Towards Informative Uncertainty Measures for MRI Segmentation in Clinical Practice: Application to Multiple Sclerosis

Nataliia Molchanova[1,2,3] Vatsal Raina[2,4] Andrey Malinin[5] Francesco La Rosa[6]
Henning Muller[2] Mark Gales[4] Cristina Granziera[7] Mara Graziani[2] Meritxell Bach Cuadra[1,3]
[1] University of Lausanne and Lausanne University Hospital, Switzerland, [2] University of Applied Sciences of Western Switzerland (HES-SO), Switzerland, [3] CIBM Center for Biomedical Imaging, Switzerland,
[4] ALTA Institute, University of Cambridge, UK, [5] Shifts Project, Finland,
[6] Icahn School of Medicine at Mount Sinai, USA, [7] University Hospital Basel, Switzerland.

1 Introduction

For white matter lesion (WML) segmentation in magnetic resonance imaging of multiple sclerosis patients, both detection and delineation quality is relevant and can affect clinical decisions [10, 1]. Uncertainty quantification can serve as a proxy for the degree of trustworthiness of deep-learning model predictions. This work studies the ability of different voxel- and lesion-scale uncertainty measures to capture lesion segmentation and detection errors, respectively. Our main contributions are (i) proposing a new measure of lesion-scale uncertainty based on structural information rather than voxel uncertainties; (ii) extending an error retention curves (RC) [5] analysis framework for the lesion-scale uncertainty measures evaluation.

2 Materials and Methods

We use deep ensembles for uncertainty quantification [3]. Treating the segmentation task as a classification of each voxel, we use six uncertainty measures to estimate total, data, and knowledge uncertainty for each voxel. We compare the ability of these measures to capture model errors in segmentation using Dice score RC (DSC-RC) [5, 6, 7]. For lesion-scale uncertainty estimation, we use a previously proposed method of averaging voxel uncertainties across the lesion region [4], and the proposed detection disagreement uncertainty (DDU) measure [8] based only on structural predictions. DDU is defined as one minus agreement, where agreement is an average across ensemble members intersection over union between the lesion regions predicted by the ensemble and by ensemble members. We extend the RC definition to the lesion scale, proposing a lesion positive predictive value RC (LPPV-RC), to quantify how well different lesion-scale measures capture errors related to lesion detection. We use a U-net architecture widely investigated for the particular task [2, 9, 4, 6]. Each of the five ensemble members is trained on 40 pairs of FLAIR scans and consensus ground truths masks [6]. Uncertainty measures comparison is done under the domain shift (different scanners, medical centers, MS stages, and others) on a test set of 99 subjects from two different medical centers.

3 Results and Discussion

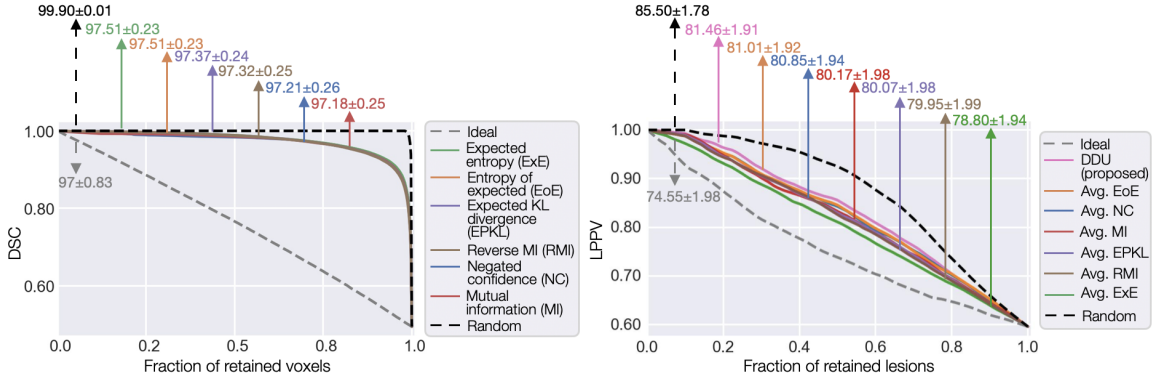

Figure 1: The average across the out-of-domain test set DSC-RC and LPPV-RC and areas under the corresponding curves: DSC-AUC $\cdot 100(\uparrow)$ and LPPV-AUC $\cdot 100(\uparrow)$, with bootstrapped standard errors.

All voxel-scale measures are comparable in DSC-AUC, with the entropy-based measures (ExE and EoE) having the highest DSC-AUC. Nevertheless, aggregation of ExE loses informativeness for lesion detection, showing the lowest LPPV-AUC. The proposed DDU shows LPPV-AUC significantly higher than all measures, except Avg. EoE (Pvalue=0.052), using one-sided paired Wilcoxon tests with a significance level of 0.01. The observed change in the ranking of measures on the lesion scale indicates that a naive aggregation of voxel uncertainties by averaging across the lesion region does not guarantee optimal lesion-scale uncertainty measure construction. The proposed analysis derives assumptions about uncertainty measures that would be more informative for clinicians, *i.e.* the ones reflecting an increased likelihood of erroneous predictions. It is yet important to verify in practice if introducing uncertainty maps to clinicians can speed up or simplify the correction process of predicted WML masks.

