Exploiting XAI maps to improve MS lesion segmentation and detection in MRI

Federico Spagnolo[1,2,3,4][0000−0001−9606−0400], Nataliia Molchanova[4,5][0000−0002−7211−8863], Mario Ocampo Pineda[1,2,3][0000−0003−2239−7355], Lester Melie-Garcia[1,2,3][0000−0001−5602−8916], Meritxell Bach Cuadra[5,6][0000−0003−2730−4285], Cristina Granziera[1,2,3][0000−0002−4917−8761], Vincent Andrearczyk[4][0000−0003−0793−5821], and Adrien Depeursinge[4,7][0000−0002−2362−0304]✉

[1] Translational Imaging in Neurology (ThINk) Basel, Department of Medicine and Biomedical Engineering, 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] MedGIFT, Institute of Informatics, School of Management, HES-SO Valais-Wallis University of Applied Sciences and Arts Western Switzerland, Sierre, Switzerland
[5] CIBM Center for Biomedical Imaging, Lausanne, Switzerland
[6] Radiology Department, Lausanne University Hospital (CHUV) and University of Lausanne, Lausanne, Switzerland
[7] Nuclear Medicine and Molecular Imaging Department, Lausanne University Hospital (CHUV), Lausanne, Switzerland
✉adrien.depeursinge@hevs.ch

Abstract. To date, several methods have been developed to explain deep learning algorithms for classification tasks. Recently, an adaptation of two of such methods has been proposed to generate instance-level explainable maps in a semantic segmentation scenario, such as multiple sclerosis (MS) lesion segmentation. In the mentioned work, a 3D U-Net was trained and tested for MS lesion segmentation, yielding an F1 score of 0.7006, and a positive predictive value (PPV) of 0.6265. The distribution of values in explainable maps exposed some differences between maps of true and false positive (TP/FP) examples. Inspired by those results, we explore in this paper the use of characteristics of lesion-specific saliency maps to refine segmentation and detection scores. We generate around 21000 maps from as many TP/FP lesions in a batch of 72 patients (training set) and 4868 from the 37 patients in the test set. 93 radiomic features extracted from the first set of maps were used to train a logistic regression model and classify TP versus FP. On the test set, F1 score and PPV were improved by a large margin when compared to the initial model, reaching 0.7450 and 0.7817, with 95% confidence intervals of [0.7358, 0.7547] and [0.7679, 0.7962], respectively. These results suggest that saliency maps can be used to refine prediction scores, boosting a model's performances.

Keywords: XAI · radiomics · segmentation · multiple sclerosis.

1 Introduction

Multiple sclerosis (MS) is a demyelinating and autoimmune disease of the central nervous system, which increasingly affects the quality of life of relatively young people [1]. A crucial magnetic resonance imaging (MRI) biomarker in diagnosing and monitoring MS is the presence of plaques (or lesions) in the white matter (WM), which are visible on fluid attenuated inversion recovery (FLAIR) and T1-weighted contrasts, such as the magnetisation-prepared rapid gradient echo (MPRAGE) [2–4].

Manual or semi-automatic annotation of such lesions is a tedious process, which has been automated in many tools based on deep learning (DL) [6, 5]. The "black box" nature of standard DL models [7] and the lack of clinical validation [8] have jeopardized the clinical integration of these tools. In this sense, research in explainable AI (XAI) could result as decisive to better understand and optimize the architecture and the performances of DL models [9]. An exhaustive review of XAI models and applications published before 2023 can be found in [10].

However, XAI methods were not designed for segmentation tasks and, to date, no ad-hoc methods were capable to do so [11, 12]. To this end, two methods were recently developed in [14], which adapt SmoothGrad [13] and Grad-CAM++ [15] to provide instance-level explanation maps. Therefore, these two methods can generate separate (i.e., lesion-specific) explainable maps for distinct instances of a class, e.g., MS plaques. This is important to understand which parts of the image were responsible for the segmentation of a specific targeted lesion. The distribution of maximum and minimum values in explainable (saliency) maps generated with the adapted SmoothGrad for true positive (TP) predictions was compared to that of false positives (FP), false negatives (FN), and true negatives (TN). The first two groups were defined as having, respectively, a non-zero and zero overlap with ground truth (GT). FN predictions were determined as GT segmentations with zero overlap with the predicted lesions. TN examples were obtained by randomly sampling ten sphere-like shapes (with the average lesion volume of the test set), which have zero overlap with the previous groups and are located in patients' brain and skull. The study suggested that maximum and minimum values of XAI maps (with respect to FLAIR) of the four groups present different distributions, as shown in Fig. 1.

