Voxel-wise Comparison with *a-contrario* Analysis for Automated Segmentation of Multiple Sclerosis Lesions from Multimodal MRI

Francesca Galassi[1], Olivier Commowick[1], Emmanuel Vallee[2], and Christian Barillot[1]

[1] INRIA, CNRS, INSERM, IRISA, VisAGeS, Rennes, France
[2] FMRIB, NDCN, University of Oxford, UK
francesca.galassi@inria.fr

Abstract. We introduce a new framework for the automated and unsupervised segmentation of Multiple Sclerosis lesions from multimodal Magnetic Resonance images. It relies on a voxel-wise approach to detect local white matter abnormalities, with an *a-contrario* analysis, which takes into account local information. First, a voxel-wise comparison of multimodal patient images to a set of controls is performed. Then, region-based probabilities are estimated using an *a-contrario* approach. Finally, correction for multiple testing is performed. Validation was undertaken on a multi-site clinical dataset of 53 MS patients with various number and volume of lesions. We showed that the proposed framework outperforms the widely used FDR-correction for this type of analysis, particularly for low lesion loads.

Keywords: multiple sclerosis, voxel-wise comparison, a-contrario

1 Introduction

Multiple Sclerosis (MS) is a chronic inflammatory-demyelinating disease of the central nervous system [1]. Magnetic Resonance Imaging (MRI) is fundamental in MS to characterize and quantify MS lesions. The number and volume of lesions are used for MS diagnosis, to track its progression and to evaluate treatments [2]. Conventional MRI in MS usually consists in Fluid-Attenuated Inversion Recovery (FLAIR), T2-weighted (T2-w) and T1-weighted (T1-w) images. Accurate identification of MS lesions in MR images is extremely difficult due to variability in lesion location, size and shape, in addition to anatomical variability between subjects. Since manual segmentation requires expert knowledge, it is time consuming and prone to intra- and inter-expert variability, several methods have been proposed to automatically segment lesions [1]. In order to reduce false lesion detections, segmentation algorithms have to integrate complementary information from multimodal data. Although many solutions have been proposed, e.g. 3-class tissue classification and Machine Learning (ML) approaches [1], the challenge remains to provide segmentation techniques that work regardless of

the type of MS lesion or MRI protocol.
MS lesion segmentation algorithms are generally prone to detection of false positives, especially voxel-wise approaches, where inference is performed directly on the voxel-wise probabilities. We propose to tackle this problem by replacing classical methods for correction for multiple testings, e.g. Bonferroni and FDR-correction, with a locally multivariate inference: the *a-contrario* analysis [3].
We present a novel framework for the automated segmentation of MS lesions from multimodal MRI, based on a comparison at the voxel level between a patient and a model of healthy controls with an *a-contrario* approach. In Section 2, we present the steps of the proposed framework and the evaluation metrics. Then, in Section 3 we illustrate the experiments, performed on a multi-site clinical dataset. Finally, we discuss the results and conclude in Section 4.

2 Material and methods

2.1 MS lesion detection framework

The *a-contrario* approach The *a-contrario* approach is a locally multivariate procedure which uses the size of a local excursion set as statistic [3]. An a-contrario framework was previously presented to extract patterns of abnormal perfusions in individual patients [4]. Its general steps can be summarized as follows: *i)* a voxel-wise probability map is computed under a background model (i.e. the null-hypothesis in statistical decision theory [5]), *ii)* a locally multivariate probability is estimated, and *iii)* a correction for multiple testing is performed. We propose to apply the *a-contrario* approach to the segmentation of MS lesions from multimodal MRI as follows.

i) Voxel-wise probability map. In [6], a general methodology for the comparison, at a voxel level, of a patient model with a group of models was presented. We adopted a similar approach to compute the input voxel-wise probability map of the *a-contrario* analysis. Precisely, at a given voxel, we compared an intensity vector $V_P \in \mathbb{R}^h$, where h is the dimension and P indicates the patient, with a set of intensity vectors V_j from the control group, with $j = 1, \ldots, N$ controls. These intensity vectors were created from the image modalities (i.e. in our workflow we used FLAIR and T2-w modalities). The group of controls is assumed to follow a multivariate Normal distribution $\mathcal{N}(\overline{V}, \Sigma_V)$, where $\overline{V}$ and Σ_V denote respectively the average and covariance matrix of the control group. Thus, the difference statistic between V_P and $\overline{V}$ can be computed as a Mahalanobis distance $d^2(V_P) = (V_P - \overline{V})^T \Sigma_V^{-1} (V_P - \overline{V})$. $d^2(V_P)$ varies between zero and infinity, with smaller values if the patient vector more likely belongs to $\mathcal{N}(\overline{V}, \Sigma_V)$. The test p-value can be computed as:

$$p(V_P) = 1 - F_{h,N-h}(d^2(V_P)) \tag{1}$$

where $F_{h,N-h}$ is the cumulative distribution function of a Fisher distribution with parameters h and $N - h$. The obtained p-value map was employed as the input for the region-based probabilities estimation.

