Diagnostic and Interventional Imaging 101 (2020) 795–802

Original article/*Computer developments*

# Artificial intelligence to predict clinical disability in patients with multiple sclerosis using FLAIR MRI

P. Roca[a,*], A. Attye[b,c], L. Colas[d], A. Tucholka[a], P. Rubini[a], S. Cackowski[e], J. Ding[d], J.-F. Budzik[d], F. Renard[f,g], S. Doyle[a], E.L. Barbier[e], I. Bousaid[h], R. Casey[i,j,k,l], S. Vukusic[i,j,k,l], N. Lassau[m,n], S. Verclytte[d], F. Cotton[k,o,p], On behalf of OFSEP Investigators: [1]

[a] *Pixyl, Research and Development Laboratory, 38000 Grenoble, France*
[b] *Grenoble Alpes University, 38000 Grenoble, France*
[c] *Sydney Imaging Lab, Sydney University, 2006 Sydney, NSW, Australia*
[d] *Imaging Department, Lille Catholic Hospitals, Lille Catholic University, 59000 Lille, France*
[e] *University Grenoble Alpes, Inserm, U1216, Grenoble Institute Neurosciences, 38000 Grenoble, France*
[f] *University Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France*
[g] *University Grenoble Alpes, AGEIS, 38000 Grenoble, France*
[h] *Direction Transformation Numérique et Systèmes d'Information, Institut Gustave Roussy, 94805 Villejuif, France*
[i] *Department of Neurology–Multiple Sclerosis, Pathologies de la myéline et neuro-inflammation, Hôpital Pierre Wertheimer, Hospices Civils de Lyon, 69500 Bron, France*
[j] *Université Claude Bernard Lyon 1, Université de Lyon, 69622 Villeurbanne, France*
[k] *Observatoire Français de la Sclérose en Plaques, Centre de Recherche en Neurosciences de Lyon, INSERM 1028 et CNRS UMR 5292, 69003 Lyon, France*
[l] *Eugène Devic EDMUS Foundation Against Multiple Sclerosis, 69500 Bron, France*
[m] *Radiology Department, Institut Gustave Roussy, 94805 Villejuif, France*
[n] *BIOMAPS, UMR1281, Université Paris-Saclay, Inserm, CNRS, CEA, Laboratoire d'Imagerie Biomédicale Multimodale Paris-Saclay, 94800 Villejuif, France*
[o] *Department of Radiology, Centre Hospitalier Lyon-Sud, Hospices Civils de Lyon, 69310 Pierre-Bénite, France*
[p] *CREATIS, CNRS UMR 5220, INSERM U1044, 69622 Villeurbanne, France*

## ARTICLE INFO

*Keywords:*
Artificial intelligence
Machine learning
Multiple sclerosis
Disability prediction
Magnetic resonance imaging (MRI)

## ABSTRACT

*Purpose:* The purpose of this study was to create an algorithm that combines multiple machine-learning techniques to predict the expanded disability status scale (EDSS) score of patients with multiple sclerosis at two years solely based on age, sex and fluid attenuated inversion recovery (FLAIR) MRI data.
*Materials and methods:* Our algorithm combined several complementary predictors: a pure deep learning predictor based on a convolutional neural network (CNN) that learns from the images, as well as classical machine-learning predictors based on random forest regressors and manifold learning trained using the location of lesion load with respect to white matter tracts. The aggregation of the predictors was done through a weighted average taking into account prediction errors for different EDSS ranges. The training dataset consisted of 971 multiple sclerosis patients from the "Observatoire français de la sclérose en plaques" (OFSEP) cohort with initial FLAIR MRI and corresponding EDSS score at two years. A test dataset (475 subjects) was provided without an EDSS score. Ten percent of the training dataset was used for validation.

*Abbreviations:* 2D, Two-dimensional; 3D, Three-dimensional; AI, Artificial intelligence; CNN, Convolutional neural network; EDSS, Expanded disability status scale; FLAIR, Fluid attenuated inversion recovery; MNI, Montreal Neurological Institute; MRI, Magnetic resonance imaging; MS, Multiple sclerosis; MSE, Mean square error; OFSEP, Observatoire français de la sclérose en plaques; UMAP, Uniform manifold approximation and projection; WMH, White matter hyperintensities.

* Corresponding author at: Pixyl, Research and Development Laboratory, 38000 Grenoble, France.
*E-mail address:* contact@pixyl.ai (P. Roca).

[1] On behalf of OFSEP Investigators: Steering Committee B. Brochet (Centre hospitalier universitaire de Bordeaux, Hôpital Pellegrin, Service de neurologie, Bordeaux, France), R. Casey (Observatoire français de la sclérose en plaques (OFSEP), Centre de coordination national, Lyon/Bron, France), F. Cotton (Hospices civils de Lyon, Hôpital Lyon sud, Service d'imagerie médicale et interventionnelle, Lyon/Pierre-Bénite, France), J. De Sèze (Hôpitaux universitaire de Strasbourg, Hôpital de Hautepierre, Service des maladies inflammatoires du système nerveux – neurologie, Strasbourg, France), P. Douek (Union pour la lutte contre la sclérose en plaques (UNISEP), Ivry-sur-Seine, France), F. Guillemin (CIC 1433 Epidémiologie Clinique, Centre hospitalier régional universitaire de Nancy, Inserm et Université de Lorraine, Nancy, France), D. Laplaud (Centre hospitalier universitaire de Nantes, Hôpital nord Laennec, Service de neurologie, Nantes/Saint-Herblain, France), C. Lebrun-Frenay (Centre hospitalier universitaire de Nice, Université Nice Côte d'Azur, Hôpital Pasteur, Service de neurologie, Nice, France), L. Mansuy (Hospices civils de Lyon, Département de la recherche clinique et de l'innovation, Lyon, France), T. Moreau (Centre hospitalier universitaire Dijon Bourgogne, Hôpital François Mitterrand, Service de neurologie, maladies inflammatoires du système nerveux et neurologie générale, Dijon, France), J. Olaiz (Université Claude Bernard Lyon 1, Lyon ingéniérie projets, Lyon, France), J. Pelletier (Assistance publique des hôpitaux de Marseille, Centre hospitalier de la Timone, Service de neurologie et unité neuro-vasculaire, Marseille, France), C. Rigaud-Bully (Fondation Eugène Devic EDMUS contre la sclérose en plaques, Lyon, France), B. Stankoff (Assistance publique des hôpitaux de Paris, Hôpital Saint-Antoine, Service de neurologie, Paris, France),

https://doi.org/10.1016/j.diii.2020.05.009

*Results:* Our algorithm predicted EDSS score in patients with multiple sclerosis and achieved a MSE = 2.2 with the validation dataset and a MSE = 3 (mean EDSS error = 1.7) with the test dataset.
*Conclusion:* Our method predicts two-year clinical disability in patients with multiple sclerosis with a mean EDSS score error of 1.7, using FLAIR sequence and basic patient demographics. This supports the use of our model to predict EDSS score progression. These promising results should be further validated on an external validation cohort.


