A More Interpretable Classifier for Multiple Sclerosis

Valentine Wargnier-Dauchelle, Thomas Grenier, Françoise Durand-Dubief, François Cotton, Michaël Sdika

A MORE INTERPRETABLE CLASSIFIER FOR MULTIPLE SCLEROSIS

V. Wargnier-Dauchelle[1] *T. Grenier*[1] *F. Durand-Dubief* [1,3] *F. Cotton*[1,2] *M. Sdika*[1]

[1] Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1206, F-69100, Lyon, France
[2] Service de Radiologie, Centre Hospitalier Lyon-Sud, Hospices Civils de Lyon, Pierre-Bénite, France
[3] Service de Neurologie A, Hôpital Neurologique, Hospices Civils de Lyon, Bron, France

ABSTRACT

Over the past years, deep learning proved its effectiveness in medical imaging for diagnosis or segmentation. Nevertheless, to be fully integrated in clinics, these methods must both reach good performances and convince area practitioners about their interpretability. Thus, an interpretable model should make its decision on clinical relevant information as a domain expert would. With this purpose, we propose a more interpretable classifier focusing on the most widespread autoimmune neuroinflammatory disease: multiple sclerosis. This disease is characterized by brain lesions visible on MRI (Magnetic Resonance Images) on which diagnosis is based. Using Integrated Gradients attributions, we show that the utilization of brain tissue probability maps instead of raw MR images as deep network input reaches a more accurate and interpretable classifier with decision highly based on lesions.

***Index Terms*—** Multiple sclerosis, deep learning, attribution maps, MRI, tissue probability maps

1. INTRODUCTION

Multiple sclerosis (MS) is the most widespread disabling neurological disease of young adults. In 2020, around 2.5 millions cases were reported in the world. It is a central nervous system inflammatory and demyelinating disease which causes motor, sensory, visual, etc. disorders. The MS diagnosis is based, in addition to symptoms, on the McDonald criteria which evaluates spatial and temporal spread of lesions visible on magnetic resonance images (MRI). Thus, lesions detection is the gold standard in MS diagnosis. As such, many automatic methods have been proposed in recent years for MS lesions segmentation as referenced in Danelakis et al. survey [1]. Some are based on registration to a reference atlas, some include a features extraction step followed by clustering, region growing, etc. Convolutional neural networks (CNN) have also produced interesting results for MS lesions segmentation [2, 3].

Although deep learning has shown its strong potential in medical imaging, the "black box" nature of deep learning is still an obstacle for clinical practitioner who need to have confidence in the proposed automatic decision. The large number of parameters and the non-linearity of neural networks make their decision difficult to understand and interpret.

Attribution maps computation is recent technique to interpret deep networks decision [4, 5]. These heatmaps indicate the positive or negative relevance of each voxel in the input image for the classification. There are useful in a medical context to validate a network. Indeed, diseases often generate anatomical abnormalities which should match with attribution relevant voxels, and especially in MS which is characterized by brain lesions [6].

In this work we propose to use the three tissue probability maps, cerebrospinal (CSF), grey matter (GM) and white matter (WM) ones, as input of a deep network for the classification of MS vs healthy subjects. Using this input for classification leads to 1/ a better accuracy, 2/ a more interpretable output as given by attribution maps, 3/ a less prone to overfitting model thanks to this strong input normalization.

The paper is organized as follows: in section 2 we introduce our methodology. Then, section 3 presents the experiments and section 4 the results. Finally, section 5 concludes the paper.

2. METHODOLOGY

In this work, we compared the performances and interpretability of our proposed classifier input namely CSF, GM and WM segmentation probability maps as generated by FSL FAST [7] to raw MRI image. We evaluate the interpretability of our trained MS vs healthy subjects classifier using attribution maps.

2.1. Tissue probability maps

MRI intensities are not absolute values: two MR acquisitions of the same patient will have intensity variations due to field strength, acquisition protocol, scanner brand, MR artifacts, etc. This is a problem for machine learning as the classification decision can be based on the acquisition signatures of the different datasets used and not only on the pathology. To cope

Fig. 1: From left to right: a brain MR T1 image (axial view) and the corresponding CSF, GM and WM probability maps.

with this problem, we choose a strong normalization protocol: replacing MR image by probability maps of three tissue classes: CSF, GM and WM.

