold, not the ΔEDSS model.

Table 2. *Cont.*

Author & Reference	Cohort, Study Design, Follow-Up and DMT Exposure	Prediction Outcome, Definition and Horizon	Input Data	MRI Preprocessing and Segmentation	Feature Extraction and Quantification	Data Processing and Validation	Model/ Architecture	Performance Metrics	Key MRI Predictors/Contribution	Methodological Considerations
RRMS-to-SPMS Conversion										
Rocca et al. (2021)—Clinical Worsening Prediction [104]	Prospective single-centre cohort of 233 patients with MS (157 RRMS, 59 SPMS, 17 PPMS), median follow-up 6.4 years. Most patients received DMTs; 111/233 changed treatment during follow-up.	Three-month relapse-free confirmed EDSS worsening in all MS patients (105/233) and RRMS-to-SPMS conversion defined by ≥1-year irreversible relapse-independent EDSS worsening (26/157).	Baseline demographic/clinical variables and quantitative structural/resting-state MRI features; follow-up DMT change was also included in the reported RF models.	Baseline 3T structural and resting-state MRI with lesion-volume quantification, brain volumetry, and network analysis.	Functional connectivity networks, grey-matter covariance networks, lesion volumes, and global brain-volume metrics.	Whole-dataset logistic screening of 62 candidate predictors followed by RF with bootstrap/OOB evaluation. Screening was not nested within OOB resampling; no independent external validation.	Random forest	Complete RF: disability worsening OOB AUC 0.76 (95% CI 0.69–0.82), Brier 0.199; SPMS conversion OOB AUC 0.84 (0.76–0.91), Brier 0.111.	Lower global/GM volume and default-mode, sensorimotor and frontoparietal network abnormalities contributed to worsening prediction; normalized GM volume and sensorimotor-network atrophy were retained for SPMS conversion.	Whole-dataset outcome screening preceded RF/OOB evaluation, which the authors acknowledged could cause double-dipping and overfitting. Only 26 RRMS patients converted to SPMS. More importantly, DMT change measured during follow-up was included in the final models, so performance was not based solely on baseline predictors.
Pisani et al., (2021)—Predicting Conversion to SPMS [105]	Retrospective single-centre cohort of 262 RRMS patients followed for a mean 9.55 years (range 6.79–13.13); 69 converted to SPMS. All initially received first-line DMT and 114 later switched to second-line therapy.	SPMS conversion during long-term follow-up, predicted from data collected from diagnosis through year 2. Progressive disease was defined by continuous disability accumulation independent of relapse and confirmed 12 months later.	Age, sex, EDSS, early relapses and baseline/2-year MRI measures of cortical and WM lesions, cortical thickness, cerebellar volume and spinal-cord lesions. The upstream RSF was additionally adjusted for second-line treatment switching, but treatment was not included in the final SP-RiSc.	Serial 1.5T brain and spinal-cord MRI was acquired using 3D MP-RAGE, FLAIR, DIR, PD/T2, and STIR sequences, with standardized patient positioning for longitudinal imaging. Detailed lesion segmentation and volumetric-processing procedures were not described in the main article.	Twelve candidate variables entered RSF; seven with minimal depth <4.21 were combined into SP-RiSc: cortical-thickness change, cerebellar cortical-volume change, baseline and new cortical lesions, age, baseline EDSS and baseline WM lesion number.	Random 80:20 split (219 development; 43 held-out test). Feature selection, score construction and the 17.7 ROC/Youden cut-off were derived in development data; the fixed cut-off was then applied to the test set.	Random survival forest followed by the weighted SP-RiSc score; Cox models and a score excluding gray-matter measures were comparators.	Held-out test: accuracy 88%, sensitivity 92%, specificity 87%, PPV 75% and NPV 96%. The test set contained 13 converters and 30 nonconverters.	Early cortical thinning, cerebellar cortical-volume loss and cortical-lesion burden were major predictors. Excluding GM measures reduced test accuracy to 74% and specificity to 67%.	A locked cut-off was tested in a separate same-centre holdout, but only 13 test patients converted. Binary test metrics did not clearly account for variable follow-up, and the upstream RSF adjusted for treatment switching during the observation period. No external cohort was tested.

Table 2. *Cont.*

Author & Reference	Cohort, Study Design, Follow-Up and DMT Exposure	Prediction Outcome, Definition and Horizon	Input Data	MRI Preprocessing and Segmentation	Feature Extraction and Quantification	Data Processing and Validation	Model/ Architecture	Performance Metrics	Key MRI Predictors/Contribution	Methodological Considerations
R. Seccia et al. 2020—Predictions on MS Course [106]	Retrospective single-centre real-world cohort of 1624 MS outpatients contributing 18,574 visits from 1978–2018; 207 were recorded in the SP phase. Mean observation was six years. Treatment indicators included first-/second-line DMTs, immunosuppressants and other therapies.	Visit-level prediction of neurologist-recorded RRMS-to-SPMS transition within 180, 360 or 720 days. Disease stage could be revised at later visits; no standardized SPMS definition or confirmation interval was reported.	Up to 21 demographic, EDSS/relapse, treatment, CSF and categorical routine-MRI variables. Visit-oriented models used one visit; the LSTM used the patient's preceding visit sequence.	No raw MRI was analysed. MRI information consisted of routine categorical indicators of enhancing lesions and lesion location; scanner, segmentation, harmonisation, QC and reader-reliability details were not reported.	No imaging feature extraction was performed. Feature-saving retained 21 predictors after complete-case deletion; record-saving removed Status T1/T2 and oligoclonal bands, retaining 18 predictors and more visits. No imputation was used.	Balanced bootstrap subsampling addressed severe class imbalance. Visit-oriented models used patient-level leave-one-group-out CV; nesting of grid-search tuning was unclear. The LSTM used one stratified train–test split of unreported size and trial-and-error architecture selection. No external validation was performed.	Visit-oriented: Random Forest, nonlinear SVM, AdaBoost and KNN. History-oriented: 10-cell LSTM, sigmoid feed-forward output and dropout 0.2.	Visit models achieved recall 64.9–100%, but precision was only 2.0–9.3%. The 720-day record-saving LSTM achieved accuracy 98.0%, recall 67.3%, precision 42.7% and specificity 98.5%. AUROC, calibration, confidence intervals and external performance were not reported.	Using longitudinal patient history substantially improved precision compared with single-visit models, particularly at 720 days, although recall was lower. Retaining more visits generally helped more than retaining Status T1/T2 and oligoclonal-band variables.	Transition-labelled visits were extremely rare (~1%), so high accuracy mainly reflected the large non-transitioning class and positive-prediction precision remained limited. SPMS status was routinely clinician-recorded and could later be revised. Visit models used patient-level LOGO, but the LSTM used only one train–test split.

2D, two-dimensional; 3D, three-dimensional; DGM, deep grey matter; 9HPT, Nine-Hole Peg Test; ANTs, Advanced Normalization Tools; AUC, area under the curve; AUROC, area under the receiver operating characteristic curve; BPF, brain parenchymal fraction; CART, classification and regression tree; CATE, conditional average treatment effect; CCDP24, 24-week confirmed composite disability progression; CDA, confirmed disability accumulation; CDMS, clinically definite multiple sclerosis; CI, confidence interval; CIS, clinically isolated syndrome; CNN, convolutional neural network; CSF, cerebrospinal fluid; CV, cross-validation; DBM, deformation-based morphometry; DIR, double inversion recovery; DIS, dissemination in space; DL, deep learning; DMD, disease-modifying drug; DMT, disease-modifying therapy; DOR, diagnostic odds ratio; EDSS, Expanded Disability Status Scale; EHR, electronic health record; EPTS, evoked-potential time series; FLAIR, fluid-attenuated inversion recovery; FSL, FMRIB Software Library; GM, grey matter; HR, hazard ratio; ICA, independent component analysis; KFS, Kurtzke Functional Systems Score; kNN/KNN, k-nearest neighbours; LASSO, least absolute shrinkage and selection operator; LSTM, long short-term memory; MAE, mean absolute error; MALF, multi-atlas label fusion; MCC, Matthews correlation coefficient; MCMC, Markov chain Monte Carlo; MEP, motor-evoked potential; MICE, multiple imputation by chained equations; ML, machine learning; MLP, multilayer perceptron; MNI, Montreal Neurological Institute; MRI, magnetic resonance imaging; MS, multiple sclerosis; MSE, mean squared error; NAWM, normal-appearing white matter; NEDA-3, no evidence of disease activity-3; NE-T2, new/enlarging T2 lesions; NPV, negative predictive value; OCB, oligoclonal bands; OCT, optical coherence tomography; OOB, out-of-bag; PCA, principal component analysis; PD, proton density; PDDS, Patient-Determined Disease Steps; PIRA, progression independent of relapse activity; PIRMA, progression independent of relapse and MRI activity; PPMS, primary progressive multiple sclerosis; PPV, positive predictive value; PR-AUC, area under the precision–recall curve; qMRI, quantitative magnetic resonance imaging; RBF, radial basis function; ReLU, rectified linear unit; RF, random forest; RFE, recursive feature elimination; RFECV, recursive feature elimination with cross-validation; RMSE, root-mean-square error; ROC, receiver operating characteristic; ROC-AUC, area under the receiver operating characteristic curve; ROI, region of interest; RRMS, relapsing-remitting multiple sclerosis; RSF, random survival forest; SD, standard deviation; SDMT, Symbol Digit Modalities Test; SEL, slowly expanding lesion; sGFAP, serum glial fibrillary acidic protein; SHAP, SHapley Additive exPlanations; SMOTE, Synthetic Minority Over-sampling Technique; sNfL, serum neurofilament light chain; SPMS, secondary progressive multiple sclerosis; SP-RiSc, secondary progressive risk score; SuStaIn, Subtype and Staging Inference; SVC, support vector classifier; SVM, support vector machine; T25FW, Timed 25-Foot Walk; T25WT, Timed 25-Foot Walk Test; UMAP, Uniform Manifold Approximation and Projection; VFT, verbal fluency test; WML, white matter lesion; XGBoost, eXtreme Gradient Boosting; tARCS, Total Audio Recorded Cognitive Screen score; LCModel, Linear Combination Model.

7. Emerging Directions in AI for MRI Analysis

Emerging AI methods are beginning to reshape MRI analysis in MS by moving beyond narrowly task-specific models toward reusable representations, collaborative training strategies, and more interpretable systems. This shift is especially relevant in MS, where scanner variability, modest cohort sizes, and the need for longitudinal monitoring have traditionally limited robustness and generalizability. Recent MS-focused reviews also note that many methods remain early-stage, but the field is clearly moving toward larger, more transferable, and clinically deployable frameworks [107].

One important direction is the rise of foundation models for medical imaging. In general terms, foundation models aim to learn broad reusable image representations that can later be adapted to multiple downstream tasks rather than retrained from scratch for each one. In MRI, Sun et al. reported a foundation model for motion correction, resolution enhancement, denoising, and harmonization, trained on a large and diverse set of images and shown to handle pathological brains including MS. Their model also improved downstream segmentation, registration, and diagnostic tasks, suggesting that foundation-style models may become valuable early-stage components in MS MRI pipelines as universal preprocessors or representation backbones [108]. A more recent example is BrainIAC, a general-purpose brain MRI foundation model trained with contrastive self-supervised learning and validated across seven downstream applications, showing how large pretraining quantities can support transferable latent representations for later task adaptation [109].