## 1. Introduction

Multiple sclerosis (MS) is a chronic inflammatory demyelinating disease of the central nervous system, which remains the leading cause of non-traumatic disability in young people and is associated with a high economic burden on society partly due to the high cost of the available treatments [1,2]. In most patients with MS, the initial phase of the disease consists of reversible episodes of neurological deficits and over time, the development of permanent neurological deficits and progression of clinical disability [3]. Correctly predicting short-term outcome in patients with MS is an important issue as this could help identify patients who may benefit from a more aggressive treatment.

So far, knowledge of natural disability evolution of MS is mainly based on cohort studies and focused on long-term clinical progression. As a consequence, baseline factors strongly predictive of worsening disability have not yet been fully identified [4,5]. Advanced statistical modeling using support vector machine and random forest has been recently applied on 1582 patients to predict short-term expanded disability status scale (EDSS) score progression after 2 years from a comprehensive list of baseline factors [6]. These factors included clinical factors (such as age, gender, ethnicity, number of relapses 1 and 3 years prior to study, disease duration, prior treatment, EDSS score) and imaging factors (number of lesions, lesion volume and brain parenchymal fraction) [6]. Nevertheless, the predictor models showed poor discriminating capabilities so that there is a need for alternate predictors [6].

Artificial intelligence (AI) has demonstrated utility in the identification of abnormalities on imaging studies [7–11]. However, the capabilities of AI as directly applied to fluid-attenuated inversion recovery (FLAIR) MRI data have received little attention so far because of the well-known "clinico-radiological paradox". Indeed, the clinical course of MS based on the burden of lesions is known to be unpredictable [12–14]. It is not clear whether this paradox relies on a lack of information, for example regarding the gray matter MS

S. Vukusic (Hospices civils de Lyon, Hôpital Pierre Wertheimer, Service de neurologie A, Lyon/Bron, France), , Investigators R. Marignier (Hospices civils de Lyon, Hôpital Pierre Wertheimer, Service de neurologie A, Lyon/Bron, France), M. Debouverie (Centre hospitalier régional universitaire de Nancy, Hôpital central, Service de neurologie, Nancy, France), G. Edan (Centre hospitalier universitaire de Rennes, Hôpital Pontchaillou, Service de neurologie, Rennes, France), J. Ciron (Centre hospitalier universitaire de Toulouse, Hôpital Purpan, Service de neurologie inflammatoire et neuro-oncologie, Toulouse, France), A. Ruet (Centre hospitalier universitaire de Bordeaux, Hôpital Pellegrin, Service de neurologie, Bordeaux, France), N. Collongues (Hôpitaux universitaire de Strasbourg, Hôpital de Hautepierre, Service des maladies inflammatoires du système nerveux – neurologie, Strasbourg, France), C. Lubetzki (Assistance publique des hôpitaux de Paris, Hôpital de la Pitié-Salpêtrière, Service de neurologie, Paris, France), P. Vermersch (Centre hospitalier universitaire de Lille, Hôpital Salengro, Service de neurologie D, Lille, France), P. Labauge (Centre hospitalier universitaire de Montpellier, Hôpital Gui de Chauliac, Service de neurologie, Montpellier, France), G. Defer (Centre hospitalier universitaire de Caen Normandie, Service de neurologie, Hôpital Côte de Nacre, Caen, France), M. Cohen (Centre hospitalier universitaire de Nice, Université Nice Côte d'Azur, Hôpital Pasteur, Service de neurologie, Nice, France), A. Fromont (Centre hospitalier universitaire Dijon Bourgogne, Hôpital François Mitterrand, Service de neurologie, maladies inflammatoires du système nerveux et neurologie générale, Dijon, France), S. Wiertlewsky (Centre hospitalier universitaire de Nantes, Hôpital nord Laennec, Service de neurologie, Nantes/Saint-Herblain, France), E. Berger (Centre hospitalier régional universitaire de Besançon, Hôpital Jean Minjoz, Service de neurologie, Besançon, France), P. Clavelou (Centre hospitalier universitaire de Clermont-Ferrand, Hôpital Gabriel-Montpied, Service de neurologie, Clermont-Ferrand, France), B. Audoin (Assistance publique des hôpitaux de Marseille, Centre hospitalier de la Timone, Service de neurologie et unité neuro-vasculaire, Marseille, France), C. Giannesini (Assistance publique des hôpitaux de Paris, Hôpital Saint-Antoine, Service de neurologie, Paris, France), O. Gout (Fondation Adolphe de Rothschild de l'œil et du cerveau, Service de neurologie, Paris, France), E. Thouvenot (Centre hospitalier universitaire de Nîmes, Hôpital Carémeau, Service de neurologie, Nîmes, France), O. Heinzlef (Centre hospitalier intercommunal de Poissy Saint-Germain-en-Laye, Service de neurologie, Poissy, France), A. Al-Khedr (Centre hospitalier universitaire d'Amiens Picardie, Site sud, Service de neurologie, Amiens, France), B. Bourre (Centre hospitalier universitaire Rouen Normandie, Hôpital Charles-Nicolle, Service de neurologie, Rouen, France), O. Casez (Centre hospitalier universitaire Grenoble-Alpes, Site nord, Service de neurologie, Grenoble/La Tronche, France), P. Cabre (Centre hospitalier universitaire de Martinique, Hôpital Pierre Zobda-Quitman, Service de Neurologie, Fort-de-France, France), A. Montcuquet (Centre hospitalier universitaire Limoges, Hôpital Dupuytren, Service de neurologie, Limoges, France), A. Créange (Assistance publique des hôpitaux de Paris, Hôpital Henri Mondor, Service de neurologie, Créteil, France), J.-P. Camdessanché (Centre hospitalier universitaire de Saint-Étienne, Hôpital Nord, Service de neurologie, Saint-Étienne, France), J. Faure (Centre hospitalier universitaire de Reims, Hôpital Maison-Blanche, Service de neurologie, Reims, France), A. Maurousset (Centre hospitalier régional universitaire de Tours, Hôpital Bretonneau, Service de neurologie, Tours, France), I. Patry (Centre hospitalier sud francilien, Service de neurologie, Corbeil-Essonnes, France), K. Hankiewicz (Centre hospitalier de Saint-Denis, Hôpital Casanova, Service de neurologie, Saint-Denis, France), C. Pottier (Centre hospitalier de Pontoise, Service de neurologie, Pontoise, France), N. Maubeuge (Centre hospitalier universitaire de Poitiers, Site de la Milétrie, Service de neurologie, Poitiers, France), C. Labeyrie (Assistance publique des hôpitaux de Paris, Hôpital Bicêtre, Service de neurologie, Le Kremlin-Bicêtre, France), C. Nifle (Centre hospitalier de Versailles, Hôpital André-Mignot, Service de neurologie, Le Chesnay, France), , Imaging group R. Ameli (Hospices civils de Lyon, Service de radiologie, Lyon, France), R. Anxionnat (CHU Nancy, Service de radiologie, Nancy, France), A. Attye (CHU de Grenoble, Service de radiologie, Grenoble, France), E. Bannier (Institut de Recherche en Informatique et Systèmes Aléatoires, Rennes, France), C. Barillot (INRIA, Rennes, France), D. Ben Salem (CHU Brest, Service de radiologie, Brest, France), M.-P. Boncoeur-Martel (CHU Limoges, Service de radiologie, Limoges, France), F. Bonneville (CHU Toulouse Purpan, Service de radiologie, Toulouse, France), C. Boutet (CHU Saint-Etienne, Service de radiologie, Saint-Etienne, France), J.-C. Brisset (Median technologies, Valbonne, France), F. Cervenanski (CREATIS, Villeurbanne, France), B. Claise (CHU Clermont-Ferrand, Service de radiologie, Clermont-Ferrand, France), O. Commowick (NRIA, Rennes, France), J.-M. Constans (CHU Amiens–Picardie, Service de radiologie, Amiens, France), P. Dardel (CH Chambéry, Service de radiologie, Chambéry, France), H. Desal (CHU Nantes, Service de radiologie, Nantes, France), Vincent Dousset (CHU Bordeaux, Service de radiologie, Bordeaux, France), F. Durand-Dubief (Hospices civils de Lyon, Service de Neurologie, Lyon, France), J.-C. Ferre (CHU Rennes, Service de radiologie, Rennes, France), E. Gerardin (CHU Rouen, Service de radiologie, Rouen, France), T. Glattard (CREATIS, Villeurbanne, France), S. Grand (CHU de Grenoble, Service de radiologie, Grenoble, France), T. Grenier (CREATIS, Villeurbanne, France), R. Guillevin (CHR Poitiers, Service de radiologie, Poitiers, France), C. Guttmann (Harvard Medical School, Boston, USA), A. Krainik (CHU Grenoble Alpes, Service de radiologie, Grenoble, France), S. Kremer (CHU Strasbourg, Service de radiologie, Strasbourg, France), S. Lion (Centre de coordination national de l'OFSEP, Lyon/Bron, France), N. Menjot de Champfleur (CHU Montpellier, Service de radiologie, Montpellier, France), L. Mondot (CHU Nice, Service de radiologie, Nice, France), O. Outteryck (CHRU Lille, Consultations de neurologie D, Lille, France), N. Pyatigorskaya (ICM, Service de radiologie, Paris, France), J.-P. Pruvo (CHRU Lille, Service de radiologie, Lille, France), S. Rabaste (Hospices civils de Lyon, Service de radiologie, Lyon, France), J.-P. Ranjeva (APHM - CHU Marseille Timone, Service de radiologie, Marseille, France), J.-A. Roch (Hôpital privé Jean Mermoz, Service de radiologie, Lyon, France), J.C. Sadik (Fondation A. de Rothschild, Service de radiologie, Paris, France), D. Sappey-Marinier (Hospices civils de Lyon, Service de radiologie, Lyon, France), J. Savatovsky (Fondation A. de Rothschild, Service de radiologie, Paris, France), J.-Y. Tanguy (CH Angers, Service de radiologie, Angers, France), A. Tourbah (Hôpital Raymond Poincaré, Service de Neurologie, Garches, France), T. Tourdias (CHU Bordeaux, Service de radiologie, Bordeaux, France),

