A COMBINED RADIOMICS AND MACHINE LEARNING APPROACH TO OVECOME THE CLINICO-RADIOLOGICAL PARADOX IN MULTIPLE SCLEROSIS

MATERIALS AND METHODS

MRI data acquisition

Site 1 (MS Center of the University of Naples "Federico II")

3T Magnetom Trio scanner (Siemens Healthineers), equipped with an 8-channel head coil, with the following protocols:

Protocol 1 (361 subjects)

- 3D T1-weighted Magnetization Prepared Rapid Acquisition Gradient Echo sequence (MPRAGE; TR=1900 ms; TE=3.4 ms; TI=900 ms; Flip Angle=9°; resolution=1x1x1 mm^3; 160 axial slices)

- 2D T2-weighted Fluid Attenuated Inversion Recovery sequence (FLAIR; TR=8500 ms; TE=106 ms; TI=2500 ms; Flip Angle=150°; voxel size=0.9x0.9x4 mm^3; 25 axial slices)

Protocol 2 (102 subjects)

- 3D T1-weighted Magnetization Prepared Rapid Acquisition Gradient Echo sequence (MPRAGE; TR=2500 ms; TE=2.8 ms; TI=900 ms; Flip Angle=9°; resolution=1x1x1 mm^3; 160 axial slices)

- 3D T2-weighted Fluid Attenuated Inversion Recovery sequence (FLAIR; TR=6000 ms; TE=396 ms; TI=2200 ms; Flip Angle=120°; voxel size=1x1x1 mm^3; 160 sagittal slices)

Protocol 3 (37 subjects)

- 3D T1-weighted Magnetization Prepared Rapid Acquisition Gradient Echo sequence (MPRAGE; TR=3000 ms; TE=2.4 ms; TI=1000 ms; Flip Angle=7°; resolution=0.8x0.8x0.8 mm^3; 224 sagittal slices)

- 3D T2-weighted Fluid Attenuated Inversion Recovery sequence (FLAIR; TR=6000 ms; TE=404 ms; TI=2200 ms; Flip Angle=120°; voxel size=1x1x1 mm^3; 160 sagittal slices)

<u>Site 2 (Human Neuroscience Department of the University of Rome "Sapienza")</u>

3T Magnetom Verio scanner (Siemens Healthineers), equipped with a 12-channel head coil, with the following protocol:

- 3D T1-weighted Magnetization Prepared Rapid Acquisition Gradient Echo sequence (MPRAGE; TR=1900 ms; TE=2.93 ms; TI=900 ms; Flip Angle=9°; resolution=0.5x0.5x1 mm^3;176 sagittal slices)

- 2D T2-weighted Fluid Attenuated Inversion Recovery sequence (FLAIR; TR=9000 ms; TE=94 ms; TI=2500 ms; Flip Angle=150°; voxel size=0.5x0.5x5 mm^3; 25 axial slices).

MRI data processing

Initially, in order to take into account possible differences in terms of spatial resolution and orientation, T1-weighted and FLAIR images were automatically reoriented and resampled to 1mm isotropic resolution by rigidly aligning them to corresponding templates in the MNI space using the Statistical Parametric Mapping software package (SPM12, http://www.fil.ion.ucl.ac.uk/spm).

Demyelinating lesions were automatically segmented on FLAIR images using the lesion prediction algorithm[1] implemented in the Lesion Segmentation Tool (LST) toolbox v3.0.0 (www.statistical-modelling.de/lst.html) for SPM. Lesion probability maps where then used to fill

lesions in T1-weighted images for subsequent processing steps via LST's default lesion filling procedure, and binarized (thresholding at 0.5 probability) to compute T2-LL.