Closely related to this is self-supervised and contrastive learning, which reduce dependence on expensive manual labels by learning useful representations from unlabeled images. A systematic review of self-supervised learning in medical imaging found that self-supervised pretraining generally improves performance over purely supervised training and grouped current methods into contrastive, generative, innate-relationship, and self-prediction strategies [110]. This is particularly relevant to MS, where lesion labels, progression labels, and harmonized multicenter annotations are costly to obtain [111]. The BrainIAC study is also important in this context because it shows how contrastive SSL can be scaled to large brain MRI collections and then adapted to downstream tasks, providing a concrete example of how self-supervised learning and foundation modelling are beginning to converge [109].

Another key direction is federated learning and privacy-preserving AI. In MS, Hindawi et al. demonstrated a real-world, multicenter federated learning framework for lesion segmentation using nnU-Net, trained across three sites and 512 MRI cases without sharing raw patient data. Their study is important not only because it showed technical feasibility, but also because it addressed the practical regulatory and governance barriers that often prevent centralized pooling of MS MRI data. For future progression modelling, this kind of distributed training is likely to be increasingly important, particularly when multi-site robustness is needed [112].

Explainable AI is also emerging as a central requirement for clinical translation. In progression modelling, Campanioni et al. used baseline MRI and SHAP-based analysis to predict MS trajectory descriptors while identifying which features most strongly influenced model output [102]. At the segmentation level, Spagnolo et al. proposed instance-level explanation maps for MS lesion segmentation by extending SmoothGrad and Grad-CAM++ and showed that the model relied more on FLAIR than MPRAGE, with lesion-border context contributing to detection [113]. These studies are useful because they move beyond generic black box prediction and begin to connect AI decisions back to anatomically and clinically interpretable MRI patterns.

Large-scale and multi-center MRI resources are another major driver of progress. MindGlide showed how DL can repurpose large routine MRI archives by learning to extract brain region and WM lesion volumes from single contrast MRI; it was trained on 4247 scans from 2934 MS patients across 592 scanners and externally validated on 14,952 scans from 1001 patients [114]. At a broader disease-modelling level, Eshaghi et al. used MRI-based unsupervised ML to identify three data-driven MS subtypes from a training cohort of 6322 patients, with an independent validation cohort of 3068 patients, and showed that these subtypes differed in disability progression and treatment response [115]. Together, these examples show that larger and more heterogeneous MRI collections are not only improving validation but are also enabling new biological and prognostic questions to be addressed.

Overall, these emerging directions suggest that the next phase of AI for MS MRI analysis will move beyond isolated single-task models toward more robust and clinically useful frameworks. For MS specifically, the most important near-term advances are likely to come from integrating these developments across the full MRI pipeline, from preprocessing and biomarker extraction to prognosis and treatment-response modelling.

8. Discussion

The central finding of this review is that MRI-based prediction of MS progression should be considered as a complete analytical pipeline rather than as the performance of a single algorithm. Image quality, spatial alignment, intensity standardization, harmonization, segmentation, feature construction, data processing, and outcome modelling are closely connected. Errors or variability introduced at an earlier stage may affect lesion or tissue measurements and can subsequently influence the predictive model. This issue is particularly relevant in multicenter and longitudinal studies, where differences in scanners, acquisition protocols, and image-processing procedures may resemble disease-related change. Reported model performance should therefore be interpreted in relation to the complete MRI and modelling workflow rather than the final classifier or neural network alone.

The reviewed studies also demonstrate that the prognostic information contained in MRI extends beyond conventional lesion count and total lesion volume. These measures remain useful indicators of inflammatory lesion burden, but lesion shape and spatial distribution provided additional information for predicting conversion from CIS to MS or CDMS [60–62]. Grey-matter and deep-grey-matter measures also contributed to prediction across CIS conversion, PIRA/disability progression, and RRMS-to-SPMS transition [85,91,93,96,104,105]. Radiomic and other high-dimensional MRI features contributed to disability-worsening prediction [14,94,95], while learned MRI representations also supported future lesion-activity and disability prediction [40,70,72,79,82,90]. These findings are consistent with the clinico-radiological paradox, because lesions with similar total burden may differ in their anatomical location, structural characteristics, and clinical effect [8]. The selection of imaging biomarkers should therefore be related to the intended prediction outcome rather than based only on increasing the number of available imaging features.

The evidence did not show that increasing model complexity consistently improved prediction. Classical ML approaches remained effective when MRI information was represented by structured and interpretable variables, including lesion geometry, lesion burden, regional measures, radiomics, and combined imaging and clinical variables. DL approaches were more directly suited to settings in which prognostic representations were learned directly from MRI volumes or other high-dimensional imaging inputs [40,70,72,79,90,101]. These approaches address different modelling requirements and should not be compared only according to the highest reported AUC or accuracy. More informative comparisons require evaluation on similar populations and outcomes and comparison with simpler clin-

ical or lesion-based models. This is particularly important for multimodal models because improvements in predictive performance may partly arise from age, baseline disability, treatment information, or other clinical variables rather than from MRI alone.

A major challenge was the heterogeneity of outcomes described as MS progression. The included studies addressed six main prediction settings: CIS-to-MS or CDMS conversion, future patient-level lesion or MRI disease activity, PIRA/PIRMA, disability worsening or confirmed progression, future EDSS or disability status, and RRMS-to-SPMS transition. These outcomes represent different aspects of disease evolution and are not directly interchangeable. Formal composite endpoints also differed in purpose. Yoo et al. used CIS-to-MS conversion according to the McDonald diagnostic criteria [89], whereas Shen et al. used 24-week confirmed composite disability progression based on EDSS, T25FW, and 9HPT [96]. These predefined diagnostic or disability composites were distinguished from generic non-separable disease-activity composites when determining eligibility. Even within disability-related studies, operational definitions differed substantially. Poretto et al. used non-confirmed PIRA at Month 36 [76], Ianniello et al. required PIRA to be confirmed after 24 weeks and sustained for at least 12 months [92], and Cagol et al. used relapse-independent EDSS worsening confirmed for at least six months [93]. In contrast, Aerts et al. evaluated unconfirmed EDSS worsening [94]. These differences in confirmation requirements, relapse independence, and follow-up duration limit direct comparison of model performance across studies. Direct AI-based prediction of RRMS-to-SPMS transition remained uncommon and was identified in Rocca et al. [104], Pisani et al. [105], and Seccia et al. [106]. Consequently, similar performance measures may represent substantially different prediction tasks, emphasizing the need for consistent outcome definitions and prediction horizons.

At the level of the included studies, validation design remained one of the most important methodological limitations. Cross-validation was widely used, but its implementation differed considerably across studies. Table 2 also shows that feature selection, imputation, preprocessing, hyperparameter selection, and other data-dependent procedures were not always clearly confined to the training data. In some studies, predictor screening or feature selection was performed before cross-validation or before internal evaluation. In others, the relationship between preprocessing, model selection, and the reported validation procedure was unclear. These practices can lead to optimistic performance estimates because information from the evaluation data may indirectly influence model development. Patient-level separation and clear placement of all data-dependent procedures within the training process are therefore essential when evaluating MRI-based progression models.

External validation was also limited compared with internal evaluation. Several studies used independent cohorts or second-centre datasets, but many relied entirely on cross-validation, internal holdout testing, or out-of-bag evaluation. The methodological information summarized in Table 2 further indicates that some studies repeated model development or feature selection in an independent cohort rather than applying a completely fixed model. Such studies provide useful evidence of reproducibility, but they do not provide the same evidence as direct evaluation of a previously developed model in an independent population. Furthermore, reductions in performance were observed in some external evaluations, illustrating the challenge of maintaining predictive performance across scanners, sites, and populations. Generalizability therefore remains insufficiently established for many of the reviewed approaches.

Differences in study populations and treatment exposure provide an additional source of heterogeneity. The included studies ranged from clinical-trial and registry populations to single-centre observational cohorts, and DMT exposure was handled inconsistently. Some studies included treatment-related variables in the prediction model, whereas others

included treated populations without explicitly modelling treatment effects, and several studies did not report treatment information. This issue is particularly relevant to PIRA-related prediction. Poretto et al. incorporated treatment lag and DMT efficacy among the candidate predictors [76], whereas treatment use was not incorporated into the prediction model in Coll et al. [72]. Differences in DMT exposure and treatment changes may therefore influence subsequent progression trajectories and complicate interpretation of the independent contribution of MRI biomarkers. The available reporting also provided limited evidence for assessing generalizability across pediatric populations and across racially or ethnically diverse groups. These factors should be considered when determining whether a model developed in one cohort can be applied to a broader MS population.

Some included publications also analyzed overlapping patient populations, which should be considered when interpreting the overall evidence. Aksoy et al., Haouam and Benmalek, and Rasouli et al. were based on the same or substantially overlapping Mexican CIS source cohort, although the analytic samples, preprocessing, and modelling approaches differed between reports [62,78,86]. Similarly, Yoo et al. (2016) and Yoo et al. (2017) analyzed the same 140-patient cohort, with the later study extending the earlier model by incorporating additional clinical and MRI variables [77,89]. These studies were therefore treated as separate modelling reports but not as independent patient cohorts or independent replications of predictive performance. Consequently, concordant findings across these publications should not be interpreted as equivalent to replication across independently recruited populations, and participant numbers from overlapping reports should not be summed as unique evidence.

Sample size and the number of progression events also require careful consideration. Cohort sizes varied markedly across the reviewed studies, ranging from 40 patients in Yılmaz Acar et al. [82] to 1817 participants in Durso-Finley et al. [79]. Such differences limit direct comparison of reported predictive performance across studies. Several studies also evaluated large numbers of imaging features or complex models despite relatively small numbers of patients or progression events. A large total cohort does not necessarily provide strong model-development evidence when the outcome event is uncommon, and small event numbers may increase uncertainty in performance estimates. More generally, limited event numbers combined with multiple predictors, feature-selection procedures, or model configurations may produce unstable internal estimates that require evaluation in larger independent cohorts.

The evaluation measures reported across studies were similarly heterogeneous. Accuracy, sensitivity, specificity, and AUC were frequently used for classification outcomes, whereas regression and survival models required different measures. Roca et al. evaluated continuous two-year EDSS prediction using MSE [101], while Cagol et al. [93] and Coll et al. [72] used concordance-based measures for time-to-PIRA prediction. Calibration was reported much less frequently. A model may distinguish higher-risk from lower-risk patients while still providing inaccurate estimates of individual risk. The absence of calibration, confidence intervals, or other measures of uncertainty in many studies therefore limits interpretation of the clinical reliability of reported performance. This limitation is particularly important in studies with small event numbers or substantial class imbalance.

