A comparison of Machine Learning approaches for classifying Multiple Sclerosis courses using MRSI and brain segmentations

Adrian Ion-Mărgineanu[1,2,3], Gabriel Kocevar[1], Claudio Stamile[1,2,3], Diana M Sima[2,3,4], Françoise Durand-Dubief[1,5], Sabine Van Huffel[2,3], and Dominique Sappey-Marinier[1,6]

[1] CREATIS CNRS UMR5220 & INSERM U1206; Université de Lyon, Université Claude Bernard-Lyon 1, INSA-Lyon, Villeurbanne, France
[2] KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Leuven, Belgium
[3] imec, Leuven, Belgium
[4] ico**metrix**, R&D department, Leuven, Belgium
[5] Service de Neurologie A, Hôpital Neurologique, Hospices Civils de Lyon, Bron, France
[6] CERMEP - Imagerie du Vivant, Université de Lyon, Bron, France
adrian@esat.kuleuven.be

Abstract. The objective of this paper is to classify Multiple Sclerosis courses using features extracted from Magnetic Resonance Spectroscopic Imaging (MRSI) combined with brain tissue segmentations of gray matter, white matter, and lesions. To this purpose we trained several classifiers, ranging from simple (i.e. Linear Discriminant Analysis) to state-of-the-art (i.e. Convolutional Neural Networks). We investigate four binary classification tasks and report maximum values of Area Under receiver operating characteristic Curve between 68% and 95%. Our best results were found after training Support Vector Machines with gaussian kernel on MRSI features combined with brain tissue segmentation features.

Keywords: machine learning, convolutional neural networks, multiple sclerosis, magnetic resonance spectroscopic imaging, brain segmentation

1 Introduction

Multiple sclerosis (MS) is an inflammatory disorder of the brain and spinal cord [1], affecting approximately 2.5 million people worldwide.

The majority of MS patients (85%) usually experience a first attack defined as Clinically Isolated Syndrome (CIS), and will develop a relapsing-remitting (RR) form [2]. Two thirds of the RR patients will develop a secondary progressive (SP) form, while the other third will follow a benign course [3]. The rest of MS patients (15%) will start directly with a primary progressive (PP) form.

The criteria to diagnose MS forms were originally formulated by McDonald in 2001 [4] and revised by Polman in 2005 [5] and 2011 [6]. They all rely on using

conventional magnetic resonance imaging techniques (MRI), such as T1 and FLAIR, due to high sensitivity in visualizing MS lesions. More recently [7], ^{1}H-Magnetic Resonance Spectroscopic Imaging (MRSI) has been shown to provide a better understanding of the pathological mechanisms of MS.

The objective of this study is to fully explore the potential of MRSI for automatic classification of MS courses. To this purpose we use four different machine learning approaches to classify individual spectroscopic voxels inside the brain. We start by using simple machine learning methods (i.e. Linear Discriminant Analysis (LDA)) trained on low-level features commonly used in MRSI, and advance up to state-of-the-art methods (e.g. Convolutional Neural Networks (CNN)) trained on high-level MRSI features.

2 Materials and Methods

2.1 Patient population

This longitudinal study includes 87 MS patients who were scanned multiple times over several years between 2006 and 2012. Diagnosis and disease course were established according to the McDonald criteria [4, 8]. This study was approved by the local ethics committee (CPP Sud-Est IV) and the French national agency for medicine and health products safety (ANSM), and written informed consents were obtained from all patients prior to study initiation. More details for each MS group can be found in Table 1.

	CIS	RR	PP	SP
Number of patients	12	30	17	28
Total number of scans	60	212	117	192
Total number of voxels	5916	18682	10830	17377

Table 1. MS population details

2.2 Magnetic Resonance data acquisition and processing

All patients underwent magnetic resonance (MR) examination using a 1.5 Tesla MR system (Sonata Siemens, Erlangen, Germany) and an 8 elements phased-array head-coil.

