Longitudinal FDG-PET features for the classification of Alzheimer's Disease

Filipa Rodrigues[1], Margarida Silveira[2] and the Alzheimer's Disease Neuroimaging Initiative

Abstract— **The majority of existing computer-aided diagnosis (CAD) schemes for Alzheimer's disease (AD) rely on the analysis of biomarkers at a single time-point, ignoring the progressive nature of the disorder. Recently, a method was proposed by Gray et al. [1] for the multi-region analysis of longitudinal fluorodeoxyglucose positron emission tomography (FDG-PET) images which reported classification improvements by using regional signal intensities combined with regional change over a 12 month period. In this paper we extend the approach proposed in [1] to the analysis of the entire brain pattern. Compared to [1], our method uses voxel-wise differences and avoids segmentation of the images into regions of interest. For our study, FDG-PET scans at the baseline and at 12-month follow-up of cognitively normal (CN), mild cognitive impairment (MCI) and AD subjects were retrieved from the Alzheimer's disease neuroimaging initiative (ADNI) database. For both AD and MCI identification, the best classification results were achieved by combining cross-sectional and longitudinal information rather than using only the cross-sectional data. Furthermore, the longitudinal voxel-based analysis outperformed multi-region analysis.**

I. INTRODUCTION

Alzheimer's disease (AD) is a neurodegenerative disorder and the most common cause of dementia in the elderly. AD is characterized by long-term memory loss along with the progressive decline of other cognitive functions and a cure has not been found yet [2]. The onset of AD is typically preceded by an intermediate stage known as mild cognitive impairment (MCI) [3]. Since AD is a progressive disorder, the diagnosis of the MCI stage is crucial in order to enable earlier therapeutic intervention which would offer greater protection against further neuronal damage. Fluorodeoxiglucose positron emission tomography (FDG-PET) has been reported as a powerful MCI and AD biomarker. Several FDG-PET studies have shown that metabolic patterns and its progression along time differ from cognitively normal (CN), MCI and AD subjects. For example, a reduced FDG-PET metabolism within specific brain regions in AD and MCI patients compared to CN individuals was reported in [4] and [5]. In their 3-year longitudinal study, de Leon et al. [6] concluded that metabolic reduction along time could predict the conversion from CN to MCI or AD. In 2010 Chen et al. [7], compared the decline of brain metabolic rate over 12 months in CN, MCI and AD patients reporting significant group differences. Despite of the findings in [6] and [7], when the utility of incorporating longitudinal information was tested by Hinrichs et al. [8] in their multimodality approach, the results obtained suggested that longitudinal data alone are not sufficient to accurately identify AD. Two approaches using FDG-PET data were considered (voxel-wise temporal differences and voxel-wise temporal ratio) and both performed worse than cross-sectional data. In 2012, Gray et al. [1] proposed a multi-region approach to evaluate the combination of cross-sectional with longitudinal FDG-PET imaging data. They reported a classification accuracy improvement by using regional signal intensities combined with regional change over a 12 month period. The aim of the present work is to extend the method proposed in [1] to the analysis of the whole brain voxel pattern for the classification of CN vs AD and CN vs MCI. Since our method uses a voxel-wise approach, the procedure of segmenting the brain into regions of interest is avoided.

*This work was supported by FCT project ADIAR PTDC/SAU-ENB/114606/2009 and LARSyS-PEst-OE/EEI/LA0009/2013. Data used in the preparation of this article were obtained from the ADNI database (http://www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report.

[1]Faculdade de Ciências, Universidade de Lisboa, Portugal

[2]Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, Portugal

II. MATERIALS AND METHODS

A. Subjects

The data used in this study were retrieved from the Alzheimer's disease neuroimaging initiative (ADNI) database (http://adni.loni.usc.edu/). The ADNI study enrolls CN, MCI and AD subjects and its primary goal is to develop imaging, clinical, genetic and biochemical biomarkers for early detection and tracking of AD. For our study, data from participants with baseline and 12-month follow-up FDG-PET scans were used. Imaging data from 223 subjects were used, 66 were from CN subjects, 109 from MCI subjects and 48 from AD subjects. Some demographic and clinical information (CDR - Clinical Dementia Rating, MMSE - Mini-Mental State Examination) at baseline of the study population is presented in TABLE I. Age and gender of the different clinical groups do not vary significantly (p-value$>$0.05) according to the t-test performed between the different group pairs (CN vs AD, CN vs MCI, MCI vs AD).