**Table 1**
MRI characteristics of the different datasets.

| MRI characteristic | Training set | Validation set | Test set |
|---|---|---|---|
| Three-dimensional | 557 (557/856; 65%) | 67 (67/96; 70%) | 410 (410/475; 86%) |
| Two-dimensional | 299 (299/856; 35%) | 29 (29/96; 30%) | 65 (65/475; 14%) |
| 3T | 445 (445/856; 52%) | 58 (58/96; 60%) | 237 (237/475; 50%) |
| 1.5T | 411 (411/856; 48%) | 38 (38/96; 40%) | 238 (238/475; 50%) |
| Siemens/Philips/GE/Canon (%) | 42/45.1/12.4/0.5 | 40/49/11/0 | 44/41/14.6/0.4 |
| Period of MR image acquisition (years) | 2008–2017 | 2009–2017 | 2015–2019 |
| Dataset count[a] | 856 | 96 | 475 |

Siemens: Siemens Healthineers; Philips: Philips Healthcare; GE: General Electric Healthcare; Canon: Canon Medical Systems.
[a] 19 subjects were excluded from training and validation sets due to poor image quality or small field of view.

injuries, or due to the absence of appropriate tools to analyze the white matter spatial distribution of MS lesions.

The purpose of this study was to create an algorithm that combines multiple machine-learning techniques with the ability to predict EDSS score of patients with MS, based on age, sex and FLAIR MRI data.

## 2. Materials and methods

### 2.1. Study population

The FLAIR MRI data were provided as part of the "Multiple Sclerosis" challenge organized during the 2019 edition of the Journées françaises de radiologie, which is the annual meeting of the French Society of Radiology (Société française de radiologie). Two training datasets of patients with MS with initial FLAIR MRI and EDSS score at two years were used. A first dataset (DS1) included 480 subjects and a second one (DS2) 491 subjects. A third new test dataset without EDSS values (DS3) of 475 subjects was the reference to evaluate the exactness of the model. Datasets DS1, DS2, and DS3 were part of the OFSEP ("Observatoire français de la sclérose en Plaques") cohort, registered on clinicaltrials.gov (NCT02889965) and compliant with French data confidentiality regulations.

The MRI characteristics of the different datasets are summarized in Table 1. This multi-centric dataset originated from 37 institutions in 13 French cities and contained a variety of FLAIR sequences (*i.e.*, two-dimensional [2D] or three-dimensional [3D] acquisition, sagittal/axial or coronal planes, contrast-enhanced or not, and various imaging parameters) acquired using various MRI units (Siemens Healthineers, General Electric Healthcare, Philips Healthcare, Canon Medical Systems) and magnetic fields (1.5 T or 3 T).