MRI acquisition Conventional MRI protocol consisted of a 3 dimensional T1-weighted (magnetization prepared rapid gradient echo-MPRAGE) sequence with repetition time/echo time/time for inversion TR/TE/TI=1970/3.93/1100 ms, flip angle=15°, matrix size=256×256, field of view (FOV)=256×256mm, slice thickness=1mm, voxel size=1×1×1mm, and a fluid attenuated inversion recovery (FLAIR) sequence with TR/TE/TI=8000/105/2200ms, flip angle=150°, matrix size= 192×256, FOV=240×240 mm, slice thickness=3mm, voxel size=0.9 × 0.9 × 3mm.

MRSI acquisition MRSI data was acquired from one slice of 1.5 cm thickness, placed above the corpus callosum and along the anterior commissure - posterior commissure (AC-PC) axis, encompassing the centrum semioval region. A point-resolved spectroscopic sequence (PRESS) with TR/TE=1690/135ms was used to select a volume of interest (VOI) of $105\times105\times15\text{mm}^3$ during the acquisition of 24×24 (interpolated to 32×32) phase-encodings over a FOV of 240×240 mm^2.

MRI processing Three tissues of the brain, gray matter (GM), white matter (WM), and lesions, were segmented based on T1 and FLAIR, using the MSmetrix software [9] developed by ico**metrix** (Leuven, Belgium).

MRSI processing MRSI data processing was performed using SPID [10] in MatLab 2015a (MathWorks, Natick, MA, USA). Three metabolites well-studied in MS, *N*-acetyl-aspartate (NAA), Choline (Cho), and Creatine (Cre), were quantified with AQSES [10](**A**utomated **Q**uantitation of **S**hort **E**cho time MR **S**pectra), using a synthetic basis set which incorporates prior knowledge of the individual metabolites. Maximum-phase finite impulse response filtering was included in the AQSES procedure for residual water suppression, with a filter length of 50 and spectral range from 1.7 to 4.2 ppm.

Quality control First, we removed a band of two voxels at the outer edges of each VOI in order to avoid chemical shift displacement artifacts and lipid contamination artifacts. Second, for each voxel inside a grid, we performed three outlier detections, corresponding to each metabolite, using the median absolute deviation filtering. Final selection includes voxels with a maximum Cramer Rao Lower Bound of 20% for each metabolite, preserved by all three outlier detection mechanisms. In the end, average voxel exclusion rate was 31% $\pm$ 6% standard deviation, and only 2 out of 581 spectroscopy grids had an exclusion rate higher than 50%.

2.3 Classification tasks and performance measures

We study four binary classification tasks, relevant from a clinical point of view: CIS vs. RR, CIS vs. PP, RR vs. PP, and RR vs. SP. For each task we set the less represented class between the two to be the positive class, or the class of interest. Therefore, we set the positive class to CIS, CIS, PP, and SP, corresponding to each task. When classifying, we perform a 2-fold stratified cross-validation at the patient level, meaning that each patient will be assigned once to training, and once to testing. The training dataset includes all voxels from all patients assigned to training. When testing, a voxel will be assigned to one of the two classes. For each grid, we compute the probability to be assigned to the positive class by measuring the percentage of voxels assigned to the positive class.

We compute and report three performance measures widely used in classification: AUC (**A**rea **U**nder receiver operating characteristic (ROC) **C**urve),

sensitivity, and specificity. The last two measures were computed for the optimal operating point of the ROC curve. Using the general formulation of the confu-

Confusion matrix		predicted condition	
		predicted negative	predicted positive
true condition	condition negative	True Negative (TN)	False Positive (FP)
	condition positive	False Negative (FN)	True Positive (TP)

Table 2. General confusion matrix.

sion matrix from Table 2, sensitivity, or true positive rate (TPR), is defined as $\frac{TP}{TP+FN}$. Specificity, or true negative rate (TNR), is defined as $\frac{TN}{TN+FP}$.

The ROC curve can be created when the classification model gives probability values of test points belonging to the positive class, by plotting Sensitivity (y-axis) against 1-Specificity (x-axis) at various probability thresholds. A random classifier has an AUC of 0.5 or 50%, while a perfect classifier will have an AUC of 1 or 100%.