B. Image Processing

1) Registration: In order to make voxel-wise comparisons, all images were warped into the MNI152 standard space. The warping procedure used PET and MRI images, both of which had undergone a sequence of preprocessing steps by ADNI researchers to eliminate meaningless differences. Firstly, the brain tissue in all MR images was extracted (skull-stripping) and then segmented into white-matter (WM)

TABLE I

DEMOGRAPHIC AND CLINICAL INFORMATION OF THE STUDY POPULATION (MEAN ± STANDARD DEVIATION)

Group	CN	MCI	AD
No. of subjects	66	109	48
Gender (M/F)	41/25	69/40	27/21
Age, mean ± SD	75.9 ± 4.5	74.8 ± 7.1	76.5 ± 6.7
CDR, mean ± SD	0.0 ± 0.0	0.5 ± 0.0	0.8 ± 0.2
MMSE, mean ± SD	29.1 ± 1.0	27.2 ± 1.6	23.4 ± 2.0

and gray-matter (GM). The extraction of brain tissue was performed with FreeSurfer [9] and tissue classification was conducted with SPM8 [10]. Secondly, all PET images were co-registered with the corresponding skull-stripped MR images using SPM8. In order to conduct these co-registrations rigid-body transformations (6 degrees of freedom) and an objective function based on the "sharpness" of the normalized mutual information between the two images was used [11]. Thirdly, the MR images acquired at baseline and month 12 for each subject were non-linearly registered into a subject-specific template using the DARTEL toolbox from SPM8 [12]. Finally, all MR images taken during the screening visit were non-linearly registered to an inter-subject template using DARTEL. The template was then mapped to the MNI-ICBM 152 non-linear symmetric atlas (version 2009a) [13] using an afine transformation. After completing the above steps, the original PET and MR images were resampled into the MNI152 standard space with a $1.5\times1.5\times1.5$ mm resolution using the appropriate composition of transformations. The final images were represented by a $121\times145\times121$ matrix.

2) FDG-PET Normalization: Intensity normalization of FDG-PET images is usually performed relative to the cerebral global mean. However, this procedure requires that the cerebral global mean does not vary significantly between the subjects under study, which usually does not occur in neurodegenarative disorders such as AD. Hence, this procedure conducts to an attenuation of the differences between clinical groups, since the intensity signal of FDG-PET images from patients are artificially scaled up while those from healthy individuals are scaled down. Therefore, this type of normalization results in an apparent hypometabolism for healthy subjects in regions that are known to be relatively preserved in AD, such as the brainstem, cerebellum, basal ganglia and sensorimotor cortex [14]. Recent studies suggest that using these preserved regions instead of using the cerebral global mean for normalization leads to an improvement of the clinical groups discrimination [15]. However, these methods require that the regions are defined *a priori* without knowing if in fact these regions are the most preserved in the group of subjects under study. In 2009, Yakushev et al. [16] proposed a different approach to define a reference cluster for normalization. This method relies on defining a reference cluster *a posteriori*, and consists of 2 iterations. In the first iteration, a cerebral mean global normalization is performed. Then, in the second iteration a t-test is conducted in order to find the apparently hypermetabolic (p-value$<$0.05) regions in the patient group compared to the healthy one, and these regions are chosen to be the reference cluster. This procedure relies on the principle that the apparently hypermetabolic regions found in the pathological group compared to the healthy control one are in fact preserved regions. In our work, intensity normalization of the FDG-PET images was conducted performing this reference cluster method.