Reproducibility of imaging features also remains important. The included radiomics studies demonstrated different levels of reporting for feature extraction, preprocessing, feature reduction, harmonization, and standardization [14,92,93]. Explicit IBSI-related reporting was not consistent across the included progression-prediction studies. The use of PyRadiomics alone should not be interpreted as evidence of complete standardization. More consistent reporting of feature-extraction settings and reproducibility procedures would improve comparison of radiomic findings across studies. Similarly, lesion masks

and tissue measures should not automatically be considered error-free predictors, because segmentation performance may vary across scanners and methods [13,49,52]. The reliability of downstream prediction therefore remains dependent on the stability of imaging biomarkers entering the model.

Interpretability methods such as feature-importance analysis, SHAP, regional relevance maps, and saliency approaches provided useful information about the imaging and clinical variables contributing to prediction [62,72,84,93,100]. However, identification of an influential feature or anatomical region does not by itself demonstrate that the association is biologically stable or clinically generalizable. Interpretability findings should therefore be considered together with validation design, cohort size, imaging reproducibility, and consistency across independent populations. This is particularly relevant when high-dimensional imaging features are reduced to a small number of apparently important predictors.

8.1. Methodological Quality, Risk of Bias, and Applicability of the Prediction Evidence

The PROBAST + AI assessment provides an additional perspective on the prediction findings synthesized in this review. The detailed model- and evaluation-stage assessments are presented in Table 3, and the corresponding study-level summary across the six MS progression-prediction categories is shown in Table 4. Among the 37 included studies, model Development was rated as High methodological concern in 35 studies and Unclear in two, whereas Evaluation was rated as High risk of bias in all 37 studies. Applicability showed a substantially different pattern, with Low concern in 32 studies, Unclear concern in two, and High concern in three. These findings should not be interpreted as indicating that the reported prediction results lack value. Several studies demonstrated encouraging discrimination and showed that MRI-derived information can contribute to prediction of future MS-related outcomes. Rather, the PROBAST + AI findings indicate the methodological conditions that influence how confidently those performance estimates can be expected to reproduce in new patients.

Development, Evaluation, and Applicability address different aspects of the evidence. Development reflects how appropriately a model was constructed, Evaluation reflects the reliability of its performance assessment, and Applicability reflects alignment with the review question. This distinction is important because a clinically relevant model may still require stronger development or evaluation procedures.

Several practices observed across the reviewed studies provide a useful foundation for future research. Predictor information was commonly collected before the future outcome, MRI or MRI-derived variables were incorporated directly into prediction, and many studies used cross-validation, holdout testing, out-of-bag evaluation, or independent cohorts rather than reporting training performance alone. Regularisation, dropout, early stopping, patient-level data separation, nested cross-validation, multicentre datasets, and independent-cohort assessment were also used in parts of the literature. These are important strengths. The assessment nevertheless shows that a Low-concern Development or Low-risk Evaluation judgement generally requires these safeguards to be applied consistently across the complete modelling pipeline, rather than relying on one methodological strength in isolation.

The clearest opportunities for improvement were concentrated in the analysis and evaluation process. Future studies should plan the sample size and number of outcome events in relation to the number of candidate predictors and model complexity. They should also document participant exclusions and missing-data handling and maintain strict patient-level separation between model development and evaluation. Importantly, all data-dependent operations, including imputation, normalization or harmonization, feature selection, predictor screening, and hyperparameter optimization, should be fitted using

the training data only. When resampling is used, these procedures should be repeated within each training fold rather than performed once before cross-validation. The reviewed literature already recognizes this issue: feature selection, imputation, preprocessing, and hyperparameter selection were not always clearly confined to the training data.

Class imbalance also requires careful treatment. Oversampling, undersampling, SMOTE, balanced bootstrapping, or class weighting can be useful during model development, particularly for uncommon progression outcomes. Future studies should clearly report where these procedures are applied and evaluate the final model under the outcome distribution expected in the target population. Where imbalance correction changes predicted probabilities, recalibration should also be considered. Similarly, external validation should ideally evaluate a fully specified and locked model in an independent population without repeating feature selection, tuning, or model fitting. Redevelopment in a second cohort provides useful evidence of reproducibility but should be distinguished from validation of the original fixed model. This distinction was relevant in the reviewed literature, where some independent cohorts were used for renewed model development rather than direct testing of a previously fixed model.

A particularly important gap was the completeness of performance evaluation. Discrimination and classification measures such as AUC, C-index, accuracy, sensitivity, specificity, precision, and F1 score were commonly reported, whereas calibration was much less consistently assessed. For individualized prognosis, these properties are complementary: discrimination describes how well a model separates patients with different outcomes, whereas calibration determines whether predicted risks agree with observed risks. Future studies should therefore report calibration together with discrimination, appropriate uncertainty estimates, and, where clinically relevant, measures of clinical utility. Predictor- and outcome-assessment independence should also be reported more explicitly. Where a variable used as a predictor is closely related to, or contributes to, the definition of the later outcome, its role should be clearly justified because this can affect the independence of prediction and outcome assessment.

Overall, the PROBAST + AI assessment indicates that the current evidence is highly relevant to MRI-based MS progression prediction, as reflected by the predominantly Low applicability concern. The main opportunity is therefore to strengthen how promising models are developed and evaluated so that their reported predictive value can be established more confidently across independent populations. More rigorous and transparent validation will help translate the prognostic information already demonstrated in this field into reproducible and clinically useful prediction tools.

8.2. Limitations

This systematic review has some limitations. The included studies were heterogeneous in study population, MRI representation, prediction outcome and definition, follow-up duration, modelling approach, and reported performance measures. This heterogeneity limited direct comparison between studies and prevented meaningful quantitative pooling of predictive performance. Methodological reporting was also incomplete in several primary studies, particularly for preprocessing and feature selection within validation procedures, missing-data handling, class-imbalance management, calibration, treatment exposure, and external validation. These reporting gaps also affected the PROBAST + AI assessment; where the information required for a signalling question was not reported in the article or available Supplementary Material, it was recorded as No Information rather than inferred. Therefore, some uncertainty in the assessment reflects limitations in reporting as well as limitations in study design. Although PROBAST + AI provides structured criteria and guidance, some domain-level and overall judgements require reviewer

interpretation; consequently, individual ratings in Tables 3 and 4 may vary to some extent between assessors. This potential variability was considered when interpreting the overall assessment findings. Finally, the review was not prospectively registered, and no separate review protocol was prepared.

Table 3. Detailed PROBAST + AI assessment of model development quality, model evaluation risk of bias, and applicability for the included studies and distinct model or evaluation stages.

No.	Study/Model/Stage	Assessment Type	Dev D1	Dev D2	Dev D3	Dev D4	Overall Dev	Eval D1	Eval D2	Eval D3	Eval D4	Overall Eval RoB	App D1	App D2	App D3	Overall App
CIS-to-MS/CDMS conversion																
1a	Tayyab et al. (2023)—RFexclude [85]	Development + evaluation	L	L	H	H	H	L	L	H	H	H	L	L	L	L
1b	Tayyab et al. (2023)—RFnaive [85]	Development + evaluation	L	L	H	H	H	L	L	H	H	H	L	L	L	L
1c	Tayyab et al. (2023)—PRF [85]	Development + evaluation	L	L	H	H	H	L	L	H	H	H	L	L	L	L
2	Aksoy et al. (2025)—SeruNet-MS [86]	Development + evaluation	U	L	U	H	H	U	L	U	H	H	L	L	L	L
3	Wottschel et al. (2019)—SVM-RFE [87]	Development + evaluation	L	L	L	H	H	L	L	L	H	H	L	L	L	L
4	Yoo et al. (2016)—CNN with EDT + pretraining [77]	Development + evaluation	U	L	L	H	H	U	L	L	H	H	L	L	L	L
5a	Wottschel et al. (2015)—SVM/fRFE, 1-year CDMS [88]	Development + evaluation	H	L	L	H	H	H	L	L	H	H	L	L	L	L
5b	Wottschel et al. (2015)—SVM/fRFE, 3-year CDMS [88]	Development + evaluation	H	L	L	H	H	H	L	L	H	H	L	L	L	L
6	Yoo et al. (2017)—CNN with EDT + pretraining + user-defined MRI/clinical features [89]	Development + evaluation	H	L	H	H	H	H	L	H	H	H	L	L	L	L
7	Haouam & Benmalek (2024)—Random Forest (final model; common pipeline with XGBoost/DT/LR) [78]	Development + evaluation	U	U	U	H	H	U	U	U	H	H	L	U	U	U
8a	Bendfeldt et al. (2019)—nonlinear RBF-SVM lesion-geometry models [61]	Development + evaluation	L	L	L	H	H	L	L	L	H	H	L	L	L	L
8b	Bendfeldt et al. (2019)—linear SVM cortical-GM analysis [61]	Development + evaluation	L	L	L	H	H	L	L	L	H	H	L	L	L	L
9	Rasouli et al. (2024)—XGBoost [62]	Development + evaluation	L	L	H	H	H	L	L	H	H	H	L	L	L	L
10	Zhang et al. (2019)—oblique random forest [60]	Development + evaluation	L	L	L	H	H	L	L	L	H	H	L	L	L	L

Table 3. *Cont.*

No.	Study/Model/Stage	Assessment Type	Dev D1	Dev D2	Dev D3	Dev D4	Overall Dev	Eval D1	Eval D2	Eval D3	Eval D4	Overall Eval RoB	App D1	App D2	App D3	Overall App
							Future lesion or MRI disease activity									
11	Yılmaz Acar et al. (2022)—3D-DWT + SVM/ML classifiers [82]	Development + evaluation	U	L	H	H	H	U	L	H	H	H	L	L	L	L
12	Mohammadi Sepahvand et al. (2019)—3D CNN with baseline MRI ± lesion labels [90]	Development + evaluation	U	L	L	H	H	U	L	L	H	H	L	L	L	L
13	Durso-Finley et al. (2022)—multi-head ResNet/TARNET lesion-count model [79]	Development + evaluation	U	L	U	U	U	U	L	U	H	H	L	L	L	L
							PIRA/PIRMA prediction									
14	Marastoni et al. (2025)—MRI/CSF RF + Cox/ROC [91]	Development + evaluation	U	L	U	H	H	U	L	U	H	H	H	L	L	H
15a	Ianniello et al. (2026)—Naïve Bayes, 24-month PIRA [92]	Development + evaluation	H	L	H	H	H	H	L	H	H	H	L	L	L	L
15b	Ianniello et al. (2026)—Naïve Bayes, Month-24 landmark → Month-36 PIRA [92]	Development + evaluation	H	L	H	H	H	H	L	H	H	H	L	L	L	L
16a	Cagol et al. (2026)—Cohort 1 Boruta/RF survival/classification [93]	Development + evaluation	U	L	L	H	H	U	L	L	H	H	L	L	L	L
16b	Cagol et al. (2026)—Cohort 2 independent redevelopment Boruta/RF [93]	Development + evaluation	U	L	L	H	H	U	L	L	H	H	L	L	L	L
16c	Cagol et al. (2026)—4-variable ridge-Cox PIRA risk score [93]	Development + evaluation	U	L	L	H	H	U	L	L	H	H	L	L	L	L
17a	Coll et al. (2025)—EfficientNet discrete-time survival model [72]	Development + evaluation	U	L	L	H	H	U	L	L	H	H	L	L	L	L
17b	Coll et al. (2025)—locked external MS PATHS evaluation [72]	Evaluation only						U	L	H	H	H	L	L	L	L
18	Poretto et al. (2025)—Random Forest with randomized nested grid-search CV [76]	Development + evaluation	L	U	H	H	H	L	U	H	H	H	L	L	L	L