2.4 Feature extraction models

Model nr.1 (M1) We use the absolute values of the complex frequency spectrum cut by a pass-band filter between 1.2 and 4.2 ppm, so that we retain the most useful information. In order to have a perfect alignment of all spectra for all patients, we detect the highest peak in the low frequencies (NAA) and shift to the NAA peak of a randomly assigned reference voxel. In this case, each voxel is represented by the filtered frequency vector, which has 81 points. We normalize each vector to its L_2-norm.

Model nr.2 (M2) We use the three quantified metabolite concentrations (NAA, Cho, Cre) to compute three ratios: NAA/Cho, NAA/Cre, and Cho/Cre. Mean values and standard deviations for each MS group can be found in Table 3.

	CIS	RR	PP	SP
NAA/Cho	2.21 (0.24)	2.02 (0.25)	1.83 (0.18)	1.86 (0.32)
NAA/Cre	1.36 (0.1)	1.35 (0.11)	1.27 (0.11)	1.22 (0.12)
Cho/Cre	0.63 (0.07)	0.69 (0.08)	0.72 (0.1)	0.69 (0.1)

Table 3. MS population: metabolite ratios - mean (standard deviation).

Model nr.3 (M3) For each voxel, we measure the percentage of each tissue of the brain (GM, WM, lesions). In this case, each voxel is represented by 6 features: three metabolic ratios and three tissues percentages.

Model nr.4 (M4) For each voxel, we compute the spectrogram of its time-domain signal. First, we interpolate the time-domain signal to 1024 points. We compute the spectrogram using a moving window of 128 points, with an overlap of 112 points. In the end, each voxel will be represented by a 128×57 image. These values have been especially selected such that the final image is large enough to be used as input in CNNs.

2.5 Classifiers

For each classification task and for each of the first three feature extraction models, we used three supervised classifiers: (1) LDA [11] without adjusting for class unbalance, (2) Random Forest [12] (RF) with 1000 trees, adjusted for class unbalance by setting the *class_weight* parameter to *balanced_subsample*, and (3) Support Vector Machines with radial basis function (SVM-rbf) [13], adjusted for class unbalance by setting the *class_weight* parameter to *balanced*, and tuned the misclassification cost "C" by selecting its optimal value out of four values (0.1, 1, 10, and 100) over a 5-fold cross-validation loop. The *gamma* parameter was set to *auto*. All classifiers were built in Python 2.7.11 with scikit-learn 0.17.1 [14]. Feature scaling was learned using the training set and applied on both training and test sets, only for the second and third model.

For the last feature extraction model and for each classification task, we built a CNN inspired by [15] using the Keras package [16] based on Theano [17]. Our architecture consists of 8 weighted layers: 6 convolutional (conv) and 2 fully connected (FC). All convolutional layers have a receptive field of 3×3 and the *border_mode* parameter set to 'same'. All weighted layers are equipped with the rectification non-linearity (ReLU). Spatial pooling is carried out by 3 max-pooling (MP) layers over a 2×2 window with stride 2. The first FC layer has 64 channels, while the second one has only 2, because it performs the two-class classification. The final layer is the sigmoid layer. To regularise the training, we used a Dropout layer (D) between the two FC layers, with ratio set to 0.8. A simplified version of our architecture is (conv-conv-MP-conv-conv-MP-conv-conv-MP-FC(64)-D(0.8)-FC(2)-Sigmoid). When training each CNN, we used the 'adadelta' optimizer, the 'categorical_crossentropy' loss function, and we split the training dataset into 70-30 training-validation data. We stopped training after 200 epochs, and for each classification task, validation accuracy was at a stable value over 85%, signalling that training was performed correctly.

3 Results and Discussion

All performance measures can be found in Table 4. Maximum AUC values for each classification task are highlighted in gray.