C. Feature Selection

Two different types of features were tested for classification. One of the sets was based on a multi-region approach, using only the mean intensity of brain regions as features. The other set included intensity of all brain voxels as features. These two different approaches are explained below:

1) Multi-region analysis: In order to obtain regional features, an atlas developed by Desikan et al. [17] was used to automatically segment the whole brain into 69 anatomical regions. The atlas was used to segment both baseline and 12-month follow up images. The mean of the signal intensity of each region was extracted for the two imaging time-points. Furthermore, the regional change over a 12 month follow-up was determined. Two different sets of features were used for classification. One set contained only cross-sectional information (signal intensities of the baseline or at 12-month follow-up images), and the other set contained the cross-sectional combined with the longitudinal information (regional change over 12-month follow-up) by simple concatenation. A 10-fold cross validation strategy was used to assess the classification accuracy, i.e., the dataset was split, 90% was used for training and 10% was used for testing. This process was repeated 10 times in order to avoid any bias due to the random partition of the dataset in cross-validation.

2) Voxel-based analysis: For the voxel-based analysis, the signal intensity of intracerebral voxels were used as features, summing up a total of 486353 voxels. In order to select the most discriminative voxels between clinical groups, a feature ranking method was applied. The rank score was the t-test performed between two different pairs of clinical groups (CN vs MCI and CN vs AD). The t-value obtained reflects the rate of atrophy of the AD and MCI patients in each voxel compared to the CN. Voxels with larger absolute t-values correspond to higher atrophy rates thus with higher discriminative power. For our classification experiment, sets with varying number of features were used, starting with 1000 features (the 1000 voxels with higher absolute t-values), and in each iteration 1000 features were added, going up to a maximum of 15000 features. Similarly to what was done in the multi-region analysis, two types of features were used. In one of the sets, the features were simply the voxels selected by our feature selection method in a single time-point imaging data (baseline or 12-month). The second set also included the longitudinal information, i.e., the change over the follow-up period of the 15000 features selected in a single-time point. The longitudinal change was obtained by subtracting the intensity of each voxel at the baseline from the intensity of the correspondent

voxel at month 12. Moreover, in order to select the most discriminative longitudinal features from the 15000 features selected in the single time-point, an additional step was performed. Assuming that voxels with higher differences along the follow period correspond to brain regions with higher atrophy rate due to the progress of the disease, the longitudinal differences of each feature were sorted in descending order. Hence, in each iteration, voxels with higher differences between the two-time points were selected first for the classification. In this set each subject had 2000 (1000 cross-sectional features plus 1000 longitudinal features) up to a maximum of 30000 features for classification. The cross-sectional and longitudinal information were combined by simple concatenation of features. Again a 10-fold cross validation strategy was performed to assess the classification accuracy, and the result was averaged over the 10 folds. This process was repeated 10 times.

D. Classification

A support vector machine (SVM) was used as classifier implemented on LIBSVM matlab toolbox [18]. For the multi-region analysis, since the number of features is relatively well balanced with the number of subjects, a radial basis functions (RBF) kernel was used along with the optimization of the C (range: 10E-6 to 10E2) and γ (range: 10E-5 to 5E-2) parameters by performing a grid-search using 5-fold cross validation within the training set. For the voxel-based features, since the number of features were much larger that the number of subjects, a linear kernel was chosen. The C parameter value (range: 10E-6 to 10E2) of the SVM algorithm was also tuned using a 5-fold nested cross validation strategy.

III. Results

A. Classification Experiments

In Fig.1, the results of the voxel-based approach for the classification of CN vs AD are presented. The upper plot shows the classification accuracy using the signal intensity of the voxels at the baseline and the classification accuracy using the signal intensity of the voxels at the baseline combined with the change of the signal intensity over a 12-month period. As one can observe, better results are achieved by combining cross-sectional and longitudinal information. According to the t-test, these two results are significantly different (p-value$<$0.05). The bottom plot exhibits the classification accuracy obtained when using the signal intensities of the voxels at 12-month follow-up images as features, and the classification accuracy using this information combined with the change over the follow-up period. In this case, the results of the t-test performed to the differences in the classification results using these two sets of features show that these are not statistically different (p-value$>$0.05).