### 2.2. MS disability prediction

Different complementary strategies were combined. They included intensity bias field correction, FLAIR normalization to a customized brain template, data augmentation, tract-based lesion load computation, pre-training, ensemble aggregation of a pure deep learning model [15] and predictor models using "hand-crafted features" based on a priori anatomical knowledge, and parallel deployment on the Pixyl Cloud Infrastructure. Fig. 1 presents the flowchart of our pipeline.

### 2.3. Preprocessing and training/validation split

FLAIR images were first corrected for inhomogeneities using the N4 algorithm [16] and registered to a common home-made FLAIR template provided by the Montreal Neurological Institute (MNI) using linear and nonlinear registrations of the ANTS library [17]. After this step, FLAIR images were normalized to zero mean and unit variance and resized to 148 × 148 × 154 voxels. The FLAIR

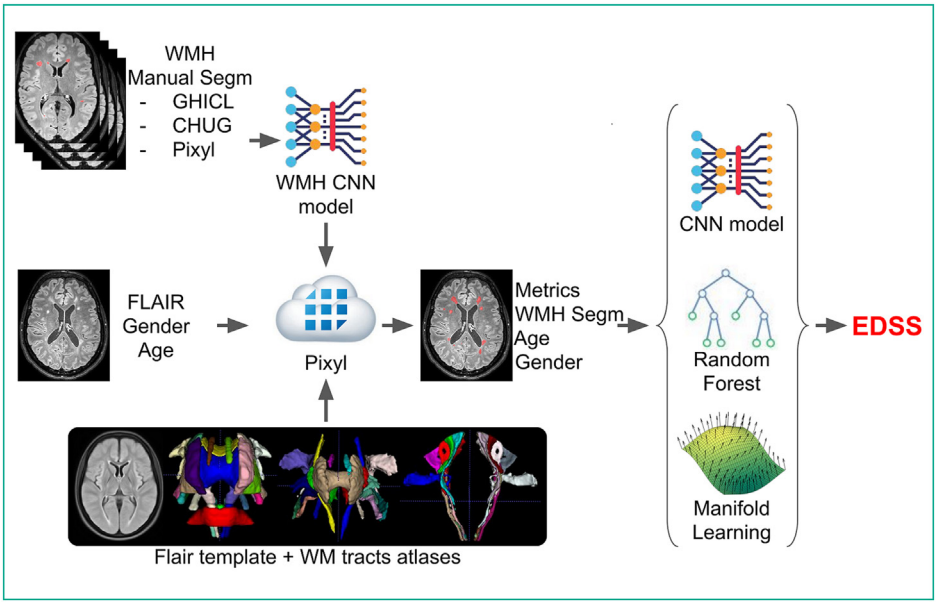


**Fig. 1.** Flowchart of the prediction pipeline. CNN: convolutional neural network; EDSS: expanded disability status scale; FLAIR: fluid attenuated inversion recovery; GHICL: Groupement des Hôpitaux de l'Institut Catholique de Lille; CHUG: Centre Hospitalier Universitaire de Grenoble.

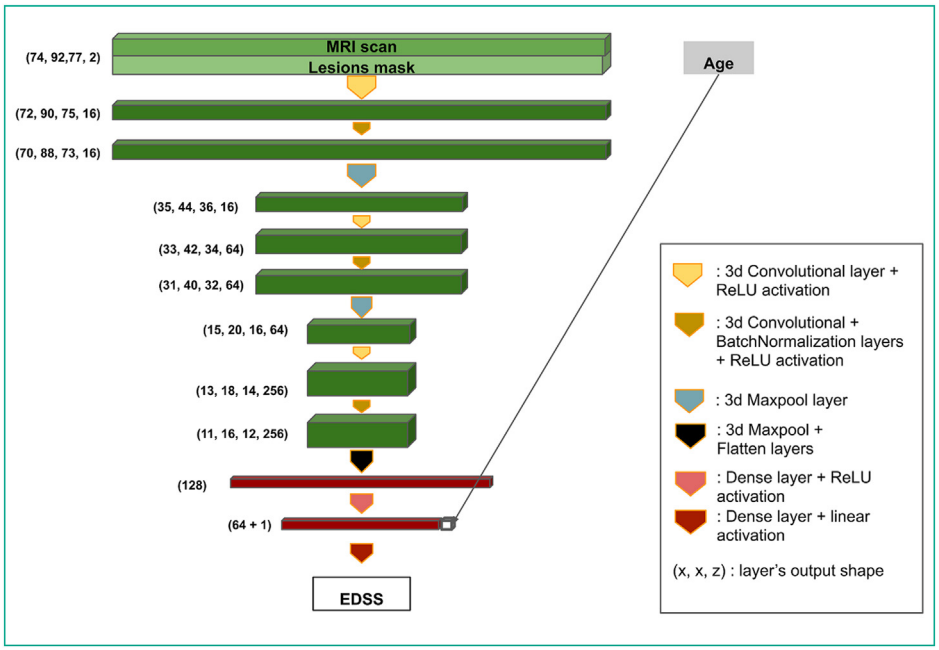


**Fig. 2.** Drawing shows the architecture of the convolutional neural network predictor. EDSS = expanded disability status scale;.

template was built using a subset of 195 3D FLAIR images from DS1 first registered to the MNI space using an affine transform by two observers (A.T., P. Ro.) [18].

In order to train and validate the predictor models, we divided the union of DS1 and DS2 into a training set (90%) and a validation set (10%) in a stratified way guaranteeing that each subgroup follows the same EDSS distribution.

### 2.4. Deep learning predictor

To facilitate the prediction task, we divided the problem into two steps. First, we segmented white matter hyperintensities (WMH) from linearly normalized FLAIR MRI, leading to a lesion map in normalized space. Second, we predicted the two-year EDSS score from age, normalized FLAIR MRI and lesion map.

WMH were segmented using the Pixyl.Neuro CE-marked solution (https://pixyl.ai/). This solution used a convolutional neural network (CNN) based on a multi-level patch-based series of convolutions and max pools in TensorFlow. The CNN was pre-trained on hundreds of FLAIR images from multiple MRI manufacturers, labeled by experts, augmented using noise, inhomogeneities and geometric deformations. The CNN was retrained using an additional dataset of 29 FLAIR images of MS patients from the "Groupement des Hôpitaux de l'Institut Catholique de Lille" manually segmented by three expert radiologists (S.V., J. D., L. C.).

