NeuroImage 265 (2023) 119787

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Single-timepoint low-dimensional characterization and classification of acute versus chronic multiple sclerosis lesions using machine learning

Bastien Caba[a,*], Alexandre Cafaro[e], Aurélien Lombard[e], Douglas L. Arnold[b,c], Colm Elliott[c], Dawei Liu[a], Xiaotong Jiang[a], Arie Gafson[a], Elizabeth Fisher[a], Shibeshih Mitiku Belachew[a], Nikos Paragios[d,e]

[a] Biogen Digital Health, Biogen, Cambridge, MA, USA
[b] Montreal Neurological Institute, McGill University, Montreal, QC, Canada
[c] NeuroRx Research, Montreal, QC, Canada
[d] CentraleSupélec, University of Paris-Saclay, Gif-sur-Yvette, France
[e] TheraPanacea, Paris, France

ARTICLE INFO

Keywords:
Multiple sclerosis
Acute lesions
Radiomics
Image inpainting
Feature selection
Machine learning

ABSTRACT

Multiple sclerosis (MS) is a chronic inflammatory and neurodegenerative disease characterized by the appearance of focal lesions across the central nervous system. The discrimination of acute from chronic MS lesions may yield novel biomarkers of inflammatory disease activity which may support patient management in the clinical setting and provide endpoints in clinical trials. On a single timepoint and in the absence of a prior reference scan, existing methods for acute lesion detection rely on the segmentation of hyperintense foci on post-gadolinium T1-weighted magnetic resonance imaging (MRI), which may underestimate recent acute lesion activity. In this paper, we aim to improve the sensitivity of acute MS lesion detection in the single-timepoint setting, by developing a novel machine learning approach for the automatic detection of acute MS lesions, using single-timepoint conventional non-contrast T1- and T2-weighted brain MRI. The MRI input data are supplemented via the use of a convolutional neural network generating "lesion-free" reconstructions from original "lesion-present" scans using image inpainting. A multi-objective statistical ranking module evaluates the relevance of textural radiomic features from the core and periphery of lesion sites, compared within "lesion-free" versus "lesion-present" image pairs. Then, an ensemble classifier is optimized through a recursive loop seeking consensus both in the feature space (via a greedy feature-pruning approach) and in the classifier space (via model selection repeated after each pruning operation). This leads to the identification of a compact textural signature characterizing lesion phenotype. On the *patch-level* task of acute versus chronic MS lesion classification, our method achieves a balanced accuracy in the range of 74.3–74.6% on fully external validation cohorts.

1. Introduction

Multiple sclerosis (MS) is a chronic immune-mediated neurodegenerative disease affecting the central nervous system. Its pathological hallmark is the accumulation of demyelinated lesions or *plaques*, detectable on conventional T2-weighted magnetic resonance imaging (MRI) as areas of white matter hyperintensity (WMH) relative to the normal-appearing white matter (NAWM) (Traboulsee and Li, 2006). The delineation of WMHs therefore reveals the spatial distribution of MS damage but does not discriminate lesions that have recently formed and may be undergoing active demyelination (known as *acute*) from lesions that have been present for some time (known as *chronic*[1]). The detection and quantification of *acute* lesions is relevant to the clinical management of patients with MS, in whom evidence of recent disease activity may guide treatment decisions such as switching anti-inflammatory

[1] In this context, chronic lesions may or may not harbor chronic active demyelination.

Abbreviations: BBB, blood brain barrier; CNN, convolutional neural network; CSF, cerebro-spinal fluid; DMT, disease-modifying therapy; EDSS, expanded disability status scale; GAN, generative adversarial network; Gd+, gadolinium enhancement; ML, machine learning; MRI, magnetic resonance imaging; MS, multiple sclerosis; NAWM, normal-appearing white matter; NET2, new or enlarging T2 MS lesion; RFE, recursive feature elimination; ROI, region of interest; TA, texture analysis; WMH, white matter hyperintensity.

* Corresponding author at: Bastien Caba, Biogen Inc. 225 Binney Street Cambridge, MA 02142.
E-mail address: bastien.caba@biogen.com (B. Caba).

https://doi.org/10.1016/j.neuroimage.2022.119787.
Received 20 September 2022; Received in revised form 16 November 2022; Accepted 2 December 2022
Available online 5 December 2022.

disease-modifying therapies (DMTs) and may support short- and long-term disease prognostication (Wattjes et al., 2021). In clinical trials evaluating MS DMTs, evidence of recent acute lesion activity is also used to define inclusion criteria and enrichment strategies, while measurements of *ongoing* acute lesion activity are commonly used as endpoints (Calabresi et al., 2014; Kapoor et al., 2018; Wattjes et al., 2021).

The formation of new lesions is generally accompanied by a transient period of blood brain barrier (BBB) breakdown and acute inflammatory activity (Kappos et al., 1999). Following the intravenous administration of a gadolinium chelate contrast agent, areas of BBB breakdown can be detected as hyperintensities on post-contrast T1-weighted MRI. The delineation of areas of gadolinium enhancement (Gd+) thus allows for the detection of acute lesions at a single timepoint; however, due to the relatively short half-life of BBB disruption and Gd+ persistence (Cotton et al., 2003; Guttmann et al., 2016), acute lesion detection based on this method severely underestimates acute pathology. This method is also invasive and potentially nephrotoxic (Perazella, 2008). Repeated gadolinium-based contrast agent administrations may also result in gadolinium accumulation in the human brain, the clinical implications of which are currently poorly understood (Guo et al., 2018).

Beyond Gd+ status, acute lesions can also be detected as focal areas of new or substantially enlarging T2-weighted (NET2) lesions via the comparison of a present T2-weighted MRI scan with a previously acquired reference scan (Altay et al., 2013). Notably, when a previous MRI scan is available from up to 24 weeks prior, NET2 counts are generally 3-fold to 5-fold higher than the single-timepoint detection of Gd+ lesion counts (Hauser et al., 2017; Moraal et al., 2010). The requirement of a reference scan to identify NET2 can, however, delay diagnosis or treatment decisions and incurs an economic burden (Kobelt et al., 2017). Furthermore, by construction, the identification of NET2 only allows detection of the subset of acute lesions that emerge from the NAWM and fails to detect acute activity within existing WMHs. Since acute lesions may also form in pre-existing lesion locations (as demonstrated by the detection of Gd+ foci within prior WMHs), some degree of acute lesion activity may be missed using NET2 detection techniques. In particular, acute lesions occurring in areas of pre-existing WMH which are captured outside of their Gd+ phase will thus be missed under the conventional definition of “MRI-acute” lesions.

The hypothesis that “MRI-acute” lesions may underestimate the extent of pathologically “active” demyelination is further supported by the high prevalence of “active” plaques observed in pathology studies reporting 70–80% of patients with at least one “active” plaque between age 35–50 years-old (Frischer et al., 2015), which exceeds the prevalence of “MRI-acute” lesions generally reported in age-matched clinical trial populations – albeit MS trials typically involve an acute lesion enrichment selection bias inherent to disease activity eligibility criteria.

2. Objectives

These limitations highlight the need for a single-timepoint acute lesion detection method that could identify the totality of recent acute MS lesion activity (including gadolinium-negative lesions in areas of pre-existing lesions). This work seeks to augment the sensitivity of cross-sectional methods of acute lesion detection in MS by capturing MRI biomarker signatures identified within ground truth examples of both Gd+ and NET2 lesions, from non-contrast conventional T1- and T2-weighted MRI only. In this work, ground truth NET2 lesions were defined based on the comparison of two T2-weighted MRI scans acquired at most 24 weeks apart.

Using intensity-based radiomic features from the core and periphery of lesion sites, compared with their counterparts on a “lesion-free” synthetic image, we identify patterns of MRI signals that are discriminative of recent acute MS lesion activity and reflective of ongoing demyelination. For notational compactness, we will denote this set of radiomic features enriched with lesion-free information as “*α-radiomics*”. The relevant patterns of *α-radiomics* are interpreted by an ensemble of machine learning (ML) algorithms trained to discriminate acute from chronic MS lesions. The resulting ensemble predicts the label (i.e., acute/chronic) of an individual MRI voxel using MRI signals aggregated across a cubic patch centered on that voxel.

The methodological contributions of this work are as follows:

- Using a convolutional neural network (CNN) for image inpainting, we generate synthetic “lesion-free” T1- and T2-weighted brain MRI scans from original “lesion-present" scans, and compare textural information contained in “lesion-present” versus “lesion-free” scans.
- We design a patch sampling strategy combined with a simple adaptive region of interest (ROI) definition rule, which allows us to segment a “core” and “periphery” regions from lesion sites that may correspond to either focal or confluent MS lesions.
- We present an end-to-end greedy iterative feature-pruning algorithm for feature selection (i.e., imaging biomarker discovery) exploiting textural characteristics within lesions and their periphery, coupled with an ensemble classifier optimization module applied at each feature pruning step.

3. Backgrounds

3.1. Texture analysis in MS

Several studies have investigated the utility of texture analysis (TA) on MRI in MS. It is hypothesized that pathological processes in MS induce structural changes expressed on the nanometer to micrometer scale, which manifest as voxel pattern changes on conventional MRI images (Zhang et al., 2011). It follows that the detection of textural biomarkers associated with specific types of biological activity may support the identification of MS lesions from single-timepoint conventional MRI. The feasibility of this approach is further strengthened by previous work showing that TA on MRI is relatively unaffected by the image acquisition variables (Mayerhoefer et al., 2009; Savio et al., 2010) and as such offers robust biomarkers suitable for use in routine clinical care of patients with MS, where image acquisition parameters may vary.

Specifically, TA has been investigated for its ability to discriminate MS plaques from the NAWM using conventional cross-sectional T1- and T2-weighted MRI (Harrison et al., 2010; Zhang et al., 2008), to discriminate Gd+ lesions from chronic lesions using T2-weighted MRI (Michoux et al., 2015; Yu et al., 1999), and for the precise delineation of Gd+ lesions using post-contrast T1-weighted MRI (Karimaghaloo et al., 2013). Furthermore, TA has been leveraged to characterize acute lesions, by spatially distinguishing a demyelinating core from a border of inflammation (Drabycz and Mitchell, 2008), measuring tissue injury in acute foci (Zhang et al., 2009) and by predicting the persistence of T1 black holes beyond acute onset (Zhang et al., 2011). Building upon these prior studies, our work leverages radiomics analysis, which consists of mining a large number of quantitative imaging variables (Zwanenburg et al., 2020). Radiomics analysis has been suggested to hold great promise in the transition towards large-scale medical imaging studies, whereby the abundance of visual evidence can be leveraged by ML techniques to address clinical challenges (Gillies et al., 2016; Lambin et al., 2012) – in this context, our work also builds upon prior applications of ML to augment lesion segmentation in MS including for WMHs (Fartaria et al., 2016; Zeng et al., 2020; Zhong et al., 2014), Gd+ (Coronado et al., 2021; Gaj et al., 2021; Narayana et al., 2020), and NET2 lesions (Elliott et al., 2013; Salem et al., 2018).

3.2. Prior work

In this study, NET2 lesions are defined as those WMHs that are less than 24 weeks old. As such, the task of acute versus chronic MS lesion classification can be seen as a binarized version of the task of MS lesion age estimation. In this context, Sweeney et al. (Sweeney et al., 2021) recently developed a ML-based algorithm for MS lesion age prediction

Table 1
Key population, lesion, and imaging statistics of the participants pooled across the ADVANCE, ASCEND, and DECIDE trials. (*) Exceptionally, for 2 ADVANCE subjects and 13 DECIDE subjects, 1.0T MRI scans were acquired.

Cohort		ADVANCE	ASCEND	DECIDE
MS Stage		RRMS	SPMS	RRMS
EDSS Range		[{0:5}]	[{3:6.5}]	[{0:5}]
Subjects		**1397**	**814**	**1713**
Studies (with acute lesion activity)		3971 (1612)	3695 (769)	3601 (958)
Lesion volume per	Acute	0.34 ± 1.3	0.10 ± 0.50	0.14 ± 0.73
study (c{m^3})	Chronic	10 ± 12	17 ± 17	11 ± 9
Lesion count per	Acute	3 ± 7	1 ± 3	1 ± 5
study	Chronic	68 ± 48	69 ± 38	63 ± 44
MRI Protocol	T1-w MRI	3D Spoiled Gradient Echo	3D Spoiled Gradient Echo	3D Spoiled Gradient Echo
		TR = 28 - 35ms	TR = 28 - 35ms	TR = 28 - 35ms
		TE = 4 - 11ms	TE = 4 - 11ms	TE = 4 - 11ms
		Flip angle = 27 - 30°	Flip angle = 27 - 30°	Flip angle = 27 - 30°
		Resolution = 1 × 1 × 3 mm	Resolution = 0.98 × 0.98 × 3 mm	Resolution = 0.98 × 0.98 × 3 mm
		Field Strength = 1.5 T/3.0 T*	Field Strength = 1.5 T/3.0 T	Field Strength = 1.5 T/3.0 T*
	T2-w MRI	2D Fast Spin Echo	2D Fast Spin Echo	2D Fast Spin Echo
		TR = 4000 - 7720 ms	TR = 4000 - 7400 ms	TR = 4000 - 7400 ms
		TE = 56 - 93 ms	TE = 58 - 95 ms	TE = 60 - 96 ms
		Resolution = 1 × 1 × 3 mm	Resolution = 0.98 × 0.98 × 3 mm	Resolution = 0.98 × 0.98 × 3 mm
		Field Strength = 1.5 T/3.0 T*	Field Strength = 1.5 T/3.0 T	Field Strength = 1.5 T/3.0 T*

Note: In the rows related to lesion volume and count, we are reporting the mean ± standard deviation of lesion statistics across studies.
Abbreviations: EDSS, Expanded Disability Status Scale; MRI, magnetic resonance imaging; MS, multiple sclerosis; RR, relapsing-remitting; SP, secondary progressive; TR, repetition time; TE, echo time.

from a single MRI study, using radiomic features from T1-weighted (pre- and post-contrast), T2-weighted, FLAIR and quantitative susceptibility mapping (QSM) MRI. They analyzed 53 discrete MS lesion samples and reported mean and median absolute errors of 7.23 months (95% CI: [6.98, 13.43]) and 5.98 months (95% CI: [5.26, 13.25]), respectively. Although this prior work is relevant to our objectives, our approach differs in that we leverage examples of both discrete *and confluent* MS lesions on a larger scale ($\sim 10^4$ samples), while restricting data input to MRI sequences readily available in the clinic (non-contrast T1- and T2-weighted MRI).

4. Materials

4.1. MRI data

Brain MRI scans from three large-scale double-blind phase 3 pivotal trials were retrospectively analyzed, including T1- and T2-weighted brain MRI scans from ADVANCE (1512 subjects with relapsing-remitting MS; NCT00906399), ASCEND (886 subjects with secondary progressive MS; NCT01416181), and DECIDE (1841 subjects with relapsing-remitting MS; NCT01064401). Study details and MRI acquisition parameters for these trials have been reported previously (Calabresi et al., 2014; Kapoor et al., 2018; Kappos et al., 2015) and are summarized in Table 1. For each participant, one baseline study and a series of follow-up studies were conducted; each subject could thus contribute multiple observations to our analysis. Each study was associated with at least one T2-weighted MRI scan as well as T1-weighted MRI scans pre- and post-gadolinium injection. For ADVANCE, the follow-up MRI scans were acquired at 24-, 48-, and 96-weeks post-baseline; for ASCEND at 24, 48, 72, 96, 108, and 156 weeks, and for DECIDE at 24 and 96 weeks. A subset ($n = 142$) of participants from ADVANCE had follow-up scans every 4 weeks up to 24 weeks post-baseline. In total, 11,267 pairs of T1- and T2-weighted brain MRI scans were analyzed in this study.

4.2. Ground truth

In the context of the original clinical trial, MRI scans were analyzed by a central reading center (NeuroRx Research, Montreal, QC, Canada), as follows: Gd+ lesions were manually segmented at each timepoint of MRI acquisition by consensus of two trained experts while WMHs were segmented using a semi-automated method including a manual verification and correction step (Francis, 2010). This process yielded two binary maps per study, denoting WMHs and Gd+ lesions respectively (see Fig. 1).

From the WMH masks, the NET2 lesion mask detected in a present scan relative to a prior reference scan was determined automatically by considering the difference between the sets of WMH voxels detected across these time-points, where artifactual differences due to segmentation variability or imperfect registration were automatically excluded. These NET2 lesion masks were manually reviewed and corrected where necessary (see Fig. 2). Importantly, a new WMH was labeled as NET2 if it was detected via the comparison of scans acquired at most 24 weeks apart. Ultimately, acute lesions were defined as the union between the Gd+ and NET2 masks, within the bounds of the WMH mask. Conversely, chronic lesions were defined as the non-acute section of the WMH mask (see Fig. 3).

5. Methods

5.1. MRI pre-processing

Across all scans, spatial coherence was enforced via 6-parameter rigid registration (translation and rotation in 3D space) to the 2009a Non-linear Symmetric atlas (Fonov et al., 2009) in the ICBM-152 reference space, followed by a resampling operation to isotropic voxel spacing (1 × 1 × 1 mm) via trilinear interpolation, yielding final scan dimensions of 256 × 256 × 180 voxels. Additionally, each scan underwent N3 bias-field correction (Sled et al., 1998), followed by an intensity standardization procedure enforcing that NAWM voxels should have zero mean and unit-variance. The NAWM was defined as the set of white matter voxels without WMH; to mitigate the risk of contamination of the normalization parameters by peri-lesion abnormalities, NAWM voxels located within 2 mm of a WMH were excluded from the computation of the normalization parameters.