Table 3. *Cont.*

No.	Study/Model/Stage	Assessment Type	Dev D1	Dev D2	Dev D3	Dev D4	Overall Dev	Eval D1	Eval D2	Eval D3	Eval D4	Overall Eval RoB	App D1	App D2	App D3	Overall App
Disability worsening or confirmed progression																
19	Aerts et al. (2026)—multimodal MRI/EPTS ML [94]	Development + evaluation	H	U	H	H	H	H	U	H	H	H	L	L	L	L
20	Wang et al. (2025)—DP combined nomogram [95]	Development + evaluation	H	L	H	H	H	H	L	H	H	H	L	L	L	L
21a	Shen et al. (2025)—CCDP24 RSF/CNN [96]	Development + evaluation	U	L	H	H	H	U	L	H	H	H	L	L	L	L
21b	Shen et al. (2025)—sustained T25FW progression [96]	Development + evaluation	U	L	H	H	H	U	L	H	H	H	L	L	L	L
22	Bejarano et al. (2011)—confirmed disability-progression NNet [97]	Development + evaluation	U	L	H	H	H	U	L	H	H	H	L	L	L	L
23a	Storelli et al. (2022)—EDSS-based 3D CNN, internal validation [40]	Development + evaluation	U	L	U	H	H	U	L	U	H	H	L	L	L	L
23b	Storelli et al. (2022)—independent scanner holdout [40]	Evaluation only						U	L	U	H	H	L	L	L	L
24a	Zhao et al. (2020)—nested CLIMB internal evaluation [98]	Development + evaluation	U	U	H	H	H	U	U	H	H	H	L	L	L	L
24b	Zhao et al. (2020)—EPIC external evaluation [98]	Evaluation only						U	U	H	H	H	L	L	L	L
25	Taloni et al. (2022)—slice-based ResNet50 EDSS discriminators [71]	Development + evaluation	U	L	L	H	H	U	L	L	H	H	L	L	L	L
26	Zhao et al. (2017)—SVM/LR CLIMB models [74]	Development + evaluation	U	U	U	H	H	U	U	U	H	H	L	L	L	L
27	Khan et al. (2025)—FLAIR radiomics ML [14]	Development + evaluation	H	L	U	H	H	H	L	U	H	H	L	L	L	L
28a	Andorra et al. (2024)—Sys4MS Random Forest models [99]	Development + evaluation	U	L	U	H	H	U	L	U	H	H	L	L	L	L

Table 3. *Cont.*

No.	Study/Model/Stage	Assessment Type	Dev D1	Dev D2	Dev D3	Dev D4	Overall Dev	Eval D1	Eval D2	Eval D3	Eval D4	Overall Eval RoB	App D1	App D2	App D3	Overall App
28b	Andorra et al. (2024)—Barcelona cohort redevelopment [99]	Development + evaluation	U	U	U	H	H	U	U	U	H	H	L	L	L	L
29	Pellegrini et al. (2020)—six prognostic modelling strategies [100]	Development + evaluation	H	U	H	U	H	H	U	H	H	H	L	L	L	L
30	Tousignant et al. (2019)—3D CNN MRI ± lesion masks [70]	Development + evaluation	H	L	L	H	H	H	L	L	H	H	L	L	L	L
	Future EDSS, Disability Status, or Severity Trajectory															
31	Roca et al. (2020)—multimodel EDSS regression ensemble [101]	Development + evaluation	L	L	U	U	U	L	L	U	H	H	L	L	L	L
32a	Campanioni et al. (2024)—β1/β2 trajectory models [102]	Development + evaluation	L	U	U	U	U	L	U	U	H	H	L	L	L	L
32b	Campanioni et al. (2024)—future EDSS(t) classifiers [102]	Development + evaluation	L	U	U	H	H	L	U	U	H	H	L	L	L	L
33a	Kırkbir et al. (2025)—GridSearch LSTM with all-variable input [83]	Development + evaluation	H	H	U	H	H	H	H	U	H	H	L	L	L	L
33b	Kırkbir et al. (2025)—final feature-selected LSTM [83]	Development + evaluation	H	U	U	H	H	H	U	U	H	H	L	H	L	H
34	Al-Iedani et al. (2024)—5-year ΔEDSS penalised regression [103]	Development + evaluation	U	L	L	H	H	U	L	L	H	H	L	L	L	L
	RRMS-to-SPMS Conversion															
35	Rocca et al. (2021)—RF SPMS-conversion model [104]	Development + evaluation	U	H	U	H	H	U	H	U	H	H	L	H	L	H
36	Pisani et al. (2021)—RSF/SP-RiSc [105]	Development + evaluation	U	L	U	H	H	U	L	U	H	H	L	L	L	L
37a	Seccia et al. (2020)—visit-oriented, feature-saving [106]	Development + evaluation	H	U	H	H	H	H	U	H	H	H	L	L	U	U
37b	Seccia et al. (2020)—visit-oriented, record-saving [106]	Development + evaluation	U	U	H	H	H	U	U	H	H	H	L	L	U	U

Table 3. *Cont.*

No.	Study/Model/Stage	Assessment Type	Dev D1	Dev D2	Dev D3	Dev D4	Overall Dev	Eval D1	Eval D2	Eval D3	Eval D4	Overall Eval RoB	App D1	App D2	App D3	Overall App
37c	Seccia et al. (2020)—history-oriented LSTM, feature-saving [106]	Development + evaluation	H	U	H	H	H	H	U	H	H	H	L	L	U	U
37d	Seccia et al. (2020)—history-oriented LSTM, record-saving [106]	Development + evaluation	U	U	H	H	H	U	U	H	H	H	L	L	U	U

Dev, model development quality; Eval, model evaluation risk of bias; App, applicability to the review question; D1, participants and data sources; D2, predictors; D3, outcome; D4, analysis; L, Low concern/risk; U, Unclear concern/risk; H, High concern/risk. Background colours indicate the PROBAST + AI judgements: green = Low, yellow = Unclear, red = High, and grey = Evaluation only. Separate rows are presented where materially different models, prediction horizons, redevelopment procedures, or validation stages required separate assessment. Applicability is assessed for D1–D3 only.

Table 4. Study-level summary of PROBAST + AI overall judgements for the studies presented in Table 3, by MS progression-prediction outcome category.

Prediction Category	Studies Assessed	Development Quality Low	Development Quality Unclear	Development Quality High	Evaluation RoB Low	Evaluation RoB Unclear	Evaluation RoB High	Applicability Low	Applicability Unclear	Applicability High
CIS → MS/CDMS conversion	10	0	0	10	0	0	10	9	1	0
Future lesion or MRI disease activity	3	0	1	2	0	0	3	3	0	0
PIRA/PIRMA prediction	5	0	0	5	0	0	5	4	0	1
Disability worsening or confirmed progression	12	0	0	12	0	0	12	12	0	0
Future EDSS/ disability status/ severity trajectory	4	0	1	3	0	0	4	3	0	1
RRMS → SPMS conversion	3	0	0	3	0	0	3	1	1	1
Overall	**37**	**0**	**2**	**35**	**0**	**0**	**37**	**32**	**2**	**3**

9. Conclusions

This review shows that MRI-based prediction of MS progression is best considered as an end-to-end measurement and modelling pipeline rather than as a comparison of algorithms alone. Across the 37 included studies, prognostic information was derived from lesion geometry and spatial distribution, radiomics, tissue atrophy, longitudinal imaging changes, and multimodal combinations of imaging and clinical variables. Classical ML and DL each showed value depending on MRI representation, prediction outcome, and study design. PROBAST + AI assessment rated model Development as High concern in 35 studies and Unclear in two, Evaluation as High risk of bias in all 37 studies, and Applicability as Low concern in 32, Unclear in two, and High in three. This pattern indicates that most studies addressed populations, MRI predictors, and progression outcomes relevant to the review question, while stronger methodological safeguards are needed to establish reproducible predictive performance. Future studies should prioritize adequately sized multicenter cohorts, training-only data-dependent processing, transparent handling of missing data and class imbalance, calibration and uncertainty assessment, and independent evaluation of locked models. These steps will help translate the promising MRI-based prognostic evidence into more reliable tools for individualized MS risk assessment and monitoring.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app16188955/s1, Table S1: Complete Literature Search Strategies.

Author Contributions: U.V.: Writing—review & editing, Writing—original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. A.W.: Writing—review & editing, Supervision, Formal analysis, Conceptualization. W.S.: Writing—review & editing, Supervision, Formal analysis, Conceptualization. E.K.: Writing—review & editing, Formal analysis, Conceptualization. H.V.D.-M.: Funding acquisition, Writing—review & editing, Supervision, Formal analysis, Conceptualization. S.H.: Writing—review & editing, Supervision, Formal analysis, Conceptualization. C.S.: Writing—review & editing, Formal analysis, Data curation, Conceptualization. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Vision Research Foundation, Health Research Council of New Zealand's project 21/144, the Marsden Fund Project 22-UOA-120, Royal Society Catalyst: Seeding General Project 25-UOA-035-CSG, the MedTech Research Translator—Research Acceleration Programme (RAP) Stage I Project (3735793), the Hebei Provincial Science and Technology Program—International Science and Technology Cooperation Project (25292001D), the Kānoa Regional Development Unit New Zealand and New Zealand Multiple Sclerosis Research Trust.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

2D	Two-dimensional	MEP	Motor-Evoked Potential
3D	Three-dimensional	MICE	Multiple Imputation By Chained Equations
9HPT	Nine-Hole Peg Test	MLP	Multilayer Perceptron
AI	Artificial Intelligence	MNI	Montreal Neurological Institute
ANTs	Advanced Normalization Tools	MPRAGE	Magnetization-Prepared Rapid Gradient-Echo

