Beware of feature selection bias! Example on Alzheimer's disease classification from diffusion MRI

Junhao Wen, Jorge Samper-González, Alexandre M Routier, Simona Bottani, Stanley Durrleman, Ninon Burgos, Olivier Colliot

Beware of feature selection bias! Example on Alzheimer's disease classification from diffusion MRI

Junhao Wen[1], Jorge Samper-González[1], Alexandre Routier[1], Simona Bottani[1], Stanley Durrleman[1], Ninon Burgos[1], Olivier Colliot[1,2]

[1]ARAMIS Lab, ICM, Inserm U1127, CNRS UMR 7225, Sorbonne University, Inria, Paris, France
[2]AP-HP, Departments of Neuroradiology and Neurology, Pitié-Salpêtrière Hospital, Paris, France

Introduction

Various studies leveraged machine learning and diffusion tensor imaging (DTI) techniques for Alzheimer's disease (AD) classification, often reporting high classification accuracies. However, several of these studies used inadequate validation of feature selection (FS) procedures, leading to potentially over-optimistic results (Graña et al., 2011; Mesrob et al., 2012; O'Dwyer et al., 2012). In previous works (Samper-González et al., 2018, Wen et al., 2018), we proposed a reproducible framework for automatic classification of AD from MRI (T1-weighted and diffusion) and PET data. Here, we applied this framework to diffusion MRI to study the potential bias due to improper FS.

Methods

The framework comprises different components. First, tools were developed to automatically convert original ADNI diffusion MRI into BIDS format. Secondly, an image preprocessing pipeline was implemented, including steps for susceptibility-induced distortions, eddy current-induced distortions and head motion corrections. FA and MD maps were generated by fitting DTI model. FA and MD maps were nonlinearly registered onto the John Hopkins University (JHU) atlas template. We then extracted the voxel-based features, from which FA and MD maps were masked using the GM+WM binarized maps. Classification was performed using a linear support vector machine (SVM) from scikit-learn. A repeated holdout CV (250 runs of stratified random splits with 20% of the data used for testing) with a 10-fold inner grid search for hyperparameter optimization was performed.

After the feature extraction, both non-nested FS and nested FS strategies were embedded into the current classification framework, as shown in Figure 1. More precisely, the non-nested FS was performed with the entire dataset and totally independent from the cross-validation (CV) procedure. On the contrary, a nested FS is a procedure blind to the test data and incorporated into the nested CV (Maggipinto et al., 2017, Kriegeskorte et al., 2009; Rathore et al., 2017). Two FS algorithms were studied: i) filter-based univariate method (ANOVA) and ii) wrapper-based multivariate method (SVM-RFE). For both methods, we tested varying numbers of selected features (1% of the total number of features and then from 10% to 100%, increasing by 10% at each step).

Experiments were performed with 46 AD patients and 46 cognitively normal (CN) subjects.

Results

All classification results are shown in Figure 2. The non-nested FS gave vastly over-optimistic results, from 5 up to 40 percentage points increase in balanced accuracy. For instance for FA, the balanced accuracy was 0.99 with non-nested SVM-RFE and 0.75 with nested SVM-RFE. For ANOVA, the best performance was obtained with the first 1% most informative voxels for non-nested approach (0.78 for FA and 0.83 for MD), and with all available voxels for nested approach (0.71 for FA and 0.76 for MD). For SVM-RFE, the best performance was achieved with the first 10% most informative voxels for non-nested

approach (0.99 for FA and 0.83 for MD), and with the first 70% most informative voxels with FA (0.75) and the first 1% most informative voxels with MD (0.77) for nested approach. Compared to non-FS case, the nested ANOVA FS did not give better performance. The nested SVM-RFE slightly improved the performance compared to non-FS: balanced accuracy increased from 0.71 (non-FS) to 0.75 (nested FS) for FA and 0.76 (non-FS) to 0.77 (nested FS) for MD.

Conclusions

We demonstrated that inadequate evaluation of FS strategies lead to highly over-optimistic classification performances. The bias is particularly severe for multivariate FS methods. Such approach being unfortunately still too common, we believe it is important to raise awareness about this issue in the community. The code will be made publicly available at the time of the conference at https://gitlab.icm-institute.org/aramislab/AD-ML.

References

Graña, M., Termenon, M., Savio, A., Gonzalez-Pinto, A., Echeveste, J., Pérez, J.M., Besga, A., 2011. Computer aided diagnosis system for Alzheimer disease using brain diffusion tensor imaging features selected by Pearson's correlation. Neurosci. Lett. 502, 225–229.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., Baker, C.I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. 12, 535–540.

Maggipinto, T., Bellotti, R., Amoroso, N., Diacono, D., Donvito, G., Lella, E., Monaco, A., Antonella Scelsi, M., Tangaro, S., 2017. DTI measurements for Alzheimer's classification. Phys. Med. Biol. 62, 2361–2375.

Mesrob, L., Sarazin, M., Hahn-Barma, V., Souza, L.C. de, Dubois, B., Gallinari, P., Kinkingnéhun, S., 2012. DTI and Structural MRI Classification in Alzheimer's Disease. AMI 02, 12–20.

O'Dwyer, L., Lamberton, F., Bokde, A.L.W., Ewers, M., Faluyi, Y.O., Tanner, C., Mazoyer, B., O'Neill, D., Bartley, M., Collins, D.R., Coughlan, T., Prvulovic, D., Hampel, H., 2012. Using support vector machines with multiple indices of diffusion for automated classification of mild cognitive impairment. PLoS One 7, e32441.

