Reproducible evaluation of Alzheimer's Disease classification from MRI and PET data

Jorge Samper-Gonzalez, Simona Bottani, Ninon Burgos, Sabrina Fontanella, Pascal Lu, Arnaud Marcoux, Alexandre Routier, Jérémy Guillon, Michael Bacci, Junhao Wen, et al.

Reproducible evaluation of Alzheimer's Disease classification from MRI and PET data

Jorge Samper-González[1,2], Simona Bottani[1,2], Ninon Burgos[1,2], Sabrina Fontanella[1,2], Pascal Lu[1,2], Arnaud Marcoux[1,2], Alexandre Routier[1,2], Jérémy Guillon[1,2], Michael Bacci[1,2], Junhao Wen[1,2], Anne Bertrand[2,3,4], Hugo Bertin[5], Marie-Odile Habert[5,6], Stanley Durrleman[1,2], Theodoros Evgeniou[7], Olivier Colliot[1,2, 8]

[1]Inria Paris, Aramis project-team, Paris, France
[2]Sorbonne Universités, UPMC Univ Paris 06, Inserm, CNRS, Institut du cerveau et la moelle (ICM) - Hôpital Pitié-Salpêtrière, Paris, France
[3]Sorbonne Universités, UPMC Univ Paris 06, Inserm, CNRS, Institut du cerveau et la moelle (ICM), Assistance Publique-Hôpitaux de Paris (AP-HP) - Hôpital Pitié-Salpêtrière, Paris, France
[4]AP-HP, Hôpital Saint Antoine, Department of Radiology, Paris, France
[5]Laboratoire d'Imagerie Biomédicale, Sorbonne Universités, UPMC Univ Paris 06, Inserm U 1146, CNRS UMR 7371, Paris, France
[6]AP-HP, Hôpital Pitié-Salpêtrière, Department of Nuclear Medicine, Paris, France
[7]INSEAD, Fontainebleau, France
[8]AP-HP, Department of Neuroradiology, Pitié-Salpêtrière Hospital, Paris, France

Keywords:
- Data organization
- Machine Learning
- Workflows
- Degenerative Disease
- STRUCTURAL MRI
- Positron Emission Tomography (PET)

Introduction

Various publications have proposed machine learning approaches to classify and predict Alzheimer's disease (AD) from neuroimaging data (e.g. Rathore et al, 2017; Jie et al, 2015; Cuingnet et al, 2013; Young et al, 2013; Fan et al, 2008; Klöppel et al, 2008). The vast majority make use of the Alzheimer's Disease Neuroimaging Initiative (ADNI) public dataset. However, such studies usually differ in terms of: i) subsets of subjects; ii) image processing pipelines; iii) feature extraction and selection; iv) machine learning algorithms; v) cross-validation procedures and vi) reported evaluation metrics. These differences make it, in practice, impossible to determine which methods perform the best and difficult to assess which contributions provide a real classification improvement, e.g. a specific image processing or classification algorithm. We propose a framework for the reproducible evaluation of machine learning approaches in AD. The main contributions are a framework for management of three public datasets: ADNI, the Australian Imaging Biomarker and Lifestyle study (AIBL) and the Open Access Series of Imaging Studies (OASIS) and a modular set of preprocessing pipelines, feature extraction and classification methods,

together with an evaluation framework that provides a baseline for benchmarking the different components. The present work extends that of (Samper-González et al, 2017) by including more datasets (AIBL, OASIS), more feature types and more classification algorithms. We demonstrate the use of the framework for comparison of different classifiers, features and imaging modalities.

Methods

Despite their incontestable value, AD public datasets such as ADNI and AIBL do not rely on community standards for data organization and lack a clear structure. This poses a significant setback on their immediate use. We provide code that performs the conversion of the data as they were downloaded for ADNI, AIBL and OASIS into Brain Imaging Data Structure (BIDS) format (Gorgolewski et al., 2016), which is a community standard. This allows direct reproducibility by other groups without having to redistribute the dataset. Tools for subject selection according to imaging modalities, duration of follow up and diagnoses are provided.

A T1 MRI processing pipeline was implemented using SPM, involving tissue segmentation, a DARTEL group template creation, registration to MNI space, optional smoothing and regional parcellation. For PET images, a pipeline performing an optional partial volume correction (PVC) step, spatial normalization, computation of standardized uptake value ratio (SUVR) maps and parcellation was developed. A BIDS-inspired standardized structure was defined for the pipelines' outputs.

We proposed an evaluation framework consisting of three layers: i) an input to select the imaging modality and the features (regions or voxels); ii) a cross validation method (we performed 250 runs of stratified random splits with 70% as training set); iii) a classification algorithm (SVM, L2 Logistic regression and Random Forest). Accuracy, balanced accuracy, AUC, sensitivity, specificity and subjects predicted class are reported.

Results

We found that FDG PET provides better classification results than T1 MRI for all the tasks and features tested, and that random forest systematically performs worse than SVM and L2 logistic regression (Fig 1). All the voxel-based classifications results are shown in Fig 2. We observed that ADNI trained SVM classifiers generalize well when tested on AIBL and OASIS datasets. Of note, they perform better than those trained on AIBL and OASIS, probably because of ADNI's larger number of patients.

Conclusions

We proposed a framework for the evaluation of machine learning algorithms that could prove a useful tool for improving comparability and reproducibility in AD classification. The new version of the code will be made publicly available at the time of the conference at https://gitlab.icm-institute.org/aramislab/AD-ML.