Radiomic features have been extensively used in radiology to perform a quantitative analysis of medical images [25–27]. A recent work [28] showed that the outputs of an automatic segmentation model can be used to determine the region of interest (ROI) to extract radiomic features from PET images of intraprostatic cancer lesions.

Inspired by these results, in this work we investigate the discriminatory power of radiomic features extracted from XAI maps to distinguish observations from the first two groups (TP and FP), aiming at improving the model's performance. Our main hypothesis is that XAI maps contain specific signatures that are distinctive of TP versus FP. This is based on experiments in [14], which highlight the fact that TPs, FPs, and FNs presented a different distribution of values in

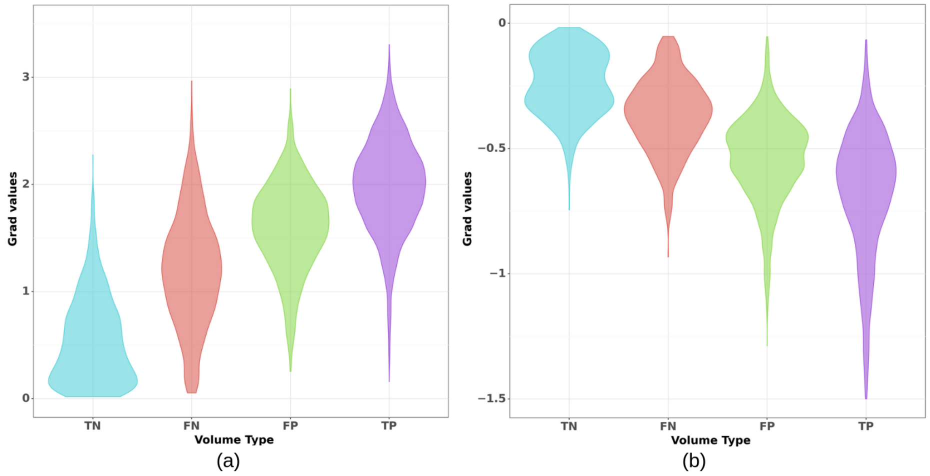


Fig. 1. Violin plots representing the distribution of saliency maps maximum (a) and minimum (b) values. The four distributions refer to true negative (TN), false negative (FN), false positive (FP) and true positive (TP) volumes. Figure retrieved from [14].

XAI maps. The quantitative nature of those saliency maps leads us to test the aforementioned hypothesis.

2 Material and methods

2.1 Dataset and model

We used 4023 FLAIR and MPRAGE MRI scans (SIEMENS Avanto/Espree/Symphony 1.5T and Prisma/Skyra/Verio/MAGNETOM Vida 3T, 1mm isotropic) from 687 patients diagnosed with MS (age=45.2±12.2, 433 females, Expanded Disability Status Scale median of 2.5 [0-9]), acquired at the University Hospital of Basel, Switzerland [29]. WM lesions annotation was performed by three expert clinicians at baseline and follow-ups. Data were randomly split into training, validation and test sets, containing 560, 90 and 37 patients with 3369, 553 and 101 visits respectively (training/validation set's and test set's mean lesions number of 52.9±36.4 and 42.3±21.4 per patient). Datasets from a same patient were ensured to be included in the same split.

Images were pre-processed by registering FLAIR images to MPRAGE space with the *elastix* toolbox [20, 21], correcting for bias field inhomogeneity [22], and standardizing intensities using z-score.

With the described data, a 3D U-Net [16] was trained and tested to segment MS plaques, using patches of dimensions 96^3 and a linear combination of normalized Dice [17] and blob [18] loss. This last choice was made to minimize the impact of instance imbalance within a class and bias towards the occurrence of positive class [19]. The trained model achieved a Dice score of 0.60, and a normalized Dice score of 0.71 on the test set, for what concerns lesion segmentation. The model predicted 3050 TP, 1818 FP, and 789 FN examples, reaching an F1 score of 0.7006 and a positive predictive value (PPV) of 0.6265.