References

[1] C. Hemond and R. Bakshi. Magnetic resonance imaging in multiple sclerosis. *Cold Spring Harbor Perspectives in Medicine*, 8(5), May 2018.

[2] F. La Rosa, A. Abdulkadir, M. J. Fartaria, R. Rahmanzadeh, P.-J. Lu, R. Galbusera, M. Barakovic, J.-P. Thiran, C. Granziera, and M. B. Cuadra. Multiple sclerosis cortical and wm lesion segmentation at 3t mri: a deep learning method based on flair and mp2rage. *NeuroImage: Clinical*, 27:102335, 2020.

[3] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[4] B. Lambert, F. Forbes, A. Tucholka, S. Doyle, and M. Dojat. Multi-Scale Evaluation of Uncertainty Quantification Techniques for Deep Learning based MRI Segmentation. In *ISMRM-ESMRMB & ISMRT 2022 - 31st Joint Annual Meeting International Society for Magnetic Resonance in Medecine*, pages 1–3, London, United Kingdom, May 2022.

[5] A. Malinin. *Uncertainty estimation in deep learning with application to spoken language assessment*. PhD thesis, University of Cambridge, United Kingdom, 2019.

[6] A. Malinin, A. Athanasopoulos, M. Barakovic, M. B. Cuadra, M. J. F. Gales, C. Granziera, M. Graziani, N. Kartashev, K. Kyriakopoulos, P.-J. Lu, N. Molchanova, A. Nikitakis, V. Raina, F. La Rosa, E. Sivena, V. Tsarsitalidis, E. Tsompopoulou, and E. Volf. Shifts 2.0: Extending the dataset of real distributional shifts, 2022.

[7] R. Mehta, A. Filos, U. Baid, C. Sako, R. McKinley, M. Rebsamen, K. Dätwyler, R. Meier, P. Radojewski, G. K. Murugesan, S. Nalawade, C. Ganesh, B. Wagner, F. F. Yu, B. Fei, A. J. Madhuranthakam, J. A. Maldjian, L. Daza, C. Gómez, P. Arbeláez, C. Dai, S. Wang, H. Reynaud, Y. Mo, E. Angelini, Y. Guo, W. Bai, S. Banerjee, L. Pei, M. AK, S. Rosas-González, I. Zemmoura, C. Tauber, M. H. Vu, T. Nyholm, T. Löfstedt, L. M. Ballestar, V. Vilaplana, H. McHugh, G. Maso Talou, A. Wang, J. Patel, K. Chang, K. Hoebel, M. Gidwani, N. Arun, S. Gupta, M. Aggarwal, P. Singh, E. R. Gerstner, J. Kalpathy-Cramer, N. Boutry, A. Huard, L. Vidyaratne, M. M. Rahman, K. M. Iftekharuddin, J. Chazalon, E. Puybareau, G. Tochon, J. Ma, M. Cabezas, X. Llado, A. Oliver, L. Valencia, S. Valverde, M. Amian, M. Soltaninejad, A. Myronenko, A. Hatamizadeh, X. Feng, Q. Dou, N. Tustison, C. Meyer, N. A. Shah, S. Talbar, M.-A. Weber, A. Mahajan, A. Jakab, R. Wiest, H. M. Fathallah-Shaykh, A. Nazeri, M. Milchenko, D. Marcus, A. Kotrotsou, R. Colen, J. Freymann, J. Kirby, C. Davatzikos, B. Menze, S. Bakas, Y. Gal, and T. Arbel. Qu-brats: Miccai brats 2020 challenge on quantifying uncertainty in brain tumor segmentation – analysis of ranking scores and benchmarking results. *Machine Learning for Biomedical Imaging*, 1:1–54, 2022.

[8] N. Molchanova, V. Raina, A. Malinin, F. La Rosa, H. Muller, M. Gales, C. Granziera, M. Graziani, and M. B. Cuadra. Novel structural-scale uncertainty measures and error retention curves: application to multiple sclerosis. *Accepted to IEEE ISBI 2023*, 2022.

[9] T. Nair, D. Precup, D. Arnold, and T. Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical Image Analysis*, 59:101557, January 2020.

[10] A. J. Thompson, B. L. Banwell, F. Barkhof, W. M. Carroll, T. Coetzee, G. Comi, J. Correale, F. Fazekas, M. Filippi, M. S. Freedman, K. Fujihara, S. L. Galetta, H. P. Hartung, L. Kappos, F. D. Lublin, R. A. Marrie, A. E. Miller, D. H. Miller, X. Montalban, E. M. Mowry, P. S. Sorensen, M. Tintoré, A. L. Traboulsee, M. Trojano, B. M. J. Uitdehaag, S. Vukusic, E. Waubant, B. G. Weinshenker, S. C. Reingold, and J. A. Cohen. Diagnosis of multiple sclerosis: 2017 revisions of the mcdonald criteria. *The Lancet Neurology*, 17(2):162–173, 2018.