App	Applicability to the Review Question	MRI	Magnetic Resonance Imaging
AUC	Area Under the Curve	MRIQC	MRI Quality Control
AUROC	Area Under the Receiver Operating Characteristic Curve	MS	Multiple Sclerosis
BET	Brain Extraction Tool	MSE	Mean Squared Error
BIANCA	Brain Intensity AbNormality Classification Algorithm	N3	Nonparametric Nonuniform Intensity Normalization
BIANCA-MS	BIANCA for Multiple Sclerosis	NAWM	Normal-Appearing White Matter
BPF	Brain Parenchymal Fraction	NEDA-3	No Evidence of Disease Activity-3
CART	Classification and Regression Tree	NE-T2	New/Enlarging T2 Lesions
CATE	Conditional Average Treatment Effect	NGTDM	Neighboring Grey Tone Difference Matrix
CCDP24	24-Week Confirmed Composite Disability Progression	NPV	Negative Predictive Value
CDA	Confirmed Disability Accumulation	OCB	Oligoclonal Bands
CDMS	Clinically Definite MS	OCT	Optical Coherence Tomography
CI	Confidence Interval	OOB	Out-of-Bag
CIS	Clinically Isolated Syndrome	PCA	Principal Component Analysis
CNN	Convolutional Neural Networks	PD	Proton Density
CSF	Cerebrospinal Fluid	PDDS	Patient-Determined Disease Steps
CV	Cross-Validation	PIRA	Progression Independent of Relapse Activity
D1	Participants And Data Sources	PIRMA	Progression Independent of Relapse and MRI Activity
D2	Predictors	PPMS	Primary Progressive MS
D3	Outcome	PPV	Positive Predictive Value
D4	Analysis	PPVL	Positive Predictive Value per Lesion
DBM	Deformation-Based Morphometry	PR-AUC	Area Under the Precision–Recall Curve
Dev	Model Development Quality	PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
DGM	Deep Grey Matter	PSNR	Peak Signal-to-Noise Ratio
dDLR	DL-Based Reconstruction	QC	Quality Control
DIR	Double Inversion Recovery	qMRI	Quantitative Magnetic Resonance Imaging
DIS	Dissemination in Space	RBF	Radial Basis Function
DL	Deep Learning	ReLU	Rectified Linear Unit
DMD	Disease-Modifying Drug	RF	Random Forest
DMT	Disease-Modifying Therapy	RFE	Recursive Feature Elimination
DOR	Diagnostic Odds Ratio	RFECV	Recursive Feature Elimination with Cross-Validation
EDSS	Expanded Disability Status Scale	RMSE	Root-Mean-Square Error
EHR	Electronic Health Record	ROC	Receiver Operating Characteristic

EPTS	Evoked-Potential Time Series	ROC-AUC	Area Under the Receiver Operating Characteristic Curve
Eval	Model Evaluation Risk of Bias	ROI	Region of Interest
FLAIR	Fluid-Attenuated Inversion Recovery	RRMS	Relapsing Remitting MS
FSL	FMRIB Software Library	RSF	Random Survival Forest
GLCM	Grey-Level Co-Occurrence Matrix	SD	Standard Deviation
GLDM	Grey-Level Dependence Matrix	SDMT	Symbol Digit Modalities Test
GLRLM	Grey-Level Run Length Matrix	SEL	Slowly Expanding Lesion
GLSZM	Grey-Level Size Zone Matrix	sGFAP	Serum Glial Fibrillary Acidic Protein
GM	Grey Matter	SHAP	SHapley Additive exPlanations
H	High concern/risk	SMOTE	Synthetic Minority Over-sampling Technique
HR	Hazard Ratio	sNfL	Serum Neurofilament Light Chain
HD-BET	High-Definition Brain Extraction Tool	SPIN	Statistical Principles of Image Normalization
IBSI	Image Biomarker Standardization Initiative	SPMS	Secondary Progressive MS
ICA	Independent Component Analysis	SP-RiSc	Secondary Progressive Risk Score
ICC	Intraclass Correlation Coefficient	SSIM	Structural Similarity Index Measure
ICBM	International Consortium for Brain Mapping	SuStaIn	Subtype and Staging Inference
KFS	Kurtzke Functional Systems Score	SVC	Support Vector Classifier
k-NN	K-Nearest Neighbor	SVM	Support Vector Machine
L	Low concern/risk	SyN	Symmetric Normalization
LCModel	Linear Combination Model	T25FW	Timed 25-Foot Walk
LGA	Lesion Growth Algorithm	T25WT	Timed 25-Foot Walk Test
LASSO	Least Absolute Shrinkage and Selection Operator	tARCS	Total Audio Recorded Cognitive Screen score
LSTM	Long Short-Term Memory	TV	Total-Variation
LoR	Logistic Regression	U	Unclear concern/risk
LST	Lesion Segmentation Tool	UMAP	Uniform Manifold Approximation and Projection
MAE	Mean Absolute Error	VFT	Verbal Fluency Test
MALF	Multi-Atlas Label Fusion	WML	White-Matter Lesion
MCC	Matthews Correlation Coefficient	WM	White Matter
MCMC	Markov Chain Monte Carlo	XGBoost	eXtreme Gradient Boosting

References

1. Ghasemi, N.; Razavi, S.; Nikzad, E. Multiple Sclerosis: Pathogenesis, Symptoms, Diagnoses and Cell-Based Therapy. *Cell J.* **2017**, *19*, 1–10. [CrossRef] [PubMed]
2. Reeve, K.; On, B.I.; Havla, J.; Burns, J.; Gosteli-Peter, M.A.; Alabsawi, A.; Alayash, Z.; Götschi, A.; Seibold, H.; Mansmann, U.; et al. Prognostic models for predicting clinical disease progression, worsening and activity in people with multiple sclerosis. *Cochrane Database Syst. Rev.* **2023**, *9*, CD013606. [CrossRef] [PubMed]

3. Katz Sand, I. Classification, diagnosis, and differential diagnosis of multiple sclerosis. *Curr. Opin. Neurol.* **2015**, *28*, 193. [CrossRef] [PubMed]
4. Oh, J.; Alikhani, K.; Bruno, T.; Devonshire, V.; Giacomini, P.S.; Giuliani, F.; Nakhaipour, H.R.; Schecter, R.; Larochelle, C. Diagnosis and Management of Secondary-Progressive Multiple Sclerosis: Time for Change. *Neurodegener. Dis. Manag.* **2019**, *9*, 301–317. [CrossRef] [PubMed]
5. Hechenberger, S.; Helmlinger, B.; Ropele, S.; Pirpamer, L.; Bachmaier, G.; Damulina, A.; Pichler, A.; Khalil, M.; Enzinger, C.; Pinter, D. Information processing speed as a prognostic marker of physical impairment and progression in patients with multiple sclerosis. *Mult. Scler. Relat. Disord.* **2022**, *57*, 103353. [CrossRef] [PubMed]
6. Pretorius, P.M.; Quaghebeur, G. The Role of MRI in the Diagnosis of MS. *Clin. Radiol.* **2003**, *58*, 434–448. [CrossRef] [PubMed]
7. Deloire, M.S.A.; Ruet, A.; Hamel, D.; Bonnet, M.; Dousset, V.; Brochet, B. MRI predictors of cognitive outcome in early multiple sclerosis. *Neurology* **2011**, *76*, 1161–1167. [CrossRef] [PubMed]
8. Barkhof, F. The clinico-radiological paradox in multiple sclerosis revisited. *Curr. Opin. Neurol.* **2002**, *15*, 239. [CrossRef] [PubMed]
9. Rocca, M.A.; Preziosa, P.; Barkhof, F.; Brownlee, W.; Calabrese, M.; De Stefano, N.; Granziera, C.; Ropele, S.; Toosy, A.T.; Vidal-Jordana, À.; et al. Current and future role of MRI in the diagnosis and prognosis of multiple sclerosis. *Lancet Reg. Health Eur.* **2024**, *44*, 100978. [CrossRef] [PubMed]
10. Narayana, P.A.; Coronado, I.; Sujit, S.J.; Wolinsky, J.S.; Lublin, F.D.; Gabr, R.E. Deep Learning for Predicting Enhancing Lesions in Multiple Sclerosis from Noncontrast MRI. *Radiology* **2020**, *294*, 398–404. [CrossRef] [PubMed]
11. Elhaie, M.; Shahbazi-Gahrouei, D. Sequence-level machine learning assessment of MRI protocols in classifying multiple Sclerosis lesions. *J. Radiat. Res. Appl. Sci.* **2025**, *18*, 101860. [CrossRef]
12. De Stefano, N.; Battaglini, M.; Pareto, D.; Cortese, R.; Zhang, J.; Oesingmann, N.; Prados, F.; Rocca, M.A.; Valsasina, P.; Vrenken, H.; et al. MAGNIMS recommendations for harmonization of MRI data in MS multicenter studies. *Neuroimage Clin.* **2022**, *34*, 102972. [CrossRef] [PubMed]
13. Zeng, C.; Gu, L.; Liu, Z.; Zhao, S. Review of Deep Learning Approaches for the Segmentation of Multiple Sclerosis Lesions on Brain MRI. *Front. Neuroinform* **2020**, *14*, 610967. [CrossRef] [PubMed]
14. Khan, H.; Woodruff, H.C.; Giraldo, D.L.; Werthen-Brabants, L.; Mali, S.A.; Amirrajab, S.; De Brouwer, E.; Popescu, V.; Van Wijmeersch, B.; Gerlach, O.; et al. Leveraging hand-crafted radiomics on multicenter FLAIR MRI for predicting disability worsening in people with multiple sclerosis. *Front. Neurosci.* **2025**, *19*, 1610401. [CrossRef] [PubMed]
15. Vázquez-Marrufo, M.; Sarrias-Arrabal, E.; García-Torres, M.; Martín-Clemente, R.; Izquierdo, G. A systematic review of the application of machine-learning algorithms in multiple sclerosis. *Neurología* **2023**, *38*, 577–590. [CrossRef] [PubMed]
16. Yousef, H.; Malagurski Tortei, B.; Castiglione, F. Predicting multiple sclerosis disease progression and outcomes with machine learning and MRI-based biomarkers: A review. *J. Neurol.* **2024**, *271*, 6543–6572. [CrossRef] [PubMed]
17. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
18. Moons, K.G.M.; Damen, J.A.A.; Kaul, T.; Hooft, L.; Navarro, C.A.; Dhiman, P.; Beam, A.L.; Calster, B.V.; Celi, L.A.; Denaxas, S.; et al. PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* **2025**, *388*, e082505. [CrossRef] [PubMed]
19. Zhang, L.; Wu, M.; Wang, L.; Steffens, D.C.; Potter, G.G.; Liu, M. Iterative Learning for Joint Image Denoising and Motion Artifact Correction of 3D Brain MRI. *arXiv* **2024**, arXiv:2403.08162. [CrossRef]
20. Havsteen, I.; Ohlhues, A.; Madsen, K.H.; Nybing, J.D.; Christensen, H.; Christensen, A. Are Movement Artifacts in Magnetic Resonance Imaging a Real Problem?—A Narrative Review. *Front. Neurol.* **2017**, *8*, 232. [CrossRef] [PubMed]
21. Liu, J.; Beck, E.S.; Filippini, S.; van Gelderen, P.; de Zwart, J.A.; Norato, G.; Sati, P.; Al-Louzi, O.; Kolb, H.; Donadieu, M.; et al. Navigator-Guided Motion and B0 Correction of T2*-weighted MRI Improves Multiple Sclerosis Cortical Lesion Detection. *Investig. Radiol.* **2021**, *56*, 409–416. [CrossRef] [PubMed]
22. Begnoche, J.P.; Schilling, K.G.; Boyd, B.D.; Cai, L.Y.; Taylor, W.D.; Landman, B.A. EPI susceptibility correction introduces significant differences far from local areas of high distortion. *Magn. Reson. Imaging* **2022**, *92*, 1–9. [CrossRef] [PubMed]
23. Yarach, U.; Luengviriya, C.; Danishad, K.A.; Stucht, D.; Godenschweger, F.; Schulze, P.; Speck, O. Correction of Gradient Nonlinearity Artifacts in Prospective Motion Correction for 7T MRI. *Magn. Reson. Med.* **2015**, *73*, 1562–1569. [CrossRef] [PubMed]
24. Helms, G.; Dathe, H.; Kallenberg, K.; Dechent, P. High-resolution maps of magnetization transfer with inherent correction for RF inhomogeneity and T1 relaxation obtained from 3D FLASH MRI. *Magn. Reson. Med.* **2008**, *60*, 1396–1407. [CrossRef] [PubMed]
25. Muro, I.; Isoiwa, T.; Shibukawa, S.; Usui, K.; Otsuka, Y. Using Deep Learning to Simultaneously Reduce Noise and Motion Artifacts in Brain MR Imaging. *Magn. Reson. Med. Sci.* **2025**, *24*, mp.2024-0098. [CrossRef] [PubMed]
26. Kidoh, M.; Shinoda, K.; Kitajima, M.; Isogawa, K.; Nambu, M.; Uetani, H.; Morita, K.; Nakaura, T.; Tateishi, M.; Yamashita, Y.; et al. Deep Learning Based Noise Reduction for Brain MR Imaging: Tests on Phantoms and Healthy Volunteers. *Magn. Reson. Med. Sci.* **2020**, *19*, 195–206. [CrossRef] [PubMed]