Fig.2 presents the same results as Fig.1 but now for the CN vs MCI classification. The upper plot represents the results using the signal intensity of the voxels of the baseline images and results of the combination of this information with the change of the signal intensity over 12-month. Again,

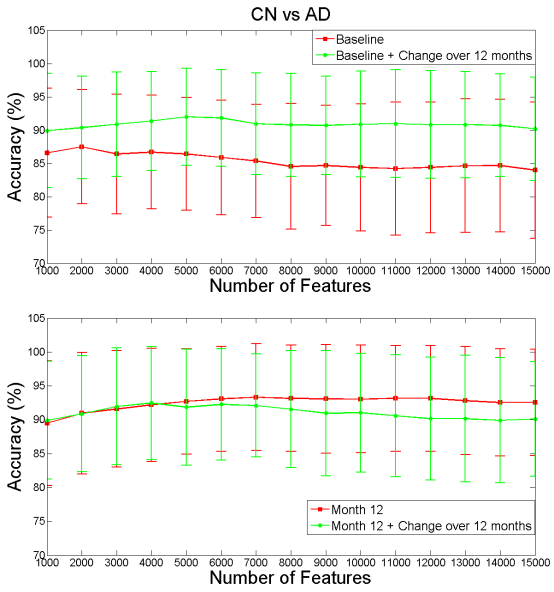

Fig. 1. Classification accuracy for CN vs AD as a function of the number of features. The averaged accuracy and the standard deviation were obtained using 10-fold cross validation repeated 10 times.

higher classification accuracy is achieved when combing the cross sectional and longitudinal information, and these differences are statistically significant according to t-test (p-value$<$0.05). The bottom plot shows the classification results when using the imaging data at the 12-month, and also this information combined with the change over the 12-month follow-up period. The differences of the results using these two different sets of features are statistically significant (p-value$<$0.05).

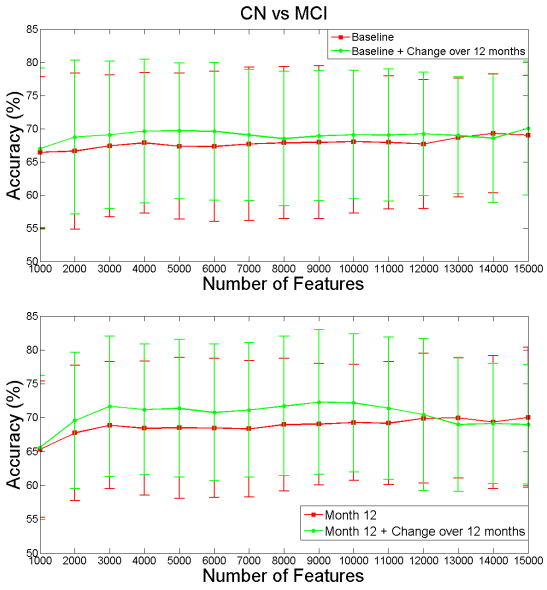

Fig. 2. Classification accuracy for CN vs MCI as a function of the number of features. The averaged accuracy and the standard deviation were obtained using 10-fold cross validation repeated 10 times.

Classification results using the multi-region approach are displayed in Table II. In order to compare this method

with the voxel-based approach, the results obtained using voxels as features are also shown in Table II. These results were obtained by performing an optimization of the number of features within the training set adopting a 10-fold nested cross validation strategy and performing 10 iterations. For both approaches, the combination of single time and longitudinal data led to higher classification results (with statistical significance, p-value$<$0.05), in all cases except CN vs AD at month 12. Furthermore, the voxel-based approach outperforms the multi-region method for all the classification experiments. The best results were achieved for the classification of CN vs AD performing a voxel-based approach.

TABLE II

CLASSIFICATION ACCURACY (MEAN $\pm$ STANDARD DEVIATION) OVER 10 10-FOLD CROSS VALIDATION RUNS, USING DIFFERENT FEATURES.