2.2 Instance-level saliency

Following [14], we referred to the lesion domain as Ω: a subset of the image domain Γ, such that $\Omega \subset \Gamma \subset \mathbb{Z}^D$. Each lesion domain presents a cardinality $|\Omega|$, which is the number of voxels within the lesion. For a given Ω, the generation of explainable maps followed these steps:

1. Injecting Gaussian noise $\mathcal{N}(0, \sigma)$ with standard deviation $\sigma = 0.05$ to obtain N noisy versions of the input,
2. Generating a collection of saliency maps for all output voxels in the domain Ω of the lesion,
3. Determining the voxel-wise maximum with sign from this collection of maps,
4. Repeating steps 1-3 and combining these $N = 50$ saliency maps to obtain a single one for the target lesion.

The computation of instance-level saliency maps $M_{\Omega}^{\text{gradient}}[\boldsymbol{v}] \in \mathbb{R}$ is summarized in Eq.(1). Originally, separate saliency maps were obtained for each input modality (FLAIR and MPRAGE) to differentiate their respective contribution. For this work we selected maps with gradients computed with respect to FLAIR, based on findings reported in [14]. The first is that MPRAGE was shown to have a lower contribution to the segmentation of lesions. Secondly, gradients with respect to MPRAGE were less sensitive to the targeted groups (TP, FP, etc.).

$$M_{\Omega}^{\text{gradient}}[\boldsymbol{v}] = \frac{1}{N} \sum_{n=1}^{N} D^n_{argmax_{\boldsymbol{v}'}|D^n_{\boldsymbol{v}'}|}, \text{ where } D^n_{\boldsymbol{v}'} = \frac{\partial y(x_n)[\boldsymbol{v}']}{\partial x_n[\boldsymbol{v}]} \tag{1}$$

2.3 Saliency domain-shift verification between training and test sets

Such explainable maps were generated for TP and FP predictions in the whole test set (3050 and 1818) and a subset S of the training set (containing 11569 TP and 9434 FP from 217 and 439 visits). The mean, maximum and minimum values (and their standard deviation) of maps in the training set were compared with those in the test set, to exclude possible domain shifts. Such shifts could jeopardize the ability of the radiomics-based model to distinguish between TP and FP on the test set due to changes in the saliency maps. A Mann-Whitney U test was run over the distributions of the three measurements in the two groups.

2.4 Radiomics feature extraction

Standardized (z-score) explainable maps and the predicted binary lesion masks from S were used as input to a radiomic features extractor. The binary segmentation of each lesion was dilated to ensure that the computation of features was determined by the saliency values within the ROI (i.e., the lesion domain Ω and in its neighborhood). This is needed to avoid the exclusion of negative saliency map values, as described in [14]. An illustration is provided by the block diagram in Fig. 2.

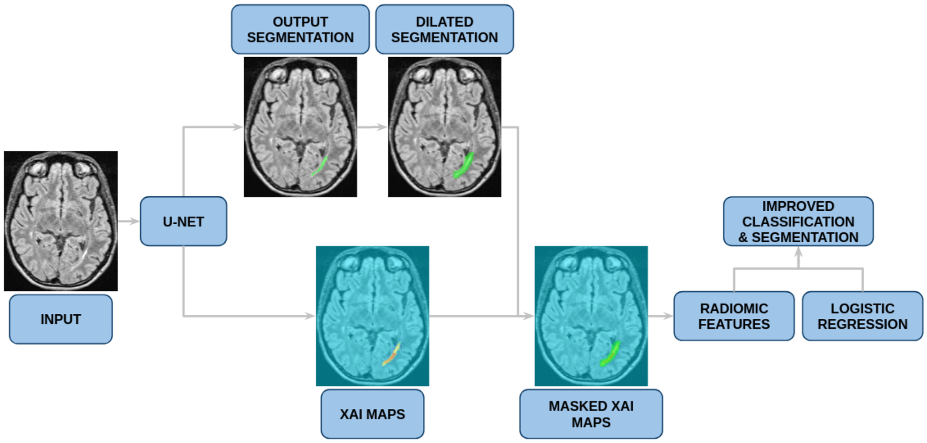


Fig. 2. Block diagram describing how an XAI map is used to extract radiomic features. In this example, a true positive lesion is shown in the axial plane.