5.2. Data supplement via MS lesion inpainting

Our dataset of T1- and T2-weighted brain MRI scans was supplemented with synthetic lesion-free equivalents of each scan, generated via a CNN-based inpainting model. This model performs an image-to-image translation task whereby the MRI intensity profile within the

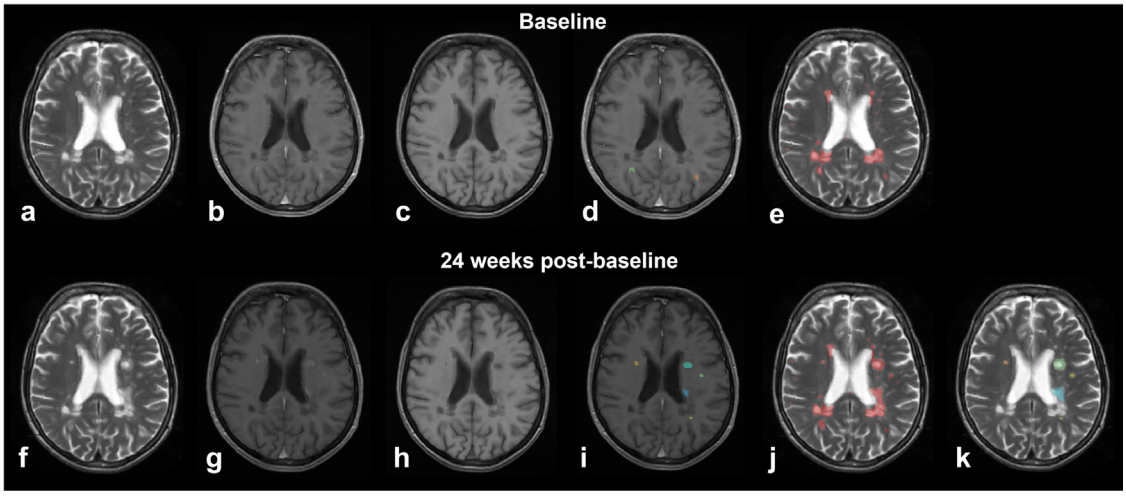

Fig. 1. Example of MRI scans and MS lesion masks available for an arbitrarily selected subject from the ADVANCE cohort. (a) T2-weighted MRI at baseline; (b) Contrast-enhanced T1-weighted MRI at baseline; (c) Non-contrast T1-weighted MRI at baseline; (d) Gd+ lesion masks at baseline; (e) WMH mask at baseline; (f) T2-weighted MRI at week 24 post-baseline; (g) Contrast-enhanced T1-weighted MRI at week 24 post-baseline; (h) Non-contrast T1-weighted MRI at week 24 post-baseline; (i) Gd+ lesion masks at week 24 post-baseline; (j) WMH mask at week 24 post-baseline; (k) NET2 lesion masks at week 24 post-baseline, relative to baseline.
Abbreviations: Gd+, gadolinium enhancement; MRI, magnetic resonance imaging; MS, multiple sclerosis; NET2, new or substantially enlarging T2-weighted; WMH, white matter hyperintensity.

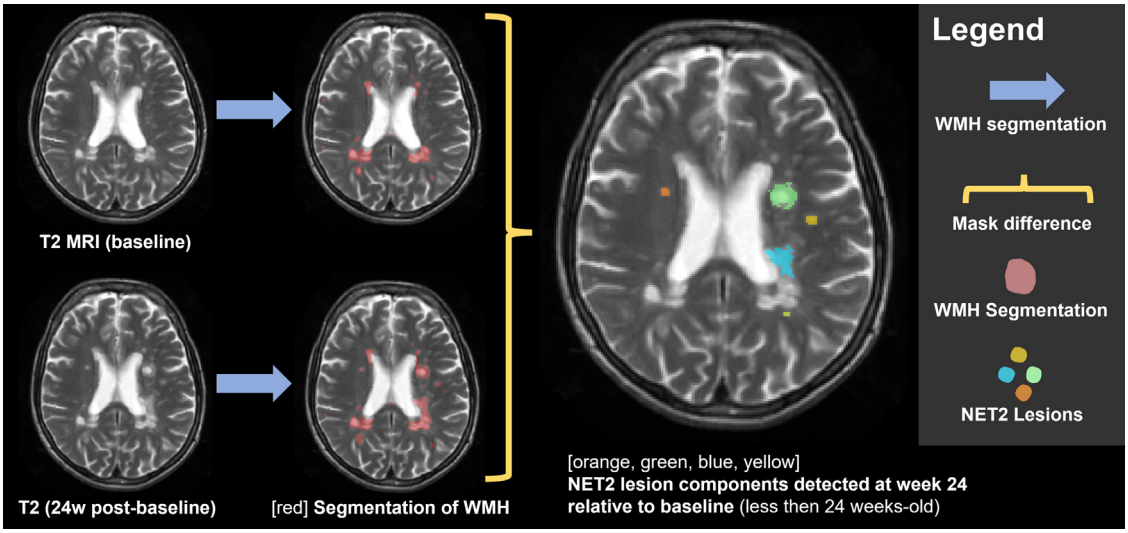

Fig. 2. Definition of ground truth NET2 lesion mask. A new WMH was labeled as acute if it arose within the previous 24weeks. The WMH region is first segmented in each timepoint of T2-weighted MRI scan acquisition. For each post-baseline timepoint t (e.g., 24 weeks post-baseline), the NET2 mask is the set of voxels that are labeled as WMH at that timepoint t and were not labeled as WMH in a previous timepoint $t-1$ acquired at most 24 weeks prior to t. After their automatic detection, NET2 masks were manually corrected by trained MRI readers.
Abbreviations: Gd+, gadolinium enhancement; MRI, magnetic resonance imaging; NET2, new or substantially enlarging T2-weighted; w, week; WMH, white matter hyperintensity.

WMH region of an input MRI scan is replaced with a synthetic inpainting mimicking normal-appearing tissue. During both training and inference, WMH masks were isotropically dilated by 1 mm, which ensures that both the lesion foci and any immediate peri-plaque abnormality are inpainted.

5.2.1. *Architecture*

The 2D inpainting method proposed by Yu et al. (Yu et al., 2019) was adapted to the 3D setting via multi-view ensembling. Specifically, we trained one model per anatomical view (axial, coronal, sagittal) and aggregated predictions across views via voxel-wise averaging at test time. Each 2D inpainting model was based on a generative adversarial network (GAN) (Goodfellow et al., 2020) composed of two encoder-decoder generator blocks referred to as the *coarse* and *refinement* blocks (see Fig. 4), followed by a discriminator block. Each generator block implemented gated convolutions (Liu et al., 2018) to restrict the encoding-decoding process to information contained outside of the region to be inpainted; thus, each generator block produces a synthetic full-brain MRI image, whereby the intensity profile in the inpainted region is predicted using information outside of that region. Specifically, the refinement generator block included an attention module guiding the encoding process. The contextual attention module originally used by Yu et al. (Yu et al., 2018) was replaced with a recursive self-attention module. This module allows for MRI signals distant from the lesion site to be integrated into the prediction of a hypothetical normal-appearing tissue profile replacing lesion tissue (Zhang et al., 2019). The combined use

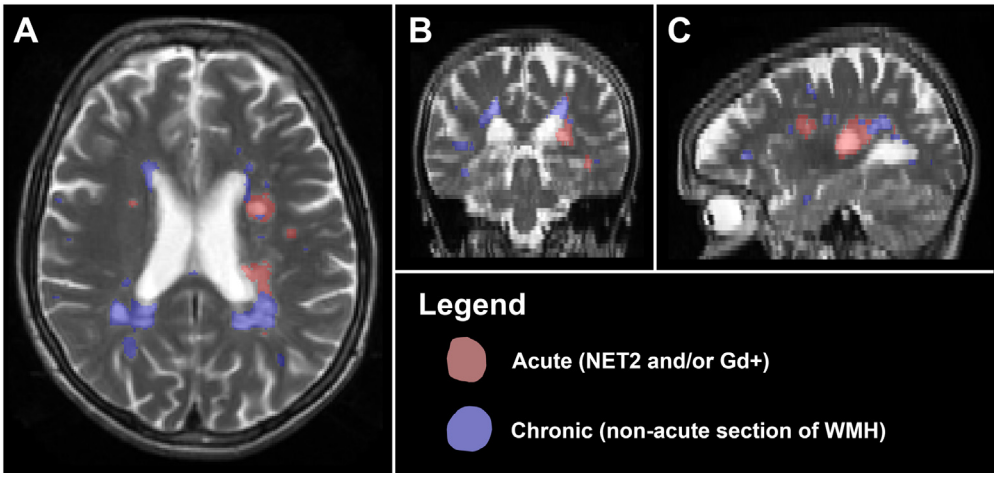

Fig. 3. Axial (A), coronal (B), and sagittal (C) views of a T2-weighted brain MRI scan randomly selected from the ADVANCE cohort, showing the acute (red) and chronic (blue) ground truth segmentation maps. We observe isolated chronic and acute foci, as well as contiguous acute and chronic lesions.
Abbreviations: Gd+, gadolinium enhancement; MRI, magnetic resonance imaging; NET2, new or substantially enlarging T2-weighted; WMH, white matter hyperintensity.

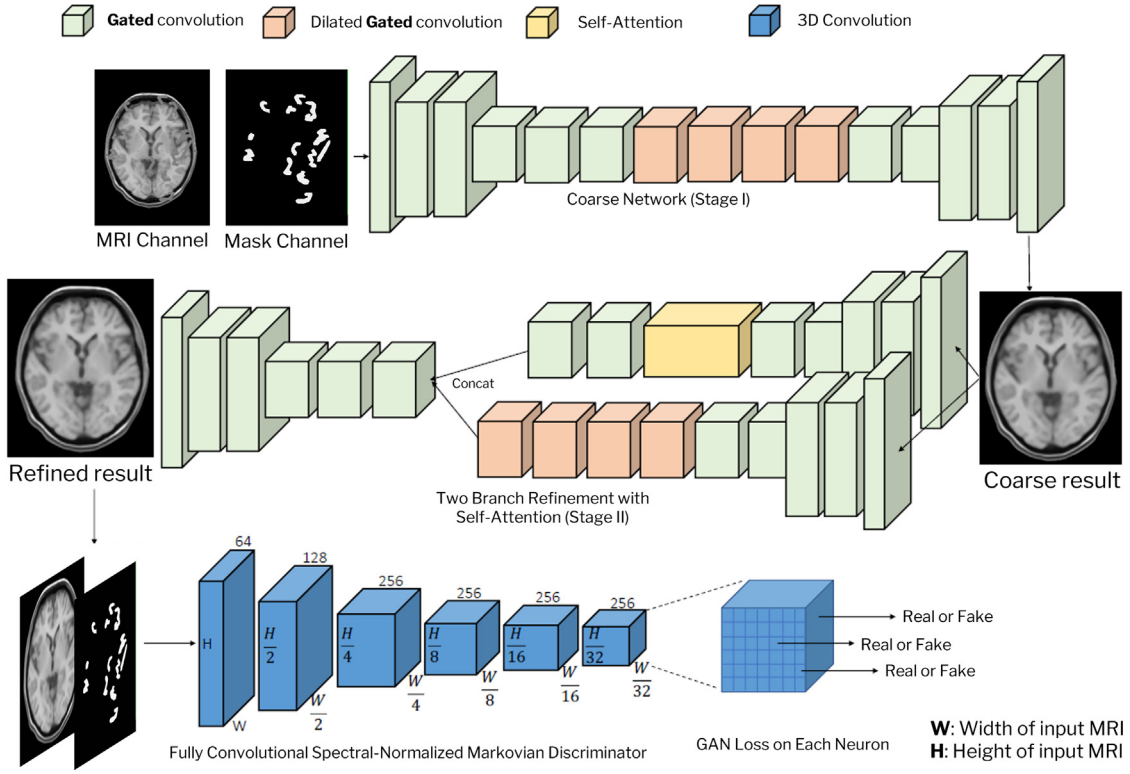

Fig. 4. Architecture of our CNN model for MS lesion inpainting, adapted from (Yu et al., 2019). During training, free-form synthetic lesion masks are created outside of the WMH masks while during inference, WMH masks are fed into the model. The generator block consists of a sequence of coarse and refinement autoencoders. The output of the generator block ("Refined result") is the inpainted "lesion-free" image extracted during inference. During training, the output of the generator block is subsequently fed into a discriminator model, which encodes it into a set of latent features. In parallel, the discriminator model also encodes a randomly selected, non-inpainted MRI image containing healthy tissue within the white matter region delimited by the free-form synthetic lesion mask. For each latent feature (i.e., neuron of the last feature map), the probability for that feature to be derived from a real (non-inpainted) image is estimated. This estimated probability is leveraged to train the model, via the discriminator loss term L_G. Overall, a high-performing generator can "fool" the discriminator into believing it is creating real images.
Abbreviations: GAN, generative adversarial network; MRI, magnetic resonance imaging; MS, multiple sclerosis; WMH, white matter hyperintensity.

of a discriminator loss (which encourages realistic inpainting profiles) and recursive attention mechanism (which facilitates the integration of MRI signals collected across brain locations) differentiates our approach from other state-of-the-art methods for MS lesion inpainting from brain MRI (Manjón et al., 2020).

Furthermore, anatomical consistency was imposed by ensembling models built on four reference anatomical spaces. These spaces were defined by randomly selecting four T1-weighted MRI scans from the baseline studies of the subset of ADVANCE participants exhibiting a low MS lesion volume ($\leq$ 100 m{m^3). Each scan thus selected defines one reference anatomical space, and the process of affine registration to each reference space is illustrated in the appendix in Fig. 16. In total, 12 models were trained, corresponding to one model per anatomical view (axial, coronal, and sagittal) for each anatomical template. During inference, each input MRI scan was first projected to each one of the four reference spaces (see Fig. 16), which was achieved using a linear registration algorithm implemented within the CE-marked and FDA-approved ART-Plan software (TheraPanacea, Paris, France) (Ferrante et al., 2019). The syn-

thetic full-brain MRI images produced in each one of the four reference spaces were subsequently projected back to the source space and predictions were aggregated via voxel-wise averaging in the native space. A sample inpainting result is shown in Fig. 9 as part of the Results section.

5.2.2. Optimization

A subset of the participants, 80%, pooled from the ADVANCE and ASCEND trials, were randomly selected for training the inpainting model; the remaining 20% were used for validation. A set of 80,000 2D slices were extracted from the white matter of all training scans across the axial, sagittal, and coronal planes. Prior to slice extraction, each MRI scan underwent intensity normalization based on histogram matching (Nyú and Udupa, 1999) performed separately on T1- and T2-weighted scans, using intensity statistics from the brain region and excluding lesion areas (and their associated 2 mm peri-plaque margin). Histogram-matched intensities were then mapped linearly to the range (–1 to 1) via min-max normalization.

Let x_i denote a sample in a set of n 2D images extracted from a collection of 3D T1- or T2-weighted brain MRI scans and let $\mathbf{m}_i$ denote a binary mask defining the region to be inpainted in sample x_i. The mask $\mathbf{m}_i$ was randomly generated via the procedure originally proposed in (Yu et al., 2019) and designed to simulate brush strokes, whereby stroke length and brush width parameters were optimized to produce free-form shapes reproducing the visual aspect of MS lesions geometry. Since the task of the inpainting model is to produce a lesion-free tissue profile replacing an MS lesion, the mask $\mathbf{m}_i$ was constrained to white matter regions located at least 5 mm away from any WMH. Then, x_i was preprocessed such that voxels from x_i contained within $\mathbf{m}_i$ were set to a constant value of 0 to avoid contributing to the feedforward pass.

We seek to produce a predicted inpainting $\hat{x}_i$ that minimizes the L1 distance between x_i and $\hat{x}_i$ in both the inpainted region $\mathbf{m}_i$ as well as (trivially) in the non-inpainted region. As x_i is fed into our network, we denote the output of the coarse generator block as $\hat{x}_i^c$ and the output of the refinement generator block as $\hat{x}_i^r$. The reconstruction loss L_M in the masked inpainting region $\mathbf{m}_i$ and the loss L_U in the unmasked region are expressed below, where $\odot$ denotes voxel-wise multiplication. The network was further supervised by minimizing the GAN loss L_G defined below, where D denotes the discriminator loss function. The total model loss is defined as $L_{total} := \alpha L_M + \beta L_U + \gamma L_G$; empirical results motivated the choice of $\alpha = \beta = \gamma = 1$, as proposed in (Yu et al., 2019). Model parameters were trained via the Adam optimizer with a learning rate of 1×10^{-4}, using a batch size of eight. Training was stopped early if the global L1 loss on the validation set did not decrease for two consecutive epochs. The model was implemented in PyTorch and took 14 h to train on a GTX-1080 GPU.

$$L_M := \sum_{\hat{x}_i \in \{\hat{x}_i^c, \hat{x}_i^r\}} \mathbb{E}\left[\left\|\left(x_i - \hat{x}_i\right) \odot \mathbf{m_i}\right\|_1\right]$$

$$L_U := \sum_{\hat{x}_i \in \{\hat{x}_i^c, \hat{x}_i^r\}} \mathbb{E}\left[\left\|\left(x_i - \hat{x}_i\right) \odot \left(1 - \mathbf{m_i}\right)\right\|_1\right]$$

$$L_G := -\mathbb{E}\left\{D\left(\hat{x}_i^r,\ \mathbf{m_i}\right)\right\}$$

5.3. Patch extraction

5.3.1. Motivation for a patch-based approach

In early stages of MS, brain WMH regions can generally be separated into foci via connected-component analysis, whereby one prediction (acute or chronic) can be produced for each discrete lesion. In contrast, in later stages of MS, the confluence of multiple lesions accumulated over time makes it difficult to isolate individual lesions (Dworkin et al., 2018). This problem precludes a lesion-level approach and instead prompts a voxel-level analysis. Thus, in this work, the class of each WMH voxel is predicted using information contained within a cubic *patch* of dimensions 15 × 15 × 15 mm centered on that voxel, whereby patch dimensions were selected by computing the smallest possible patch size satisfying the condition that at least 90% of all acute lesions in our dataset could be fully contained within the patch. Patch-level information is extracted across the original and inpainted T1- and T2- weighted MRIs; each patch is associated with the WMH mask detected in that patch (see Fig. 6).