For CIS vs. RR we obtain a maximum AUC of 77% when combining metabolite ratios with GM, WM, and lesions percentage. The increase in AUC for both SVM-rbf and RF is higher than 10% when we compare M3 to M1 or M2, therefore we can safely conclude that adding GM, WM, and lesions percentage, is indeed

Percentage [%]		M1			M2			M3			M4
		LDA	RF	SVM-rbf	LDA	RF	SVM-rbf	LDA	RF	SVM-rbf	CNN
CIS vs. RR	AUC	65	50	63	53	55	66	63	76	77	71
	Sensitivity	0	0	38	2	0	13	2	28	25	17
	Specificity	100	100	83	100	100	99	100	96	100	98
CIS vs. PP	AUC	89	92	88	87	90	90	88	91	95	83
	Sensitivity	68	68	63	67	72	78	65	77	83	73
	Specificity	93	95	94	91	90	89	91	87	90	82
RR vs. PP	AUC	66	62	68	64	64	68	55	54	57	68
	Sensitivity	21	17	50	29	37	56	0	0	0	28
	Specificity	93	94	78	87	82	76	100	100	100	92
RR vs. SP	AUC	72	72	73	73	71	72	73	71	71	69
	Sensitivity	60	54	57	40	43	48	51	38	29	56
	Specificity	75	84	77	90	86	81	82	92	97	75

Table 4. AUC, Sensitivity, and Specificity values for all classifiers, feature extraction models (M1-M4), and classification tasks.

beneficial when classifying CIS vs. RR courses. This is most probably due to the fact that RR patients have more lesions than CIS patients. It is worth mentioning that the CNN, which takes as input only the MRSI spectrogram, performs better than all other classifiers based on spectroscopic features.

For CIS vs. PP we obtain a maximum AUC of 95% when combining metabolite ratios with GM, WM, and lesion percentages in each voxel. The increase in AUC for SVM-rbf is higher than 5% when we compare M3 to M1 or M2. This task is not too interesting from the medical point of view, because we know that PP patients have a more aggressive form of MS and a higher lesion load than CIS patients. Our results confirm the clinical background and provide an accurate classification with high sensitivity for PP.

For RR vs. PP we obtain the lowest AUC value of the four classification tasks, only 68%. It is interesting to see that adding GM, WM, and lesion percentages did not improve the results, but on the contrary. This indicates an opposing effect between brain segmentation percentages and metabolic ratios. Another interesting fact is that maximum results obtained with M1, M2, or M4, are exactly the same, indicating that spectroscopy is not sensitive enough to classify these two MS courses.

For RR vs. SP we obtain a maximum AUC value of 73%, if we use M1, M2, or M3. There are two main observations to be made: (1) LDA trained on metabolic ratios can be regarded as the best classifier for this task, due to a simple feature extraction model and high computational speed, and (2) adding brain segmentation percentages did not improve the results.

To our knowledge, there are only two other studies which report classification results between MS courses, and both are based on diffusion MRI. Muthuraman et al. [18] report almost a perfect accuracy of 97% for 20 CIS vs. 33 RR patients, and Kocevar et al. [19] report F1-scores of 91.8% for 12 CIS vs. 24 RR patients, 75.6% for 24 RR vs. 17 PP patients, and 85.5% for 24 RR vs. 24 SP patients.

These results show that features extracted from diffusion MRI are clearly better than MRSI features at discriminating MS courses.

The main goal of this study was to compare different levels of extracting information from the MRSI voxels. To that extent, at the low-level we used only 3 metabolite ratios, at the mid-level we used the entire absolute frequency spectrum of 81 points, and at the high-level we used the MRSI spectrograms, of size 128×57. To boost the low-level features, we added the brain tissue segmentations percentages of WM, GM, and lesions. We used spectrograms as input to state of the art classifiers (e.g. CNNs), and compared the results with widely used machine learning algorithms (e.g. LDA, RF, SVM-rbf) trained on features commonly used in MRSI. We observe that results obtained with CNNs are not significantly worse or better than the rest. Thus, it means that there is an inherent limitation of our particular MRSI protocol to classify MS courses.

Our results show that combining low-level MRSI features with brain tissue segmentations percentages can improve classification between the least aggressive MS course (CIS) and the moderate-severe courses (RR and PP). However, there are obvious limitations on any level of the MRSI features when classifying moderate (RR) from severe MS courses (PP and SP). In the future we will incorporate diffusion MRI features and perform multi-class classification.