ii) Region-based probabilities. The uncorrected p-value map was partitioned into regions, namely a grid of spheres of radius r centered at each voxel. A set of uncorrected p-value thresholds $p = \{p_1, ..., p_T\}$ was defined i.e. a set of decision thresholds. For a threshold p_i, the p-value map was thresholded to produce a binary map referred to as a *rare event* map. For each region s, the number of *rare events* occurring at a level p_i was computed and denoted as k_s. Hence, the probability π_i^s of having k_s or more *rare events* was calculated from the tails of the binomial distribution:

$$\pi_i^s = P(X \geq k_s), \quad \text{with} \quad X \sim B(n, p_i) \tag{2}$$

where n is the total number of voxels in the sphere s, i.e. the number of tests. The probability π_i^s associated to a region s was then assigned to its center voxel. Of all region-based probabilities, only the minimum probability over all p-value thresholds p_i, $min(\pi_i^s)$, was retained per voxel.

iii) Correction for multiple testing. The probability map from step *ii)* was then corrected for multiple testing. The probability map was converted to a *Number of False Alarms* (NFA) map, i.e. the number of false detections in the background, as:

$$\text{NFA}_s = N_s T min(\pi_i^s) \tag{3}$$

where N_s and T are the total number of regions and p-value thresholds, respectively. Last, the NFA map was thresholded so that regions with NFA $> \epsilon$ were discarded to obtain $\epsilon-$significant regions, where ϵ is the detection threshold.

Post-processing After the *a-contrario* analysis, the segmentation outcome may still include false positives due to e.g. registration errors, noise and artifacts. A few post-processing steps were therefore performed to reduce these false detections. A candidate lesion was discarded if one of the following conditions was verified: *i)* it did not belong to an hyper-intensities mask, *ii)* it was not sufficiently located in the white matter, *iii)* its size was lower than 3mm^3. The hyper-intensities mask was computed by performing Otsu's thresholding [7] on the product of the T2-w and FLAIR images of a subject [8]. The white matter probability map was calculated from the control subjects and then thresholded at 0.7 to obtain a mask.

2.2 Dataset and Pre-processing

MS patients. We evaluated the proposed method on the MICCAI 2016 MS lesion segmentation challenge dataset [9]. It included 53 images of patients suffering from MS (15 training images and 38 testing images; evaluation on the testing images can be performed by submission to the evaluation platform[1]). They were acquired in four different sites (Siemens 3T Verio, Siemens Aera 1.5T, Philips

[1] https://portal.fli-iam.irisa.fr/msseg-challenge/overview

3T Ingenia, GE 3D Discovery). The MR imaging protocol included 3D T1-w, T2-w and 3D FLAIR anatomical images. More details on the imaging protocol are available on the challenge website[1]. For each subject, manual delineations of MS lesions from seven trained radiologists were provided; the ground truth was computed from the seven independent manual segmentations using LOP STAPLE [10].

Group of controls. 20 MRI datasets of healthy subjects were acquired on a Siemens 3T Verio scanner. The MR imaging protocol included: 3D T1-w (matrix size: 256x256x160, resolution: 1x1x1 mm^3); T2-w (matrix size: 192x256x44, resolution: 1x1x3 mm^3); 3D FLAIR (matrix size: 256x256x160, resolution: 1x1x1 mm^3).

Pre-processing. MR images were denoised [11], rigidly registered towards T1-w images [12], skull-stripped [13] and bias corrected [14]. The proposed framework relies on a voxel-wise comparison of a patient to a set of controls. Hence, it requires that patient and controls images are in the same coordinates system, i.e. corresponding voxels describe the same spatial position, and corresponding anatomical tissues show the same intensity profile. A template image was generated from the set of controls images by applying a method derived from [15], which constructs an unbiased atlas representing the average intensity and shape of a number of images. Patient images were registered to the template image using a linear registration, based on a block-matching algorithm [12], followed by a dense non-linear registration [16]. In order to reduce inter-subject variability, intensities were normalized using k-means [17].