A common approach to segment brain MRI in tissue classes is to model the intensity using a Gaussian mixture model and the voxel label interaction with a Markov Random Field (MRF). Formally, it is defined by:

$$P(y_i|x_{N_i},\theta) = \sum_{l\in L} g(y_i,\theta_l)P(l|y_i,x_{N_i},\theta) \qquad (1)$$

where y_i is the intensity of the i^{th} voxel, L the set of tissue class labels, g a Gaussian with mean and variance $\theta_l = (\mu_l,\sigma_l)$, x_{N_i} the class labels of the neighbors voxels N_i and where θ includes both the θ_l and the artefacts parameters such as the bias field parameters. The model parameters are estimated iteratively with the Expectation Maximization algorithm.

Thus, in this model, each tissue (CSF/GM/WM) is represented as a class and its intensity has a Gaussian distribution. The conditional probability maps $P(l|y_i,x_{N_i},\theta)$ (see Fig. 1) are then given as the three channel input to a deep classifier. As these maps are supposedly noise, bias field inhomogeneity and tissue contrast free, so should also be the classification. Moreover, it allows the investigation of the network decision with respect to the tissue class which have a medical signification. More details on the probability maps computation used in this work can be found in [7].

2.2. Attribution maps

Attribution maps allow to investigate the output of a deep model. Here, they are used to show that the network decision is based on relevant MS clinical features: MS lesions.

We use Integrated Gradients method [4] to compute this heatmaps as it respects two important axioms: sensitivity and implementation invariance. Indeed, it is defined along the i^{th} dimension of the input as:

$$IntGrads_i(x) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x-x'))}{\partial x_i} d\alpha \quad (2)$$

where x is the input, x' is the baseline and F represents the network.

Thereby, if the baseline and the input differ in one feature and have different network output, then attribution map of the input has a non-zero attribution value associated to this feature regardless of the network used.

3. EXPERIMENTS

MS vs healthy subject images classification is compared using either a CNN classifier with the MR image (C-MRI) or the three tissue probability maps (C-PMAPS) as input. Classification performance is evaluated using accuracy on two test datasets. The effects of lesions inpainting, filling MS lesions area with WM intensity [8], on classification probability and attributions has also been used to assess the lesions contribution in the decision. As lesions are MS characteristic, we expect that lesions inpainted images to be seen as less pathological by the network.

3.1. Data

Three T1 MRI datasets are used in our experiments: the public IXI healthy database[1], the OFSEP/EDMUS MS dataset[2] from the "Observatoire français de la sclérose en plaques", MS the french registry [9, 10], and the MICCAI 2016 MS-SEG challenge dataset [11]. Division in training/validation/test sets are indicated in Table 1.

Table 1: T1 MRI datasets. H refers to healthy dataset.

Dataset	N_{train}	N_{val}	N_{test}	H/MS	Annoted
IXI	400	130	50	H	No
OFSEP	383	97	30	MS	No
MICCAI	0	0	52	MS	Yes

Attribution maps are generated only on MICCAI images. As the MICCAI FLAIR lesions used appear larger than T1 lesions and are not used to compute segmentation metrics, there is no detrimental impact in our experiments (inpainting in section 3 or attributions statistics in section 4.2 and 4.4).

Data are acquired in different centers with Philips, Siemens, General Electric, 3T and 1.5T scanners. All volumes are preprocessed using FSL FLIRT affine registration on MNI atlas T1 MRI [12, 13], HD-BET brain extraction [14] and N4 bias field correction [15]. Final image size is $91 \times 109 \times 91$ with a 2mm voxel size.

3.2. Implementation details

Our network was implemented using Pytorch. We used a 3D 70x70 PatchGan [16] as classifier, trained with Adadelta optimizer [17] and cross entropy loss. This CNN is defined

[1] http://brain-development.org/ixi-dataset

[2] http://www.ofsep.org

as $C64$-$C128$-$C256$-$C512$ where Ck denotes a Convolution-BatchNorm-LeakyReLU (slope 0.2) layer with k filters, except for the first layer on which no BatchNorm is applied. At the end, a convolution is applied to obtain a 1-dimensional output. Moreover, we performed data augmentation with brightness variation, elastic deformation and mirroring along sagittal plane. For attribution maps generation, we use Captum library[3]. Integrated Gradients was used with a zero-constant baseline.