Rathore, S., Habes, M., Iftikhar, M.A., Shacklett, A., Davatzikos, C., 2017. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. Neuroimage 155, 530–548.

Samper-González, J., Burgos, N., Bottani, S., Fontanella, S., Lu, P., Marcoux, A., Routier, A., Guillon, J., Bacci, M., Wen, J., Bertrand, A., Bertin, H., Habert, M.O., Durrleman, S., Evgeniou, T., Colliot, O., Alzheimer's Disease Neuroimaging Initiative, Australian Imaging Biomarkers and Lifestyle flagship study of ageing, 2018. Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data. 183, pp 504-521, Neuroimage. https://doi.org/10.1016/j.neuroimage.2018.08.042

Wen, J., Samper-González, J., Bottani, S., Routier, A., Burgos, N., Jacquemont, T., Fontanella, S., Durrleman, S., Bertrand, A., Colliot, O., 2018. Using Diffusion MRI for Classification and Prediction of Alzheimer's Disease: A Reproducible Study. Alzheimer's & Dementia: The Journal of the Alzheimer's Association. 2018 July 1;14(7):P891-2.

Figures

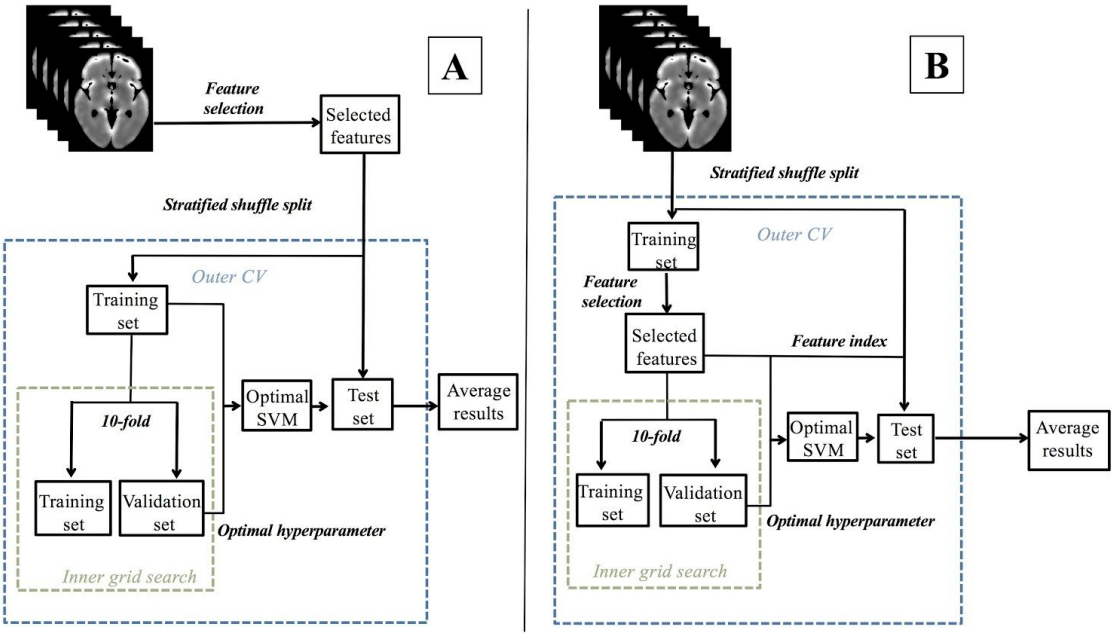

Figure 1. Flow chart of feature selection procedure. For cross-validation (CV), a repeated holdout CV (250 runs of stratified random splits with 20% of the data used for testing) with a 10-fold inner grid search for hyperparameter optimization was performed. (A) Non-nested feature selection; (B) Nested feature selection.

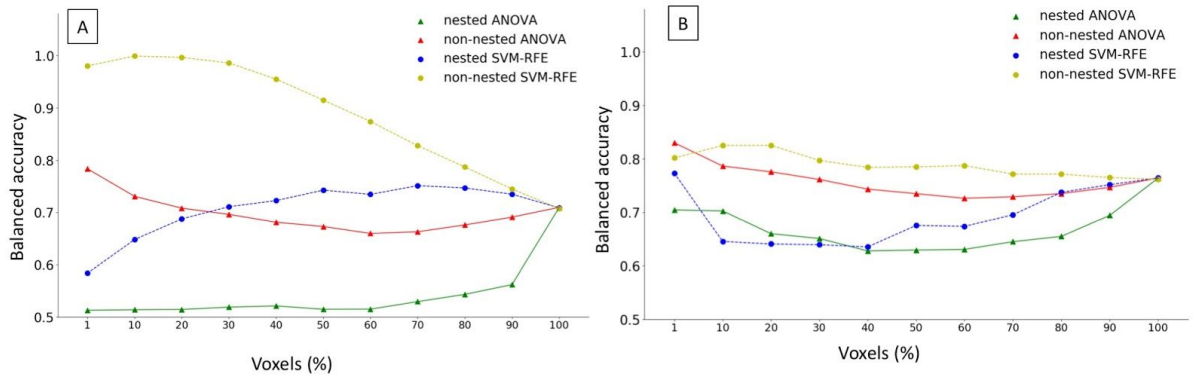

Figure 2. Balanced accuracy of CN vs AD obtained varying the number of voxels for ANOVA and SVM-RFE approaches. (A) GM+WM-FA feature; (B) GM+WM-MD feature.