We used the library *pyradiomics* [23], with $binWidth = 10$ and $sigma = [1, 2, 3]$. From the pool of 107 available features, those related to *shape* were not considered relevant as depending on the initial segmentation of Ω, resulting in 93 target features.

2.5 TP/FP prediction refinement

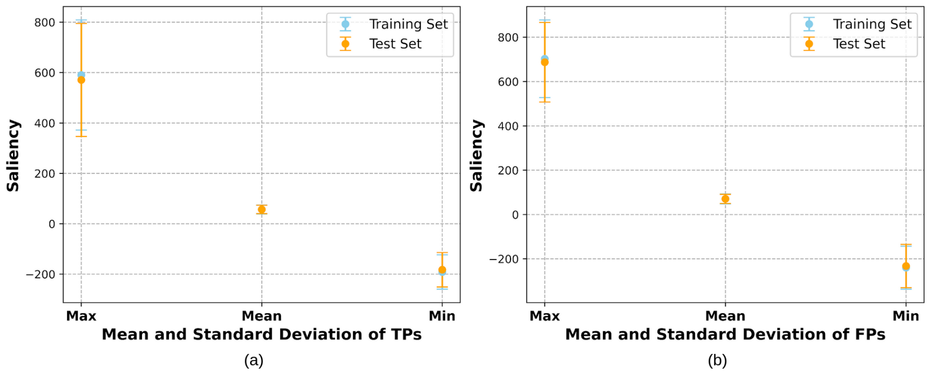
The radiomic features of saliency maps from the training set were standardized and used to train a logistic regression (LR) model using *scikit-learn* [24], with the following parameters: maximum number of iterations: 10000, solver: *liblinear*, penalty: *L1*, class weights: 0.29, 0.71. The class weights were selected to reflect the proportion of TP and FP examples per visit in the training set S. To investigate feature importance, we compared the normalized (0-1) regression coefficients of each feature.

The trained model was applied to the entire test set. Following a bootstrapping approach with 1000 iterations, we computed the mean and 95% confidence interval (CI) of the F1 score and PPV. The updated number of TP, FP and FN predictions was also derived for comparison with the performance of the initial U-Net model. The confusion matrix, F1 score, and PPV reported by the U-Net represent the best scores achieved on the entire test set.

3 Results

3.1 Domain-shift of saliency between training and test sets

The comparison between saliency maps computed on the subset S of the training set and the test set is summarized in Fig. 3. For all the metrics, mean values from both sets fell into the interval $\pm\sigma$ (one standard deviation), associated to non-significant differences.



(a)

(b)

Fig. 3. Comparison between mean, maximum and minimum values of saliency maps computed on the training and test set, for TP (a) and FP (b) examples.

3.2 Refinement of TP/FP predictions

Testing the LR model with bootstrapping we obtained an F1 score of 0.7450 with a 95% CI of [0.7358, 0.7547], and a PPV of 0.7817 with a 95% CI of [0.7679, 0.7962]. The F1 score and PPV of the non-refined U-Net were 0.7006 and 0.6265, respectively. A comprehensive performance comparison is presented in Table 1.

Table 1. Comparing MS lesion detection performance between the original U-Net and its refined version using logistic regression (LR) relying on radiomic features from XAI maps. We report the number of true positives (TPs), false positives (FPs), false negatives (FNs), and the mean and confidence interval (CI) of F1 score and positive predictive value (PPV).

	U-Net only	**U-Net + saliency [95% CI]**
TPs	3050	2732
FPs	1818	763
FNs	789	1107
F1 score	0.7006	0.7450 [0.7358, 0.7547]
PPV	0.6265	0.7817 [0.7679, 0.7962]

Examples of a slice in the sagittal plane of the 3D XAI maps generated from a TP and a FP are illustrated in Fig. 4. The LR output probability (relative to the TP class) for the TP lesion was 0.9398, and 0.0232 for the FP. The example FP candidate is located at the boundary between WM and cortex, and corresponds to one of the brain sulci.