To predict the EDSS score, a 3D-CNN composed of three convolutional blocks, each corresponding to a succession of two 3D convolutional layers followed by 3D max-pooling layer was developed (Fig. 2). A ReLU activation was added after each convolutional layer and batch-normalization was used after the second convolutional layer. After extraction of the features we added a succession of dense layers. Patient age was added as a new feature in the last layer as it is one of the most relevant features for EDSS score prediction and it would help increase algorithm performance. Finally, this densely connected layer of 65 features predicted the EDSS score. We addressed the EDSS score prediction from a regression point of view as the EDSS scores are ordered by disability severity. Weights were initialized using a truncated normal distribution centered on 0 with a standard deviation of 0.02. The model was trained on batches of size 16, using Adam optimizer with a learning rate of 10e-3 to minimize the mean squared error (MSE) loss function.

Two instances of the model were trained: one using the FLAIR and lesion map linearly normalized to the MNI space which takes into account brain atrophy specific to each subject, and a second one based on the non-linear registration computed previously less sensitive to atrophy. Indeed, the non-linear registration allows a better matching of anatomical structures, but can mask relevant differences, particularly those associated with brain atrophy.

### 2.5. Classical machine learning predictors using anatomical knowledge

A dimension reduction was performed using handcrafted features summarizing the impact of lesions on the brain network through tract-based lesion load computation, and more general volumetric measures. Quantitative analysis of white matter lesion burden was performed in 60 tracts of interest from the ICBM-DTI-81 white matter labels [19–21] and the sensorimotor tracts atlases [22] in MNI space using nonlinear registration. In addition, measures of whole-brain lesion load and volume of the lateral ventricles were performed. These volumetric measures, combined with age, gender and 3D/2D nature of FLAIR sequence constituted 65 features used to train two additional EDSS score predictors.

The first predictor used random forest regressors from scikit-learn [23,24] with 200 estimators and three samples minimum per leaf to reduce overfitting. The random forest regressor was trained twice, firstly on the whole training dataset (R̈F single¨) then on a subset containing 3D FLAIR images only. These two models were combined in a unique predictor ("RF dual") using the 3D nature of the input data (Fig. 3).

A second complementary predictor using manifold learning was built using the uniform manifold approximation and projection (UMAP) algorithm, chosen for its good property to preserve the global structure of the data [25]. Then the EDSS score was predicted in a 2D reduced space by a local interpolation of the targets associated with the nearest neighbors in the training set.

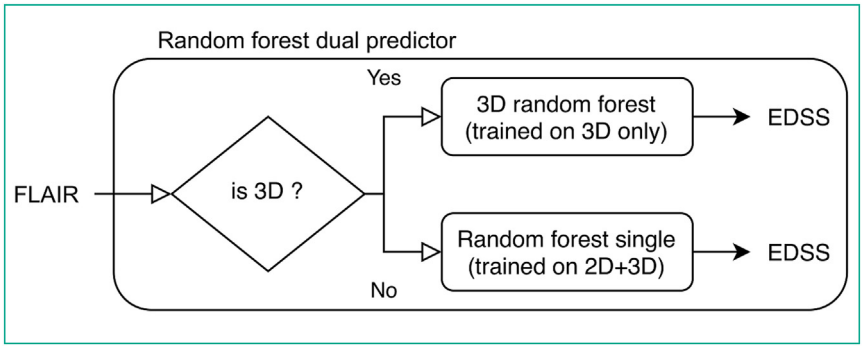


**Fig. 3.** Flowchart associated with the random forest dual predictor. Depending on the three-dimensional (3D) nature of the input data, the model uses either a model trained exclusively on 3D FLAIR or one trained on both 3D and 2D FLAIR to predict the EDSS score. FLAIR: fluid attenuated inversion recovery; EDSS: expanded disability status scale.

### 2.6. Ensemble aggregation and implementation

To evaluate the performance of each predictor, we classified EDSS scores in ten groups according to the EDSS integer part: group $p = 0 : EDSS < 1$, group $p = 1 : 1 \leq EDSS < 2, \ldots,$ group $p = 9 : 9 \leq EDSS < 10$, group $p = 10 : EDSS = 10$. For each predictor $k$ the mean square error across each EDSS group ($MSE_k(p)$, *for* $p = 1 \cdot \cdot 10$) was computed on the validation dataset. The EDSS score predictors were then aggregated using a weighted average where the weights associated with each predictor relied on its performance on the validation dataset as follows:

For each subject $x$:

$$edss_{agg}(x) = \sum_k w_k(floor(edss_k(x)))\, edss_k(x) \times \left(1 / \sum_k w_k(floor(edss_k(x)))\right)$$

where $edss_k$ is the EDSS score predicted by the kth predictor, and $w_k(p)$, the weight associated with this predictor for the EDSS group $p$, is equal to the inverse of $MSE_k(p)$ presented previously. In order to study the contribution of this aggregated predictor compared to age only, an additional predictor using Ridge linear regression based on age was built.

To provide the prediction in DS3 within two hours, we integrated all processing steps, from preprocessing to ensemble aggregation, in an automated pipeline that predicted the EDSS score from a raw FLAIR sequence, as well as age and gender. By having a stand-alone pipeline, we were able to use Pixyl's infrastructure to run all the analyses in parallel.

## 3. Results

### 3.1. Datasets characteristics

In addition to the heterogeneity in MRI quality, the EDSS score distribution was very unbalanced (Fig. 4). There were more low scores than high scores (81% of EDSS scores $\leq 4$) and > 22% of the samples corresponded to an EDSS score of 0. In addition, integer scores were over-represented (75% of EDSS > 0) by comparison with non-integer scores (25%).

### 3.2. Score, ranking and predictor performances

We achieved a MSE score of 3 (associated with an estimated mean EDSS score error of 1.7) on this new dataset of 475 subjects and submitted the results in one hour and a half, scoring first in the data challenge. Fig. 5 presents the most informative features, with the associated measure of importance given by the random forest model trained on 3D FLAIR.

Our different predictors demonstrated similar performance in terms of global MSE on the validation dataset (Table 2), with the random forest model specific to 3D/2D data ranking first with a

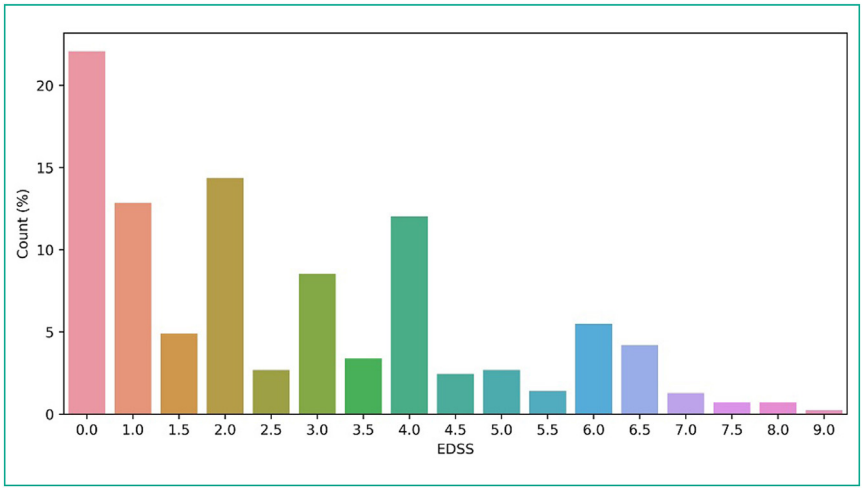


**Fig. 4.** Histogram of expanded disability status scale scores in the training set, reflecting the unbalanced nature of EDSS distribution. There are more low scores than high scores (81% of EDSS scores inferior or equal to 4) and more than 22% of the samples correspond to an EDSS score of 0. Integer (1, 2, 3, etc.) scores are more represented (75% of EDSS > 0) than non-integer (1.5, 2.5, etc.) scores (25%), this could be due to a human bias towards integer scores when scoring. EDSS: expanded disability status scale.