Group		CN/AD	CN/MCI
Multi-Region Analysis	Baseline	81.1 $\pm$ 11.1	68.5$\pm$ 9.5
	Baseline + Change	83.3 $\pm$ 9.7	68.9 $\pm$9.7
	12-Month	87.4 $\pm$ 9.8	65.1 $\pm$11.3
	12-Month + Change	87.8 $\pm$ 9.1	65.6 $\pm$9.6
Voxel-Based Analysis	Baseline	84.2 $\pm$ 10.0	68.1 $\pm$ 10.6
	Baseline + Change	**91.2** $\pm$ **8.0**	69.3 $\pm$ 10.9
	12-Month	92.8 $\pm$ 6.3	69.7 $\pm$ 10.6
	12-Month + Change	**92.6** $\pm$ **6.7**	70.2 $\pm$ 9.0

B. Selected Features

The most discriminative features (selected with our feature ranking method more than 75% of the times) both for CN vs AD and CN vs MCI, included regions in inferior temporal gyrus, temporal fusiform cortex, temporal pole and parahippocampal gyrus. These results are in agreement with previous FDG-PET studies (for example, [5], [6]) that reported the temporal area as one of the most affected by the disease.

IV. DISCUSSION AND CONCLUSION

The combination of cross-sectional and longitudinal information led to higher classification accuracy compared with using only the cross-sectional data, for both AD and MCI identification. Furthermore, the best results were achieved using the signal intensity of the voxels as features instead of using the mean signal intensity of the anatomical regions. Similarly to the work of Gray et al. [1], our results suggest that longitudinal data can provide useful complementary information to improve the classification accuracy. Additionally, our voxel-based method outperforms (92.6% for CN vs AD) their multi-region approach (88.4% for CN vs AD) using data from the same database, and does not require segmentation of the images into regions of interest.

Despite these encouraging results, more suitable methods for longitudinal feature selection and extraction should be further investigated.

REFERENCES

[1] K. R. Gray, R. Wolz, R. Heckemann, P. Aljabar, A. Hammers, and D. Rueckert, "Multi-region analysis of longitudinal FDG-PET for the classification of Alzheimer's disease." *Neuroimage*, vol. 60, no. 1, pp. 221–9, Mar. 2012.

[2] C. Ballard, S. Gauthier, A. Corbett, C. Brayne, D. Aarsland, and E. Jones, "Alzheimer's disease." *Lancet*, vol. 377, no. 9770, pp. 1019–31, Mar. 2011.

[3] J. Golomb, A. Kluger, and S. Ferris, "Mild cognitive impairment: historical development and summary of research." *Dialogues Clin. Neurosci.*, pp. 351–367, 2004.

[4] M. Ewers, M. Brendel, A. Rizk-Jackson, A. Rominger, P. Bartenstein, N. Schuff, and M. W. Weiner, "Reduced FDG-PET brain metabolism and executive function predict clinical progression in elderly healthy subjects." *NeuroImage Clin.*, vol. 4, pp. 45–52, Jan. 2013.

[5] J. B. S. Langbaum, K. Chen, W. Lee, C. Reschke, D. Bandy, A. S. Fleisher, G. E. Alexander, N. L. Foster, M. W. Weiner, R. a. Koeppe, W. J. Jagust, and E. M. Reiman, "Categorical and correlational analyses of baseline fluorodeoxyglucose positron emission tomography images from the Alzheimer's Disease Neuroimaging Initiative (ADNI)." *Neuroimage*, vol. 45, no. 4, pp. 1107–16, May 2009.

[6] M. J. D. Leon, A. Convit, O. T. Wolf, C. Y. Tarshish, S. Desanti, H. Rusinek, W. Tsui, E. Kandil, A. J. Scherer, A. Roche, A. Imossi, E. Thorn, M. Bobinski, C. Caraos, P. Lesbre, D. Schlyer, J. Poirier, B. Reisberg, and J. Fowler, "Prediction of cognitive decline in normal elderly subjects with 2-[(18)F]fluoro-2-deoxy-D-glucose/positron-emission tomography (FDG/PET)." vol. 98, no. 19, 2001.