2.3 Evaluation of MS lesion detection

The quality of the proposed segmentation framework was assessed using three metrics:

(i) Dice Similarity Coefficient (DSC), i.e. the spatial overlap between the result R and the ground truth G:

$$DSC = 2\frac{\mid R \cap G \mid}{\mid R \mid + \mid G \mid} \tag{4}$$

(ii) Positive Predictive Value (PPV), i.e. the proportion of true positive lesions TP_R within the segmented N lesions:

$$PPV = 2\frac{TP_R}{N} \tag{5}$$

(iii) $F1$ score, i.e. the weighted average of the lesion sensitivity Se_L and the positive predictive value PPV:

$$F1 = 2\frac{Se_L PPV}{Se_L + PPV} \tag{6}$$

These two last metrics evaluated the algorithm in terms of detection of individual lesions, independently of their contour quality i.e. at the lesion level and not at the voxel level.

Comparison with False Discovery Rate correction. Inference in voxel-wise comparison approaches is generally performed directly on the p-value map by applying a False Discovery Rate (FDR) correction for multiple comparison [6]. The widely applied Benjamini-Hochberg procedure enables controlling the expected proportion of false positives when considering all tests, e.g. it ensures that no more than a ratio $q = 5\%$ of detections are false positives [18]. For comparison with our method, we replaced the *a-contrario* analysis with the FDR correction. Hence, we applied the method in [18] to the voxel-wise probability map as obtained from step *i)*, followed by the same post-processing steps. We evaluated the outcomes using the three metrics presented above. We explored the significance of the differences in the scores obtained by the two approaches using the Wilcoxon test (a p-value $< 5\%$ was considered significant).

3 Results

3.1 Implementation and Computation Time

The framework was implemented in Python and employed in-house tools[3] for the pre-processing and post-processing steps. In the *a-contrario* framework, the radius r of a sphere was equal to two voxels, the set of p-values was $p = \left\{ 1.10^{-05}, 1.10^{-04}, 1.10^{-03} \right\}$, and ϵ=1. The computation time to process a subject on a laptop with an Intel Core i7 CPU 2.40GHz (8 cores) was approximately 10 minutes.

3.2 Evaluation of MS lesion detection

Fig. 1 shows a representative case of uncorrected p-value map from step *i)* and detected MS lesions as obtained with the proposed framework. In Fig. 2, two segmentations outcomes as obtained with the two methods, i.e. the proposed method and the FDR-corrected voxel-wise probability map, are reported. From visual inspection, it appears that both methods are capable of detecting the true lesions; however, the FDR correction approach seems to be more prone to false positives than the proposed approach.

[3] Anima: Open source software for medical image processing from the INRIA VISAGES team.

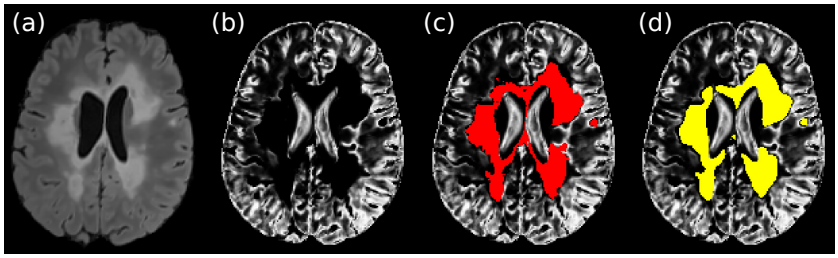

Fig. 1: (a) Original FLAIR image followed by (b) its uncorrected p-value map and superimposed MS lesion segmentations from (c) experts segmentation and (d) proposed framework.

