Effect of data leakage in brain MRI classification using 2D convolutional neural networks

Ekin Yagis[1,5,*], Selamawet Workalemahu Atnafu[2,5], Alba García Seco de Herrera[1,6], Chiara Marzi[2], Riccardo Scheda[2], Marco Giannelli[3], Carlo Tessa[4], Luca Citi[1,6], and Stefano Diciotti[2,6]

[1] *School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK.*

[2] *Department of Electrical, Electronic, and Information Engineering "Guglielmo Marconi", University of Bologna, Via dell'Università 50, 47521 Cesena, Italy.*

[3] *Unit of Medical Physics, Pisa University Hospital "Azienda Ospedaliero-Universitaria Pisana", Pisa, Italy.*

[4] *Division of Radiology, Versilia Hospital, Azienda USL Toscana Nord Ovest, Lido di Camaiore, LU, Italy.*

[5] *These authors contributed equally: Ekin Yagis and Selamawet Workalemahu Atnafu.*

[6] *These authors jointly supervised this work: Alba García Seco de Herrera, Luca Citi and Stefano Diciotti.*

[*] ekinyagis@gmail.com

In recent years, 2D convolutional neural networks (CNNs) have been extensively used to diagnose neurological diseases from magnetic resonance imaging (MRI) data due to their potential to discern subtle and intricate patterns. Despite the high performances reported in numerous studies, developing CNN models with good generalization abilities is still a challenging task due to possible data leakage introduced during cross-validation (CV). In this study, we quantitatively assessed the effect of a data leakage caused by 3D MRI data splitting based on a 2D slice-level using three 2D CNN models to classify patients with Alzheimer's disease (AD) and Parkinson's disease (PD). Our experiments showed that slicelevel CV erroneously boosted the average slice level accuracy on the test set by 30% on Open Access Series of Imaging Studies (OASIS), 29% on Alzheimer's Disease Neuroimaging Initiative (ADNI), 48% on Parkinson's Progression Markers Initiative (PPMI) and 55% on a local de-novo PD Versilia dataset. Further tests on a randomly labeled OASIS-derived dataset produced about 96% of (erroneous) accuracy (slice-level split) and 50% accuracy (subject-level split), as expected from a randomized experiment. Overall, the extent of the effect of an erroneous slice-based CV is severe, especially for small datasets.

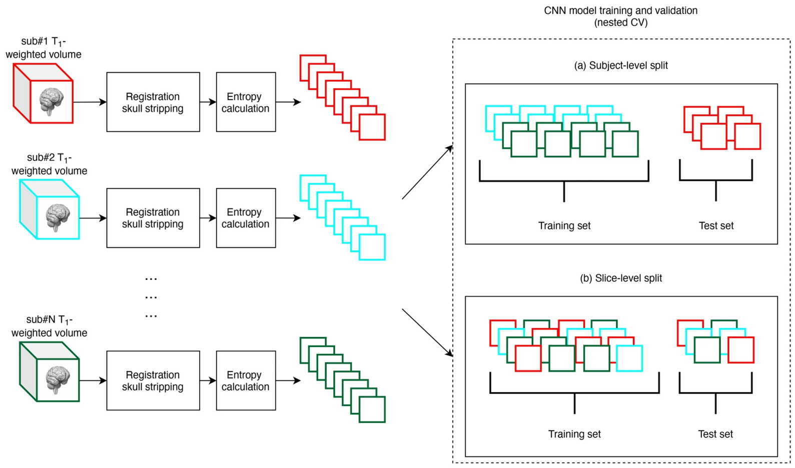
CNN model training and validation
(nested CV)
sub#1 T_1-weighted volume
Registration skull stripping
Entropy calculation
sub#2 T_1-weighted volume
Registration skull stripping
Entropy calculation
...
...
...
sub#N T_1-weighted volume
Registration skull stripping
Entropy calculation
(a) Subject-level split
Training set
Test set
(b) Slice-level split
Training set
Test set

Figure 1: Schematic diagram of the overall T1-weighted MRI data processing and validation scheme. First, a preprocessing stage included co-registration to a standard space, skull-stripping and slices selection based on entropy calculation. Ten, CNNs model's training and validation have been performed on each dataset in a nested CV loop using two diferent data split strategies: (a) subject-level split, in which all the slices of a subject have been placed either in the training or in the test set, avoiding any form of data leakage; (b) slice-level split, in which all the slices have been pooled together before CV, then split randomly into training and test set [1].

References

[1] E. Yagis *et al.*, *Sci. Rep.* **11**, 22544 (2021). DOI: 10.1038/s41598-021-01681-w