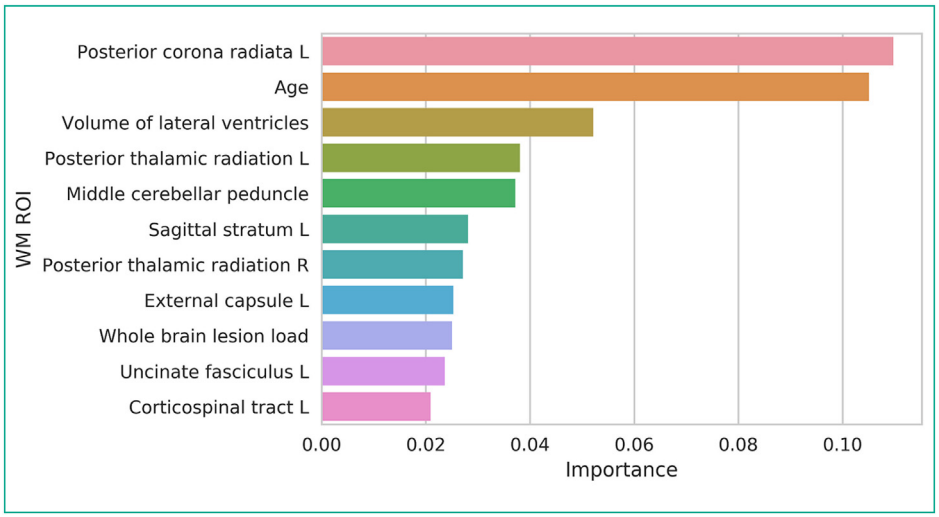


**Fig. 5.** Diagram shows the most informative features associated with the random forest single predictor model. "L" and "R" mean "Left" and "Right" respectively. WM: white matter; ROI: region of interest.

**Table 2**
Mean square error of each model with the validation set.

| Method | MSE on the validation set | MSE on the test set |
|---|---|---|
| Age-only ridge regression | 3.779 | Unknown |
| CNN with linear registration | 2.705 | Unknown |
| CNN with nonlinear registration | 2.714 | Unknown |
| Random forest dual | 2.560 | Unknown |
| Random forest single | 2.697 | Unknown |
| Manifold | 3.216 | Unknown |
| Aggregated model | 2.210 | 3 |

CNN: Convolutional Neural Network; MSE: mean square error.

MSE of 2.56. The aggregated model reached a MSE of 2.21. For comparison purposes, the Ridge regression model based on age only had a MSE of 3.8. Fig. 6 shows the example of two patients with MS for whom the aggregated model correctly predicted the EDSS scores while the Ridge regression model did not. These two patients (A and B) were 45- and 55-year-old, respectively and had close imaging characteristics at baseline but a different EDSS at two years (3 and 6.5 respectively). Our quantitative image analysis revealed differences between the two patients in terms of volume of lateral ventricles (60 mL and 84 mL for A and B respectively) and left posterior corona radiata lesion load (33% and 48% for A and B respectively) (Fig. 6). We obtained the best predic-

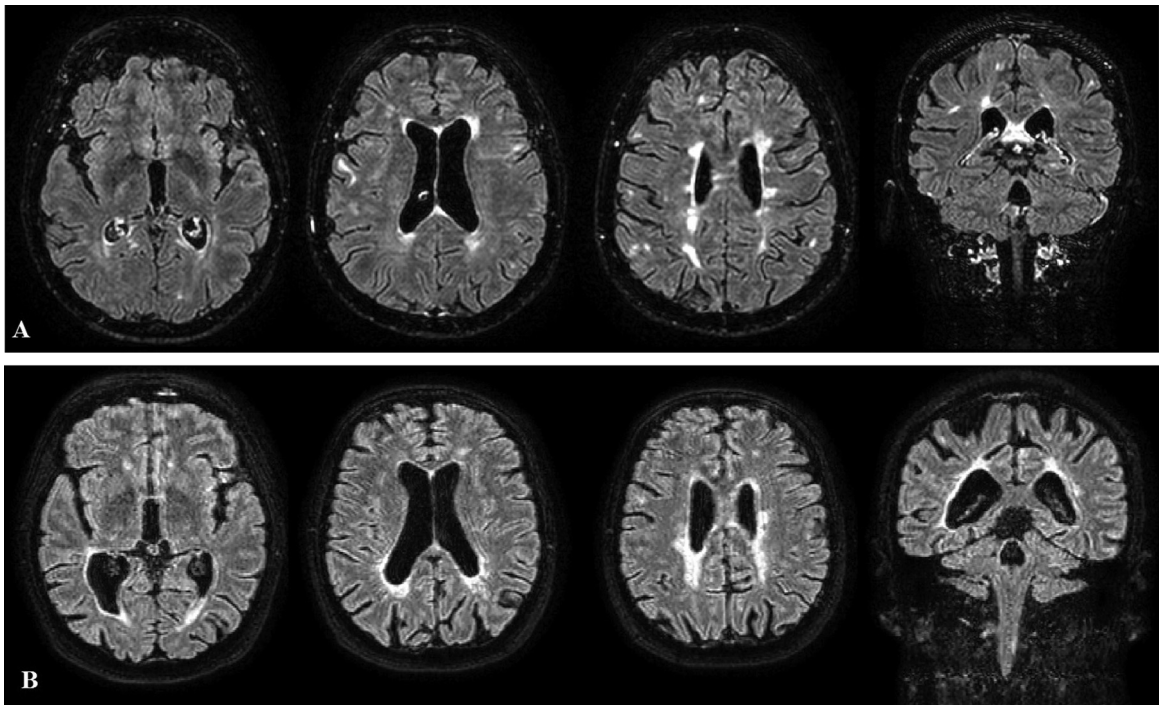


**Fig. 6.** FLAIR images of two patients with multiple sclerosis (MS) for which the aggregated model correctly predicted the expanded disability status scale scores while the Ridge regression model using age did not. A. 46-year-old woman with MS, volume of lateral ventricles = 60 mL, left posterior corona radiata lesion load = 2.5 mL (33%), EDSS at two years = 3. B. 55-year-old man with MS, volume of lateral ventricles = 84 mL, left posterior corona radiata lesion load = 3.6 (48%), EDSS at two years = 6.5. The age-only Ridge regression model predicted an EDSS of 3 and 3.5 for A and B respectively.