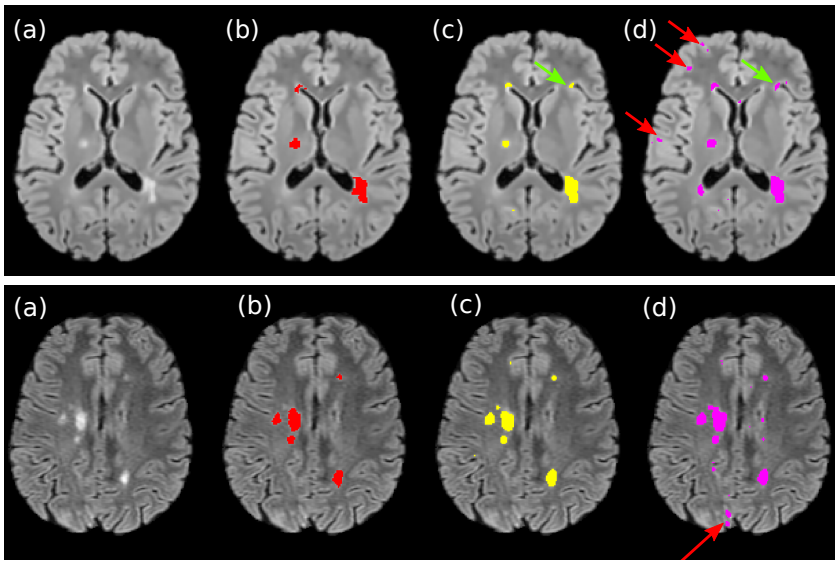

Fig. 2: (a) Original FLAIR image followed by FLAIR image and superimposed MS lesion segmentations from: (b) experts segmentation, (c) proposed framework, (d) FDR-correction. Arrow heads show some false detected lesions: green arrows for false positive on both (c) and (d), red arrows in (d) only.

For each patient and for both the methods, we computed the three evaluation metrics. The average scores are reported in Table 1, together with the outcomes of the Wilcoxon test. In Fig. 3, the three scores for the proposed method are reported for increasing Total Lesion Load (TLL). Fig. 4 shows the differences in scores per patient between the proposed framework and the FDR-correction approach for increasing TLL, where positive difference values indicate that the first outperforms the latter. The Wilcoxon test indicates that the scores are significantly different.

Generally, the proposed method outperforms the classical approach. This is particularly evident for low lesion loads, whereas the two performances tend to converge for high lesion loads. The highest improvements of the proposed method over the FDR correction approach were 36% in DSC (TTL$\approx$ 3cm^3), 73% in PPV (TTL$\approx$ 3cm^3), and 31% in F1 score (TTL$\approx$ 8cm^3). The average improvements were about 10% in DSC, 20% in PPV, and 10% in F1 score. Overall, we observed that all the scores tend to decrease with the total lesion load. This can be partially explained by the disagreement among the experts, which increases and hence becomes more relevant for a lower lesion load.

Table 1: Average scores per metric and p-value of the Wilcoxon test on corresponding sets of scores.

	DSC	**PPV**	**F1 score**
Proposed framework	0.51^*	0.56^*	0.32^*
FDR-correction	0.48	0.45	0.25
Wilcoxon p-value	0.007	$1.86.10^{-8}$	0.03

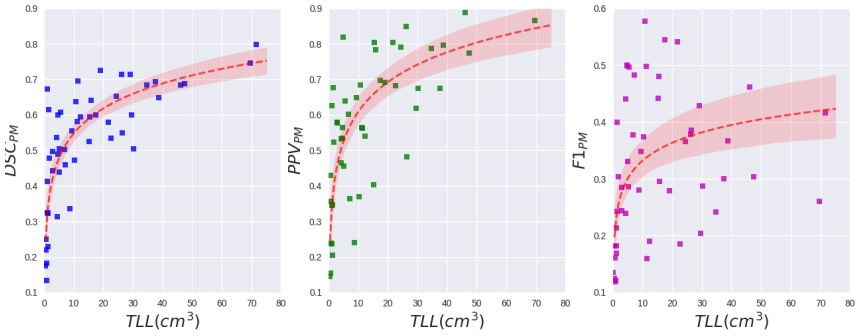

Fig. 3: Metrics as obtained with the proposed method (PM) for increasing Total Lesion Load (TLL) per patient. From the left: DSC, PPV, and F1 score. TLL varied from about 0.5cm^3 to 70cm^3. A log regression model is fitted to the data and a 95% confidence interval for that regression is shown.

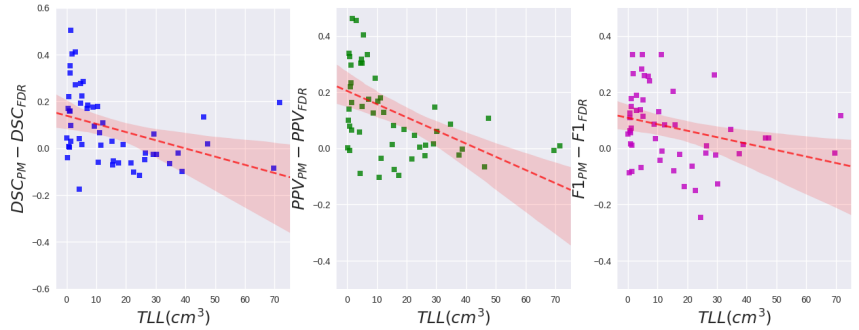

Fig. 4: The differences in scores as obtained with the two approaches for increasing Total Lesion Load. From the left: DSC, PPV, and F1 score. A linear regression model is fitted to the data and a 95% confidence interval for that regression is shown.