Filled T1-weighted volumes were processed via the segmentation pipeline implemented in the Computational Anatomy Toolbox (CAT12.6, http://www.neuro.uni-jena.de/cat) for SPM, using the default settings (http://dbm.neuro.uni-jena.de/cat12/CAT12-Manual.pdf), with extended tissue priors to ensure better classification of subcortical brain structures[2] and atlas-based parcellation of native space images into 114 brain regions defined according to an adapted version of the Automated Anatomical Labeling (AAL) atlas[3] implemented in CAT12[1]. Subsequently, whole brain volume (WBV) was computed as the sum of GM and WM binary tissue maps and GM subregions ROIs (and corresponding volumes) were obtained as the intersection between each atlas-based parcel and GM binary mask. Furthermore, WM binary map was used to obtain a normal-appearing white matter (NAWM) mask by subtracting the binary lesion map. As a quality check, the so obtained masks were visually inspected by an experienced neuroradiologist (M.Q., with more than 20 years of experience in the field of neuroimaging) to assess the accuracy of the segmentation procedure.

Finally, for each participant, total intracranial volume (TIV) was estimated using CAT12 standard procedure and brain volumes (both WBV and GM regions) were transformed into z-scores while adjusting for age, sex and TIV in order to correct for the effect of healthy aging, sex and head size.

Connectivity analysis

[1] The version of the AAL atlas implemented in CAT12 slightly differs from the original one, with cerebellar Crus I and II (both right and left) considered as a single region, thus resulting in a total of 114 (vs 116) brain parcels.

Subject-wise, for each of the 116 GM cortical/subcortical region defined in the AAL atlas, a Change in Connectivity (ChaCo) score was computed using the network Modification (NeMo) tool[4], representing an estimate of local structural disconnection caused by WM tracts disruption, as inferred from the location and load of WM lesions. Briefly, each WM lesion mask was transformed into MNI-space and referenced to a collection of 73 healthy controls whole brain tractograms in standard space to calculate a ChaCo score for each GM region, corresponding to the proportion of streamlines connecting that region that pass through the lesion mask and are therefore considered disrupted[4].

Radiomics analysis

First order and texture features were extracted from each ROI (NAWM and 114 GM regions) from the unfilled, bias field-corrected and intensity-normalized T1-weighted volumes using PyRadiomics v3.0[5]. Prior to the extraction, the images were preprocessed as follows: grey level normalization to a 0-600 range, resampling to 1x1x1 mm, ROI precrop with a 10 voxel padding for Laplacian of Gaussian (LoG) image filtering, grey level discretization (bin width= 3). All available features were obtained from the original as well as LoG (sigma= 1, 3, 5) and wavelet-filtered (all combinations of high and low pass filters on the three axes) images, to maximize information extraction. A detailed description of the radiomic features obtainable by PyRadiomics is available in the official documentation (https://pyradiomics.readthedocs.io/en/latest/features.html).

Radiomics feature stability with respect to the MRI processing pipeline was tested on a subset of 30 randomly selected subjects, on whom the entire preprocessing and extraction process was repeated three times. The intraclass correlation coefficient (ICC) was then calculated for each feature using a two-way random effect, single rater, absolute agreement model[6]. Only features with excellent stability (ICC 95% CI lower bound$\geq$ 0.90) were retained for subsequent analyses.

Machine Learning

Machine learning analyses were performed using the Weka data mining platform (v3.8.3)[7] and scikit-learn Python package[8]. Given the nature of the EDSS score, regression algorithms (Ridge Regression, Support Vector Machine, Random Forest, and Gaussian Process) were used to develop predictive models. A linear and variety of non-linear algorithms were investigated to assess differences in performance due to model architecture. Ridge Regression is a variant of multiple regression that takes into account feature collinearity by adding a degree of bias to regression estimates. A regression Support Vector Machine uses hyperplane maximal margin as its guiding principle. Compared to Ridge Regression, it can be non-linear if an appropriate kernel (e.g. Radial Basis Function) is used. A Random Forest Regression is another nonlinear model, based on bootstrap aggregation of data used to train a large number of decision trees. A Gaussian Process regressor is a third nonlinear algorithm. It is non-sparse (i.e. requires the entire train set information to perform the prediction) and performs a probabilistic (Gaussian) prediction new data.