5.3.2. Patch sampling strategy

Patches were extracted from randomly sampled voxel locations across the WMH regions of all available brain scans, following a spatially uniform distribution (see Fig. 5). This was constrained by a set of patch exclusion criteria eliminating small (< 9 m m^3) lesion foci as well as patches on the boundary between acute and chronic MS lesions. In the specific case of patches extracted from discrete lesion foci that could be fully contained within the patch, the center of the patch was corrected to co-localize with the center of mass of those foci. These exclusion and correction rules are further detailed in the appendix and favor patches that are non-equivocal examples of either acute or chronic MS activity, which reduces label noise and as such facilitates learning. During model validation and testing, these exclusion criteria were maintained. For patches satisfying these criteria, the label of the patch was that of its central voxel.

5.3.3. Class balancing across patches

The resulting dataset of patches was class-balanced by undersampling the majority chronic class. Specifically, the set of chronic patches extracted across the pooled population of participants was such that the count of patches across participants would reflect the extent of chronic lesion burden across those participants. As an example, if the chronic lesion volume detected in participant *A* is equal to twice the volume of chronic lesion detected in participant *B*, then the final set of chronic patches will contain twice as many chronic patches from participant *A* than patches from participant *B*. In addition, the distribution of lesion volumes contained within chronic patches was matched to the distribution of lesion volumes in acute patches. This ensures that the acute versus chronic discrimination process is driven by differences in textural patterns of MRI intensity, as opposed to variations in lesion volume that may otherwise be captured through texture features via volume-confounding effects (Jensen et al., 2021). Key statistics of the resulting class-balanced datasets of patches are presented in Table 2 as part of the Results section.

5.3.4. Regions of interest: core and periphery

In each patch, two regions of interest (ROIs) were segmented: the *core* and *periphery*. Specifically, we define the core of the patch as the section of the WMH contained within the *focus* region, as illustrated in Fig. 6, whereby the *focus* region is defined as a binary ball of radius 4 mm centered on the patch.[2] Consequently, the *core* ROI is equivalent to the lesion mask for patches centered on focal WMH components of maximum 3D diameter ≤ 8 mm (in the ADVANCE cohort, approximately 70% of acute MS lesions satisfy this criterion). For larger WMH components, the *focus* region can be used to focus on the relevant subsection of the larger lesion component. Thus, the *core* ROI can be used to dynamically adapt to different spatial extents of WMH and supports the use of our method on both discrete and confluent MS lesions.

In addition to the *core* ROI, we defined the *periphery* ROI of the patch as the set of voxels around the *core* that are at a distance ≤ 3 mm away from the edge of the *core*. Importantly, the *periphery* ROI does not systematically co-localize with the peri-plaque region. Indeed, the nature of the tissue contained within the periphery ROI may vary depending on the location of the patch within the brain, relative to other MS lesions.

[2] To the best of our knowledge, there currently exists no objective criteria for defining the spatial extent of the core of a MS lesion. The 4 mm criterion used in our work was selected empirically in the context of maximizing the accuracy of our overall framework.

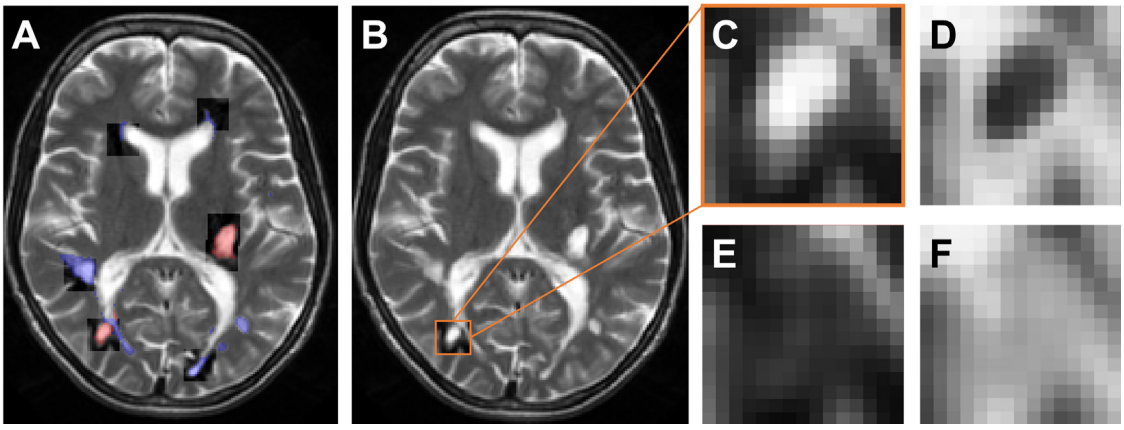

Fig. 5. Overview of the patch sampling and extraction procedure. (A) Axial view of T2-weighted brain MRI showing the acute (red) and chronic (blue) ground truth segmentation maps, along with a selection of patches extracted from the scan. (B, C) Close-up view of one T2-weighted MRI patch. Since the central voxel of this patch is labeled as acute, the patch will be labeled as acute. Note in (A) that a few chronic voxels were also captured within the bounds of the acute patch. (D) T1-weighted MRI patch. (E) Inpainted T2-weighted MRI patch associated with C. (F) Inpainted T1-weighted MRI patch associated with D.
Abbreviation: MRI, magnetic resonance imaging.

Table 2
Number of patches extracted from the ADVANCE, ASCEND, and DECIDE trials, stratified by lesion label. Acute patches are further stratified by lesion age and whether the lesion co-localized with Gd+. The acute and chronic datasets of patches were class-balanced.

Cohort	Acute	Chronic	NET2 <48 weeks	NET2 <24 weeks		NET2 <12 weeks		NET2 <4 weeks		Total NET2 Gd+
			Gd+	Gd+	Gd-	Gd+	Gd-	Gd+	Gd-	
ADVANCE	20,080	20,080	1795	7167	10,987	NA	NA	95	36	9057
ASCEND	5159	5159	157	2246	2073	489	194	NA	NA	2892
DECIDE	10,888	10,888	2936	2785	5167	NA	NA	NA	NA	5721

Notes: An interstudy period of 12 weeks occurred only in the ASCEND trial between 96- and 108-weeks post-baseline, while an interstudy period of 4 weeks occurred only among a subset of participants in the ADVANCE trial, from baseline to 24 weeks post-baseline. It is thus not possible to identify which NET2 lesions from ASCEND and DECIDE emerged within 4 weeks prior to scan acquisition, which is indicated as "NA". A lesion detected as new via the comparison of scans taken more than 24 weeks apart, but showing gadolinium enhancement at the latest timepoint, was considered acute.
Abbreviations: NA, not applicable; NET2, new or substantially enlarging T2-weighted.

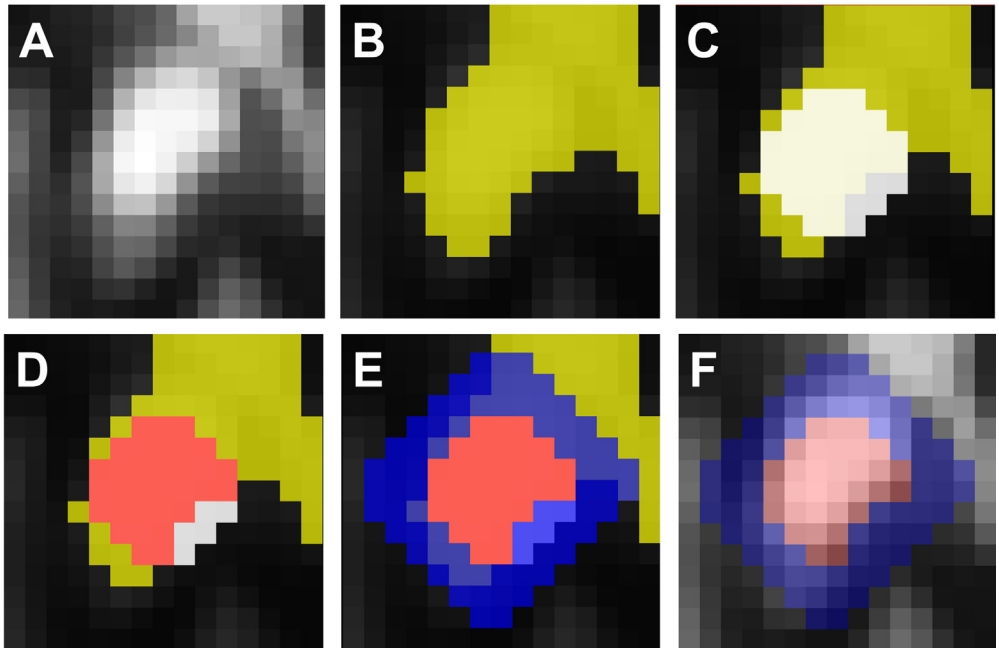

Fig. 6. Definition of the core and periphery of patches. (A) T2-weighted MRI patch. (B) WMH mask (shown in yellow). (C) Focus region (shown in white) superimposed on the WMH mask. The focus region is a binary ball containing the set of voxels located ≤ 4 mm away from the central voxel of the patch. (D) Core region (shown in red) defined as the intersection of the WMH mask and the focus region. (E) Periphery region (shown in blue) defined as the set of voxels located within 3 mm from the edge of the core region, outside of the core region. (F) Core (red) and periphery (blue) regions, within which radiomic features are computed.
Abbreviations: MRI, magnetic resonance imaging; WMH, white matter hyperintensity.

The periphery ROI may thus contain both peri-plaque tissue (which may include white matter tissue and/or CSF from CSF-rich locations such as the ventricles) as well as lesion tissue from neighboring (and possibly confluent) lesions (see Fig. 6). By allowing the periphery ROI to include neighboring lesion tissue (as opposed to restricting it to the non-lesion peri-plaque tissue), we circumvent the issue of an empty periphery, which would otherwise occur when a patch is sampled within the core of a large confluent lesion mass. Overall, the core and periphery ROIs represent two non-overlapping binary masks adapted to the WMH contained in each patch (see Fig. 6).

5.4. Radiomics computation

A set of 88 textural and first order radiomic features (Van Griethuysen et al., 2017) were extracted separately from the *core* and *periphery* ROIs for each MRI sequence (T1- and T2-weighted) as well as for their inpainted counterparts, yielding a set of 704 radiomic features per patch. The intensity discretization for computing texture features was performed using a fixed bin size (as opposed to a fixed bin count), which has been shown to improve reproducibility in TA from MRI (Duron et al., 2019). Bin width was set to 0.4, which yielded approximately 30 intensity bins per ROI and was motivated by prior studies that have shown good reproducibility for a bin count of 32 on MRI (Carré et al., 2020). No patch-level intensity normalization was applied; we instead relied on the scan-level intensity normalization scheme described in Section 5.1 "MRI Pre-Processing".

5.5. Biomarker identification and classification pipeline

The dataset of pooled participants from the ADVANCE trial was split 80:20 into training and validation sets, while ensuring that the distribution of classes in each subset was balanced. Patches extracted from participants from the ASCEND and DECIDE trials constituted two independent testing sets. Radiomic features were normalized via z-scoring, whereby normalization parameters were computed from the training set and subsequently applied to the training, validation, and testing sets. The training set was used as input to a classification pipeline identifying the optimal combination of a feature space and a classification model for the successful discrimination of acute versus chronic patches. This classification pipeline was inspired by (Chassagnon et al., 2020) and consists of an initial feature ranking pipeline, followed by a continuous loop of recursive feature elimination and ensemble classification model optimization (see Fig. 7). The feature ranking, ensemble classification, and modified recursive feature elimination (mod-RFE) modules are defined in the following sections.

5.5.1. Feature ranking

Let $\mathcal{X}_N$ denote the set of $N = 704$ radiomic features extracted from each patch sample. A feature selection algorithm was designed to reduce the dimensionality of this input space by removing irrelevant and/or redundant variables. This has the potential to reduce computation time, improve classification performance and generalizability via the elimination of noisy features, and facilitate the interpretation of the final model (Cai et al., 2018; Kuhn and Johnson, 2013). Specifically, we perform feature selection via two distinct steps. First, a feature ranking module orders the radiomic features from most to least predictive. Then, starting from a feature subspace $\mathcal{X}_K$ comprising the top-K most-relevant features, a modified recursive feature elimination (mod-RFE) process is conducted, whereby the size of the feature subspace is iteratively decremented by eliminating the least useful features at each step.

The *feature ranking* step is supported by a supervised scoring metric $p(x_i)$ denoting the *prevalence* of a feature x_i (indexed by $i \in [1, 2, \ldots, N]$) from $\mathcal{X}_N$ and defined as the sum of two metrics inspired by the *filter* and *embedded* methods of feature selection. The *filter* method selects features based on their score in univariate statistical tests for their correlation with the prediction target. In this context, we may measure the linear dependence between a feature and the patch label (e.g., using point biserial correlation) or their monotonic relationship (e.g., using Kendall's Tau) as well as more complex relationships (via information-theoretic measures such as mutual information). However, this *univariate* filter method is limited in that it may select important but correlated (and therefore redundant) predictors. In contrast, the *embedded* method is a multivariate approach traditionally implemented through the application of classification models endowed with built-in feature selection properties, such as $L1$-regularized linear models. Whilst the *multivariate* embedded method is more robust to the presence of correlation across features, it is also limited to the detection of linear relationships between features and the patch label. It may also select features in a model-dependent fashion and yield feature scores (i.e., model weights) that are dependent on the choice of regularization strategy and magnitude. Importantly, to tackle the task of feature *ranking*, as opposed to feature *selection*, we adapt the traditional *embedded* paradigm by considering the coefficients learned by $L2$-regularized linear models and inspect the relative weight given to each feature, rather than select the subset of features with non-zero weights. Similarly, for the *filter* metric, features were ranked based on their absolute correlation score with the target, rather than selected based on a p-value significance threshold. By combining both a filter and an embedded method to feature selection, we aim to mitigate the respective limitations of each approach and combine their strengths.

To evaluate the *filter* and *embedded* metrics, we randomly generated 100 subsets from the training dataset of patches, whereby each subset contained 80% of all training patches, without repeating elements. To compute the *embedded* metric, we trained a variety of linear and tree-based ML classifiers on each one of the 100 generated subsets to discriminate acute from chronic patches. These classifiers included a linear Support Vector Machine with L2 regularization $C = 0.25$ (chosen via cross-validation over the training set using subject-level splits, using balanced classification accuracy as objective measure), Logistic Regression with L2 regularization $C = 1.0$ (*idem*), as well as tree-based models, including a Decision Tree of depth 3 optimizing the Gini impurity, and boosted ensembles of decision trees, including AdaBoost with 30 boosting rounds and a learning rate of 1.0, and XGBoost with 30 boosting rounds and a learning rate $\eta = 1.0$. For each one of the 100 subsets, model coefficients were extracted from each trained linear classifier and feature importance values were extracted from each tree-based classifier (e.g., Gini impurity for the Decision Tree, or *gain* for XGBoost). These were used to rank all features from highest to lowest relative importance. From this ranking, the 5% features with the highest score were selected. The *embedded* metric score of each selected feature was then incremented by 1. This procedure was repeated for each one of the 100 subsets and is illustrated in Fig. 8. Ultimately, the *embedded* metric score for feature x_i was defined as the number of splits in which x_i was selected, summed across linear and tree-based models. Consequently, the *embedded* metric score is a natural number comprised between 0 (i.e., x_i was never selected) and 500 (i.e., x_i was selected by each of the 5 classification models in each one of the 100 splits).

Similarly, for each one of the 100 subsets, linear and non-linear correlation metrics were evaluated to quantify the degree of association between each radiomic feature and the patch label. Metrics included Point Biserial correlation, Kendall's τ, ANOVA F-value, χ^2, and mutual information. For each metric, features were ranked in ascending correlation magnitude. The *filter* metric score of a feature x_i was then defined as the number of splits in which x_i was selected across correlation metrics. Again, the *filter* metric score is a natural number comprised between 0 (i.e., x_i was never selected) and 500 (i.e., x_i was selected by each of the five correlation metrics in each one of the 100 splits). Importantly, by selecting a fixed number of features for each split and for each scoring method, we allow for conceptually different feature importance metrics to be transformed to a common space of natural numbers in the range [0:500]. The resulting *embedded* and *filter* scores are then each mapped to $[0 : 1]$ via min-max normalization across features, yielding relative

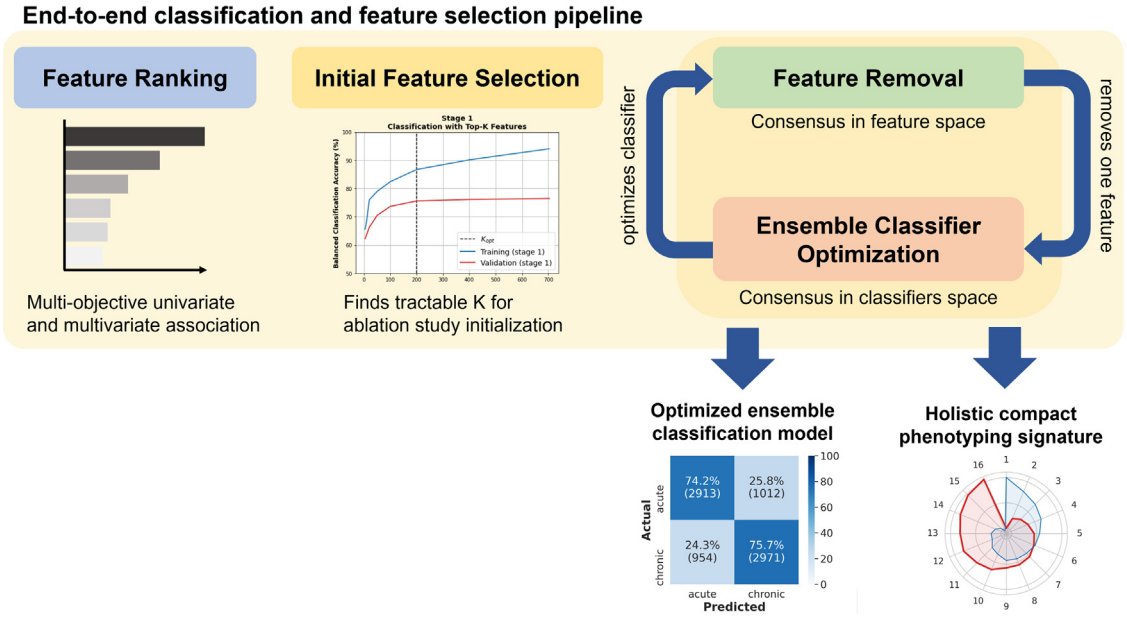

Fig. 7. Diagram of our end-to-end feature selection and classification pipeline, yielding an optimized classification model coupled with a low-dimensional α-radiomics signature.