4 Conclusions

In this paper we performed four binary classification tasks for discriminating between MS courses. We report AUC, sensitivity, and specificity values, after training simple and complex classifiers on four different types of features. We show that combining metabolic ratios with brain tissue segmentation percentages can improve classification results between CIS and RR or PP patients. Our best results are always obtained with SVM-rbf, so we can safely conclude that building complex architectures of convolutional neural networks do not add any improvement over classical machine learning methods.

Acknowledgments. This work was funded by European project EU MC ITN TRANSACT 2012 (no. 316679) and the ERC Advanced Grant BIOTENSORS nr.339804. EU: The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.

References

1. Compston, A., Coles, A.: Multiple sclerosis. The Lancet 372(9648), 1502–1518 (Oct 2008)
2. Miller, D.H., Chard, D.T., Ciccarelli, O.: Clinically isolated syndromes. The Lancet Neurology 11(2), 157–169 (2012)

3. Scalfari, A., Neuhaus, A., Degenhardt, A., Rice, G.P., Muraro, P.A., Daumer, M., Ebers, G.C.: The natural history of multiple sclerosis, a geographically based study 10: relapses and long-term disability. Brain 133(7), 1914–1929 (2010)
4. McDonald, W.I., Compston, A., Edan, G., Goodkin, D., Hartung, H.P., Lublin, F.D., McFarland, H.F., Paty, D.W., Polman, C.H., Reingold, S.C., et al.: Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. Annals of neurology 50(1), 121–127 (2001)
5. Polman, C.H., Reingold, S.C., Edan, G., Filippi, M., Hartung, H.P., Kappos, L., Lublin, F.D., Metz, L.M., McFarland, H.F., O'Connor, P.W., et al.: Diagnostic criteria for multiple sclerosis: 2005 revisions to the McDonald Criteria. Annals of neurology 58(6), 840–846 (2005)
6. Polman, C.H., Reingold, S.C., Banwell, B., Clanet, M., Cohen, J.A., Filippi, M., Fujihara, K., Havrdova, E., Hutchinson, M., Kappos, L., et al.: Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald Criteria. Annals of neurology 69(2), 292–302 (2011)
7. Rovira, À., Auger, C., Alonso, J.: Magnetic resonance monitoring of lesion evolution in multiple sclerosis. Therapeutic advances in neurological disorders 6(5), 298–310 (2013)
8. Lublin, F.D., Reingold, S.C., et al.: Defining the clinical course of multiple sclerosis results of an international survey. Neurology 46(4), 907–911 (1996)
9. Jain, S., Sima, D.M., Ribbens, A., Cambron, M., Maertens, A., Van Hecke, W., De Mey, J., Barkhof, F., Steenwijk, M.D., Daams, M., et al.: Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images. NeuroImage: Clinical 8, 367–375 (2015)
10. Poullet, J.B.: Quantification and classification of magnetic resonance spectroscopic data for brain tumor diagnosis. Katholic University of Leuven (2008)
11. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of eugenics 7(2), 179–188 (1936)
12. Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)
13. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning 20(3), 273–297 (1995)
14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
16. Chollet, F.: Keras. `https://github.com/fchollet/keras` (2015)
17. Theano Development Team: Theano: A Python framework for fast computation of mathematical expressions. arXiv e-prints abs/1605.02688 (May 2016), `http://arxiv.org/abs/1605.02688`
18. Muthuraman, M., Fleischer, V., Kolber, P., Luessi, F., Zipp, F., Groppa, S.: Structural brain network characteristics can differentiate cis from early rrms. Frontiers in neuroscience 10 (2016)
19. Kocevar, G., Stamile, C., Hannoun, S., Cotton, F., Vukusic, S., Durand-Dubief, F., Sappey-Marinier, D.: Graph Theory-Based Brain Connectivity for Automatic Classification of Multiple Sclerosis Clinical Courses. Frontiers in Neuroscience 10, 478 (2016)