The Site 1 cohort was randomly split in training (80% of subjects) and test (20% of subjects) sets for model tuning and testing, respectively, while the Site 2 cohort was exclusively used as an external test set. A MinMax standardization scaler (0-1 range) was fit on the numerical features of the training data and used to transform both training and test sets, as to avoid any information leak from the first to the latter. Categorical variables (k values) were converted to k-1 indicator Boolean ones.

On the training set, clinico-demographic (age, sex, disease duration, DD, disease course), textural and other MRI-derived (T2-LL, WBV, volumes and ChaCo scores for each GM region) variables underwent multiple feature selection steps after the above-mentioned removal of unstable features. First, low variance (0.01 threshold) parameters were removed, as they can be considered as not informative. Similarly, after calculating a pairwise correlation matrix, highly colinear (> 0.8) features were removed. Then, LASSO regression, using

the EDSS score as the dependent variable, was used to remove features whose coefficients shrank to 0. Finally, the Weka correlation-based subset evaluator was employed to identify the best feature subset for EDSS score prediction.

The resulting dataset was used to train the four ML regression algorithms, whose tuning and initial performance evaluation was performed using 10-fold cross-validation in the training cohort. Each final model was then assessed on the previously unseen cases of both the internal and external test sets.

RESULTS

MRI data analysis.

Visual inspection of the automatically obtained ROIs revealed high accuracy of the segmentation procedure, with no need to drop subjects or manually adjust segmentation masks due to image processing errors.

ML predictive models

Ridge Regression

weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4

weights:

-0.9817 * 6_gm1x1_wavelet-HLL_firstorder_Median +

-0.9857 * 78_gm1x1_original_firstorder_Energy +

-0.7439 * 109_gm1x1_wavelet-HLL_glszm_SizeZoneNonUniformity +

-1.3637 * 71_gm1x1_wavelet-HLL_glcm_Imc1 +

0.6393 * 101_gm1x1_log-sigma-1-0-mm-3D_gldm_SmallDependenceLowGrayLevelEmphasis +

-1.0431 * 91_gm1x1_log-sigma-1-0-mm-3D_firstorder_Median +

0.5635 * 41_gm1x1_log-sigma-1-0-mm-3D_glcm_Correlation +

0.8914 * Age +

1.8747 * Course_SP +

5.1675

ROIs anatomical labels (according to[3]): 6, Right Frontal Superior Orbital Cortex; 41, Left Amygdala; 71, Left Caudate Nucleus; 78, Right Thalamus; 91, Left Cerebellar Crus; 101, Left Cerebellar Lobule VIII; 109, Cerebellar Vermis (Lobules IV-V).

Gaussian Process

weka.classifiers.functions.GaussianProcesses -L 1.0 -N 2 -K "weka.classifiers.functions.supportVector.RBFKernel -C 250007 -G 0.01" -S 1

Support Vector Machine

weka.classifiers.functions.SMOreg -C 2.0 -N 2 -I "weka.classifiers.functions.supportVector.RegSMOImproved -T 0.001 -V -P 1.0E-12 -L 0.001 -W 1" -K "weka.classifiers.functions.supportVector.RBFKernel -C 250007 -G 0.01"

Random Forest

weka.classifiers.trees.RandomForest -P 80 -attribute-importance -I 400 -num-slots 1 -K 3 -M 1.0 -V 0.001 -S 1

TABLES

Table 1. Results of the post-hoc pairwise comparisons between different models' MAE in the internal test set.

Model (I)	Model (J)	Mean Difference (I-J)	Standard Error	*p*-value*	95% CI for Difference*	
					Lower Bound	Upper Bound
RR	**GP**	-0.149	0.042	0.004	-0.263	-0.036
	SVM	-0.029	0.033	1.000	-0.119	0.061
	RF	-0.016	0.026	1.000	-0.085	0.054
GP	**RR**	0.149	0.042	0.004	0.036	0.263
	SVM	0.120	0.022	0.000	0.062	0.178
	RF	0.133	0.043	0.014	0.018	0.249
SVM	**RR**	0.029	0.033	1.000	-0.061	0.119
	GP	-0.120	0.022	0.000	-0.178	-0.062
	RF	0.014	0.034	1.000	-0.078	0.106
RF	**RR**	0.016	0.026	1.000	-0.054	0.085
	GP	-0.133	0.043	0.014	-0.249	-0.018
	SVM	-0.014	0.034	1.000	-0.106	0.078

Based on estimated marginal means.