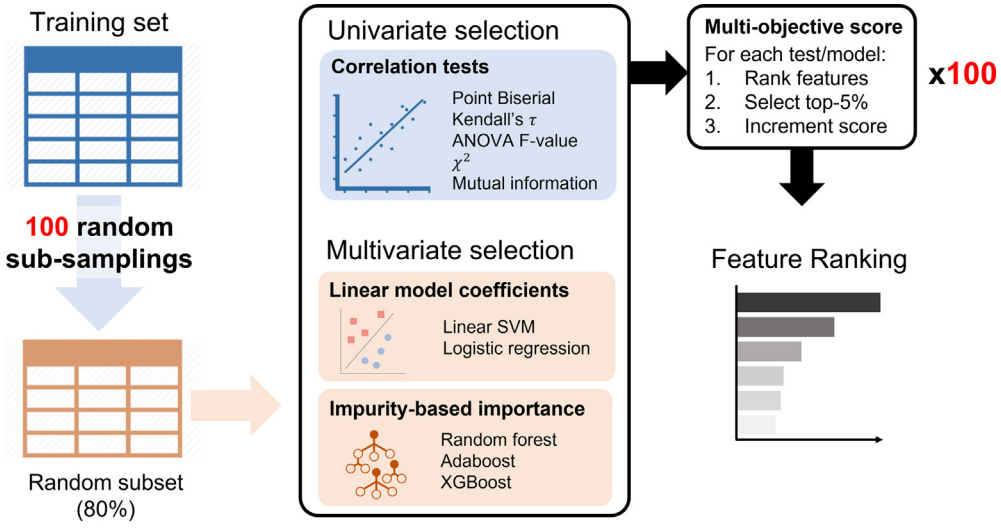

Fig. 8. Diagram showing the our multi-objective feature ranking pipeline, combining the respective strengths of filter and embedded approaches to features selection. Abbreviations: SVM, support vector machine.

metrics which are ultimately summed together to give the prevalence score $p(x_i)$. By summing together min-max normalized metrics (instead of summing up ranks across metrics, for example), the definition of $p(x_i)$ conserves the magnitude of relative discrepancies between features as measured by the *embedded* and *filter* metrics.

Finally, features were ranked using their prevalence score. The top-K most informative features could then be retrieved to construct the set of features $\mathcal{X}_K$. We specifically enforced that, for each radiomic variable retrieved in this way, the measures extracted from both the original and inpainted MRI data should be selected. This was enforced by iterating from the best-ranked variable to the worst-ranked variable and populating the K-dimensional subspace of selected features with both the current-rank variable as well as its original or inpainted counterpart, via a "tag-along" strategy. This ensures that, although original and inpainted MRI features were scored separately using p, each set $\mathcal{X}_K$ consists of pairs of original and inpainted texture radiomics such that $\mathcal{X}_K$ defines an α-radiomics signature.

5.5.2. Classification

The *ensemble classification module* then yields, from a dataset containing a subset $\mathcal{X}_K$ of top-K radiomic features, an ensemble classifier $C_{\mathcal{X}_K}$ discriminating acute from chronic patches. Specifically, $C_{\mathcal{X}_K}$ was constructed by ensembling five base classifiers. The set of candidates from which these five models were selected included K-Nearest Neighbor, Linear Support Vector Machine (SVM), Polynomial SVM, Radial Basis Function SVM, Decision Tree, Random Forest, Adaboost, XGBoost, HistGradBoost, Gaussian Naive Bayes, Quadratic Discriminant Analysis and Multi-Layer Perceptron.

First, the hyperparameters of each model taken from the candidates' pool were tuned via an extensive deterministic grid search, cross-validated three times using subject-level splits. Second, tuned models from the pool were evaluated via 5-fold cross-validations. Models were ranked in ascending order of balanced classification accuracy and the top-5 were selected, excluding models exhibiting a train-validation performance gap exceeding 35% in balanced accuracy. Third, the predic-

tions of the top-5 classifiers were combined under different ensembling strategies including hard voting, accuracy-weighted hard voting (whereby the vote of each base model was weighted by its cross-validated validation accuracy), soft voting, accuracy-weighted soft voting, and stacking. In stacking, a logistic regressor was trained to map the probabilistic output of the top-5 base estimators to the patch class (acute/chronic). The best-performing ensemble model $C_{\mathcal{X}_K}$ was selected via evaluation on the validation set from the ADVANCE trial, using balanced classification accuracy as the objective metric.

5.5.3. Recursive feature elimination

The *ensemble classification module* was applied for different values of K from 4 up to N, which yielded a series of trained classifiers $\{C_{\mathcal{X}_4}, \ldots, C_{\mathcal{X}_N}\}$. Among this set of classifiers and their associated feature spaces, the smallest feature space size K^* associated with classifier $C_{\mathcal{X}_{K^*}}$ yielding a validation accuracy comprised within a reasonable margin of the best accuracy achieved across all trained classifiers was selected as the initial feature space for the mod-RFE pipeline (i.e., the second stage of feature selection). Importantly, the computational complexity of mod-RFE exceeds an arithmetic sum up to the size K of the initial feature space, thus initializing mod-RFE with K^* features, rather than N features, improves the tractability of this last feature selection stage.

Specifically, our *mod-RFE* module integrates the *ensemble classification* module within a continuous optimization loop of greedy RFE. Starting from a feature space of size K^*, the classification module first yields an optimal classifier $C_{\mathcal{X}_{K^*}}$. Then, we iteratively loop over all pairs of radiomic features (x_i^{ori}, x_i^{inp}) in $\mathcal{X}_{K^*}$, where x_i^{ori} and x_i^{inp} denote features corresponding to a radiomic variable x_i measured from original and inpainted MRI data, respectively. In each iteration from this loop, we trained $C_{\mathcal{X}_{K^*}}$ on a feature space of size $K-2$ in which (x_i^{ori}, x_i^{inp}) were removed. The performance of each newly derived classification model trained in the absence of (x_i^{ori}, x_i^{inp}) was evaluated via 5-fold cross-validation. This was repeated for each of the $K^*/2$ pairs of features in $\mathcal{X}_{K^*}$. After $K^*/2$ experiments, the pair of features whose removal caused the least decrease in balanced accuracy was removed, yielding a feature space of size K^*-2, and a new classifier $C_{\mathcal{X}_{K^*-2}}$ was defined. This process was repeated recursively, pruning the feature space by two features in each iteration, terminating at a feature space of size 2.

To limit the computational cost of this approach, a random subset of 20% of the samples from the training set of patches was selected at each recursive loop, from which the impact of the removal of each features pair (x_i^{ori}, x_i^{inp}) was evaluated. Finally, the trace of validation accuracy against feature space sizes was plotted in Stage 2 of Fig. 10, from which an α-radiomics signature was selected via visual inspection, to identify the smallest feature space size maintaining near-optimal classification accuracy on the validation set.

5.6. Code and data availability statement

Requests for data should be submitted via the Biogen Clinical Data Request Portal (www.biogenclinicaldatarequest.com). To gain access, data requestors will need to sign a data-sharing agreement. Data are made available for 1 year on a secure platform. The code for this paper is proprietarily owned by Biogen and cannot be shared.

5.7. Ethics statement

All patients provided written informed consent to participate in the clinical studies, which included consenting to future use of their study data for medical and pharmaceutical research, such as this post-hoc analysis.

Table 3
Recall and precision statistics for the acute and chronic classes of patches for our ensemble classifier trained in the subspace of **32** α-radiomic features.

Dataset	Class	Recall (%)	Precision (%)
ADVANCE (training)	**Acute**	82.9	82.1
	Chronic	82.0	82.7
ADVANCE (validation)	**Acute**	74.7	76.4
	Chronic	77.0	75.2
ASCEND (testing)	**Acute**	71.1	76.4
	Chronic	78.1	73.0
DECIDE (testing)	**Acute**	75.9	73.6
	Chronic	72.7	75.1

6. Results

6.1. MS lesion inpainting

A sample inpainting result is shown in Fig. 9. The performance of our inpainting model is evaluated both quantitatively and qualitatively in the Appendix.

6.2. Patch sampling

Using the sampling strategy described in Section 5.3.2 “Patch Sampling Strategy”, we extracted a set of 40,160 patches (see Table 2) taken from 11,267 timepoints of MRI scan acquisition collected across 3,924 participants with MS.

6.3. Classification performance

The set of 704 radiomic features were ranked and a series of ensemble classifiers were constructed for values of K ranging from 4 up to 704 (see Stage 1 of Fig. 10). The optimal initial K for initializing mod-RFE was determined as $K_{opt} = 200$ (for $K > K_{opt}$, we observed an undesirable increase in approximation error and little to no increase in validation accuracy). The results of the mod-RFE pipeline (see Stage 2 of Fig. 10) led to the selection of a compact α-radiomics signature of 32 features comprising 16 unique variables each extracted from both original and inpainted MRI data. This signature contained 11 T2-based and 5 T1-based variables (nine of which were periphery-based and seven were core-based). The ensemble classifier optimized in this feature space combined a polynomial SVM of degree 3 and regularization parameter $C = 20.0$, a radial basis function SVM with $C = 10.0$, a multi-layer perceptron with one hidden layer comprising 100 units with Rectified Linear Unit (ReLu) activation, trained via Adam optimization with a learning rate of 0.001 and constrained with L2 regularization strength $\alpha = 0.1$, XGBoost optimizing the binary logistic loss via gradient-boosted decision trees of maximum depth 7 with 0.1 learning rate and 600 boosting rounds, and HistGradBoost optimizing the binary logistic loss via gradient-boosted decision trees (unconstrained in depth) with a minimum of 20 samples per leaf, with shrinkage coefficient 0.2 and L2 regularization strength 5.0. Detailed benchmarking results for all classification models evaluated in this paper are reported in Table 8 in the appendix.

The base estimators were combined via weighted soft voting (benchmarking results for other ensembling strategies evaluated in this paper are reported in Table 9 in Appendix). The final ensemble model achieved 75.8% balanced accuracy, 76.4% precision, 74.7% sensitivity, 77.0% specificity, and 83.4% ROC AUC on the validation set. The corresponding testing set metrics in the ASCEND/DECIDE cohorts were 74.6%/74.3%, 76.4%/73.6%, 71.1%/75.9%, 78.1%/72.7%, and 82.2%/82.2%, respectively (see Fig. 11 and Table 3). Lastly, NET2 lesions were further sub-classified

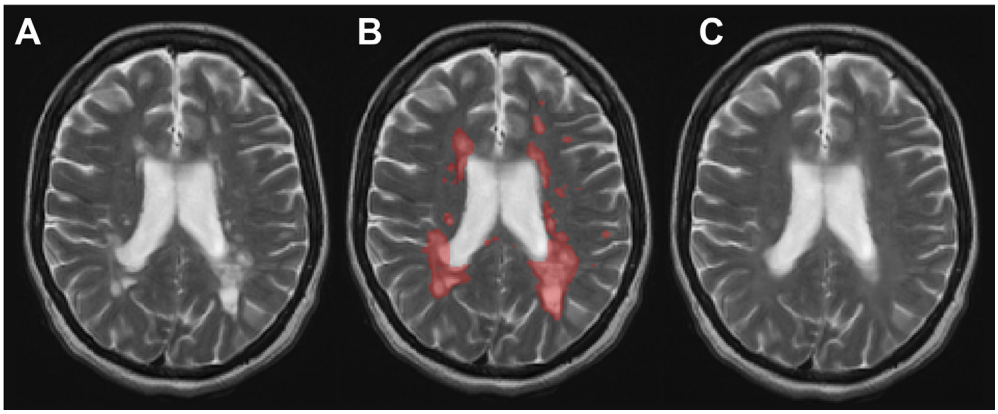

Fig. 9. Sample inpainting result on an axial slice from a T2-weighted brain MRI scan, showing the original lesion-present image (A), the dilated lesion mask (red, B), and the lesion-free inpainted image (C).
Abbreviation: MRI, magnetic resonance imaging.

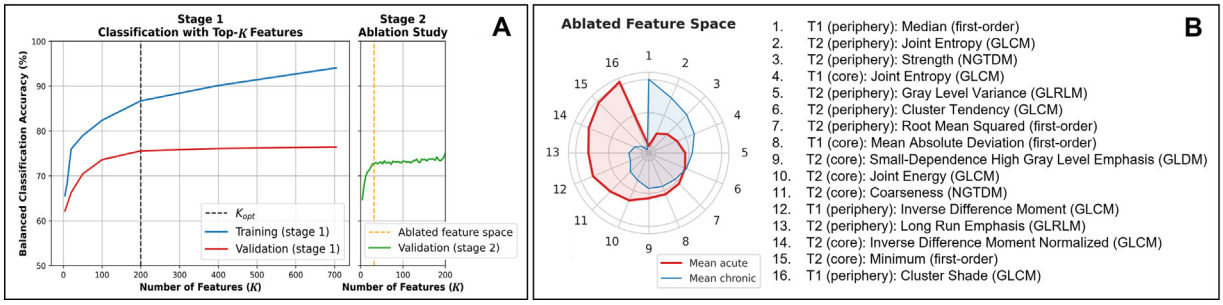

Fig. 10. Panel A: Balanced classification accuracy obtained on the training and validation sets from the ADVANCE cohort for different subsets of **K**-best radiomic features (stage 1), along with the mod-RFE curve initialized with $\mathbf{K_{opt}} = 200$ radiomic features, showing balanced classification accuracy on the validation set (stage 2). Panel B: Radar plot showing the average signature of the acute and chronic populations of patches across each one of the **16** radiomics variables selected via mod-RFE. Each variable was z-scored prior to plotting (radial ticks are spaced by 0.1 standard deviation).
Abbreviations: GLCM, gray level co-occurrence matrix; GLRLM, gray level run length matrix, NGTDM, neighboring gray tone difference matrix.

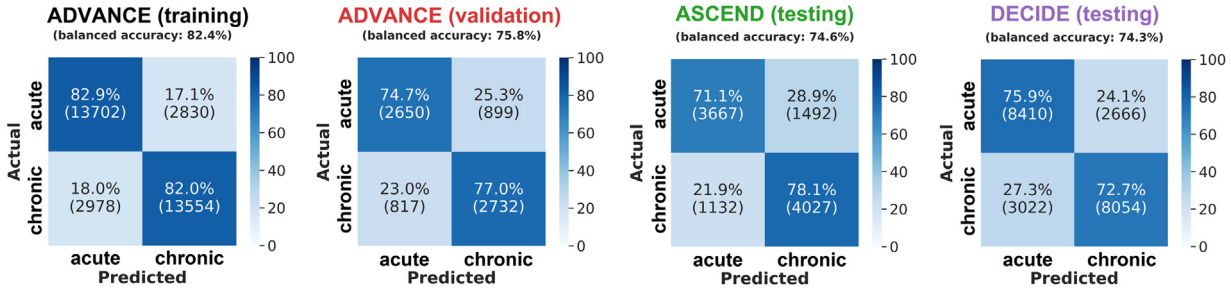

Fig. 11. Confusion matrices showing the performance of our ensemble classifier trained in the subspace of **32** α-radiomic features.

Table 4
Sensitivity (number of patches correctly detected as acute over the total number of patches with ground truth class acute) for gadolinium-enhancing and non-enhancing NET2 lesions on the validation and testing sets.

Cohort		Gadolinium-enhancing	Non-enhancing
ADVANCE	**Validation**	82.8% (1518/1834)	66.6% (1648/2474)
ASCEND	**Testing**	73.2% (2116/2892)	69.0% (1564/2267)
DECIDE	**Testing**	81.8% (4680/5721)	70.0% (3617/5167)

Abbreviation: NET2, new or substantially enlarging T2-weighted.

as gadolinium-enhancing if they overlapped by at least one voxel with a Gd+ region, or non-enhancing otherwise. The sensitivity of our classification model to each sub-population of acute patches across cohorts is reported in Table 4.

6.4. Control experiments

6.4.1. Prediction from lesion location or volume

In addition, a set of control experiments were designed to generate reference results contextualizing the performance of our classifier. To assess the effects of any potential spatial bias between acute and chronic patch samples (*experiment 1*), we applied our classification pipeline to predict the ground truth class of each patch given the 3D location of its central voxel. We achieved a balanced accuracy of 62.5% on the validation set from ADVANCE (compared to 75.8% with our α-radiomics-based approach), 63.4% on the testing set from ASCEND (compared to 74.6%), and 61.3% on the testing set from DECIDE (compared to 74.3%). Furthermore, to validate the efficacy of the volume-matching step performed between acute and chronic patches (*experiment 2*), we attempted to classify patches using the volume of their core and periphery regions. As expected, classification accuracy was close to chance (50%) across the ADVANCE (56.7%), ASCEND (52.2%), and DECIDE (53.9%) trials, thereby demonstrating appropriate elimination of the ROI volume bias.