[7] K. Chen, J. B. S. Langbaum, A. S. Fleisher, N. Ayutyanont, C. Reschke, W. Lee, X. Liu, D. Bandy, G. E. Alexander, P. M. Thompson, N. L. Foster, D. J. Harvey, M. J. de Leon, R. a. Koeppe, W. J. Jagust, M. W. Weiner, and E. M. Reiman, "Twelve-month metabolic declines in probable Alzheimer's disease and amnestic mild cognitive impairment assessed using an empirically pre-defined statistical region-of-interest: findings from the Alzheimer's Disease Neuroimaging Initiative." *Neuroimage*, vol. 51, no. 2, pp. 654–64, Jun. 2010.

[8] C. Hinrichs, V. Singh, G. Xu, and S. C. Johnson, "Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population." *Neuroimage*, vol. 55, no. 2, pp. 574–89, Mar. 2011.

[9] F. Ségonne, a. M. Dale, E. Busa, M. Glessner, D. Salat, H. K. Hahn, and B. Fischl, "A hybrid approach to the skull stripping problem in MRI." *Neuroimage*, vol. 22, no. 3, pp. 1060–75, Jul. 2004.

[10] J. Ashburner and K. J. Friston, "Unified segmentation." *Neuroimage*, vol. 26, no. 3, pp. 839–51, Jul. 2005.

[11] A. Collignon and F. Maes, "Automated multi-modality image registration based on information theory," *Inf. Process. Med. imaging*, 1995.

[12] J. Ashburner, "A fast diffeomorphic image registration algorithm." *Neuroimage*, vol. 38, no. 1, pp. 95–113, Oct. 2007.

[13] V. S. Fonov, A. C. Evans, R. C. McKinstry, C. R. Almli, and D. L. Collins, "Unbiased nonlinear average age-appropriate brain templates from birth to adulthood." *NeuroImage*, vol. 47, p. S102, Jul. 2009.

[14] K. Herholz, E. Salmon, D. Perani, J. C. Baron, V. Holthoff, L. Frölich, P. Schönknecht, K. Ito, R. Mielke, E. Kalbe, G. Zündorf, X. Delbeuck, O. Pelati, D. Anchisi, F. Fazio, N. Kerrouche, B. Desgranges, F. Eustache, B. Beuthien-Baumann, C. Menzel, J. Schröder, T. Kato, Y. Arahata, M. Henze, and W. D. Heiss, "Discrimination between Alzheimer dementia and controls by automated analysis of multicenter FDG-PET." *Neuroimage*, vol. 17, no. 1, pp. 302–16, Sep. 2002.

[15] P. Borghammer, J. Aanerud, and A. Gjedde, "Data-driven intensity normalization of PET group comparison studies is superior to global mean normalization." *Neuroimage*, vol. 46, no. 4, pp. 981–8, Jul. 2009.

[16] I. Yakushev, A. Hammers, A. Fellgiebel, I. Schmidtmann, A. Scheurich, H.-G. Buchholz, J. Peters, P. Bartenstein, K. Lieb, and M. Schreckenberger, "SPM-based count normalization provides excellent discrimination of mild Alzheimer's disease and amnestic mild cognitive impairment from healthy aging." *Neuroimage*, vol. 44, no. 1, pp. 43–50, Jan. 2009.

[17] R. S. Desikan, H. J. Cabral, F. Settecase, C. P. Hess, W. P. Dillon, C. M. Glastonbury, M. W. Weiner, N. J. Schmansky, D. H. Salat, and B. Fischl, "Automated MRI measures predict progression to Alzheimer's disease." *Neurobiol. Aging*, vol. 31, no. 8, pp. 1364–74, Aug. 2010.

[18] C. Chang and C. Lin, "LIBSVM: A Library for Support Vector Machines." pp. 1–39, 2001.