

Letter to the Editor

## On patient-level splitting, contrast-free claims and unsupported comparators in “Machine learning-based classification of multiple sclerosis lesion activity using multi-sequence MRI radiomics”

Stefania Galassi<sup>A,C,D,E,F</sup>

Neuroradiology Unit, SS. Annunziata Hospital, Cosenza, Italy

### Dear Editor,

I read with interest the article by Elhaie and colleagues describing a machine-learning model that uses multi-sequence magnetic resonance imaging radiomics to classify active and inactive multiple sclerosis lesions [1]. The topic is clinically important and the manuscript is clearly presented, but several aspects would benefit from clarification.

The first point concerns how patients and lesions were allocated between model development and evaluation. Because multiple lesions from the same person share acquisition conditions and biological context, distributing lesions from a single patient across both the training and test sets may make the results look better than they would be in practice. The safest approach is to assign each patient entirely to a single data split or to use cross-validation that groups by patient, so no individual contributes data to more than one fold [2].

There also appears to be a discrepancy between the accrual dates reported in the abstract and those in the methods. Reconciling the study window would help readers understand the timeline and any scanner or protocol changes that might affect radiomic features [1].

The work is framed as “contrast-free,” yet T1-weighted imaging (T1W) was acquired both before and after gadolinium administration, and it is not explicit which version was used for feature extraction. If post-contrast images entered the model, the “contrast-free” claim should be moderated; if only pre-contrast T1W was used (alongside

T2-weighted imaging/FLAIR/DWI/SWI), stating this plainly would avoid confusion.

The abstract also states that performance was comparable to radiologists, but no reader benchmark is shown. Unless a human-reader analysis was actually performed, that phrasing should be removed or supported with data. Given the small internal test, it would be more informative to pre-specify a single operating threshold on validation and carry it forward unchanged to the test, then report precision–recall area under the curve and predictive values at plausible prevalences to match deployment decisions [3,4]. As recently argued in related correspondence, clarity on thresholds and calibration reduces optimistic interpretations and improves reproducibility [5].

Finally, a note on preprocessing. Intensity normalisation and similar transformations should be fitted on training data only and then applied, unchanged, to validation and test. Stating this explicitly – and, if helpful, providing a sensitivity analysis – would rule out inadvertent information leakage and strengthen confidence in the results [2,4].

### Disclosures

1. Institutional review board statement: Not applicable.
2. Assistance with the article: None.
3. Financial support and sponsorship: None.
4. Conflicts of interest: None.

### Correspondence address:

Stefania Galassi, Neuroradiology Unit, SS. Annunziata Hospital, Via Felice Migliori 1, 87100 Cosenza, Italy, e-mail: [stefaniagalassi@gmail.com](mailto:stefaniagalassi@gmail.com)

### Authors' contribution:

A Study design · B Data collection · C Statistical analysis · D Data interpretation · E Manuscript preparation · F Literature search · G Funds collection

---

## References

1. Elhaie M, Etemadifar M, Rezaei Adariani A, Khorasani A, Shahbazi-Gahrouei D. Machine learning-based classification of multiple sclerosis lesion activity using multi-sequence MRI radiomics: a complete analysis of T1, T2, FLAIR, DWI, and SWI features. *Pol J Radiol* 2025; 90: e394-e403. DOI:10.5114/pjr/206986.
2. Roberts DR, Bahn V, Ciuti S, Boice MS, Elith J, Guillerá-Arroita G. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 2017; 40: 913-929.
3. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015; 10: e0118432. DOI:10.1371/journal.pone.0118432.
4. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Calster BV, et al. TRIPOD+AI: updated reporting guidance for clinical prediction models that use regression or machine learning methods. *BMJ* 2024; 385: e078378. DOI:10.1136/bmj-2023-078378.
5. Pastore EP. Correspondence regarding “Clinical parameters-based machine learning models for predicting intraoperative hemodynamic instability in hypertensive pheochromocytomas and paragangliomas patients”. *World J Urol* 2025; 43: 610. DOI:10.1007/s00345-025-06002-8.