4. RESULTS

4.1. Healthy vs MS classification performances

Classification accuracy evaluated on test sets for IXI/OFSEP databases and on MICCAI database are presented in Table 2. We note that probability maps input achieves a much better accuracy on all datasets: it is between 7.5% and 10% higher than with C-MRI.

Table 2: Classification accuracy.

	Database	
Classifier	IXI/OFSEP	MICCAI
C-MRI	87.50	84.62
C-PMAPS	**95.00**	**94.23**

4.2. MS lesions contribution to the decision

To evaluate the lesions impact in the network decision, we compare first the network output for MICCAI images before and after lesions inpainting. Ideally, the MS probability given by the network should decrease when MS lesions are inpainted. In Table 3 are reported the average difference of the log-probability ($plog_{\text{diff}}$) of the network output with and without inpainting and the number of patients with inpainted image classified as less pathological than non-inpainted one. The results show that the contribution of MS lesions in the decision is more important in C-PMAPS than in C-MRI: the log-probability difference is higher and in addition, more images are classified as less pathological after inpainting.

Table 3: Mean/standard-deviation of the classifier log-probability difference ($plog_{\text{diff}}$) with and without inpainting. "/52" column refers to the number of patients (/52) with inpainted image classified less pathological than non-inpainted.

Classifier	$plog_{\text{diff}}$	/52
C-MRI	1.20 ± 7.43	44
C-PMAPS	$\mathbf{4.04 \pm 8.68}$	**49**

[3]https://captum.ai

4.3. Cleaner attribution maps

Attribution maps were computed for C-MRI and C-PMAPS on MICCAI images (see Fig. 2). Visually, attribution maps from CM-MRI seems more noisy. This was assessed quantitatively by measuring the total variation of attribution maps: total variation is 10 times higher on average for C-MRI ($TV = 35339 \pm 4576$) than for C-PMAPS ($TV = 3326 \pm 484$ averaged on the 3 channels).

4.4. Attributions statistics

C-MRI attributions seems also less focused on lesions than C-PMAPS attributions (Fig. 2). Moreover, CSF appears to carry less information than GM or WM in C-PMAPS. Indeed, in GM, lesions are associated to MS relevance before inpainting whereas there is no relevance after lesions removal. In WM, healthy tissue brings healthy relevance unlike lesions. Therefore, with this input, the classifier seems to match with clinical knowledge.

To support this, we compute statistics on attribution maps. As inpainted image should carry less positive relevance (MS) and more negative one (healthy), we compute relative difference between attributions average for non-inpainted and inpainted images defined as:

$$\mu_{\text{Diff}} = \frac{\mu - \mu_{in}}{max(|\mu_{in}|, |\mu|)} \tag{3}$$

where μ_{in} and μ are respectively the attribution map average of the inpainted and non-inpainted image. This difference for the MICCAI dataset is reported in Table 4, for the whole image but also only in the lesions mask. We separate positive and negative relevances for a deeper analysis. For probability maps, statistics are computed on the three channels and each channel separately.

Results show that C-PMAPS is more interpretable as the difference between non-inpainted and inpainted image attributions are larger, especially in lesions. Indeed, inpainted images carry more healthiness information and non-inpainted ones more pathological information. We also notice that WM is the more important channel for healthy relevance whereas GM is the more important for MS relevance.

5. CONCLUSION AND DISCUSSION

In this paper, we proposed a more accurate and interpretable classifier based on tissue probability maps input. Indeed, results on output and attributions show that the network decision is more based on lesions with this input, which is clinically relevant. Moreover, associated attributions are less noisy and add another information on relevant information for the classifier: tissue type relevance in MS. As the classifier was trained and evaluated on data from variable scanners, we can expect a good generalization.