27. Pawar, K.; Chen, Z.; Seah, J.; Law, M.; Close, T.; Egan, G. Clinical utility of deep learning motion correction for T1 weighted MPRAGE MR images. *Eur. J. Radiol.* **2020**, *133*, 109384. [CrossRef] [PubMed]
28. Zhang, L.; Wang, X.; Rawson, M.; Balan, R.; Herskovits, E.H.; Melhem, E.R.; Chang, L.; Wang, Z.; Ernst, T. Motion Correction for Brain MRI Using Deep Learning and a Novel Hybrid Loss Function. *Algorithms* **2024**, *17*, 215. [CrossRef]
29. Sommer, K.; Saalbach, A.; Brosch, T.; Hall, C.; Cross, N.M.; Andre, J.B. Correction of Motion Artifacts Using a Multiscale Fully Convolutional Neural Network. *Am. J. Neuroradiol.* **2020**, *41*, 416–423. [CrossRef] [PubMed]
30. Zhang, Q.; Ruan, G.; Yang, W.; Liu, Y.; Zhao, K.; Feng, Q.; Chen, W.; Wu, E.X.; Feng, Y. MRI Gibbs-ringing artifact reduction by means of machine learning using convolutional neural networks. *Magn. Reson. Med.* **2019**, *82*, 2133–2145. [CrossRef] [PubMed]
31. Shinohara, R.T.; Sweeney, E.M.; Goldsmith, J.; Shiee, N.; Mateen, F.J.; Calabresi, P.A.; Jarso, S.; Pham, D.L.; Reich, D.S.; Crainiceanu, C.M.; et al. Statistical normalization techniques for magnetic resonance imaging. *Neuroimage Clin.* **2014**, *6*, 9–19. [CrossRef] [PubMed]
32. Tustison, N.J.; Avants, B.B.; Cook, P.A.; Zheng, Y.; Egan, A.; Yushkevich, P.A.; Gee, J.C. N4ITK: Improved N3 bias correction. *IEEE Trans. Med. Imaging* **2010**, *29*, 1310–1320. [CrossRef] [PubMed]
33. Roy, S.; Carass, A.; Shiee, N.; Pham, D.L.; Calabresi, P.; Reich, D.; Prince, J.L. Longitudinal intensity normalization in the presence of multiple sclerosis lesions. *Proc. IEEE Int. Symp. Biomed. Imaging* **2013**, 1384–1387. [CrossRef] [PubMed]
34. De Nunzio, G.; Cataldo, R.; Carlà, A. Robust Intensity Standardization in Brain Magnetic Resonance Images. *J. Digit. Imaging* **2015**, *28*, 727–737. [CrossRef] [PubMed]
35. Fortin, J.-P.; Cullen, N.; Sheline, Y.I.; Taylor, W.D.; Aselcioglu, I.; Cook, P.A.; Adams, P.; Cooper, C.; Fava, M.; McGrath, P.J.; et al. Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* **2018**, *167*, 104–120. [CrossRef] [PubMed]
36. Beer, J.C.; Tustison, N.J.; Cook, P.A.; Davatzikos, C.; Sheline, Y.I.; Shinohara, R.T.; Linn, K.A. Alzheimer's Disease Neuroimaging Initiative Longitudinal ComBat: A method for harmonizing longitudinal multi-scanner imaging data. *Neuroimage* **2020**, *220*, 117129. [CrossRef] [PubMed]
37. Dewey, B.E.; Zhao, C.; Reinhold, J.C.; Carass, A.; Fitzgerald, K.C.; Sotirchos, E.S.; Saidha, S.; Oh, J.; Pham, D.L.; Calabresi, P.A.; et al. DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. *Magn. Reson. Imaging* **2019**, *64*, 160–170. [CrossRef] [PubMed]
38. Schmidt, P.; Pongratz, V.; Küster, P.; Meier, D.; Wuerfel, J.; Lukas, C.; Bellenberg, B.; Zipp, F.; Groppa, S.; Sämann, P.G.; et al. Automated segmentation of changes in FLAIR-hyperintense white matter lesions in multiple sclerosis on serial magnetic resonance imaging. *Neuroimage Clin.* **2019**, *23*, 101849. [CrossRef] [PubMed]
39. Avants, B.B.; Tustison, N.J.; Song, G.; Cook, P.A.; Klein, A.; Gee, J.C. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage* **2011**, *54*, 2033–2044. [CrossRef] [PubMed]
40. Storelli, L.; Azzimonti, M.; Gueye, M.; Vizzino, C.; Preziosa, P.; Tedeschi, G.; De Stefano, N.; Pantano, P.; Filippi, M.; Rocca, M.A. A Deep Learning Approach to Predicting Disease Progression in Multiple Sclerosis Using Magnetic Resonance Imaging. *Investig. Radiol.* **2022**, *57*, 423. [CrossRef] [PubMed]
41. Balakrishnan, G.; Zhao, A.; Sabuncu, M.R.; Guttag, J.; Dalca, A.V. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *IEEE Trans. Med. Imaging* **2019**, *38*, 1788–1800. [CrossRef] [PubMed]
42. Tousignant Duran, A. Prediction of Future Multiple Sclerosis Disease Progression Using Deep Learning Analysis of MRI Data. Available online: https://escholarship.mcgill.ca/concern/theses/td96k7322 (accessed on 22 February 2026).
43. Esteban, O.; Birman, D.; Schaer, M.; Koyejo, O.O.; Poldrack, R.A.; Gorgolewski, K.J. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS ONE* **2017**, *12*, e0184661. [CrossRef] [PubMed]
44. Samani, Z.R.; Alappatt, J.A.; Parker, D.; Ismail, A.A.O.; Verma, R. QC-Automator: Deep Learning-Based Automated Quality Control for Diffusion MR Images. *Front. Neurosci.* **2020**, *13*, 1456. [CrossRef] [PubMed]
45. Ackaouy, A.; Courty, N.; Vallée, E.; Commowick, O.; Barillot, C.; Galassi, F. Unsupervised Domain Adaptation with Optimal Transport in Multi-Site Segmentation of Multiple Sclerosis Lesions from MRI Data. *Front. Comput. Neurosci.* **2020**, *14*, 19. [CrossRef] [PubMed]
46. Jain, S.; Sima, D.M.; Ribbens, A.; Cambron, M.; Maertens, A.; Van Hecke, W.; De Mey, J.; Barkhof, F.; Steenwijk, M.D.; Daams, M.; et al. Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images. *NeuroImage Clin.* **2015**, *8*, 367–375. [CrossRef] [PubMed]
47. Shiee, N.; Bazin, P.-L.; Ozturk, A.; Reich, D.S.; Calabresi, P.A.; Pham, D.L. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. *Neuroimage* **2010**, *49*, 1524–1535. [CrossRef] [PubMed]
48. Griffanti, L.; Zamboni, G.; Khan, A.; Li, L.; Bonifacio, G.; Sundaresan, V.; Schulz, U.G.; Kuker, W.; Battaglini, M.; Rothwell, P.M.; et al. BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities. *NeuroImage* **2016**, *141*, 191–205. [CrossRef] [PubMed]
49. Gentile, G.; Jenkinson, M.; Griffanti, L.; Luchetti, L.; Leoncini, M.; Inderyas, M.; Mortilla, M.; Cortese, R.; De Stefano, N.; Battaglini, M. BIANCA-MS: An optimized tool for automated multiple sclerosis lesion segmentation. *Human. Brain Mapp.* **2023**, *44*, 4893–4913. [CrossRef] [PubMed]