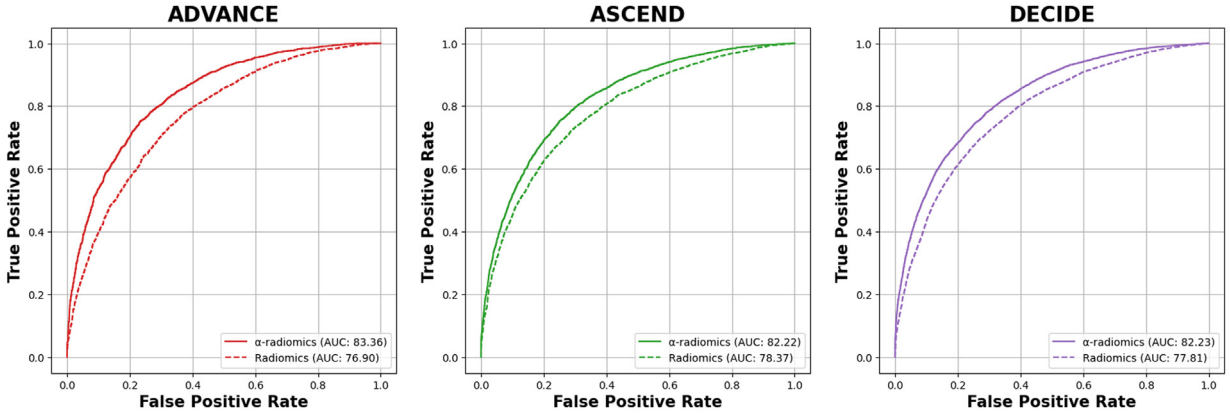

Fig. 12. Receiver-Operating Curves (ROC) comparing our proposed pipeline (leveraging α-radiomics extracted from original and inpainted MRI images) versus a pipeline leveraging radiomic features extracted from the original MRI images only.

6.4.2. *Prediction without inpainting features*

Furthermore, to assess the benefit of including inpainting features (*experiment 3*), our classification pipeline was applied to the restricted subset of radiomic features extracted from the original MRI scans. To ensure a fair comparison, a feature space of 32 radiomic features was selected via mod-RFE. The resulting model achieved a balanced accuracy of 70.1% on the validation set from ADVANCE (compared to 75.8% with inpainted information), 71.4% on ASCEND (compared to 74.6%), and 71.2% on DECIDE (compared to 74.3%). The radiomics model without inpainted features is additionally compared against our proposed α-radiomics pipeline via receiver operating curves, as shown in Fig. 12.

6.4.3. *Prediction using a simple convolutional neural network*

Lastly, we compared our α-radiomics ML classifier against a simple convolutional neural network (CNN) trained to discriminate acute from chronic patches (*experiment 4*) using the same input data as our proposed model (T1-weighted, T2-weighted, and inpainted T1- and T2-weighted patches concatenated together with the WMH mask patch). The CNN architecture consisted of two convolutional blocks followed by two fully-connected layers. The convolutional blocks comprised eight and 16 filter kernels of dimensions $3 \times 3 \times 3$ with stride 1 and 1-voxel padding, each followed by a leaky ReLu activation function and a max-pooling block of kernel size 2. The fully-connected layers contained 32 hidden units and one output unit respectively, and were connected via a leaky ReLu activation function, followed by a batch-normalization layer and a drop-out layer with $p = 0.4$. The model thus contained a total of 37,457 trainable parameters. A sigmoid activation function was applied to the output node, and the model optimized the binary cross-entropy loss using the Adam optimizer with a learning rate of 0.001. The resulting CNN was cross-validated by generating 20 train/validation splits from the original training set on the basis of 80:20. It achieved a balanced accuracy of 74.3% ($\pm$0.7) on the validation set from ADVANCE (compared to 75.8% with our α-radiomics approach), 74.9% ($\pm$0.7) on ASCEND (compared to 74.6%), and 74.3% ($\pm$0.8) on DECIDE (compared to 74.3%). Overall, our proposed α-radiomics pipeline performed similarly to this simple CNN, as demonstrated via the receiver operating curves shown in Fig. 13. The results of each control experiment are summarized together in Table 5.

6.5. *Full-Brain prediction*

Our patch-based classification framework may be adapted to the full-brain dense acute MS lesion segmentation task by independently predicting the label of each WMH voxel (*acute* or *chronic*) using information contained in its surrounding patch. Sample predictions showing true positive, false negative and false positive examples are shown in Fig. 14. In addition, the performance of our classification framework in the context of brain scans containing no ground truth acute MS lesion activity is illustrated in Fig. 15, showing varying degrees of acute "over-prediction". Future histopathology-MRI correlation analysis will be needed to determine whether some of these areas of apparent "over-prediction" may nevertheless contain foci of pathologically "active" demyelination which are currently missed by the ground truth conventional MRI definition of "acute" lesions (the limitations of which were summarized in the Introduction section).

7. Discussion

We have developed an ensemble ML algorithm discriminating acute from chronic MS lesion patches with accuracy in the range 74.3–74.6% using a compact set of 32 α-radiomic features. This algorithm demonstrated good generalization properties across different MS disease stages (relapsing-remitting MS [RRMS] in ADVANCE/DECIDE versus secondary progressive MS [SPMS] in ASCEND).

7.1. *Interpretation of radiomic features*

The potential interpretation of the 32 selected radiomic features that distinguished acute from chronic lesions was explored. In line with prior work (Zhang et al., 2008, 2009), our results suggest that acute MS lesions are associated with a coarser texture on T2-weighted MRI relative to chronic MS lesions.[3] Among the selected radiomic features, coarseness is reflected by variables such as the *coarseness* (as extracted from the Neighboring Gray Tone Difference Matrix, NGTDM) as well as the *normalized inverse difference moment* (as extracted from the Gray Level Co-occurrence Matrix, GLCM) measured in the core of T2-weighted patches. Previous studies have associated increased coarseness in MS lesions on T2-weighted MRI with higher levels of demyelination, axonal damage, and inflammation consistent with acute pathology, which suggests that our method has content validity (Zhang et al., 2013). Furthermore, it has been suggested that the increased coarseness in T2-weighted MRI texture in acute MS foci may be associated with infiltration by inflammatory cells (including macrophages, lymphocytes, and glial cells), loss of oligodendrocytes, recruitment of undifferentiated oligodendrocyte progenitors, phagocytosis of myelin proteins by macrophages (Brück et al., 1995; Pittock and Lucchinetti, 2007; Prineas et al., 1993; Zhang et al., 2009), and acute axonal pathology (Tedeschi et al., 2002).

[3] In this context, a *coarse* texture is defined as consisting of thick threads or large pieces: it is locally homogeneous and exhibits regularity over a large spatial scale. In prior studies (Zhang et al., 2008, 2009), coarseness has traditionally been measured via low-frequency spectrum energy, as measured via the Polar Stockwell Transform.

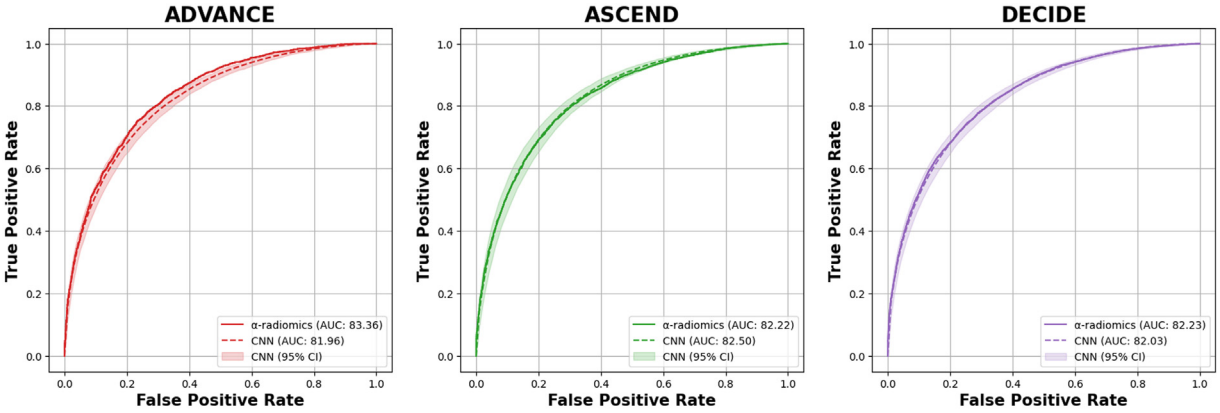

Fig. 13. Receiver-Operating Curves comparing our proposed approach (leveraging α-radiomics via an ensemble of ML classifiers) versus a simple CNN, showing no statistically significant difference in performance. The 95% confidence interval (CI, shaded) was computed by fitting 20 different models on 20 different train/validation splits from the training set of patches, and applying each resulting trained model to each test set.

Table 5
Key global classification performance statistics across the ADVANCE, ASCEND, and DECIDE trials, for different input feature spaces in the context of 4 control experiments.

Experiment #	Cohort	Balanced accuracy (%)	ROC AUC (%)
1: Location	ADVANCE (training)	67.3	73.9
	ADVANCE (validation)	62.5	66.2
	ASCEND (testing)	63.4	67.3
	DECIDE (testing)	61.3	65.4
2: Volume	ADVANCE (training)	58.5	62.9
	ADVANCE (validation)	56.7	60.0
	ASCEND (testing)	52.2	55.4
	DECIDE (training)	53.9	56.4
3: Original MRI only	ADVANCE (training)	75.9	84.9
	ADVANCE (validation)	70.1	76.9
	ASCEND (testing)	71.4	78.4
	DECIDE (training)	71.2	77.8
4: CNN	ADVANCE (training)	83.8 (±1.5)	91.2 (±1.3)
	ADVANCE (validation)	74.3 (±0.7)	82.0 (±0.9)
	ASCEND (testing)	74.9 (±0.7)	82.5 (±0.8)
	DECIDE (training)	74.3 (±0.8)	82.0 (±0.8)
Ours: α-radiomics	ADVANCE (training)	82.4	91.4
	ADVANCE (validation)	75.8	83.4
	ASCEND (testing)	74.6	82.2
	DECIDE (training)	74.3	82.2

In addition, many of the features discriminating acute from chronic lesions were found to be in the periphery of the patch, which is consistent with prior results demonstrating the presence of rich textural biomarkers within the peri-plaque tissue space in MS (Zhang et al., 2013). We observed a greater degree of T1-weighted signal hypointensity in the periphery of acute patches, along with a greater degree of T2-weighted signal inhomogeneity in this region, as measured via the *joint entropy* (GLCM) and *gray level variance* (Gray Level Run-Length Matrix) variables. This may be consistent with clearance of debris at the lesion site, which occurs predominantly at the periphery of an MS lesion (Zhou et al., 2019), as well as partial demyelination in the tissue surrounding acute foci. Nonetheless, it is important to recognize that the periphery region as defined in this study may not systematically overlap with the NAWM surrounding a focal MS lesion. Instead, it may intersect with varying degrees of WMH, depending on the proximity of that MS foci to other regions associated with MS damage. In particular, the periphery of patches extracted from confluent lesions will primarily contain neighboring lesion tissue and in contrast will contain little to no peri-plaque tissue. Consequently, we cannot exclude that the increased T2-weighted inhomogeneity and decreased median T1-weighted signal intensity observed in the periphery of acute patches could also reflect the presence of WMHs in the vicinity of acute foci.

7.2. Sensitivity to gadolinium-enhancing lesions

The sensitivity of our ensemble classifier to the detection of acute lesions at different ages was investigated by using the presence of Gd+ as a marker of lesion age. Since the timescale for persistence of BBB disruption, as measured via Gd+, has been reported to be around 1.5 to 3 weeks (Cotton et al., 2003; Guttmann et al., 2016), then Gd+ lesions can be expected to have emerged on average within 1.5 to 3 weeks prior to scan acquisition. Across all cohorts, we achieved a significantly higher (+10% improvement on average) sensitivity to Gd+ than non-enhancing NET2 lesions. Intuitively, since NET2 lesions are most distinguishable from chronic lesions shortly after their inception, before they may slowly transition towards a chronic lesion profile over time (Guttmann et al., 1995; Meier et al., 2007; Meier and Guttmann, 2003; Rovira et al., 2013), we may indeed expect that more recent NET2 may harbor maximal textural discrimination versus chronic lesions. Nevertheless, our classifier was still able to accurately discriminate acute from chronic lesions when the former was not Gd+, emphasizing that the distinguishing features selected by our algorithm are not restricted to the early-acute stage of new lesion formation.

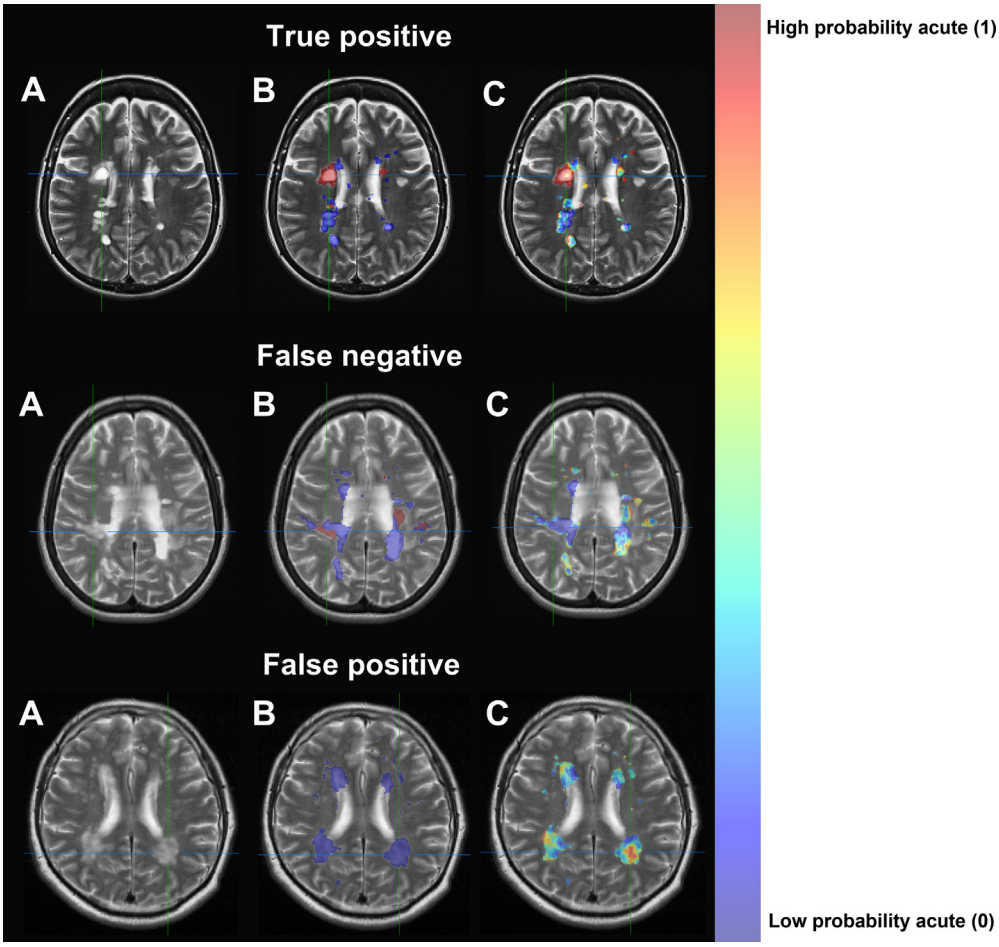

Fig. 14. Panel showing axial slices sampled from three full-brain prediction maps derived from three participants from the ASCEND test cohort, obtained by independently predicting the label of each WMH voxel using information contained in its surrounding patch. This panel illustrates, from top to bottom, an example of true positive acute MS lesion detection, an example of false negative (acute lesion incorrectly predicted as chronic) and an example of false positive (chronic lesion incorrectly predicted as acute). In each row we display, from left to right, (A) axial slice of T2-weighted brain MRI scan, (B) ground truth map of acute (red) and chronic (blue) MS lesions, and (C) probabilistic prediction map generated via the application of our patch-based classifier, where blue indicates a low predicted probability of acute MS lesion and red indicates a high predicted probability of acute MS lesion.

7.3. Analysis of control experiments

7.3.1. Location-driven classification

In *experiment 1*, the significant improvement in performance (+12%) observed in the α-radiomics classification relative to the location-driven control demonstrates the ability of our α-radiomics model to detect local textural biomarkers specifically associated with acute versus chronic MS lesion activity, beyond location-dependent textural clues revealing the anatomical context of each patch.

7.3.2. Benefit of inpainting features

In *experiment 3*, we observed a consistent 4% increase in classification accuracy under the inclusion of inpainted information, relative to using original MRI data only. Importantly, it should be observed that, since the inpainting model was trained from non-lesion white matter tissue from MS-diagnosed patients, and since the totality of the white matter in the brain of MS patients is abnormal relative to healthy subjects (Elliott et al., 2021), we expect our inpainting model to reproduce MS-associated white matter abnormalities, rather than generate a truly "healthy" white matter tissue profile. We hypothesize that the observed performance gain may result from the improved detectability of anatomical landmarks within the inpainted image, relative to the original MRI scans. In particular, the presence of a hyperintense T2-weighted signal in the periphery of a patch taken from the original MRI scan could denote either proximity to a WMH, or proximity to a region rich in cerebrospinal fluid (CSF), such as the ventricles. In contrast, after removing WMHs via inpainting, the presence of a hyperintense T2-weighted signal in the periphery of a patch acts as a specific marker of proximity to CSF-rich brain regions. This may allow for a better spatial contextualization of each patch, potentially supporting the construction of a location-dependent decision rule. Additionally, it should be observed that lesion inpainting transfers information from various brain locations into the inpainted region (via attention mechanisms) and as such expands the receptive field offered to our classifier beyond the core and periphery ROIs.