4 Conclusion

In this paper, we have proposed an automatic and unsupervised framework for the segmentation of MS lesions from multimodal MRI. It computes a voxel-wise probability map by comparing a patient with a group of controls, and it estimates locally multivariate probabilities using an *a-contrario* approach. Experiments have shown that the method outperforms the classical FDR-correction approach. Improvements increase with decreasing total lesion load, indicating that the proposed method is more specific and sensitive for patients with low lesion loads. The performance of the method relies on parameters, i.e. size of a region and set of thresholds, that must be accurately tuned on a set of cases.
Evaluation was performed on the MICCAI 2016 MS lesions segmentation challenge dataset, comprising clinical images acquired with different MR scanners

and acquisition protocols [9]. This is an important aspect when developing techniques that are meant to be employed in the clinical practice. Compared to the results from the challenge results board[1], the accuracy of the proposed framework was similar to that of the top rank strategies. Compared to other multivariate approaches, such as Machine Learning techniques, it has the clear advantage of being simple and not computationally intensive. These are important benefits, as the primary objective of the proposed framework is to assist radiologists in the clinical practice.

References

1. Garcia-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D.L., Collins, D.L. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. Med Image Anal. 17(1), 1–18 (2013)
2. Polman, C.H., et al. Diagnostic criteria for multiple sclerosis: 2005 revisions to the Mc Donald criteria. Ann Neurol 58, 840–846 (2005)
3. Robin, A., Moisan, L., Le Hgarat-Mascle, S.: An a-contrario approach for sub-pixel change detection in satellite imagery. IEEE Trans Pattern Anal Mach Intell. 32, 1977-93 (2010)
4. Maumet, C., Maurel, P., Ferré, J.C., Barillot, C.: An a contrario approach for the detection of patient-specific brain perfusion abnormalities with arterial spin labelling. NeuroImage 134, 424–433 (2016)
5. Rousseau, F., Faisan, S., Heitz, F., Armspach, J.P., Chevalier, Y., Blanc, F., de Seze, J., Rumbach, L.: An a contrario approach for change detection in 3D multimodal images: application to multiple sclerosis in MRI. IEEE International Symposium on Biomedical Imaging (ISBI), 2069–2072 (2007)
6. Commowick, O., Maarouf, A., Ferré, J.C., Ranjeva, J.P., Edan, G., Barillot C.: Diffusion MRI abnormalities detection with orientation distribution functions: a multiple sclerosis longitudinal study. Med Image Anal. 22, 114–23 (2015)
7. Otsu, N.: A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics 9, 62–66 (1979)
8. Gabr, R. E., Hasan, K. M., Haque, M. E., Nelson, F. M., Wolinsky, J. S. and Narayana, P. A.: Optimal combination of FLAIR and T2weighted MRI for improved lesion contrast in multiple sclerosis. J. Magn. Reson. Imaging 44, 1293–1300 (2016)
9. Commowick, O. et al.: Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure. bioRxiv 367557 (2018)
10. Akhondi-Asl, A., Hoyte, L., Lockhart, M. E., Warfield, S. K.: A Logarithmic Opinion Pool Based STAPLE Algorithm For The Fusion of Segmentations With Associated Reliability Weights. IEEE Trans. Med. Imaging 33, 1997–2009 (2014)
11. Coupe, P. et al.: An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images. IEEE Trans. Med. Imaging 27, 425-41 (2008)
12. Commowick, O., Wiest-Daesslé, N., Prima, S.: Block matching strategies for rigid registration of multimodal medical images. IEEE International Symposium on Biomedical Imaging (ISBI) 700–703 (2012)
13. Manjn, J. V., Coup, P.: An Online MRI Brain Volumetry System. Front. Neuroinformatics 10 (2016)
14. Tustison, N. J. et al.: N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging 29, 1310–1320 (2010)

15. Guimond, A., Meunier, J., Thirion, J.P.: Average Brain Models. Comput Vis Image Underst 77, 192–210 (2000)
16. Commowick, O., Wiest-Daesslé, N., Prima, S.: Automated diffeomorphic registration of anatomical structures with rigid parts: application to dynamic cervical MRI. MICCAI 15, 163–170 (2012)
17. Virmani, D., Taneja, S., Malhotra, G.: Normalization based K means Clustering Algorithm. arXiv:1503.00900 (2015)
18. Hochberg, Y., Tamhane, A.: Multiple Comparison Procedures. John Wiley and Sons (1987)