50. Schmidt, P.; Wink, L. LST: A lesion segmentation tool for SPM. Manual/Documentation for Version 3.0.0, 2019. Available online: https://www.statistical-modelling.de/LST_documentation.pdf (accessed on 5 September 2026).
51. Wiltgen, T.; McGinnis, J.; Schlaeger, S.; Kofler, F.; Voon, C.; Berthele, A.; Bischl, D.; Grundl, L.; Will, N.; Metz, M.; et al. LST-AI: A deep learning ensemble for accurate MS lesion segmentation. *NeuroImage Clin.* **2024**, *42*, 103611. [CrossRef] [PubMed]
52. Valverde, S.; Cabezas, M.; Roura, E.; González-Villà, S.; Pareto, D.; Vilanova, J.C.; Ramió-Torrentà, L.; Rovira, À.; Oliver, A.; Lladó, X. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *Neuroimage* **2017**, *155*, 159–168. [CrossRef] [PubMed]
53. Calabrese, M.; Rinaldi, F.; Mattisi, I.; Bernardi, V.; Favaretto, A.; Perini, P.; Gallo, P. The predictive value of gray matter atrophy in clinically isolated syndromes. *Neurology* **2011**, *77*, 257–263. [CrossRef] [PubMed]
54. Lomer, N.B.; Asalemi, K.A.; Saberi, A.; Sarlak, K. Predictors of multiple sclerosis progression: A systematic review of conventional magnetic resonance imaging studies. *PLoS ONE* **2024**, *19*, e0300415. [CrossRef] [PubMed]
55. Pérez-Miralles, F.; Sastre-Garriga, J.; Tintoré, M.; Arrambide, G.; Nos, C.; Perkal, H.; Río, J.; Edo, M.C.; Horga, A.; Castilló, J.; et al. Clinical impact of early brain atrophy in clinically isolated syndromes. *Mult. Scler.* **2013**, *19*, 1878–1886. [CrossRef] [PubMed]
56. Moccia, M.; Quarantelli, M.; Lanzillo, R.; Cocozza, S.; Carotenuto, A.; Carotenuto, B.; Alfano, B.; Prinster, A.; Triassi, M.; Nardone, A.; et al. Grey:white matter ratio at diagnosis and the risk of 10-year multiple sclerosis progression. *Eur. J. Neurol.* **2017**, *24*, 195–204. [CrossRef] [PubMed]
57. Eshaghi, A.; Prados, F.; Brownlee, W.J.; Altmann, D.R.; Tur, C.; Cardoso, M.J.; De Angelis, F.; van de Pavert, S.H.; Cawley, N.; De Stefano, N.; et al. Deep gray matter volume loss drives disability worsening in multiple sclerosis. *Ann. Neurol.* **2018**, *83*, 210–222. [CrossRef] [PubMed]
58. Hänninen, K.; Viitala, M.; Paavilainen, T.; Karhu, J.O.; Rinne, J.; Koikkalainen, J.; Lötjönen, J.; Soilu-Hänninen, M. Thalamic Atrophy Predicts 5-Year Disability Progression in Multiple Sclerosis. *Front. Neurol.* **2020**, *11*, 606. [CrossRef] [PubMed]
59. Zivadinov, R.; Bergsland, N.; Jakimovski, D.; Weinstock-Guttman, B.; Benedict, R.H.B.; Riolo, J.; Silva, D.; Dwyer, M.G. DeepGRAI Registry Study group Thalamic atrophy measured by artificial intelligence in a multicentre clinical routine real-word study is associated with disability progression. *J. Neurol. Neurosurg. Psychiatry* **2022**, *93*, 1128–1136. [CrossRef] [PubMed]
60. Zhang, H.; Alberts, E.; Pongratz, V.; Mühlau, M.; Zimmer, C.; Wiestler, B.; Eichinger, P. Predicting conversion from clinically isolated syndrome to multiple sclerosis–An imaging-based machine learning approach. *NeuroImage Clin.* **2019**, *21*, 101593. [CrossRef] [PubMed]
61. Bendfeldt, K.; Taschler, B.; Gaetano, L.; Madoerin, P.; Kuster, P.; Mueller-Lenke, N.; Amann, M.; Vrenken, H.; Wottschel, V.; Barkhof, F.; et al. MRI-based prediction of conversion from clinically isolated syndrome to clinically definite multiple sclerosis using SVM and lesion geometry. *Brain Imaging Behav.* **2019**, *13*, 1361–1374. [CrossRef] [PubMed]
62. Rasouli, S.; Dakkali, M.S.; Azarbad, R.; Ghazvini, A.; Asani, M.; Mirzaasgari, Z.; Arish, M. Predicting the conversion from clinically isolated syndrome to multiple sclerosis: An explainable machine learning approach. *Mult. Scler. Relat. Disord.* **2024**, *86*, 105614. [CrossRef] [PubMed]
63. Invitation to Join 'MRI_Dataset_Susceptibility-Based_Radiomic_Features' | Zenodo. Available online: https://zenodo.org/me/requests/c98a2a15-348d-4643-90d5-f3370a4fa4c4 (accessed on 20 August 2026).
64. Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; van Stiphout, R.G.P.M.; Granton, P.; Zegers, C.M.L.; Gillies, R.; Boellard, R.; Dekker, A.; et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **2012**, *48*, 441–446. [CrossRef] [PubMed]
65. Zwanenburg, A.; Leger, S.; Vallières, M.; Löck, S. Image biomarker standardisation initiative. *Radiology* **2020**, *295*, 328–338. [CrossRef] [PubMed]
66. IBSI. IBSI—Image Biomarker Standardisation Initiative. Available online: https://theibsi.github.io/ (accessed on 24 March 2026).
67. Peng, Y.; Zheng, Y.; Tan, Z.; Liu, J.; Xiang, Y.; Liu, H.; Dai, L.; Xie, Y.; Wang, J.; Zeng, C.; et al. Prediction of unenhanced lesion evolution in multiple sclerosis using radiomics-based models: A machine learning approach. *Mult. Scler. Relat. Disord.* **2021**, *53*, 102989. [CrossRef] [PubMed]
68. Shi, Z.; Ma, Y.; Ding, S.; Yan, Z.; Zhu, Q.; Xiong, H.; Li, C.; Xu, Y.; Tan, Z.; Yin, F.; et al. Radiomics derived from T2-FLAIR: The value of 2- and 3-classification tasks for different lesions in multiple sclerosis. *Quant. Imaging Med. Surg.* **2024**, *14*, 2049–2059. [CrossRef] [PubMed]
69. Elhaie, M.; Etemadifar, M.; Adariani, A.R.; Khorasani, A.; Shahbazi-Gahrouei, D. Machine learning-based classification of multiple sclerosis lesion activity using multi-sequence MRI radiomics: A complete analysis of T1, T2, FLAIR, DWI, and SWI features. *Pol. J. Radiol.* **2025**, *90*, e394–e403. [CrossRef] [PubMed]
70. Tousignant, A.; Lemaître, P.; Precup, D.; Arnold, D.L.; Arbel, T. Prediction of Disease Progression in Multiple Sclerosis Patients using Deep Learning Analysis of MRI Data. In *Proceedings of the 2nd International Conference on Medical Imaging with Deep Learning*; PMLR: Cambridge, MA, USA, 2019; pp. 483–492. Available online: https://proceedings.mlr.press/v102/tousignant19a.html (accessed on 22 August 2025).

71. Taloni, A.; Farrelly, F.A.; Pontillo, G.; Petsas, N.; Giannì, C.; Ruggieri, S.; Petracca, M.; Brunetti, A.; Pozzilli, C.; Pantano, P.; et al. Evaluation of Disability Progression in Multiple Sclerosis via Magnetic-Resonance-Based Deep Learning Techniques. *Int. J. Mol. Sci.* **2022**, *23*, 10651. [CrossRef] [PubMed]
72. Coll, L.; Pareto, D.; Aparicio-Serrano, F.; Otero-Romero, S.; Cobo-Calvo, A.; Reinders, E.; Alberich, M.; Arévalo, M.J.; Arrambide, G.; Auger, C.; et al. Deep learning to predict progression independent of relapse activity at a first demyelinating event. *Brain Commun.* **2025**, *7*, fcaf243. [CrossRef] [PubMed]
73. Luu, M.S.K.; Tuchinov, B.N.; Prokaeva, A.I.; Korobko, D.S.; Malkova, N.A.; Tulupov, A.A. Discovering predictive features of multiple sclerosis from clinically isolated syndrome with machine learning. *Artif. Intell. Health* **2024**, *1*, 107. [CrossRef]
74. Zhao, Y.; Healy, B.C.; Rotstein, D.; Guttmann, C.R.G.; Bakshi, R.; Weiner, H.L.; Brodley, C.E.; Chitnis, T. Exploration of machine learning techniques in predicting multiple sclerosis disease course. *PLoS ONE* **2017**, *12*, e0174866. [CrossRef] [PubMed]
75. Houshi, S.; Khodakarami, Z.; Shaygannejad, A.; Khosravi, F.; Shaygannejad, V. AI-powered disease progression prediction in multiple sclerosis using magnetic resonance imaging: A systematic review and meta-analysis. *Neurol. Res.* **2025**, *48*, 419–433. [CrossRef] [PubMed]
76. Poretto, V.; Endrizzi, W.; Betti, M.; Bovo, S.; Bellinvia, A.; Ragni, F.; Lapucci, C.; Moroni, M.; Marangoni, S.; Portaccio, E.; et al. Machine Learning Analysis Applied to Prediction of Early Progression Independent of Relapse Activity in Multiple Sclerosis Patients. *Eur. J. Neurol.* **2025**, *32*, e70417. [CrossRef] [PubMed]
77. Yoo, Y.; Tang, L.W.; Brosch, T.; Li, D.K.B.; Metz, L.; Traboulsee, A.; Tam, R. Deep Learning of Brain Lesion Patterns for Predicting Future Disease Activity in Patients with Early Symptoms of Multiple Sclerosis. In *Deep Learning and Data Labeling for Medical Applications*; Carneiro, G., Mateus, D., Peter, L., Bradley, A., Tavares, J.M.R.S., Belagiannis, V., Papa, J.P., Nascimento, J.C., Loog, M., Lu, Z., et al., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 86–94. [CrossRef]
78. Haouam, K.-D.; Benmalek, M. Machine Learning Algorithms for Early Prediction of Multiple Sclerosis Progression: A Comparative Study. *Adv. Artif. Intell. Mach. Learn.* **2024**, *4*, 2027–2051. [CrossRef]
79. Durso-Finley, J.; Falet, J.-P.; Nichyporuk, B.; Douglas, A.; Arbel, T. Personalized Prediction of Future Lesion Activity and Treatment Effect in Multiple Sclerosis from Baseline MRI. In *Proceedings of the 5th International Conference on Medical Imaging with Deep Learning*; PMLR: Cambridge, MA, USA, 2022; pp. 387–406. Available online: https://proceedings.mlr.press/v172/durso-finley22a.html (accessed on 10 March 2026).
80. Salem, M.; Ryan, M.A.; Oliver, A.; Hussain, K.F.; Lladó, X. Improving the detection of new lesions in multiple sclerosis with a cascaded 3D fully convolutional neural network approach. *Front. Neurosci.* **2022**, *16*, 1007619. [CrossRef] [PubMed]
81. Afzal, H.M.R.; Luo, S.; Ramadan, S.; Lechner-Scott, J.; Li, J. Automatic Prediction of the Conversion of Clinically Isolated Syndrome to Multiple Sclerosis Using Deep Learning. In *Proceedings of the 2018 2nd International Conference on Video and Image Processing, ICVIP '18*; Association for Computing Machinery: New York, NY, USA, 2018; pp. 231–235. [CrossRef]
82. Yılmaz Acar, Z.; Başçiftçi, F.; Ekmekci, A.H. Future activity prediction of multiple sclerosis with 3D MRI using 3D discrete wavelet transform. *Biomed. Signal Process. Control* **2022**, *78*, 103940. [CrossRef]
83. Kırkbir, İ.B.; Kurt, B.; Boz, C.; Terzi, M.; Sarı, A. Predicting 5-Year EDSS in Multiple Sclerosis with LSTM Networks: A Deep Learning Approach to Disease Progression. *J. Clin. Neurosci.* **2025**, *136*, 111218. [CrossRef] [PubMed]
84. Elliott, C.; Belachew, S.; Wolinsky, J.S.; Hauser, S.L.; Kappos, L.; Barkhof, F.; Bernasconi, C.; Fecker, J.; Model, F.; Wei, W.; et al. Chronic white matter lesion activity predicts clinical progression in primary progressive multiple sclerosis. *Brain* **2019**, *142*, 2787–2799. [CrossRef] [PubMed]
85. Tayyab, M.; Metz, L.M.; Li, D.K.B.; Kolind, S.; Carruthers, R.; Traboulsee, A.; Tam, R.C. Accounting for uncertainty in training data to improve machine learning performance in predicting new disease activity in early multiple sclerosis. *Front. Neurol.* **2023**, *14*, 1165267. [CrossRef] [PubMed]
86. Aksoy, S.; Demircioglu, P.; Bogrekci, I. SeruNet-MS: A Two-Stage Interpretable Framework for Multiple Sclerosis Risk Prediction with SHAP-Based Explainability. *Neurol. Int.* **2025**, *17*, 151. [CrossRef] [PubMed]
87. Wottschel, V.; Chard, D.T.; Enzinger, C.; Filippi, M.; Frederiksen, J.L.; Gasperini, C.; Giorgio, A.; Rocca, M.A.; Rovira, A.; De Stefano, N.; et al. SVM recursive feature elimination analyses of structural brain MRI predicts near-term relapses in patients with clinically isolated syndromes suggestive of multiple sclerosis. *NeuroImage Clin.* **2019**, *24*, 102011. [CrossRef] [PubMed]
88. Wottschel, V.; Alexander, D.C.; Kwok, P.P.; Chard, D.T.; Stromillo, M.L.; De Stefano, N.; Thompson, A.J.; Miller, D.H.; Ciccarelli, O. Predicting outcome in clinically isolated syndrome using machine learning. *NeuroImage Clin.* **2015**, *7*, 281–287. [CrossRef] [PubMed]
89. Yoo, Y.; Tang, L.Y.W.; Li, D.K.B.; Metz, L.; Kolind, S.; Traboulsee, A.L.; Tam, R.C. Deep learning of brain lesion patterns and user-defined clinical and MRI features for predicting conversion to multiple sclerosis from clinically isolated syndrome. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* **2019**, *7*, 250–259. [CrossRef]
90. Sepahvand, N.M.; Hassner, T.; Arnold, D.L.; Arbel, T. CNN Prediction of Future Disease Activity for Multiple Sclerosis Patients from Baseline MRI and Lesion Labels. In *Proceedings of the Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*;

Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 57–69. [CrossRef]
91. Marastoni, D.; Colato, E.; Foschi, M.; Tamanti, A.; Ziccardi, S.; Eccher, C.; Crescenzo, F.; Bajrami, A.; Schiavi, G.M.; Camera, V.; et al. Intrathecal Inflammatory Profile and Gray Matter Damage Predict Progression Independent of Relapse Activity in Early Multiple Sclerosis. *Neurol. Neuroimmunol. Neuroinflamm* **2025**, *12*, e200399. [CrossRef] [PubMed]
92. Ianniello, A.; Barbuti, E.; Capobianco, M.F.; Tranfa, M.; Miele, C.; Ruggieri, S.; Pontillo, G.; Cocozza, S.; Pantano, P.; Pozzilli, C.; et al. Predictors of short-term, relapse-independent progression in multiple sclerosis: A machine learning approach based on clinical data and conventional MRI-derived features. *J. Neurol. Sci.* **2026**, *483*, 125846. [CrossRef] [PubMed]
93. Assessing the Relative Importance of Imaging and Serum Biomarkers in Capturing Disability, Cognitive Impairment, and Clinical Progression in Multiple Sclerosis-Cagol-2026-Advanced Science-Wiley Online Library. Available online: https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.202512946 (accessed on 12 August 2026).
94. Aerts, S.; Werthen-Brabants, L.; Khan, H.; Giraldo, D.L.; De Brouwer, E.; Geys, L.; Popescu, V.; Sijbers, J.; Woodruff, H.C.; Dhaene, T.; et al. Combining magnetic resonance imaging and evoked potentials enhances machine learning prediction of multiple sclerosis disability worsening. *Front. Immunol.* **2026**, *17*, 1625837. [CrossRef] [PubMed]
95. Wang, Z.; Li, S.; Liu, J.; Li, Y.; Tang, M.; Wang, J.; Ren, W.; Guo, J.; Shao, Z.; Jiang, J.; et al. Nomograms integrating cortical morphometric metrics with 3D multi-parametric MRI radiomics for predicting disability progression and cognitive worsening in relapsing–remitting multiple sclerosis: A multi-center validation study. *J. Neurol.* **2025**, *272*, 792. [CrossRef] [PubMed]
96. Shen, A.A.; McLoughlin, A.; Vernon, Z.; Lin, J.; Carano, R.A.D.; Bickel, P.J.; Song, Z.; Huang, H. Assessing the role of volumetric brain information in multiple sclerosis progression. *Comput. Struct. Biotechnol. J.* **2025**, *27*, 2014–2033. [CrossRef] [PubMed]
97. Bejarano, B.; Bianco, M.; Gonzalez-Moron, D.; Sepulcre, J.; Goñi, J.; Arcocha, J.; Soto, O.; Carro, U.D.; Comi, G.; Leocani, L.; et al. Computational classifiers for predicting the short-term course of Multiple sclerosis. *BMC Neurol.* **2011**, *11*, 67. [CrossRef] [PubMed]
98. Zhao, Y.; Wang, T.; Bove, R.; Cree, B.; Henry, R.; Lokhande, H.; Polgar-Turcsanyi, M.; Anderson, M.; Bakshi, R.; Weiner, H.L.; et al. Ensemble learning predicts multiple sclerosis disease course in the SUMMIT study. *npj Digit. Med.* **2020**, *3*, 135. [CrossRef] [PubMed]
99. Andorra, M.; Freire, A.; Zubizarreta, I.; de Rosbo, N.K.; Bos, S.D.; Rinas, M.; Høgestøl, E.A.; de Rodez Benavent, S.A.; Berge, T.; Brune-Ingebretse, S.; et al. Predicting disease severity in multiple sclerosis using multimodal data and machine learning. *J. Neurol.* **2024**, *271*, 1133–1149. [CrossRef] [PubMed]
100. Pellegrini, F.; Copetti, M.; Sormani, M.P.; Bovis, F.; de Moor, C.; Debray, T.P.; Kieseier, B.C. Predicting disability progression in multiple sclerosis: Insights from advanced statistical modeling. *Mult. Scler.* **2020**, *26*, 1828–1836. [CrossRef] [PubMed]
101. Roca, P.; Attye, A.; Colas, L.; Tucholka, A.; Rubini, P.; Cackowski, S.; Ding, J.; Budzik, J.-F.; Renard, F.; Doyle, S.; et al. Artificial intelligence to predict clinical disability in patients with multiple sclerosis using FLAIR MRI. *Diagn. Interv. Imaging* **2020**, *101*, 795–802. [CrossRef] [PubMed]
102. Campanioni, S.; Veiga, C.; Prieto-González, J.M.; González-Nóvoa, J.A.; Busto, L.; Martinez, C.; Alberte-Woodward, M.; García De Soto, J.; Pouso-Diz, J.; Fernández Ceballos, M.D.L.Á.; et al. Explainable machine learning on baseline MRI predicts multiple sclerosis trajectory descriptors. *PLoS ONE* **2024**, *19*, e0306999. [CrossRef] [PubMed]
103. Al-iedani, O.; Lea, S.; Alshehri, A.; Maltby, V.E.; Saugbjerg, B.; Ramadan, S.; Lea, R.; Lechner-Scott, J. Multi-modal neuroimaging signatures predict cognitive decline in multiple sclerosis: A 5-year longitudinal study. *Mult. Scler. Relat. Disord.* **2024**, *81*, 105379. [CrossRef] [PubMed]
104. Rocca, M.A.; Valsasina, P.; Meani, A.; Pagani, E.; Cordani, C.; Cervellin, C.; Filippi, M. Network Damage Predicts Clinical Worsening in Multiple Sclerosis: A 6.4-Year Study. *Neurol. Neuroimmunol. Neuroinflamm* **2021**, *8*, e1006. [CrossRef] [PubMed]
105. Pisani, A.I.; Scalfari, A.; Crescenzo, F.; Romualdi, C.; Calabrese, M. A novel prognostic score to assess the risk of progression in relapsing−remitting multiple sclerosis patients. *Eur. J. Neurol.* **2021**, *28*, 2503–2512. [CrossRef] [PubMed]
106. Seccia, R.; Gammelli, D.; Dominici, F.; Romano, S.; Landi, A.C.; Salvetti, M.; Tacchella, A.; Zaccaria, A.; Crisanti, A.; Grassi, F.; et al. Considering patient clinical history impacts performance of machine learning models in predicting course of multiple sclerosis. *PLoS ONE* **2020**, *15*, e0230219. [CrossRef] [PubMed]
107. Falet, J.-P.R.; Nobile, S.; Szpindel, A.; Barile, B.; Kumar, A.; Durso-Finley, J.; Arbel, T.; Arnold, D.L. The role of AI for MRI-analysis in multiple sclerosis—A brief overview. *Front. Artif. Intell.* **2025**, *8*, 1478068. [CrossRef] [PubMed]
108. Sun, Y.; Wang, L.; Li, G.; Lin, W.; Wang, L. A foundation model for enhancing magnetic resonance images and downstream segmentation, registration and diagnostic tasks. *Nat. Biomed. Eng.* **2025**, *9*, 521–538. [CrossRef] [PubMed]
109. Tak, D.; Garomsa, B.A.; Zapaishchykova, A.; Chaunzwa, T.L.; Climent Pardo, J.C.; Ye, Z.; Zielke, J.; Ravipati, Y.; Pai, S.; Vajapeyam, S.; et al. A generalizable foundation model for analysis of human brain MRI. *Nat. Neurosci.* **2026**, *29*, 945–956. [CrossRef] [PubMed]
110. Huang, S.-C.; Pareek, A.; Jensen, M.; Lungren, M.P.; Yeung, S.; Chaudhari, A.S. Self-supervised learning for medical image classification: A systematic review and implementation guidelines. *npj Digit. Med.* **2023**, *6*, 74. [CrossRef] [PubMed]

111. Bai, L.; Wang, D.; Wang, H.; Barnett, M.; Cabezas, M.; Cai, W.; Calamante, F.; Kyle, K.; Liu, D.; Ly, L.; et al. Improving multiple sclerosis lesion segmentation across clinical sites: A federated learning approach with noise-resilient training. *Artif. Intell. Med.* **2024**, *152*, 102872. [CrossRef] [PubMed]
112. Hindawi, S.; Szubstarski, B.; Boernert, E.; Tackenberg, B.; Wuerfel, J. Federated learning for lesion segmentation in multiple sclerosis: A real-world multi-center feasibility study. *Front. Neurol.* **2025**, *16*, 1620469. [CrossRef] [PubMed]
113. Spagnolo, F.; Molchanova, N.; Cuadra, M.B.; Ocampo-Pineda, M.; Melie-Garcia, L.; Granziera, C.; Andrearczyk, V.; Depeursinge, A. Instance-level quantitative saliency in multiple sclerosis lesion segmentation. *Sci. Rep.* **2026**, *16*, 6917. [CrossRef] [PubMed]
114. Goebl, P.; Wingrove, J.; Abdelmannan, O.; Brito Vega, B.; Stutters, J.; Ramos, S.D.G.; Kenway, O.; Rossor, T.; Wassmer, E.; Arnold, D.L.; et al. Enabling new insights from old scans by repurposing clinical MRI archives for multiple sclerosis research. *Nat. Commun.* **2025**, *16*, 3149. [CrossRef] [PubMed]
115. Eshaghi, A.; Young, A.L.; Wijeratne, P.A.; Prados, F.; Arnold, D.L.; Narayanan, S.; Guttmann, C.R.G.; Barkhof, F.; Alexander, D.C.; Thompson, A.J.; et al. Identifying multiple sclerosis subtypes using unsupervised machine learning and MRI data. *Nat. Commun.* **2021**, *12*, 2078. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.