7.3.3. Comparison against CNN

Lastly, in *experiment 4* we observe that our α-radiomics classifier outperforms (on DECIDE) or matches (on ASCEND) the accuracy reached by a *simplistic* CNN classifier. With regards to the unsophisticated nature of the CNN evaluated in this study, these results are promising for future applications of deep learning techniques to lesion classification in MS. In fact, the CNN-based approach is particularly advantageous with respect to timing consideration, as it accelerates inference by a factor of 100 relative to our α-radiomics approach (approximately 300 patches/second on GTX-1080 GPU for the CNN, relative to 3 patches/second on AMD EPYC 7742 64-Core Processor CPU for our approach).

Nonetheless, it should be noted that in contrast with the "black-box" characteristics of a CNN-based approach, our radiomics-based method retains a higher degree of interpretability, owing to our ability to associate specific radiomic features with their biological correlates. These associations may be leveraged to independently justify acute MS lesion predictions. As an example, some MS lesions may be predicted as acute owing to their abnormally high T2-weighted signal in the periphery ROI, suggestive of peripheral edema, while others may similarly be predicted as acute because of their high T2-weighted coarseness in the core ROI, suggestive of infiltration of inflammatory cells.

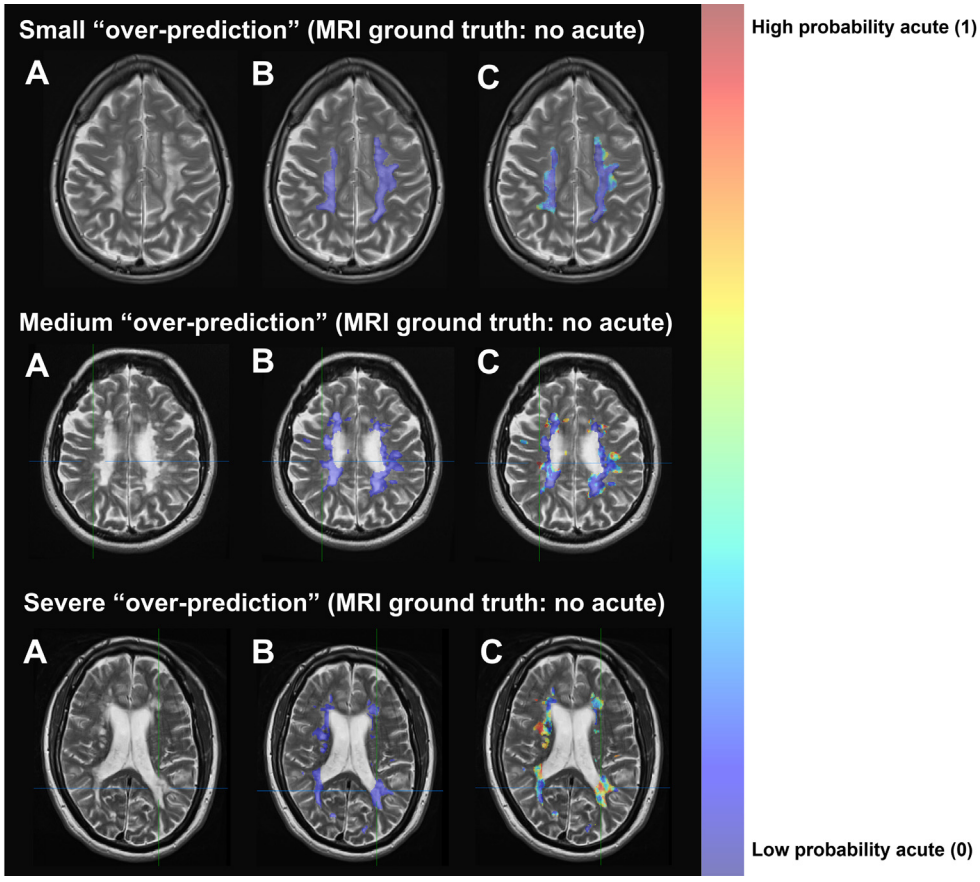

Fig. 15. Panel showing axial slices sampled from three full-brain prediction maps derived from three participants from the ASCEND test cohort for which no ground truth acute MS lesion was detected, obtained by independently predicting the label of each WMH voxel using information contained in its surrounding patch. This panel illustrates, from top to bottom, examples of small, medium and severe acute over-prediction. In each row we display, from left to right, (A) axial slice of T2-weighted brain MRI scan, (B) ground truth map of acute (red) and chronic (blue) MS lesions, and (C) probabilistic prediction map generated via the application of our patch-based classifier, where blue indicates a low predicted probability of acute MS lesion and red indicates a high predicted probability of acute MS lesion.

7.4. Limitations

7.4.1. Population bias: clinical trial participants

This study has several limitations. Importantly, models were trained and evaluated on a population of participants pooled from the placebo and treatment groups of randomized controlled clinical trials, and while treatment effects may alter the textural properties of acute and chronic MS lesions relative to a natural history population, this was not explicitly investigated. Furthermore, although the inclusion criteria for these trials spanned the MS spectrum (see Table 1), which allows our algorithm to model population variability, some differences are nonetheless likely to exist between the population considered in this study and the distribution of patients typically encountered in routine clinical settings.

7.4.2. Under-Representation of variability in MRI acquisition parameters

Furthermore, the acquisition parameters used to generate the T1- and T2-weighted MRI brain scans leveraged in this study were constrained (as verified via *dummy runs*) to similar ranges across all trials (see Table 1), yielding highly standardized images that may underrepresent the variability of image contrasts and noise conditions encountered across real-world practice. Notwithstanding, it should be noted that the compactness of the selected feature space, the inherent robustness of radiomic features to MRI acquisition parameters, and the use of an ensemble learning strategy endow our classification framework with robustness properties.

7.4.3. Limitations associated with a patch-based approach

Another important limitation of this work is its focus on a patch-level classification task guided by a handcrafted patch sampling strategy, which differs from the brain-level task of acute MS lesion detection and/or segmentation. While our approach may be adapted to the dense segmentation setting as discussed in Section 6.5 "Full-Brain Prediction", it nonetheless lacks a mechanism to ensure consistency of predictions across voxels within a discrete MS lesion component, or to enforce spatial smoothness of predictions across a confluent MS lesion mass (see Fig. 14 and Fig. 15).

Furthermore, we have chosen to facilitate model training by rejecting ambiguous patch samples (see Section 5.3.2 "Patch Sampling Strategy") and artificially class-balancing our dataset (see Section 5.3.3 "Class Balancing across Patches"). This is not representative of the brain-level task, where substantial class imbalance exists (see Table 1), and ambiguous patches cannot be avoided. To deal with class imbalance, our framework should be adapted to include classification models designed to perform well under these conditions. Additionally, we may shift our optimization objective focus from balanced accuracy (preferred in this work for its interpretability) to the F1-score metric (better suited for imbalanced datasets where positive class occurrences are rare). We hypothesize that this may contribute to reducing the occurrence of false positives, which are illustrated in Figs. 14 and 15.

7.4.4. Statistical testing

Due to the high computational cost of training our proposed pipeline, the results reported here were collected across a unique run generated from a single subject-level train/validation/test split. The lack of global cross-validation limits our ability to statistically compare different classification experiments, beyond the ROC curves shown in Figs. 12 and 13.

7.5. Translation into clinical practice

This work tackles the task of acute versus chronic MS lesion classification in the cross-sectional setting and without contrast-enhanced T1-weighted MRI. As such, it is suited for estimating the volume and spatial distribution of acute lesion burden in MS patients for whom only

one timepoint of MRI scan acquisition is available and/or for whom safety concerns related to nephrotoxicity of gadolinium injection may be relevant. In the context of a diagnostic MRI scan acquisition, our proposed method may augment the information available to the clinician, by increasing the sensitivity to acute lesion detection beyond the delineation of Gd+ foci, and as such may support the characterization of dissemination of MS lesions in both time and space on a single scan (Thompson et al., 2018). Furthermore, the detection of a high acute lesion burden from a diagnostic scan may provide motivation for selecting a treat-to-target treatment strategy, involving initiation of potent DMTs early in disease course, which may ultimately improve patient outcomes.

In the context of a clinical trial, our method may also support the implementation of inclusion/exclusion criteria based on more comprehensive measures of acute lesion burden estimated from a single screening MRI session. Lastly, our proposed algorithm may be able to detect foci of acute MS lesion activity occurring in areas of pre-existing T2 lesion and captured outside of their gadolinium enhancement phase, and as such may improve the sensitivity to acute MS lesion detection beyond current methods. All these potential claims would be subject to full clinical validation either in retrospective or prospective cohorts.

7.6. Future work

Future work should primarily focus on evolving the voxel-level MS lesion classifier to a brain-level segmentation that would augment current methods and potentially eliminate the need for gadolinium injection. This may involve integrating the set of independent voxel predictions produced by our model into a spatially smooth and consistent voxel-level prediction map, constrained within the bounds of the WMH mask. Smoothness may be enforced by post-processing the raw voxel-level prediction map generated by our current framework, via Gaussian smoothing and/or techniques derived from Markov Random Fields.

In the context of future deep learning efforts, further work may evaluate the association between textural radiomic features and latent patch embeddings produced via the feature representation mechanism inherent to CNNs. Furthermore, attention-based frameworks, or visual explanation tools such as Grad-CAM (Selvaraju et al., 2017), may be used to investigate the potential relationship between those input regions to which the CNN prediction is sensitive, and the “core” or “periphery” ROIs as defined in this work. We may also investigate the possibility of constructing a CNN producing a dense segmentation map from a multi-channel input consisting of T1- and T2-weighted MRI scans along with voxel-level feature maps computed for each one of the 32 selected radiomic features. This approach offers several benefits, such as decreased inference time, and would naturally enforce spatial consistency by producing voxel-level predictions in a globally dependent fashion. Such a model may also be informed with patient-level demographic and/or clinical disease variables.

Lastly, and most importantly, future efforts should focus on correlating predicted acute lesion burden with subsequent clinical disease outcomes. Indeed, a thorough clinical validation supported by patient-level metrics derived from a voxel-level map of predicted acute MS lesion activity may support the integration of an automatic acute MS lesion detection tool into clinical practice, by allowing inference to be drawn at the level of individual patients.

Conclusions

We have developed a ML-based ensemble classifier that can discriminate acute from chronic MS lesions using unenhanced cross-sectional T1-weighted and T2-weighted scans without the use of a previous comparative reference scan and/or gadolinium. The model leveraged a compact set of 32 α-radiomic features encoding textural patterns associated with acute versus chronic MS lesion activity. The model achieved 75.8% balanced accuracy on a validation set of RRMS subjects, which was maintained on independent test datasets comprising data from both SPMS (74.6% accuracy) and RRMS (74.3% accuracy) populations.

CRediT author statement

Bastien Caba: Conceptualization, Methodology, Software, Writing, Visualization. **Alexandre Cafaro**: Conceptualization, Methodology, Software, Writing, Visualization. **Aurélien Lombard**: Conceptualization, Methodology, Writing, Supervision. **Douglas L. Arnold**: Conceptualization, Methodology, Writing, Supervision. **Colm Elliott**: Conceptualization, Methodology, Writing, Supervision. **Dawei Liu:** Conceptualization, Methodology, Writing, Supervision. **Xiaotong Jiang:** Conceptualization, Methodology, Writing, Supervision. **Arie Gafson:** Conceptualization, Methodology, Writing, Supervision. **Elizabeth Fisher:** Conceptualization, Methodology, Writing, Supervision. **Shibeshih Mitiku Belachew:** Conceptualization, Methodology, Writing, Supervision. **Nikos Paragios:** Conceptualization, Methodology, Writing, Supervision.

Code and data availability statement

Requests for data should be submitted via the Biogen Clinical Data Request Portal (www.biogenclinicaldatarequest.com). To gain access, data requestors will need to sign a data-sharing agreement. Data are made available for 1 year on a secure platform. The code for this paper is proprietarily owned by Biogen and cannot be shared.

Funding

This work was supported by Biogen.

Disclosures of Competing Interest

Caba, Liu, Jiang, Gafson, Fisher and Belachew are employees and shareholders of Biogen. Cafaro and Lombard are employees and shareholders of Therapanacea. Elliott is an employee of NeuroRx Research. Arnold receives consulting fees from Biogen, Celgene, Frequency Therapeutics, Genentech, Merck, Novartis, Race to Erase MS, Roche, and Sanofi-Aventis, Shionogi, Xfacto Communications, grants from Immunotec and Novartis, and an equity interest in NeuroRx Research. Paragios is an employee of Therapanacea, employee of CentraleSupélec, Université Paris-Saclay, French Ministry of Higher Education and Research; holds stock options in Arterdrone and TheraPanacea; and receives compensation for editorial services from Elsevier.

Credit authorship contribution statement

Bastien Caba: Conceptualization, Methodology, Software, Visualization. **Alexandre Cafaro:** Conceptualization, Methodology, Software, Visualization. **Aurélien Lombard:** Conceptualization, Methodology, Supervision. **Douglas L. Arnold:** Conceptualization, Methodology, Supervision. **Colm Elliott:** Conceptualization, Methodology, Supervision. **Dawei Liu:** Conceptualization, Methodology, Supervision. **Xiaotong Jiang:** Conceptualization, Methodology, Supervision. **Arie Gafson:** Conceptualization, Methodology, Supervision. **Elizabeth Fisher:** Conceptualization, Methodology, Supervision. **Shibeshih Mitiku Belachew:** Conceptualization, Methodology, Supervision. **Nikos Paragios:** Conceptualization, Methodology, Supervision.

Data Availability

Data will be made available on request.

Acknowledgments

The authors thank all patients, their families, and the investigators who participated in the ADVANCE, ASCEND and DECIDE trials; and

NeuroRx Research (Montreal, QC, Canada) for the evaluation of MRI scans. Excel Medical Affairs provided editorial assistance in copyediting and styling the manuscript per journal requirements.

Appendix

A1. Exclusion criteria for patch sampling

A1.1. Definitions

Patches containing WMHs that did not exceed a volumetric threshold of 9 mm^3 (equivalent to 3 voxels in the native $1 \times 1 \times 3$ mm spacing) were ignored. This heuristic was enforced to facilitate TA, as radiomics analysis is sensitive to outlier effects for small ROI sizes (Jensen et al., 2021). Intuitively, textures characterized by patterns showing regularity over a distant spatial scale are also more difficult to detect in small ROIs.

Furthermore, let $\mathbf{r}$ denote the voxel location defining the center of a patch and let $y(\mathbf{r})$ be the label of voxel $\mathbf{r}$ (acute or chronic). Then, the volumetric proportion of the section of WMH contained in the focus region of the patch that is of class $y(\mathbf{r})$ is enforced to always exceed 80%. This heuristic guarantees that our dataset of patches does not contain ambiguous samples situated at the interface between acute and chronic tissue. On average, this exclusion criteria rejected 4.5% (±6.3%) of the WMH voxels detected in ADVANCE scans, 2.6% (±3.3%) in ASCEND, and 4.7% (±6.4%) in DECIDE.

Lastly, when a sampled voxel location was contained within a small discrete lesion, then that sampled location was corrected such as to co-localize with the center of mass of the foci. Specifically, a lesion was considered as “small" if it could be fully contained within the cubic patch of dimensions $15 \times 15 \times 15$ mm. This encourages the selection of patches centered on lesion foci, rather than on their border. In contrast, for WMH components exceeding the patch size, any voxel from the core or border of the lesion could be uniformly sampled. As previously specified, these exclusion criteria constrained the generation of both the training set of patches, as well as the validation/testing sets.

A1.2. Implications

On the full-brain voxel-level inference task in which our classification model could be applied to every voxel contained within the WMH area, we expect our classifier to reproduce the classification performance reported in this paper only within the set of WMH voxels that satisfy the above-mentioned rules. In contrast, we make no claim regarding the performance of our classification model within areas consisting of voxel locations that would fail to meet our patch inclusion criteria.

A2. Evaluation of MS lesion inpainting

A2.1. Registration to multiple atlases

Fig. 16

A2.2. Qualitative Evaluation

Our approach was visually evaluated against prior methods for MS lesion inpainting such as FSL (Oxford center for Functional MRI of the Brain [FMRIB]’s Software Library) (Battaglini et al., 2012) and SFL (Automatic Lesion Segmentation of Multiple Sclerosis [SALEM] Lesion Filling) (Valverde et al., 2014), as shown in Fig. 17. It matches or exceeds the state-of-the-art while offering improvesuppd versatility (unlike (Zhang et al., 2020), our model does not require the availability of healthy brain scans): in particular, unlike FSL and SFL our method is free from artifacts and identifiable margins around inpainted regions. Although our model does not explicitly enforce an edge prior guiding the reconstruction of anatomical structures such as the ventricles, unlike the most recently-published deep learning method based on edge prior constraints (Zhang et al., 2020); qualitative results demonstrate that we achieve continuity in the reconstruction of the periventricular space. Furthermore, a V-net segmentation model (Milletari et al., 2016) trained to detect MS lesions was unable to detect any lesion after application of the inpainting algorithm.

Table 6
Comparison of PSNR and edge preservation for different inpainting methods. In each cell we report the mean ± variance across 30 randomly selected test cases.

Metric	SFL	FSL	Our Approach
PSNR (dB)	69.5 ± 7.2	71.8 ± 6.4	76.2 ± 4.3
Edge Detection F1	0.85 ± 0.05	0.89 ± 0.04	0.94 ± 0.04

Abbreviations: PSNR, peak signal-to-noise ratio.

The effect of extending the inpainting mask beyond the boundaries of the WMH mask (Francis, 2010) was investigated. For a 1-mm margin, peri-lesional abnormalities remained visible outside of the inpainting mask, resulting in diffuse hyperintensity occasionally "bleeding into" the inpainted region. Conversely, extending the margin to 3 mm results in a large volumetric extent of tissue to be inpainted, which reduces the quality of inpainting in the core of the inpainting mask owing to the increasing distance to the non-masked region where valid information is present. The effect of different margin thicknesses is illustrated in Fig. 18. As previously reported, in the final inpainting model a margin of 1 mm was preferred, based on extensive observations suggesting that a 1-mm margin yields better inpainting results.