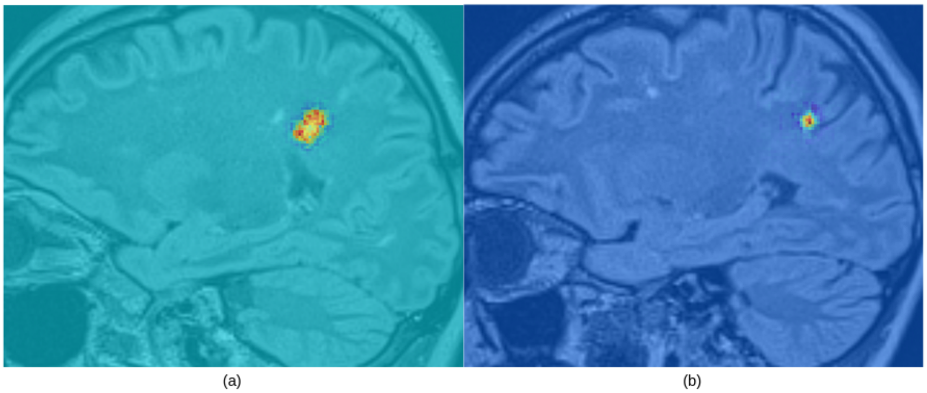


Fig. 4. An example of a slice in the sagittal plane from a saliency map computed on a true (a) and false (b) positive lesion, scoring 0.9398 and 0.0232 for the true positive class.

The importance of radiomic features used by the model is presented in Fig. 5. The most important were two features based on saliency intensity, mean absolute deviation (MAD) and Root Mean Squared (RMS), and reported a normalized coefficient of 0.89 and -1.0 respectively.

4 Discussion

This work employed radiomics to extract features from XAI (saliency) maps generated for a DL semantic segmentation model. These features were fed to a linear model (i.e., LR) to discriminate between FP and TP predictions, and improve the classification score. Our main hypothesis was that XAI maps contained specific signatures that are distinctive of TP versus FP. As a first step of our investigation, we tried using only the minimum and maximum values of saliency to discriminate between TP and FP classes. In this way, the binary classifier could reach a higher precision but a lower recall, eventually resulting in a trade off where the F1 score did not improve. A similar trend was observed when trying to adjust the segmentation threshold at the U-Net output layer. The results in Table 1 show that using radiomics on saliency maps, it was possible to refine classification scores, improving the F1 score from 0.70 to 0.75 and the PPV from 0.63 to 0.78. Confidence intervals computed for the refined model are

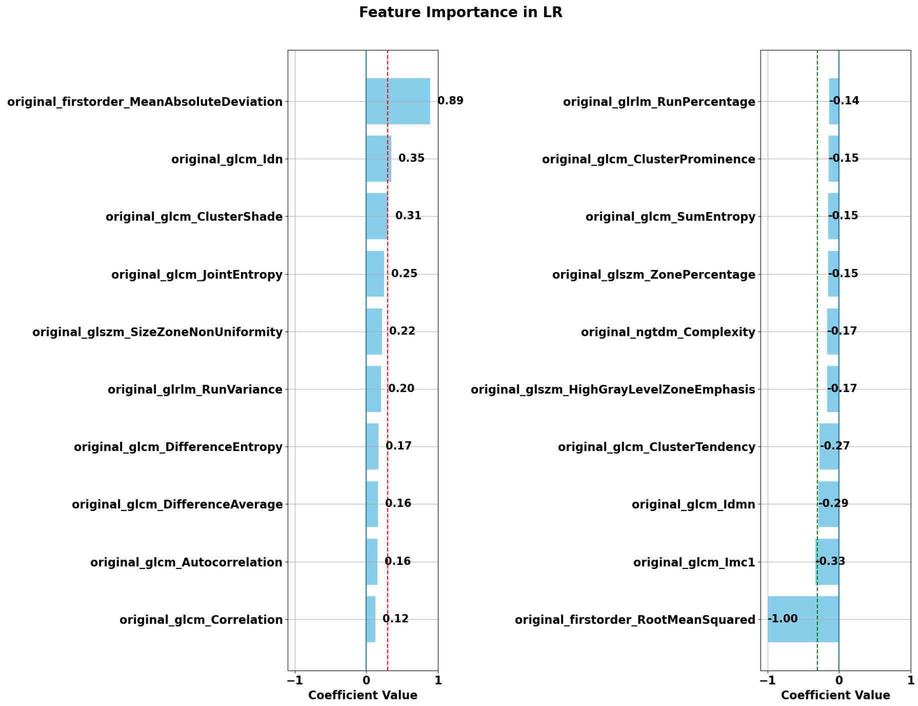


Fig. 5. Normalized radiomic features showing the highest importance (top 10 positive on the left, top 10 negative on the right), in terms of LR coefficients. The dashed red line represents a coefficient value of 0.3, the dashed green line a coefficient value of -0.3.

not including the performance values of the initial model, suggesting significance of the observed improvements.