*Adjustment for multiple comparisons: Bonferroni.

MAE: mean absolute error; CI: Confidence Interval; RR: Ridge Regression; GP: Gaussian Process; SVM: Support Vector Machine; RF: Random Forest.

Table 2. Results of the post-hoc pairwise comparisons between different models' MAE in the external test set.

Model (I)	Model (J)	Mean Difference (I-J)	Standard Error	*p*-value*	95% CI for Difference*	
					Lower Bound	Upper Bound
RR	**GP**	-0.092	0.051	0.424	-0.229	0.044
	SVM	0.043	0.036	1.000	-0.055	0.141
	RF	-0.007	0.035	1.000	-0.101	0.086
GP	**RR**	0.092	0.051	0.424	-0.044	0.229
	SVM	0.136	0.020	0.000	0.081	0.190
	RF	0.085	0.035	0.095	-0.008	0.179
SVM	**RR**	-0.043	0.036	1.000	-0.141	0.055
	GP	-0.136	0.020	0.000	-0.190	-0.081
	RF	-0.050	0.026	0.307	-0.119	0.018
RF	**RR**	0.007	0.035	1.000	-0.086	0.101
	GP	-0.085	0.035	0.095	-0.179	0.008
	SVM	0.050	0.026	0.307	-0.018	0.119

Based on estimated marginal means.

*Adjustment for multiple comparisons: Bonferroni.

MAE: mean absolute error; CI: Confidence Interval; RR: Ridge Regression; GP: Gaussian Process; SVM: Support Vector Machine; RF: Random Forest.

Table 3. Machine Learning predictive models based solely on clinico-demographic features (i.e. age and secondary progressive course). Performances of the distinct algorithms for the prediction of EDSS score in different subsets of patients are presented, along with the results of the one-way repeated measures ANOVA analysis comparing absolute errors.

Cohort	*Ridge Regression*			*Gaussian Process*			*Support Vector Machine*			*Random Forest*			*p*-value
	r	**R^2**	**MAE**	***r***	**R^2**	**MAE**	***r***	**R^2**	**MAE**	***r***	**R^2**	**MAE**	
Training	0.715	0.511	0.754	0.644	0.414	0.876	0.714	0.510	0.748	0.610	0.372	1.130	-
Internal Test	0.626	0.391	0.834	0.539	0.291	0.997	0.626	0.392	0.818	0.550	0.303	0.881	<0.001*
External Test	0.813	0.660	1.005	0.780	0.609	1.203	0.814	0.663	0.945	0.682	0.464	1.135	<0.001**

*$F(1.78, 176.06) = 20.14$; Partial $\eta^2 = 0.17$. DF were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.59$).
**$F(1.72, 176.74) = 14.13$; Partial $\eta^2 = 0.12$. DF were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 57$).

MAE: mean absolute error; DF: degrees of freedom.

References

1. Schmidt P. Bayesian inference for structured additive regression models for large-scale problems with applications to medical imaging. 2017

2. Lorio S, Fresard S, Adaszewski S, et al. New tissue priors for improved automated classification of subcortical brain structures on MRI. *NeuroImage* 2016;130:157-166

3. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* 2002;15:273-289

4. Kuceyeski A, Maruta J, Relkin N, et al. The Network Modification (NeMo) Tool: elucidating the effect of white matter integrity changes on cortical and subcortical structural connectivity. *Brain connectivity* 2013;3:451-463

5. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer research* 2017;77:e104-e107

6. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of chiropractic medicine* 2016;15:155-163

7. Frank E, Hall M, Trigg L, et al. Data mining in bioinformatics using Weka. *Bioinformatics* 2004;20:2479-2481

8. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 2011;12:2825–2830