A2.3. Quantitative Results

The evaluation procedure for inpainting evaluation established in (Zhang et al., 2020) was replicated, using peak signal-to-noise ratio (PSNR) as an evaluation metric. The quality of continuity of anatomical structures was quantified by applying an edge detection algorithm to both the original and lesion-free images, and subsequently comparing the resulting edge maps. A Canny Edge filter with $\gamma = 0.8$ (as in Zhang et al., 2020) was applied to a 5-mm periphery around each inpainted region, both within the original and inpainted images. The F1 score between the resulting binary edge maps is reported in Table 6.

Trivially, due to the presence of several MS lesions in the baseline MRI, pre-lesion ground truth is typically not available, which complicates not only model training but also model evaluation. To tackle this issue, we sampled a lesion mask $\mathbf{m}_i$ from a given subject s_i and superimposed $\mathbf{m}_i$ on the brain scan of a different subject s_j selected such that $\mathbf{m}_i$ did not co-localize with any lesion in s_j. MSE, PSNR, and edge detection analyses were conducted by comparing the inpainted region contained within $\mathbf{m}_i$ in s_j, across 30 subjects pooled from our test set, sampled equally between the ADVANCE and ASCEND populations of participants. In Table 6, we report the PSNR and F1 edge detection scores measured on T1-weighted MRI (FSL and SLF methods only support T1-weighted MRI inpainting).

A2.4. Ablation Study

Lastly, an ablation study was conducted to evaluate the effect of ensembling over models trained on different anatomical templates and views. The L1 error (voxel-wise sum of absolute intensity differences) averaged across the T1- and T2-weighted MRI scans of 150 cases sampled equally from the ADVANCE and ASCEND validation sets was computed. The L1 error is expressed relative to the range of intensity values in each scan (–1 to 1) in Table 7. Ensembling across atlases and views reduces the relative L1 error by approximately 25%, with the largest improvement being attributed to the integration of information across anatomical planes.

A3. Classification Benchmarking in the Ablated Feature Space

Tables 8 and 9

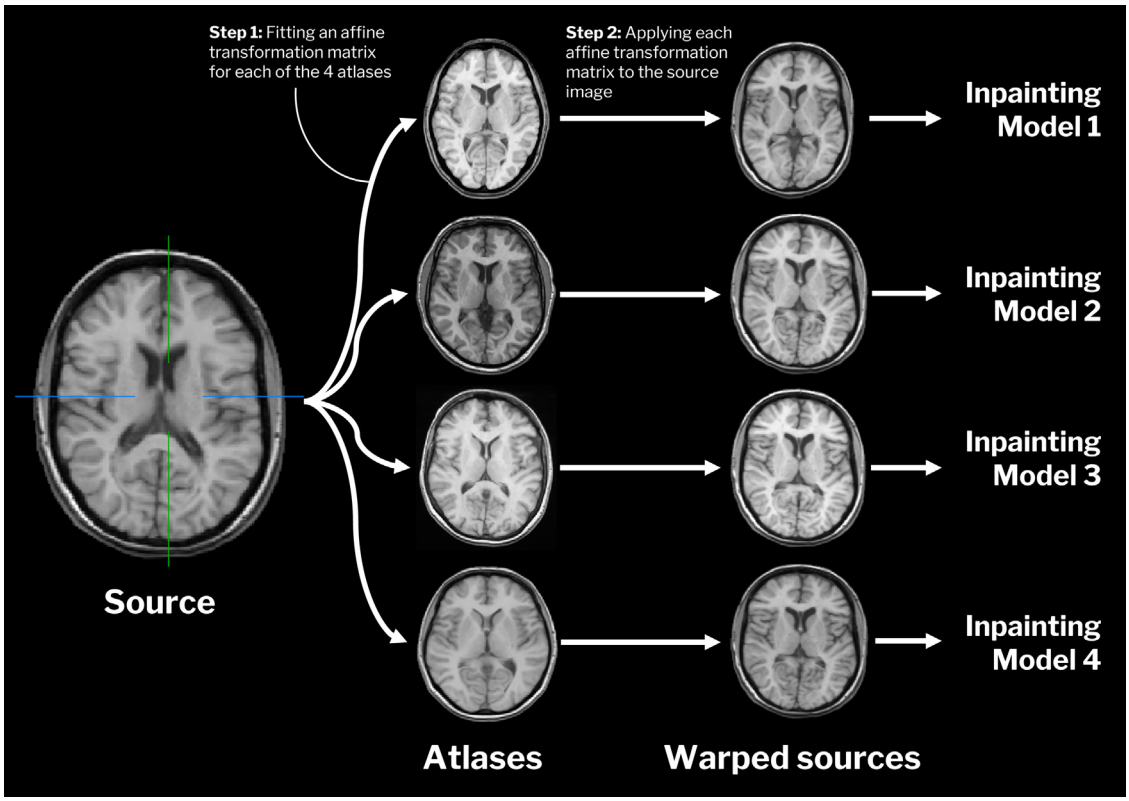

Fig. 16. Overview of our multi-atlas approach to MS lesion inpainting. Each 3D brain MRI scan is registered by an affine method to 4 common anatomical templates. For each anatomical template (and for each anatomical plane) one inpainting model is trained (yielding a total of 12 models). During inference, inpainted brains from difference atlas-specific models are mapped back to the source space by inverting the forward affine transformation, and predictions are aggregated across these models via averaging.

Fig. 17. Comparison of inpainting results on T1-hypointense lesion masks extended by a **1**-mm margin, showing, from left to right, axial slices from the original T1-weighted brain MRI, the lesion mask (red), SFL inpainting, FSL inpainting, and our inpainting.
Abbreviation: MRI, magnetic resonance imaging.

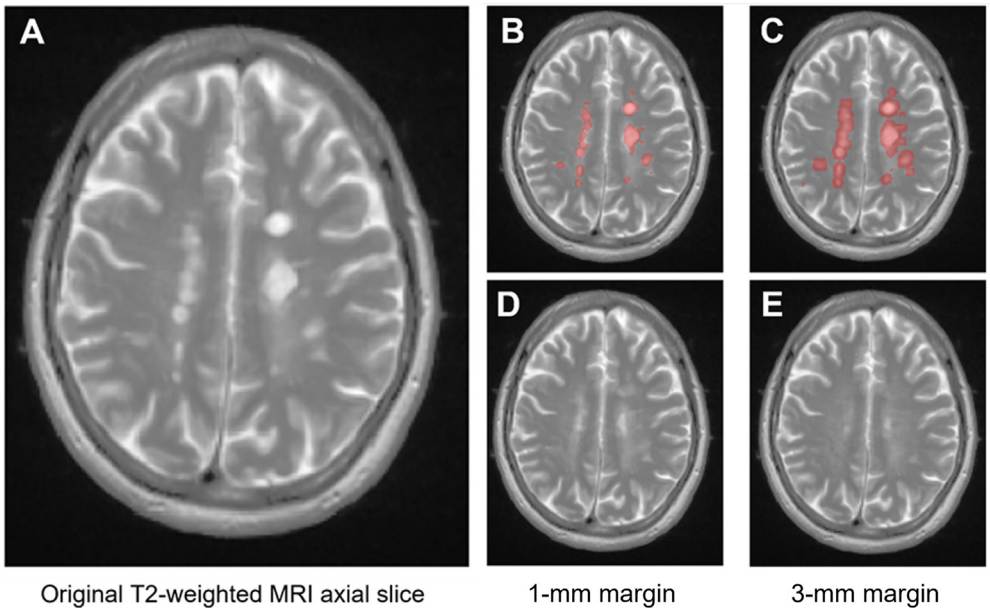

Fig. 18. Comparison of results of inpainting on T2-weighted MRI with peri-WMH margin extensions of **1** mm and **3** mm. Extended WMH masks are shown in red for margins of 1 mm and 3 mm in B and C, leading to the associated inpainted results shown in D and E, respectively.
Abbreviation: MRI, magnetic resonance imaging.

Table 7
Ablation study evaluating the effect of ensembling across atlases and views.

Model	Relative error (%)
1 Atlas, 1 Axis	2.7
4 Atlases, 1 Axis	2.5
1 Atlas, 3 Axes	2.2
4 Atlases, 3 Axes	2.0

Table 8
Classification benchmarking results showing the balanced classification accuracy obtained on the training and internal validation sets (used for hyperparameter tuning, different from the independent validation set from ADVANCE) for each base estimator within the ablated feature space (bold: five best-performing algorithms selected for ensembling). In each cell, we indicate the mean and standard deviation in accuracy across all five cross-validations runs.

Classifier	Training balanced accuracy (%)	Validation balanced accuracy (%)
Multilayer perceptron	79.5 ± 0.2	74.1 ± 0.7
RBF SVM	77.0 ± 0.5	73.6 ± 1.0
XGBoost	97.3 ± 0.3	73.0 ± 1.3
HistGradBoost	80.3 ± 0.3	72.9 ± 1.3
Polynomial SVM	75.6 ± 0.2	70.8 ± 1.6
AdaBoost	72.4 ± 0.3	69.4 ± 1.1
KNN	74.8 ± 0.2	68.0 ± 1.2
Linear SVM	68.1 ± 0.2	68.0 ± 1.2
Decision Tree	69.3 ± 0.4	68.0 ± 1.2
QDA	63.6 ± 0.4	63.1 ± 1.1
Random Forest	60.4 ± 0.4	60.2 ± 0.7
Gaussian Naive Bayes	55.3 ± 0.4	60.1 ± 0.8
Sigmoid SVM	55.6 ± 0.6	55.1 ± 0.5

Abbreviations: KNN: k-nearest neighbors; QDA: quadratic discriminant analysis; RBF: radial basis function; SVM: support vector machine.

Table 9
Classification benchmarking results showing the balanced classification accuracy obtained on the training, validation and testing sets for each ensemble classifier within the ablated feature space of 32 α-radiomic features (red: best-performing ensemble model).

Classifier	Training balanced accuracy on ADVANCE (%)	Validation balanced accuracy on ADVANCE (%)	Testing balanced accuracy on ASCEND (%)	Testing balanced accuracy on DECIDE (%)
(weighted) Hard Voting	83.9	75.1	73.9	73.8
(weighted) Soft Voting	82.4	75.8	74.6	74.3
Stacking	82.8	75.6	74.7	74.2

References

Altay, E.E., Fisher, E., Jones, S.E., Hara-Cleaver, C., Lee, J.C., Rudick, R.A., 2013. Reliability of classifying multiple sclerosis disease activity using magnetic resonance imaging in a multiple sclerosis clinic. JAMA Neurol. 70. doi:10.1001/2013.jamaneurol.211.

Battaglini, M., Jenkinson, M., De Stefano, N., 2012. Evaluating and reducing the impact of white matter lesions on brain volume measurements. Hum. Brain Mapp. 33. doi:10.1002/hbm.21344.

Brück, W., Porada, P., Poser, S., Rieckmann, P., Hanefeld, F., Kretzschmarch, H.A., Lassmann, H., 1995. Monocyte/macrophage differentiation in early multiple sclerosis lesions. Ann. Neurol. 38. doi:10.1002/ana.410380514.

Cai, J., Luo, J., Wang, S., Yang, S., 2018. Feature selection in machine learning: a new perspective. Neurocomputing 300. doi:10.1016/j.neucom.2017.11.077.

Calabresi, P.A., Kieseier, B.C., Arnold, D.L., Balcer, L.J., Boyko, A., Pelletier, J., Liu, S., Zhu, Y., Seddighzadeh, A., Hung, S., Deykin, A., 2014. Pegylated interferon beta-1a for relapsing-remitting multiple sclerosis (ADVANCE): a randomised, phase 3, double-blind study. Lancet Neurol. 13. doi:10.1016/S1474-4422(14)70068-7.

Carré, A., Klausner, G., Edjlali, M., Lerousseau, M., Briend-Diop, J., Sun, R., Ammari, S., Reuzé, S., Alvarez Andres, E., Estienne, T., Niyoteka, S., Battistella, E., Vakalopoulou, M., Dhermain, F., Paragios, N., Deutsch, E., Oppenheim, C., Pallud, J., Robert, C., 2020. Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics. Sci. Rep. 10. doi:10.1038/s41598-020-69298-z.

Chassagnon, G., Vakalopoulou, M., Battistella, E., Christodoulidis, S., Hoang-Thi, T.-N., Dangeard, S., Deutsch, E., Andre, F., Guillo, E., Halm, N., El Hajj, S., Bompard, F., Neveu, S., Hani, C., Saab, I., Campredon, A., Koulakian, H., Bennani, S., Freche, G., Barat, M., Lombard, A., Fournier, L., Monnier, H., Grand, T., Gregory, J., Nguyen, Y., Khalil, A., Mahdjoub, E., Brillet, P.-Y., Tran Ba, S., Bousson, V., Mekki, A., Carlier, R.-Y., Revel, M.-P., Paragios, N., 2020. Holistic AI-driven quantification, staging and prognosis of COVID-19 pneumonia. medRxiv.

Coronado, I., Gabr, R.E., Narayana, P.A., 2021. Deep learning segmentation of gadolinium-enhancing lesions in multiple sclerosis. Mult. Scler. J. 27. doi:10.1177/1352458520921364.

Cotton, F., Weiner, H.L., Jolesz, F.A., Guttmann, C.R.G., 2003. MRI contrast uptake in new lesions in relapsing-remitting MS followed at weekly intervals. Neurology 60. doi:10.1212/01.WNL.0000046587.83503.1E.

Drabycz, S., Mitchell, J.R., 2008. Texture quantification of medical images using a novel complex space-frequency transform. Int. J. Comput. Assist. Radiol. Surg. 3. doi:10.1007/s11548-008-0219-4.

Duron, L., Balvay, D., Perre, S.V., Bouchouicha, A., Savatovsky, J., Sadik, J.C., Thomassin-Naggara, I., Fournier, L., Lecler, A., 2019. Gray-level discretization impacts reproducible MRI radiomics texture features. PLoS ONE 14. doi:10.1371/journal.pone.0213459.

Dworkin, J.D., Linn, K.A., Oguz, I., Fleishman, G.M., Bakshi, R., Nair, G., Calabresi, P.A., Henry, R.G., Oh, J., Papinutto, N., Pelletier, D., Rooney, W., Stern, W., Sicotte, N.L., Reich, D.S., Shinohara, R.T., 2018. An automated statistical technique for counting distinct multiple sclerosis lesions. Am. J. Neuroradiol. 39. doi:10.3174/ajnr.A5556.

Elliott, C., Arnold, D.L., Collins, D.L., Arbel, T., 2013. Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain MRI. IEEE Trans. Med. Imaging 32. doi:10.1109/TMI.2013.2258403.

Elliott, C., Momayyezsiahkal, P., Arnold, D.L., Liu, D., Ke, J., Zhu, L., Zhu, B., George, I.C., Bradley, D.P., Fisher, E., Cahir-McFarland, E., Stys, P.K., Geurts, J.J.G., Franchimont, N., Gafson, A., Belachew, S., 2021. Abnormalities in normal-appearing white matter from which multiple sclerosis lesions arise. Brain Commun 3. doi:10.1093/braincomms/fcab176.

Fartaria, M.J., Bonnier, G., Roche, A., Kober, T., Meuli, R., Rotzinger, D., Frackowiak, R., Schluep, M., Du Pasquier, R., Thiran, J.P., Krueger, G., Bach Cuadra, M., Granziera, C., 2016. Automated detection of white matter and cortical lesions in early stages of multiple sclerosis. J. Magn. Reson. Imaging 43. doi:10.1002/jmri.25095.

Ferrante, E., Dokania, P.K., Silva, R.M., Paragios, N., 2019. Weakly Supervised Learning of Metric Aggregations for Deformable Image Registration. IEEE J. Biomed. Heal. Informat. 23. doi:10.1109/JBHI.2018.2869700.

Fonov, V., Evans, A., McKinstry, R., Almli, C., Collins, D., 2009. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. Neuroimage 47. doi:10.1016/s1053-8119(09)70884-5.

Francis, S.J., 2010. Thesis.

Frischer, J.M., Weigand, S.D., Guo, Y., Kale, N., Parisi, J.E., Pirko, I., Mandrekar, J., Bramow, S., Metz, I., Brück, W., Lassmann, H., Lucchinetti, C.F., 2015. Clinical and pathological insights into the dynamic nature of the white matter multiple sclerosis plaque. Ann. Neurol. 78. doi:10.1002/ana.24497.

Gaj, S., Ontaneda, D., Nakamura, K., 2021. Automatic segmentation of gadolinium-enhancing lesions in multiple sclerosis using deep learning from clinical MRI. PLoS ONE 16. doi:10.1371/journal.pone.0255939.

Gillies, R.J., Kinahan, P.E., Hricak, H., 2016. Radiomics: images are more than pictures, they are data. Radiology 278. doi:10.1148/radiol.2015151169.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. Commun. ACM 63. doi:10.1145/3422622.

Guo, B.J., Yang, Z.L., Zhang, L.J., 2018. Gadolinium deposition in brain: current scientific evidence and future perspectives. Front. Mol. Neurosci. doi:10.3389/fnmol.2018.00335.

Guttmann, C.R.G., Ahn, S.S., Hsu, L., Kikinis, R., Jolesz, F.A., 1995. The evolution of multiple sclerosis lesions on serial MR. Am. J. Neuroradiol. 16.

Guttmann, C.R.G., Rousset, M., Roch, J.A., Hannoun, S., Durand-Dubief, F., Belaroussi, B., Cavallari, M., Rabilloud, M., Sappey-Marinier, D., Vukusic, S., Cotton, F., 2016. Multiple sclerosis lesion formation and early evolution revisited: a weekly high-resolution magnetic resonance imaging study. Mult. Scler. 22. doi:10.1177/1352458515600247.

Harrison, L.C.V., Raunio, M., Holli, K.K., Luukkaala, T., Savio, S., Elovaara, I., Soimakallio, S., Eskola, H.J., Dastidar, P., 2010. MRI texture analysis in multiple sclerosis: toward a clinical analysis protocol. Acad. Radiol. 17. doi:10.1016/j.acra.2010.01.005.