In Section 3.1 we compared the maximum, minimum and mean values of XAI in the training and test set. According to this analysis, we found no evidence of domain shift between the two sets. The characteristic of having a slightly positive mean value was also preserved in both groups (TP and FP) and sets. This indicated that the radiomics model trained on saliency maps obtained on the training set will generalize to those of the test set.

In Section 3.2 we reported the features which contribute the most to discriminate between saliency maps from TP and FP examples. A strong positive coefficient for MAD reveals that the intensity variability with respect to the mean is higher for XAI maps of the TP class. This finding supports prior results obtained on maximum, minimum and mean values. A positive Inverse difference normalized (Idn) may indicate a more homogeneous texture surrounding voxels. A strong negative coefficient for RMS may describe saliency maps for the FP class as highly variable, and presenting more outliers. A negative Informational measure of correlation (Imc1) represents a higher tendency of neighbouring voxel pairs in FP cases to have similar values. Such findings suggest that saliency maps for TP examples may present a wider range of values (contrast between positive and negative regions), less outliers, and overall a more homogeneous texture.

This work has also some limitations. The application of this method is limited to identifying and ruling out FPs. A possible future improvement would be to extend the approach to FNs, and to refine the detection performance even further. In that case the issue would be to have FN candidates on the test set without using a ground truth. For this purpose, an estimation of the uncertainty [30, 31] of the model's prediction could help to target possibly missed lesions. Moreover, this study explores the refinement of predictions from a single DL model. The impact of XAI on different models should be investigated, as we would expect the initial performance of the model to be relevant. In addition, it would be important to evaluate the performances of this method on saliency maps coming from an out of domain test set, or training and testing the model on multiple datasets. Though, this would also require XAI maps to be comparable for multi-centric data which, to our knowledge, has not been investigated yet. Nevertheless, it is important to remark that the MRI data used in this study were acquired with multiple scanners. Furthermore, investigating the which and how many radiomic features are required to achieve a desired performance would be of interest.

5 Conclusion

This work demonstrated that radiomic features extracted from XAI (saliency) maps generated for a DL semantic segmentation model can be used to improve the detection performance of a segmentation model by a large margin.

Acknowledgments. This work was supported by the Hasler Foundation with the project MSxplain number 21042, the Swiss National Science Foundation (SNSF) with the project 205320_219430, and the Swiss Cancer Research foundation with the project TARGET (KFS-5549-02-2022-R). We acknowledge access to the expertise of the CIBM Center for Biomedical Imaging, a Swiss research center of excellence founded and supported by CHUV, UNIL, EPFL, UNIGE and HUG.

Disclosure of Interests. 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. Federico Spagnolo was an employee of F. Hoffmann-La Roche Ltd. The remaining authors have no competing interests to declare that are relevant to the content of this article.

References

1. Kołtuniuk, A., Pawlak, B., Krowczynska, D., Chojdak-Łukasiewicz, J.: The quality of life in patients with multiple sclerosis – Association with depressive symptoms and

physical disability: A prospective and observational study. Frontiers in Psychology 13, 1068421 (2023). https://doi.org/10.3389/fpsyg.2022.1068421