References

Cuingnet, R., Glaunès, J. A., Chupin, M., Benali, H., Colliot, O., ADNI (2013), 'Spatial and anatomical regularization of SVM: a general framework for neuroimaging data', IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, issue 3, pp. 682-696

Fan, Y., Batmanghelich, N., Clark, C.M., Davatzikos, C., ADNI (2008). 'Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline', Neuroimage, vol. 39, 1731–1743

Gorgolewski, K.J., Auer, T., Calhoun, V.D., Craddock, R.C., Das, S., Duff, E.P., Flandin, G., Ghosh, S.S., Glatard, T., Halchenko, Y.O., Handwerker, D.A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B.N., Nichols, T.E., Pellman, J., Poline, J.-B., Rokem, A., Schaefer, G., Sochat, V., Triplett, W., Turner, J.A., Varoquaux, G., Poldrack, R.A. (2016), 'The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments', Scientific Data, vol. 3, article number 160044

Jie, B., Zhang, D., Cheng, B., Shen, D., ADNI (2015). 'Manifold regularized multitask feature learning for multimodality disease classification', Hum. Brain Mapp., vol. 36, pp. 489–507.

Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack, C.R., Jr, Ashburner, J., Frackowiak, R.S.J. (2008), 'Automatic classification of MR scans in Alzheimer's disease', Brain, vol. 131, pp. 681–689

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, vol. 155, pp. 530-548

Samper-González, J., Burgos, N., Fontanella, S., Bertin, H., Habert, M.-O., Durrleman, S., Evgeniou, T., Colliot, O., ADNI (2017), 'Yet another ADNI machine learning paper? Paving the way towards fully-reproducible research on classification of Alzheimer's disease', Machine Learning in Medical Imaging, MLMI 2017, LNCS, vol. 10541, pp. 53–60

Young, J., Modat, M., Cardoso, M.J., Mendelson, A., Cash, D., Ourselin, S., ADNI (2013), 'Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment', Neuroimage Clin, vol. 2, pp. 735–745

Figures

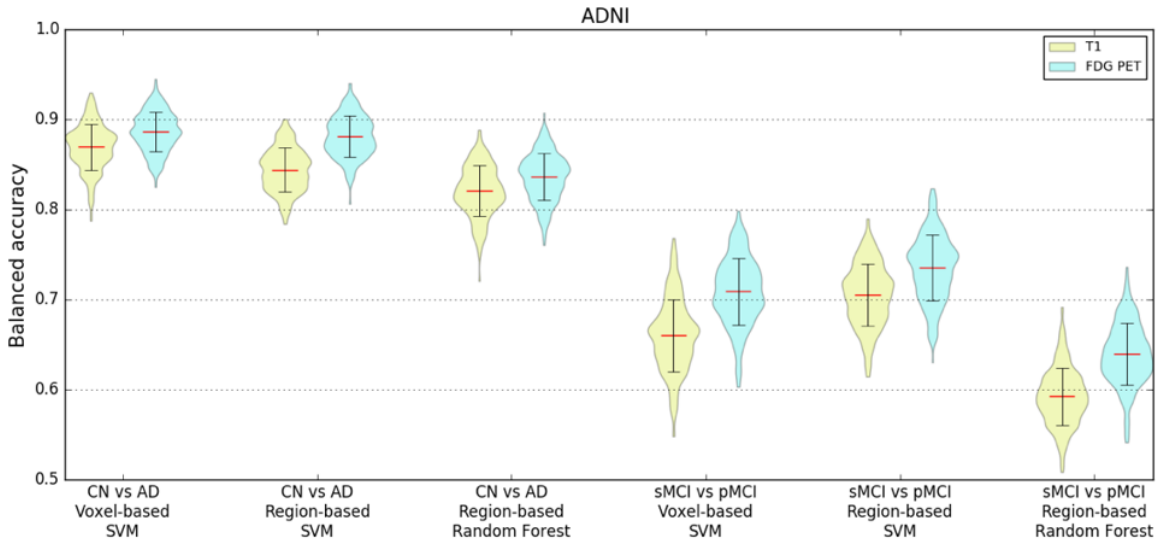

Figure 1: Distribution of the balanced accuracy obtained from the T1w MRI and FDG PET images of ADNI participants for the CN vs AD (282 vs 237) and sMCI vs pMCI (342 vs 167) tasks, using voxel-based features (smoothing 4 mm) with linear SVM, and region-based features (atlas AAL2) with linear SVM and random forest.

Dataset Training	Dataset Testing	Image Type	Task	AUC	Balanced Accuracy	Sensitivity	Specificity
ADNI	ADNI	T1w MRI	CN vs AD	0.93	0.87	0.83	0.90
			CN vs pMCI	0.81	0.74	0.86	0.62
			sMCI vs pMCI	0.73	0.66	0.58	0.74
		FDG PET	CN vs AD	0.96	0.89	0.85	0.92
			CN vs pMCI	0.85	0.77	0.85	0.69
			sMCI vs pMCI	0.78	0.71	0.65	0.77
	AIBL	T1w MRI	CN vs AD	0.93	0.88	0.86	0.89
	OASIS	T1w MRI	CN vs AD	0.82	0.76	0.73	0.80
AIBL	AIBL	T1w MRI	CN vs AD	0.92	0.79	0.61	0.97
OASIS	OASIS	T1w MRI	CN vs AD	0.75	0.70	0.72	0.67

Figure 2: Classification results obtained with voxel-based features (smoothing 4 mm) using linear SVM. When training and testing on the same dataset, the mean of the 250 runs is reported. For ADNI, 282 CN subjects, 237 AD, 342 sMCI and 167 pMCI have been used in the classification, for AIBL, 442 CN and 72 AD subjects, and for OASIS, 93 CN and 100 AD subjects.