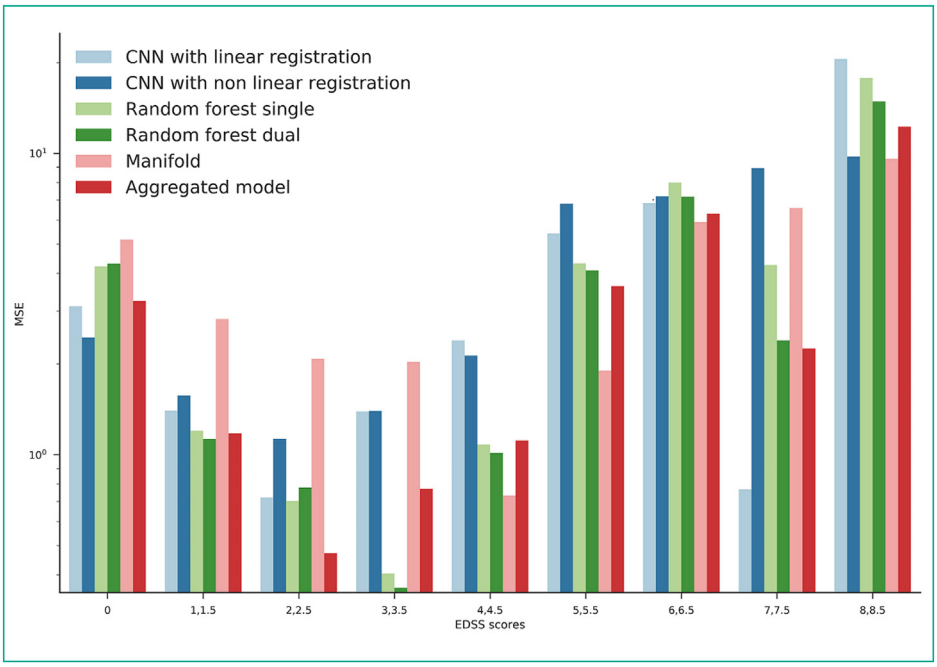


**Fig. 7.** Graph shows results of each model on the validation set: MSE for each EDSS group. The log-scale was used in order to facilitate the visualization. CNN: convolutional neural network; EDSS: expanded disability status scale; MSE: mean square error.

tion for middle EDSS scores (predictors presenting a MSE < 1.1 when 1 < EDSS score < 4.5), and the worst prediction for high EDSS scores (MSE > 9.6 for EDSS score ≥ 8). For EDSS score = 0, no model performed particularly well (MSE superior to 2.4 for all models), despite the relatively large number of training examples ($n$ = 189 corresponding to 22% of the training set). Fig. 7 shows the high variability of model performances across EDSS scores of the validation set.

## 4. Discussion

Our method can predict two-year clinical disability with a mean square EDSS score error of 3 only based on a single, baseline, routine FLAIR MRI examination with some basic clinical information, with heterogeneous imaging quality from various MRI equipments and centers. The most informative variables were the age, the volume of the lateral ventricles, and the lesion load in main white matter tracts such as corona radiata, thalamic radiation, and cerebellar peduncle.

The aggregation of complementary predictor models, through a weighted average taking into account prediction errors for different EDSS score ranges, allowed us to benefit from the strength of each predictor. Indeed, not a single predictor performs well on each one of the EDSS scores. On the contrary, the best predictor varies across EDSS score groups and as expected, the aggregated model presented shows improved performance compared to the best individual predictor. The features characterizing the impact of lesions on the brain network (using tract-based lesion loads) demonstrate their usefulness over features learned using a pure image-based deep learning approach for middle EDSS scores, reaching a very low MSE < 0.36 for EDSS of 3 and 3.5 on the validation set. We also achieve better prediction accuracy (MSE = 2.2) on this dataset compared to an age-only Ridge regression model (MSE = 3.8), highlighting the importance of imaging features in the prediction. Further studies including quantitative metrics coming from T1-weighted-based segmentation could be interesting to understand the influence of atrophy on the clinical disability.

Our study has some limitations. First, our aggregated model had difficulty to predict EDSS score = 0 at two years. The injection of a priori anatomical knowledge on brain connections was not sufficient to overcome the clinico-radiological paradox for these patients. This could be due to various factors including intra- and inter-variability when scoring EDSS, particularly with low scores [26], underestimation of damage to the normal-appearing brain tissue, neglect of spinal cord involvement, or masking effect of cortical plasticity. Second, for EDSS scores > 8, the aggregated model reached an MSE over 9 on the validation set. This result could be explained by the unbalanced EDSS score distribution of the training set which presented a limited number of examples of these high EDSS scores. A training session on a larger cohort of patients could overcome this limitation or different solutions could be tested to artificially increase the number of high EDSS scores during the training such as oversampling of high EDSS score examples or generating synthetic data. Last, the initial EDSS scores (associated with the baseline MRI examination) were not available during this challenge, thus making it impossible to estimate clinical disability progression over the follow-up. In addition, we received no information about patient treatment, so we were not able to study therapeutic effects on our disability prediction.

In conclusion, our model helps predict the EDSS score at two years for patients affected by MS by relying solely on a single FLAIR sequence and basic demographic information. This performance was achieved through the combination of multiple predictors based on images, anatomical priors, and white matter lesion load using MRI from multiple clinical centers. These promising results should be further validated on an external larger test cohort and have the potential to be highly relevant for disability prediction and the evaluation of disease-modifying treatments. This supports the use of our model to predict EDSS score progression and/or the improvement of the current prediction using additional factors such as baseline EDSS score.

## Human and animal rights

The authors declare that the work described has been carried out in accordance with the Declaration of Helsinki of the World Medical Association revised in 2013 for experiments involving humans.

## Informed consent and patient details

The authors declare that this report does not contain any personal information that could lead to the identification of the patient(s).

The authors declare that they obtained a written informed consent from the patients and/or volunteers included in the article. The authors also confirm that the personal details of the patients and/or volunteers have been removed.

## Funding

This work did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Author contributions

All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship.