2. Yang, J., Hamade, M., Wu, Q., Wang, Q., Axtell, R., Giri, S., Mao-Draayer, Y.: Current and Future Biomarkers in Multiple Sclerosis. International journal of molecular sciences 23 (2023). https://doi.org/10.3390/ijms23115877
3. Thompson, A., Banwell, B., Barkhof, F., Carroll, W., Coetzee, T., Comi, G., Correale, J., Fazekas, F., Filippi, M., Freedman, M., Fujihara, K., Galetta, S., Hartung, H.P., Kappos, L., Lublin, F., Marrie, R., Miller, A., Miller, D., Montalban, X., Cohen, J.: Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. The Lancet Neurology 17 (2017). https://doi.org/10.1016/S1474-4422(17)30470-2
4. Hemond C, C., Bakshi, R.: Magnetic Resonance Imaging in Multiple Sclerosis. Cold Spring Harbor perspectives in medicine 8, 5 (2018). https://doi.org/10.1101/cshperspect.a028969
5. Commowick, O., Comb'es, B., Cervenansky, F., Dojat, M.: Editorial: Automatic methods for multiple sclerosis new lesions detection and segmentation. Frontiers in Neuroscience 17 (2023). https://doi.org/10.3389/fnins.2023.1176625
6. Ma, Y., Zhang, C., Cabezas, M., Song, Y., Tang, Z., Liu, D., Cai, W., Barnett, M., Wang, C.: Multiple Sclerosis Lesion Analysis in Brain Magnetic Resonance Images: Techniques and Clinical Applications. IEEE Journal of Biomedical and Health Informatics PP, 1–1 (2022). https://doi.org/10.1109/JBHI.2022.3151741
7. Baselli, G., Codari, M., Sardanelli, F.: Opening the black box of machine learning in radiology: can the proximity of annotated cases be a way? European Radiology Experimental 4 (2020). https://doi.org/10.1186/s41747-020-00159-0
8. Spagnolo, F., Depeursinge, A., Schädelin, S., Akbulut, A., Müller, H., Barakovic, M., Melie-Garcia, L., Bach Cuadra, M., Granziera, C.: How far MS lesion detection and segmentation are integrated into the clinical workflow? A systematic review. NeuroImage Clinical 39, 103491 (2023). https://doi.org/10.1016/j.nicl.2023
9. Kobayashi, K., Alam, S.B.: Explainable, interpretable, and trustworthy ai for an intelligent digital twin: A case study on remaining useful life. Engineering Applications of Artificial Intelligence 129, 107620 (2024). https://doi.org/10.1016/j.engappai.2023.107620
10. Saranya, A., Subhashini, R.: A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. Decision Analytics Journal 7, 100230 (2023). https://doi.org/10.1016/j.dajour.2023.100230
11. Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., Li, M., Kalpathy-Cramer, J.: Assessing the (Un)Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. Radiology: Artificial Intelligence 3 (2021). https://doi.org/10.1148/ryai.2021200267
12. Mahapatra, D., Poellinger, A., Reyes, M.: Interpretability-Guided Inductive Bias For Deep Learning Based Medical Image. Medical Image Analysis 81, 102551 (2022). https://doi.org/10.1016/j.media.2022.102551
13. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: SmoothGrad: removing noise by adding noise. CoRR (2017).
14. Spagnolo, F., Molchanova, N., Schaer, R., Bach Cuadra, M., Ocampo Pineda, M., Melie-Garcia, L., Granziera, C., Andrearczyk, V., Depeursinge, A.: Instance-level quantitative saliency in multiple sclerosis lesion segmentation. arXiv (2024). https://doi.org/10.48550/ARXIV.2406.09335
15. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional

Networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847 (2018). https://doi.org/10.1109/WACV.2018.00097