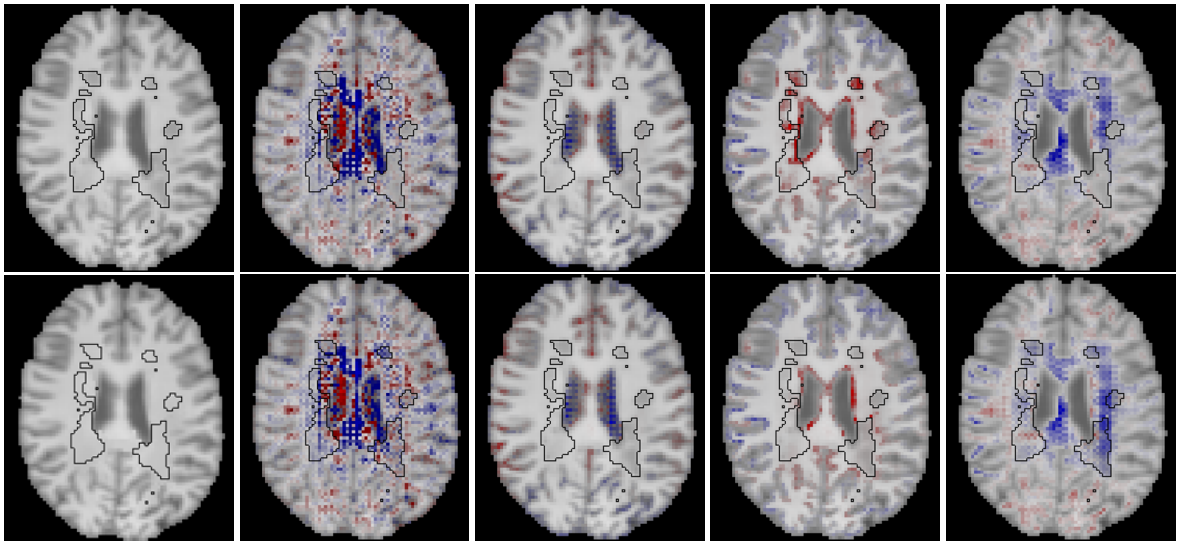

Fig. 2: Axial view attribution examples for one MICCAI patient. FLAIR lesions manual annotation is drawn in black. Blue represents healthy relevance and red MS relevance. From left to right: MR image, attribution maps for MR image input and attribution maps for probability maps (respectively for CSF, GM, WM channels). Images at the top are before inpainting and images below after.

Table 4: Mean $\pm$ standard-deviation of the relative difference (μ_{Diff}) between inpainted and non-inpainted images attributions on the whole image or within the lesions mask. "/52" columns refer to the number of patients (/52) with more MS relevance for image before inpainting than after.

	Whole image		Lesions					
			Total relevance		< 0 relevance		> 0 relevance	
Classifier	μ_{Diff}	/52	μ_{Diff}	/52	μ_{Diff}	/52	μ_{Diff}	/52
C-MRI	0.16 ± 0.31	37	0.40 ± 0.81	40	0.56 ± 0.16	**51**	-0.50 ± 0.14	1
C-PMAPS (all)	0.34 ± 0.44	44	$\mathbf{1.07 \pm 1.05}$	**44**	0.32 ± 0.37	43	0.38 ± 0.33	45
C-PMAPS (CSF)	-0.23 ± 0.33	6	-0.35 ± 0.98	19	-0.72 ± 0.51	6	-0.74 ± 0.33	9
C-PMAPS (GM)	-0.34 ± 0.49	13	0.48 ± 0.67	41	-0.73 ± 0.28	2	$\mathbf{0.75 \pm 0.26}$	**51**
C-PMAPS (WM)	$\mathbf{0.36 \pm 0.48}$	**45**	0.53 ± 0.77	42	$\mathbf{0.77 \pm 0.23}$	**51**	-0.75 ± 0.21	1

We notice that tissues probability maps can sometimes merge brain ventricles and surrounding lesions even if MRI image intensities are different. In this case, attributions show that only lesion edge has positive MS relevance whereas within lesions, relevance is more negative especially in GM. It could be a limitation for a segmentation task. The solution could be to change the tissue probability map generation in order to further separate ventricles and lesions.

Finally, our work used FLAIR lesions to compute statistics on attribution maps but our network is based on T1 images. As FLAIR lesions are larger than T1 ones, we can expect better "in lesions" statistics with a T1 lesions mask.

6. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access at http://brain-development.org/ixi-dataset. Ethical approval was not required as confirmed by the license attached with the open access data. Confidentiality and safety of OFSEP and MICCAI data are ensured by the recommendations of the French "Commission Nationale de l'Informatique et des Libertés" (CNIL), with the Reference Methodology MR-004 of the CNIL.

7. ACKNOWLEDGMENTS

This work was performed within the framework of the LABEX PRIMES (ANR-11-LABX-0063, ANR-11-IDEX-0007) in a lab member of France Life Imaging network (ANR-11-INBS-0006). We acknowledge the "Observatoire Français de la Sclérose en plaques" (OFSEP) for providing the data collected with the ANR grant ANR-10-COHO-002.

8. REFERENCES

[1] Antonios Danelakis, Theoharis Theoharis, and Dimitrios A Verganelakis, "Survey of automated multiple sclerosis lesion segmentation techniques on magnetic resonance imaging," *Computerized Medical Imaging and Graphics*, vol. 70, pp. 83–100, 2018.

[2] Richard McKinley, Tom Gundersen, Franca Wagner, Andrew Chan, Roland Wiest, and Mauricio Reyes, "Nabla-net: a deep dag-like convolutional architecture for biomedical image segmentation: application to white-matter lesion segmentation in multiple sclerosis," *MSSEG Challenge Proceedings: Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure*, p. 37, 2016.

[3] Sergi Valverde, Mariano Cabezas, Eloy Roura, Sandra González-Villà, Deborah Pareto, Joan C Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Arnau Oliver, and Xavier Lladó, "Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach," *NeuroImage*, vol. 155, pp. 159–168, 2017.

[4] Mukund Sundararajan, Ankur Taly, and Qiqi Yan, "Axiomatic attribution for deep networks," *arXiv preprint arXiv:1703.01365*, 2017.

[5] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[6] Fabian Eitel, Emily Soehler, Judith Bellmann-Strobl, Alexander U Brandt, Klemens Ruprecht, René M Giess, Joseph Kuchling, Susanna Asseyer, Martin Weygandt, John-Dylan Haynes, et al., "Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional mri using layer-wise relevance propagation," *NeuroImage: Clinical*, vol. 24, pp. 102003, 2019.

[7] Yongyue Zhang, Michael Brady, and Stephen Smith, "Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm," *IEEE transactions on medical imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[8] Michaël Sdika and Daniel Pelletier, "Nonrigid registration of multiple sclerosis brain images using lesion inpainting for morphometry or lesion mapping," *Human brain mapping*, vol. 30, no. 4, pp. 1060–1067, 2009.

[9] Sandra Vukusic, Romain Casey, Fabien Rollot, Bruno Brochet, Jean Pelletier, David-Axel Laplaud, Jérôme De Sèze, François Cotton, Thibault Moreau, Bruno Stankoff, et al., "Observatoire français de la sclérose en plaques (ofsep): A unique multimodal nationwide ms registry in france," *Multiple Sclerosis Journal*, vol. 26, no. 1, pp. 118–122, 2020.

[10] C Confavreux, DA Compston, OR Hommes, WI McDonald, and AJ Thompson, "Edmus, a european database for multiple sclerosis.," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 55, no. 8, pp. 671–676, 1992.

[11] Olivier Commowick, Audrey Istace, Michael Kain, Baptiste Laurent, Florent Leray, Mathieu Simon, Sorina Camarasu Pop, Pascal Girard, Roxana Ameli, Jean-Christophe Ferré, et al., "Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure," *Scientific reports*, vol. 8, no. 1, pp. 1–17, 2018.

[12] Mark Jenkinson and Stephen Smith, "A global optimisation method for robust affine registration of brain images," *Medical image analysis*, vol. 5, no. 2, pp. 143–156, 2001.

[13] Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," *Neuroimage*, vol. 17, no. 2, pp. 825–841, 2002.

[14] Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, et al., "Automated brain extraction of multisequence mri using artificial neural networks," *Human brain mapping*, vol. 40, no. 17, pp. 4952–4964, 2019.

[15] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee, "N4itk: improved n3 bias correction," *IEEE transactions on medical imaging*, vol. 29, no. 6, pp. 1310–1320, 2010.

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[17] Matthew D Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.