## CRediT authorship contribution statement

Pauline Roca: conceptualization, data curation, formal analysis, methodology, writing- original draft, writing- review & editing; Arnaud Attyé: investigation, writing- original draft, writing- review & editing; lucie colas: resources, writing- review & editing; Alan Tucholka: conceptualization, data curation, methodology, supervision, writing- original draft, writing- review & editing; Pascal Rubini: conceptualization, data curation, formal analysis, methodology, software, writing- original draft, writing- review & editing; Stenzel Cackowski: formal analysis, writing- original draft, writing- review & editing; Juliette Ding: resources, writing- review & editing; Jean-François Budzik: resources, writing- review & editing; Felix Renard: formal analysis, writing- original draft, writing- review & editing; Senan Doyle: resources, writing- review & editing; Emmanuel L. Barbier: resources, writing- review & editing; Imad Bousaid: investigation; Romain Casey: investigation, resources, writing- review & editing; Sandra Vukusic: investigation, resources, writing- review & editing; Nathalie Lassau: investigation, resources, writing- review & editing; Sébastien Verclytte: conceptualization, investigation, resources, writing- original draft, writing- review & editing; and François Cotton: investigation, resources, writing- review & editing.

## Acknowledgements

This work was conducted using data from the Observatoire Français de la Sclérose en Plaques (OFSEP) which is supported by a grant provided by the French State and handled by the “Agence Nationale de la Recherche” within the framework of the “Investments for the Future” program, under the reference ANR-10-COHO-002, by the Eugène Devic EDMUS Foundation against multiple sclerosis and by the ARSEP Foundation.”

## Disclosure of interest

Pauline Roca, Alan Tucholka, and Pascal Rubini are employees at Pixyl. Arnaud Attyé is a part-time consultant at Pixyl. Felix Renard has a grant from Carnot-LSI for work unrelated to the contents of this manuscript. Lucie Colas, Stenzel Cackowski, Juliette Ding, Jean-François Budzik, Emmanuel Barbier, Imad Bousaid, Romain Casey, Sandra Vukusic, Nathalie Lassau, Sébastien Verclytte, and François Cotton declare that they have no competing interest.

## References

[1] Gustavsson A, Svensson M, Jacobi F, Allgulander C, Alonso J, Beghi E, et al. Cost of disorders of the brain in Europe 2010. Eur Neuropsychopharmacol 2011;21:718–79.
[2] Chen AY, Chonghasawat AO, Leadholm KL. Multiple sclerosis: frequency, cost, and economic burden in the United States. J Clin Neurosci 2017;45:180–6.
[3] Filippi M. Multiple sclerosis. Nat Rev Dis Primer 2018;4:27.
[4] Pittock SJ, Mayr WT, McClelland RL, Jorgensen NW, Weigand SD, Noseworthy JH, et al. Change in MS-related disability in a population-based cohort: a 10-year follow-up study. Neurology 2004;62:51–9.
[5] Jokubaitis VG, Spelman T, Kalincik T, Lorscheider J, Havrdova E, Horakova D, et al. Predictors of long-term disability accrual in relapse-onset multiple sclerosis. Ann Neurol 2016;80:89–100.
[6] Pellegrini F, Copetti M, Sormani MP, Bovis F, de Moor C, Debray TP, et al. Predicting disability progression in multiple sclerosis: insights from advanced statistical modeling. Mult Scler J 2019 [135245851988734].
[7] Waymel Q, Badr S, Demondion X, Cotten A, Jacques T. Impact of the rise of artificial intelligence in radiology: what do radiologists think? Diagn Interv Imaging 2019;100:327–36.
[8] Herent P, Schmauch B, Jehanno P, Dehaene O, Saillard C, Balleyguier C, et al. Detection and characterization of MRI breast lesions using deep learning. Diagn Interv Imaging 2019;100:219–25.
[9] Couteaux V, Si-Mohamed S, Nempont O, Lefevre T, Popoff A, Pizaine G, et al. Automatic knee meniscus tear detection and orientation classification with Mask-RCNN. Diagn Interv Imaging 2019;100:235–42.
[10] Lassau N, Estienne T, de Vomecourt P, Azoulay M, Cagnol J, Garcia G, et al. Five simultaneous artificial intelligence data challenges on ultrasound, CT, and MRI. Diagn Interv Imaging 2019;100:199–209.
[11] Group SFR-IA, French Radiology Community CERF. Artificial intelligence and medical imaging 2018: French Radiology Community white paper. Diagn Interv Imaging 2018;99:727–42.
[12] Mollison D, Sellar R, Bastin M, Mollison D, Chandran S, Wardlaw J, et al. The clinico-radiological paradox of cognitive function and MRI burden of white matter lesions in people with multiple sclerosis: a systematic review and meta-analysis. PloS One 2017;12:e0177727.
[13] Altermatt A, Gaetano L, Magon S, et al. Clinical correlations of brain lesion location in multiple sclerosis: voxel-based analysis of a large clinical trial dataset. Brain Topogr 2018;31:886–94.
[14] Barkhof F. MRI in multiple sclerosis: correlation with expanded disability status scale (EDSS). Mult Scler J 1999;5:283–6.
[15] Rokach L. Ensemble-based classifiers. Artif Intell Rev 2010;33:1–39.
[16] Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 2010;29:1310–20.
[17] Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 2011;54:2033–44.
[18] Avants BB, Tustison N, Song G. Advanced normalization tools (ANTS). Insight J 2009;2:1–35.
[19] Mori S, Wakana S, Van Zijl PC, Nagae-Poetscher LM. MRI atlas of human white matter. Elsevier; 2005.
[20] Wakana S, Caprihan A, Panzenboeck MM, Fallon JH, Perry M, Gollub RL, et al. Reproducibility of quantitative tractography methods applied to cerebral white matter. Neuroimage 2007;36:630–44.
[21] Hua K, Zhang J, Wakana S, Jiang H, Li X, Reich DS, et al. Tract probability maps in stereotaxic spaces: analyses of white matter anatomy and tract-specific quantification. Neuroimage 2008;39:336–47.
[22] Archer DB, Vaillancour DE, Coombes SA. A template and probabilistic atlas of the human sensorimotor tracts using diffusion MRI. Cereb Cortex 2018;28:1685–99.
[23] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12:2825–30.
[24] Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. 2nd ed New York: Springer series in statistics; 2001.
[25] Sánchez-Rico M, Alvarado JM. A machine learning approach for studying the comorbidities of complex diagnoses. Behav Sci 2019;9:122.
[26] Goodkin DE, Cookfair D, Wende K, Bourdette D, Pullicino P, Scherokman B, et al. Inter-and intrarater scoring agreement using grades 1.0 to 3.5 of the Kurtzke expanded disability status scale (EDSS). Neurology 1992;42:859.