Abstract Number **308**

Explainability in automatic Paramagnetic Rim Lesion classification

Federico Spagnolo[1,2,3,11], Pedro M. Gordaliza[4,5], Po-Jui Lu[1,2,3], Mario Ocampo Pineda[1,2,3], Xinjie Chen[1,2,3], Matthias Weigel[1,2,3,6], Maxence Wynen[7,8], Martina Absinta[9], Pietro Maggi[8,10], Meritxell Bach Cuadra[4,5], Vincent Andrearczyk[11], Adrien Depeursinge[11,12], Cristina Granziera[1,2,3]

1. *Translational Imaging in Neurology (ThINk) Basel, Department of Biomedical Engineering, Faculty of Medicine, University Hospital Basel and University of Basel, Basel, Switzerland*
2. *Department of Neurology, University Hospital Basel, Basel, Switzerland*
3. *Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University Hospital Basel and University of Basel, Basel, Switzerland*
4. *CIBM Center for Biomedical Imaging, Lausanne, Switzerland*
5. *Radiology Department, Lausanne University Hospital (CHUV) and University of Lausanne, Lausanne, Switzerland*
6. *Division of Radiological Physics, Department of Radiology, University Hospital Basel, Basel, Switzerland*
7. *ICTEAM Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium.*
8. *Louvain Neuroinflammation Imaging Lab (NIL), Université Catholique de Louvain, Brussels, Belgium*
9. *Translational Neuropathology Unit, Division of Neuroscience, IRCCS San Raffaele Scientific Institute, Milan, Italy.*
10. *Department of Neurology, Cliniques Universitaires Saint-Luc, Université Catholique de Louvain, Brussels, Belgium*
11. *MedGIFT, Institute of Informatics, School of Management, HES-SO Valais-Wallis University of Applied Sciences and Arts Western Switzerland, Sierre, Switzerland*
12. *Nuclear Medicine and Molecular Imaging Department, Lausanne University Hospital (CHUV) and University of Lausanne, Lausanne, Switzerland*

Introduction: Paramagnetic Rim Lesions (PRLs) are dark rim-like areas of focal damage, which are visible in the brains of Multiple Sclerosis (MS) patients using susceptibility-based magnetic resonance imaging. Recent studies have used Deep Learning (DL) for their classification, but these methods are deployed as *black boxes*. Explainable methods (XAI) can help build trust in these models and facilitate their adoption in clinical practice.

Objective/Aims: To better understand what drives the decision of a DL network employed to classify PRLs.

Methods: Two trained experts annotated 384 PRLs using FLuid Attenuated Inversion Recovery (FLAIR) and T2*-w Unwrapped Phase (UP) images from 124 MS patients (77 females, age 45.0 ± 13.1, median EDSS 2.0 [1.5-4.5]) collected at the Lausanne University Hospital (Lausanne, Switzerland) and at the Universitätsspital Basel (Basel, Switzerland). Patches of 28x28x28 from both images (176x240x256 FLAIR, 256x336x384 UP) were extracted around MS lesions (which were automatically segmented and manually corrected), intensity normalised (0-1) and fed as input to train a convolutional neural network (RimNet). RimNet, trained on FLAIR only, has shown a good sensitivity but a poorer specificity compared to the use of UP. 12 patients (78 PRLs) were used for inference.

We generated an attention map for each input sequence with the XAI method *Integrated gradients*. First, the distributions of extreme values (maximum and minimum) in FLAIR- and UP-based XAI maps were compared. Then, we reported how frequently the extreme values were found in the manually segmented rims, which would mean that the model's decisions are particularly influenced by voxel values in these regions. The Wilcoxon signed-rank was adopted to compare FLAIR- and UP-based XAI maps.

Results: RimNet reached a precision/recall of 0.80/0.86, classifying 67 True Positive (TP), 16 False Positive (FP), 11 False Negative (FN), and 988 True Negative (TN) examples. UP-based maps presented negative values within the rim, and positive in its close neighbourhood. Values' distribution in FLAIR-based maps did not show a clear correspondence to the rim. A higher mean of extreme

values was found in UP-based maps ([-0.297, 0.300] for TPs, while [-0.232, 0.189] in FLAIR), with a *p*-value < 0.05 except in TNs and FPs. In the second test on TPs and FNs, minimum values of UP-based maps were found most frequently on the rim (respectively >50% and >72% of the times), with peak values of -0.286 ± 0.127 in TPs and of -0.297 ± 0.127 in FNs.

Conclusion: XAI RimNet confirms that UP's voxels are more important than FLAIR's to classify PRLs. Predictions can be attributed to dark voxels within the rim in UP and, perhaps, to darker demyelinated lesion cores in FLAIR.

Disclosures

Federico Spagnolo: was an employee of F. Hoffmann-La Roche Ltd.
Pedro M. Gordaliza: nothing to disclose.
Po-Jui Lu: nothing to disclose.
Mario Ocampo-Pineda: nothing to disclose.
Xinjie Chen: nothing to disclose.
Matthias Weigel: nothing to disclose.
Maxence Wynen: nothing to disclose.
Pietro Maggi: research activity is supported by the Fondation Charcot Stichting Research Fund 2023, the Fund for Scientific Research (F.R.S, FNRS; grant #40008331), Cliniques universitaires Saint-Luc "Fonds de Recherche Clinique" and Biogen. Received consulting honoraria from Sanofi, Biogen and Merck.
Martina Absinta: consultancy honoraria from Sanofi, GSK, Biogen, Immunic Therapeutics and Abata Therapeutics.
Meritxell Bach Cuadra: nothing to disclose.
Vincent Andrearczyk: nothing to disclose.
Adrien Depeursinge: nothing to disclose.
C. Granziera: The University Hospital Basel (USB) and the Research Center for Clinical neuroimmunology and Neuroscience (RC2NB), as the employers of Cristina Granziera, have received the following fees which were used exclusively for (research support from Siemens, GeNeuro, Genzyme-Sanofi, Biogen, Roche. They also have received advisory board and consultancy fees from Actelion, Genzyme-Sanofi, Novartis, GeNeuro, Merck, Biogen and Roche; as well as speaker fees from Genzyme-Sanofi, Novartis, GeNeuro, Merck, biogen and Roche.