16. Çiçek, O., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. arXiv (2016). https://doi.org/10.48550/ARXIV.1606.06650
17. Raina, V., Molchanova, N., Graziani, M., Malinin, A., Muller, H., Cuadra, M.B., Gales, M.: Tackling Bias in the Dice Similarity Coefficient: Introducing NDSC for White Matter Lesion Segmentation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), pp. 1–5 (2023). https://doi.org/10.1109/ISBI53787.2023.10230755
18. Kofler, F., Shit, S., Ezhov, I., Fidon, L., Horvath, I., Al-Maskari, R., Li, H., Bhatia, H., Loehr, T., Piraud, M., Erturk, A., Kirschke, J., Peeken, J., Vercauteren, T., Zimmer, C., Wiestler, B., Menze, B.: Blob loss: instance imbalance aware loss functions for semantic segmentation. arXiv (2022). https://doi.org/10.48550/ARXIV.2205.08209
19. Maier-Hein, L., Reinke, A., Christodoulou, E., Glocker, B., Godau, P., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M., Wiesenfarth, M., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Kavur, A., Rädsch, T., Tizabi, M.D., Acion, L., Antonelli, M., Jaeger, P.: Metrics reloaded: Pitfalls and recommendations for image analysis validation. arXiv (2022). https://doi.org/10.48550/arXiv.2206.01653
20. Klein, S., Staring, M., Murphy, K., Viergever, M., Pluim, J.: Elastix: A Toolbox for Intensity-Based Medical Image Registration. In IEEE transactions on medical imaging 29, 196–205 (2009). https://doi.org/10.1109/TMI.2009.2035616
21. Shamonin, D., Bron, E., Lelieveldt, B., Smits, M., Klein, S., Staring, M.: Fast Parallel Image Registration on CPU and GPU for Diagnostic Classification of Alzheimer's Disease. Frontiers in neuroinformatics 7, 50 (2014). https://doi.org/10.3389/fninf.2013.00050
22. Tustison, N., Avants, B., Cook, P., Zheng, Y., Egan, A., Yushkevich, P., Gee, J.: N4itk: improved n3 bias correction. Medical Imaging, IEEE Transactions on 29, 1310 – 1320 (2010). https://doi.org/10.1109/TMI.2010.2046908
23. van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G. H., Fillon-Robin, J. C., Pieper, S., Aerts, H. J. W. L.: Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research, 77(21), e104–e107 (2017). https://doi.org/10.1158/0008-5472.CAN-17-0339
24. 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)
25. Ibrahim, A., Primakov, S., Beuque, M., Woodruff, H.C., Halilaj, I., Wu, G., Refaee, T., Granzier, R., Widaatalla, Y., Hustinx, R., Mottaghy, FM., Lambin, P.: Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. Methods, 188, 20-29 (2021). https://doi.org/10.1016/j.ymeth.2020.05.022
26. Annunziata, S., Treglia, G.: Editorial: Radiomics and artificial intelligence in radiology and nuclear medicine. Frontiers in Medicine, 10 (2023). https://doi.org/10.3389/fmed.2023.1216434
27. Elmahdy, M., Ronnie, S.: Radiomics analysis in medical imaging research. Journal of medical radiation sciences, 70(1), 3-7 (2023). https://doi.org/10.1002/jmrs.662

28. Ghezzo, S., Mongardi, S., Bezzi, C., Samanes Gajate, A. M., Preza, E., Gotuzzo, I., Baldassi, F., Jonghi-Lavarini, L., Neri, I., Russo, T., Brembilla, G., De Cobelli, F., Scifo, P., Mapelli, P., Picchio, M.: External validation of a convolutional neural network for the automatic segmentation of intraprostatic tumor lesions on 68Ga-PSMA PET images. Frontiers in Medicine, 10 (2023). https://doi.org/10.3389/fmed.2023.1133269
29. Disanto, G., Benkert, P., Lorscheider, J., Mueller, S., Vehoff, J., Zecca, C., Ramseier, S., Achtnichts, L., Findling, O., Nedeltchev, K., Radue, E.W., Sprenger, T., Stippich, C., Derfuss, T., Louvion, J.F., Kamm, C.P., Mattle, H.P., Lotter, C., Du Pasquier, R., Schluep, M., Pot, C., Lalive, P.H., Yaldizli, Ö., Gobbi, C., Kappos, L., Kuhle, J.; SMSC Scientific Board. The Swiss Multiple Sclerosis Cohort-Study (SMSC): A Prospective Swiss Wide Investigation of Key Phases in Disease Evolution and New Treatment Options. PLoS One. 11(3) (2016). https://doi.org/10.1371/journal.pone.0152347
30. Nair, T. and Precup, D. and Arnold, D. L. and Arbel, T.: Exploring uncertainty measures in deep networks for Multiple sclerosis lesion detection and segmentation. Medical Image Analysis, 59, 101557 (2020). https://doi.org/10.1016/j.media.2019.101557
31. Molchanova, N. and Raina, V. and Malinin, A. and La Rosa, F. and Depeursinge, A. and Gales, M. and Granziera, C. and Müller, H. and Graziani, M. and Bach Cuadra, M.: Structural-Based Uncertainty in Deep Learning Across Anatomical Scales: Analysis in White Matter Lesion Segmentation, arXiv (2024). https://doi.org/10.48550/arXiv.2311.08931