Hauser, S.L., Bar-Or, A., Comi, G., Giovannoni, G., Hartung, H.-P., Hemmer, B., Lublin, F., Montalban, X., Rammohan, K.W., Selmaj, K., Traboulsee, A., Wolinsky, J.S., Arnold, D.L., Klingelschmitt, G., Masterman, D., Fontoura, P., Belachew, S., Chin, P., Mairon, N., Garren, H., Kappos, L., 2017. Ocrelizumab versus interferon beta-1a in relapsing multiple sclerosis. N. Engl. J. Med. 376. doi:10.1056/nejmoa1601277.

Jensen, L.J., Kim, D., Elgeti, T., Steffen, I.G., Hamm, B., Nagel, S.N., 2021. Stability of radiomic features across different region of interest sizes-A CT and MR phantom study. Tomography 7. doi:10.3390/tomography7020022.

Kapoor, R., Ho, P.R., Campbell, N., Chang, I., Deykin, A., Forrestal, F., Lucas, N., Yu, B., Arnold, D.L., Freedman, M.S., Goldman, M.D., Hartung, H.P., Havrdová, E.K., Jeffery, D., Miller, A., Sellebjerg, F., Cadavid, D., Mikol, D., Steiner, D., Bartholomé, E., D'Hooghe, M., Pandolfo, M., Van Wijmeersch, B., Bhan, V., Blevins, G., Brunet, D., Devonshire, V., Duquette, P., Freedman, M., Grand'Maison, F., Jacques, F., Lapierre, Y., Lee, L., Morrow, S., Yeung, M., Dufek, M., Havrdová, E.K., Kanovsky, P., Stetkarova, I., Talabova, M., Frederiksen, J., Kant, M., Petersen, T., Ravnborg, M., Sellebjerg, F., Airas, L., Elovaara, I., Eralinna, J.P., Sarasoja, T., Al Khedr, A., Brassat, D., Brochet, B., Camu, W., Debouverie, M., Laplaud, D., Lebrun Frenay, C., Pelletier, J., Vermersch, P., Vukusi, S., Baum, K., Berthele, A., Faiss, J., Flachenecker, P., Hohlfeld, R., Krumbholz, M., Lassek, C., Maeurer, M., Meuth, S., Ziemssen, T., Hardiman, O., McGuigan, C., Achiron, A., Karussis, D., Bergamaschi, R., Morra, V.B., Comi, G., Cottone, S., Grimaldi, L., Mancardi, G.L., Massacesi, L., Nocentini, U., Salvetti, M., Scarpini, E., Sola, P., Tedeschi, G., Trojano, M., Zaffaroni, M., Frequin, S., Hupperts, R., Killestein, J., Schrijver, H., Van Dijl, R., van Munster, E., Czarnecki, M., Drozdowski, W., Fryze, W., Hertmanowska, H., Ilkowski, J., Kaminska, A., Klodowska-Duda, G., Maciejowski, M., Motta, E., Podemski, R., Potemkowski, A., Rog, T., Selmaj, K., Stelmasiak, Z., Stepien, A., Tutaj, A., Zaborski, J., Boyko, A., Chefranova, Z., Evdoshenko, E., Khabirov, F., Sivertseva, S., Yakupov, E., Alvarez Cermeño, J.C., Escartin, A., Fernandez, O.F., Garcia-Merino, A., Hernandez Perez, M.A., Ayuso, G.I., Lallana, J.M., Gairin, X.M., Oreja-Guevara, C., Saiz Hinarejos, A., Gunnarsson, M., Lycke, J., Martin, C., Piehl, F., Roshanisefat, H., Sundstrom, P., Duddy, M., Gran, B., Harrower, T., Hobart, J., Kapoor, R., Lee, M., Mattison, P., Nicholas, R., Pearson, O., Rashid, W., Rog, D., Sharrack, B., Silber, E., Turner, B., Williams, A., Woolmore, J., Young, C., Bandari, D., Berger, J., Camac, A., Cohan, S., Conway, J., Edwards, K., Fabian, M., Florin, J., Freedman, S., Garwacki, D., Goldman, M., Harrison, D., Herrman, C., Huang, D., Javed, A., Jeffery, D., Kamin, S., Katsamakis, G., Khatri, B., Langer-Gould, A., Lynch, S., Mattson, D., Miller, T., Miravalle, A., Moses, H., Muley, S., Napier, J., Nielsen, A., Pachner, A., Pardo, G., Picone, M.A., Robertson, D., Royal, W., Sheppard, C., Thrower, B., Twyman, C., Waubant, E., Wendt, J., Yadav, V., Zabad, R., Zarelli, G., 2018. Effect of natalizumab on disease progression in secondary progressive multiple sclerosis (ASCEND): a phase 3, randomised, double-blind, placebo-controlled trial with an open-label extension. Lancet Neurol 17. doi:10.1016/S1474-4422(18)30069-3.

Kappos, L., Moeri, D., Radue, E.W., Schoetzau, A., Schweikert, K., Barkhof, F., Miller, D., Guttmann, C.R.G., Weiner, H.L., Gasperini, C., Filippi, M., 1999. Predictive value of gadolinium-enhanced magnetic resonance imaging for relapse rate and changes in disability or impairment in multiple sclerosis: a meta-analysis. Lancet 353. doi:10.1016/S0140-6736(98)03053-0.

Kappos, L., Wiendl, H., Selmaj, K., Arnold, D.L., Havrdova, E., Boyko, A., Kaufman, M., Rose, J., Greenberg, S., Sweetser, M., Riester, K., O'Neill, G., Elkins, J., 2015. Daclizumab HYP versus Interferon Beta-1a in Relapsing Multiple Sclerosis. N. Engl. J. Med. 373. doi:10.1056/nejmoa1501481.

Karimaghaloo, Z., Rivaz, H., Arnold, D.L., Collins, D.L., Arbel, T., 2013. Adaptive voxel, texture and temporal conditional random fields for detection of Gad-enhancing multiple sclerosis lesions in brain MRI. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) doi:10.1007/978-3-642-40760-4_68.

Kobelt, G., Thompson, A., Berg, J., Gannedahl, M., Eriksson, J., 2017. New insights into the burden and costs of multiple sclerosis in Europe. Mult. Scler. 23. doi:10.1177/1352458517694432.

Kuhn, M., Johnson, K., 2013. Applied predictive modeling, Applied Predictive Modeling. https://doi.org/10.1007/978-1-4614-6849-3

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., Aerts, H.J.W.L., 2012. Radiomics: extracting more information from medical images using advanced feature analysis. Eur. J. Cancer 48. doi:10.1016/j.ejca.2011.11.036.

Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B., 2018. Image inpainting for irregular holes using partial convolutions. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) doi:10.1007/978-3-030-01252-6_6.

Manjón, J.V., Romero, J.E., Vivo-Hernando, R., Rubio, G., Aparici, F., de la Iglesia-Vaya, M., Tourdias, T., Coupé, P., 2020. Blind mri brain lesion inpainting using deep learning. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) doi:10.1007/978-3-030-59520-3_5.

Mayerhoefer, M.E., Szomolanyi, P., Jirak, D., Materka, A., Trattnig, S., 2009. Effects of MRI acquisition parameter variations and protocol heterogeneity on the results of texture analysis and pattern discrimination: an application-oriented study. Med. Phys. 36. doi:10.1118/1.3081408.

Meier, D.S., Guttmann, C.R.G., 2003. Time-series analysis of MRI intensity patterns in multiple sclerosis. Neuroimage 20. doi:10.1016/S1053-8119(03)00354-9.

Meier, D.S., Weiner, H.L., Guttmann, C.R.G., 2007. Time-series modeling of multiple sclerosis disease activity: a promising window on disease progression and repair potential? Neurotherapeutics 4. doi:10.1016/j.nurt.2007.05.008.

Michoux, N., Guillet, A., Rommel, D., Mazzamuto, G., Sindic, C., Duprez, T., 2015. Texture analysis of T2-weighted MR images to assess acute inflammation in brain MS lesions. PLoS ONE 10. doi:10.1371/journal.pone.0145497.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016 doi:10.1109/3DV.2016.79.

Moraal, B., Van Den Elskamp, I.J., Knol, D.L., Uitdehaag, B.M.J., Geurts, J.J.G., Vrenken, H., Pouwels, P.J.W., Van Schijndel, R.A., Meier, D.S., Guttmann, C.R.G., Barkhof, F., 2010. Long-interval T2-weighted subtraction magnetic resonance imaging: a powerful new outcome measure in multiple sclerosis trials. Ann. Neurol. 67.

Narayana, P.A., Coronado, I., Sujit, S.J., Wolinsky, J.S., Lublin, F.D., Gabr, R.E., 2020. Deep learning for predicting enhancing lesions in multiple sclerosis from noncontrast MRI. Radiology 294. doi:10.1148/radiol.2019191061.

Nyú, L.G., Udupa, J.K., 1999. On standardizing the MR image intensity scale. Magn. Reson. Med. 42. doi:10.1002/(SICI)1522-2594(199912)42, 6<1072::AID-MRM11>3.0.CO;2-M.

Perazella, M., 2008. Gadolinium-contrast toxicity in patients with kidney disease: nephrotoxicity and nephrogenic systemic fibrosis. Curr. Drug Saf. 3. doi:10.2174/157488608783333989.

Pittock, S.J., Lucchinetti, C.F., 2007. The pathology of MS: new insights and potential clinical applications. Neurologist doi:10.1097/01.nrl.0000253065.31662.37.

Prineas, J.W., Barnard, R.O., Kwon, E.E., Sharer, L.R., Cho, E.-S, 1993. Multiple sclerosis: remyelination of nascent lesions: remyelination of nascent lesions. Ann. Neurol. 33. doi:10.1002/ana.410330203.

Rovira, A., Auger, C., Alonso, J., 2013. Magnetic resonance monitoring of lesion evolution in multiple sclerosis. Ther. Adv. Neurol. Disord. 6. doi:10.1177/1756285613484079.

Salem, M., Cabezas, M., Valverde, S., Pareto, D., Oliver, A., Salvi, J., Rovira, À., Lladó, X., 2018. A supervised framework with intensity subtraction and deformation field features for the detection of new T2-w lesions in multiple sclerosis. NeuroImage Clin 17. doi:10.1016/j.nicl.2017.11.015.

Savio, S.J., Harrison, L.C.V., Luukkaala, T., Heinonen, T., Dastidar, P., Soimakallio, S., Eskola, H.J., 2010. Effect of slice thickness on brain magnetic resonance image texture analysis. Biomed. Eng. Online 9. doi:10.1186/1475-925X-9-60.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision doi:10.1109/ICCV.2017.74.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in mri data. IEEE Trans. Med. Imaging 17. doi:10.1109/42.668698.

Sweeney, E.M., Nguyen, T.D., Kuceyeski, A., Ryan, S.M., Zhang, S., Zexter, L., Wang, Y., Gauthier, S.A., 2021. Estimation of Multiple Sclerosis lesion age on magnetic resonance imaging. Neuroimage 225. doi:10.1016/j.neuroimage.2020.117451.

Tedeschi, G., Bonavita, S., McFarland, H.F., Richert, N., Duyn, J.H., Frank, J.A., 2002. Proton MR spectroscopic imaging in multiple sclerosis. Neuroradiology 44. doi:10.1007/s002340100584.

Thompson, A.J., Banwell, B.L., Barkhof, F., Carroll, W.M., Coetzee, T., Comi, G., Correale, J., Fazekas, F., Filippi, M., Freedman, M.S., Fujihara, K., Galetta, S.L., Hartung, H.P., Kappos, L., Lublin, F.D., Marrie, R.A., Miller, A.E., Miller, D.H., Montalban, X., Mowry, E.M., Sorensen, P.S., Tintoré, M., Traboulsee, A.L., Trojano, M., Uitdehaag, B.M.J., Vukusic, S., Waubant, E., Weinshenker, B.G., Reingold, S.C., Cohen, J.A., 2018. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. Lancet Neurol. doi:10.1016/S1474-4422(17)30470-2.

Traboulsee, A.L., Li, D.K.B., 2006. The role of MRI in the diagnosis of multiple sclerosis. Adv. Neurol..

Valverde, S., Oliver, A., Lladó, X., 2014. A white matter lesion-filling approach to improve brain tissue volume measurements. NeuroImage Clin. 6. doi:10.1016/j.nicl.2014.08.016.

Van Griethuysen, J.J.M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G.H., Fillion-Robin, J.C., Pieper, S., Aerts, H.J.W.L., 2017. Computational radiomics system to decode the radiographic phenotype. Cancer Res. 77. doi:10.1158/0008-5472.CAN-17-0339.

Wattjes, M.P., Ciccarelli, O., Reich, D.S., Banwell, B., de Stefano, N., Enzinger, C., Fazekas, F., Filippi, M., Frederiksen, J., Gasperini, C., Hacohen, Y., Kappos, L., Li, D.K.B., Mankad, K., Montalban, X., Newsome, S.D., Oh, J., Palace, J., Rocca, M.A., Sastre-Garriga, J., Tintoré, M., Traboulsee, A., Vrenken, H., Yousry, T., Barkhof, F., Rovira, À., Rocca, M.A., Tintore, M., Rovira, A., 2021. 2021 MAGNIMS–CMSC–NAIMS consensus recommendations on the use of MRI in patients with multiple sclerosis. Lancet Neurol. doi:10.1016/S1474-4422(21)00095-8.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T., 2019. Free-form image inpainting with gated convolution. In: Proceedings of the IEEE International Conference on Computer Vision doi:10.1109/ICCV.2019.00457.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S., 2018. Generative image inpainting with contextual attention. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition doi:10.1109/CVPR.2018.00577.

Yu, O., Mauss, Y., Zollner, G., Namer, I.J., Chambron, J., 1999. Distinct patterns of active and non-active plaques using texture analysis on brain NMR images in multiple sclerosis patients: preliminary results. Magn. Reson. Imaging 17. doi:10.1016/S0730-725X(99)00062-4.

Zeng, C., Gu, L., Liu, Z., Zhao, S., 2020. Review of deep learning approaches for the segmentation of multiple sclerosis lesions on brain MRI. Front. Neuroinform. 14. doi:10.3389/fninf.2020.610967.

Zhang, H., Bakshi, R., Bagnato, F., Oguz, I., 2020. Robust multiple sclerosis lesion inpainting with edge prior. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) doi:10.1007/978-3-030-59861-7_13.

Zhang, H., Goodfellow, I., Metaxas, D., Odena, A., 2019. Self-attention generative adversarial networks. 36th International Conference on Machine Learning, ICML 2019.

Zhang, J., Tong, L., Wang, L., Li, N., 2008. Texture analysis of multiple sclerosis: a comparative study. Magn. Reson. Imaging 26. doi:10.1016/j.mri.2008.01.016.

Zhang, Y., Moore, G.R.W., Laule, C., Bjarnason, T.A., Kozlowski, P., Traboulsee, A., Li, D.K.B., 2013. Pathological correlates of magnetic resonance imaging texture heterogeneity in multiple sclerosis. Ann. Neurol. 74. doi:10.1002/ana.23867.

Zhang, Y., Traboulsee, A., Zhao, Y., Metz, L.M., Li, D.K., 2011. Texture analysis differentiates persistent and transient T1 black holes at acute onset in multiple sclerosis: a preliminary study. Mult. Scler. J. 17. doi:10.1177/1352458510395981.

Zhang, Y., Zhu, H., Mitchell, J.R., Costello, F., Metz, L.M., 2009. T2 MRI texture analysis is a sensitive measure of tissue injury and recovery resulting from acute inflammatory lesions in multiple sclerosis. Neuroimage 47. doi:10.1016/j.neuroimage.2009.03.075.

Zhong, Y., Utriainen, D., Wang, Y., Kang, Y., Haacke, E.M., 2014. Automated white matter hyperintensity detection in multiple sclerosis using 3D T2 FLAIR. Int. J. Biomed. Imaging 2014. doi:10.1155/2014/239123.

Zhou, T., Zheng, Y., Sun, L., Badea, S.R., Jin, Y., Liu, Y., Rolfe, A.J., Sun, H., Wang, X., Cheng, Z., Huang, Z., Zhao, N., Sun, X., Li, J., Fan, J., Lee, C., Megraw, T.L., Wu, W., Wang, G., Ren, Y., 2019. Microvascular endothelial cells engulf myelin debris and promote macrophage recruitment and fibrosis after neural injury. Nat. Neurosci. 22. doi:10.1038/s41593-018-0324-9.

Zwanenburg, A., Vallières, M., Abdalah, M.A., Aerts, H.J.W.L., Andrearczyk, V., Apte, A., Ashrafinia, S., Bakas, S., Beukinga, R.J., Boellaard, R., Bogowicz, M., Boldrini, L., Buvat, I., Cook, G.J.R., Davatzikos, C., Depeursinge, A., Desseroit, M.C., Dinapoli, N., Dinh, C.V., Echegaray, S., El Naqa, I., Fedorov, A.Y., Gatta, R., Gillies, R.J., Goh, V., Götz, M., Guckenberger, M., Ha, S.M., Hatt, M., Isensee, F., Lambin, P., Leger, S., Leijenaar, R.T.H., Lenkowicz, J., Lippert, F., Losnegård, A., Maier-Hein, K.H., Morin, O., Müller, H., Napel, S., Nioche, C., Orlhac, F., Pati, S., Pfaehler, E.A.G., Rahmim, A., Rao, A.U.K., Scherer, J., Siddique, M.M., Sijtsema, N.M., Socarras Fernandez, J., Spezi, E., Steenbakkers, R.J.H.M., Tanadini-Lang, S., Thorwarth, D., Troost, E.G.C., Upadhaya, T., Valentini, V., van Dijk, L.V., van Griethuysen, J., van Velden, F.H.P., Whybra, P., Richter, C., Löck, S., 2020. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 295. doi:10.1148/radiol.2020191145.