3D BRAIN IMAGE-BASED DIAGNOSIS OF ALZHEIMER'S DISEASE: BRINGING MEDICAL VISION INTO FEATURE SELECTION

Eduardo Bicacro[⋆]*, Margarida Silveira*[⋆]*, Jorge S. Marques*[⋆]*, Durval C. Costa*[†‡]*, and the Alzheimer's Disease Neuroimaging Initiative*

[⋆] Institute for Systems and Robotics, Instituto Superior Técnico, Lisbon, Portugal
[†] Nuclear Medicine, Champalimaud Clinical Center, Champalimaud Foundation, Lisbon, Portugal
[‡] HPP Medicina Molecular, SA, Porto, Portugal

ABSTRACT

Expert physicians are able to attain good Alzheimer's Disease (AD) diagnostic accuracy, relying on visual inspection of Positron Emission Tomography (PET) images only. Nevertheless, computerized methods have been implemented with similar or even better performance. We investigate the potential of the physician's experienced visual inspection to guide feature selection, in an automatic classification procedure. Eye tracking methodology is employed to obtain a model of the physician's visual behavior, which allows for the sampling of voxel intensity features that are then fed to an SVM classifier. This approach is compared with commonly used automatic feature selection alternatives. Image data were taken from the Alzheimer's Disease Neuroimaging Initiative database. The results show that the proposed approach marginally improves accuracy in AD vs. CN classification, but for MCI vs. CN and AD vs. MCI it presents lower performance.

***Index Terms*—** Computer aided diagnosis, Alzheimer's disease, Positron Emission Tomography, Feature selection, Eye tracking

1. INTRODUCTION

Alzheimer's Disease (AD) is a degenerative brain disorder and the leading cause of dementia. AD's identification and differential diagnosis is especially challenging and important in the earliest stages of the disease. The state of mild cognitive impairment (MCI), in which patients may experience some of AD's symptoms but do not meet the established criteria for a dementia diagnosis, is particularly important for clinical trials directed at delaying the onset of AD.

When PET is added to a clinical dementia evaluation, the sensitivity (sens.) of detection of histopathologically confirmed AD increases from 66% $\pm$ 17% to 91.5% $\pm$ 3.5% [1]. Interest has built up in the use of machine learning techniques to further amplify the power of PET as a diagnostic tool. In the past, images were examined by human visual inspection only, with good results in both sens. and specificity (spec.), as radiologists sought the patient's diagnosis. This interpretation was, of course, qualitative, and heavily relied on the observer's experience and training. Klöppel et al. [2] compared the accuracy (acc.) of the dementia diagnosis carried out by experienced radiologists and by a computerized method these researchers had previously developed, finding that the latter yielded better or equivalent results, besides allowing for ready exchange of the diagnostic tool between different centres and despite not needing any expert knowledge.

Several approaches to AD's computer aided diagnosis (CAD) have already been performed, using mostly the voxel intensity of PET or MRI 3D images as features [3, 4, 5, 6, 7]. In this framework, the number of available features is some orders of magnitude larger than the number of available image examples, a situation commonly referred to as "curse of dimensionality". The advantages of reducing the number of features before any attempt of classification is essentially three-fold: the prediction performance of the classifier is improved, the classification step becomes faster and more cost-effective, and finally, a better understanding of the underlying process behind the generation of the data is possible [8]. With this aim, several automatic techniques were employed, such as Pearson's correlation coefficient ranking [5], spatial consistency criteria [5, 6, 7], principal component analysis [9], and recursive feature elimination [5]. The majority of these studies employed Support Vector Machines (SVMs) in the classification step.

In this article, we present a medically informed approach to feature selection, in the CAD of AD using FDG-PET images, based on an expert physician's eye tracking procedure. We compare this approach to other feature selection methods, namely the use of ROIs and two automatic filter methods: point biserial correlation coefficient (PBCC) and mutual information (MI) ranking. The technique of eye tracking allows for an insight into the feature selection process that is *de facto* carried out by an expert physician himself, when observing FDG-PET images. This information might be of even more value than the physician's conscious interpretation of that process. These strategies are used to perform AD vs. Cognitively Normal (CN), MCI vs. CN and AD vs. MCI classifications, using SVM as final classifier, with both linear and RBF kernels. The data set herein employed is from a multicenter world-renowned study - Alzheimer's Disease Neuroimaging Initiative (ADNI).

2. METHODS

2.1. Information selection through eye tracking

Human vision is a piecemeal process relying on the perceptual integration of small regions to construct a coherent representation of the whole [10]. The instrumental tool for this selection process is the movement of the eyes, with which the person brings a particular portion of the visible field of view into high resolution, allowing for

the focus of concentration on the object or region of interest [10]. Herein, the T120® model from Tobii™, was used for eye tracking purposes. This model allowed for the presentation of the stimuli to be done on a $17''$ flat screen, measuring gaze point with an acc. of about 0.5 degrees. One physician with great experience in the clinical use of FDG-PET was asked to examine FDG-PET image volumes, as he would do in clinical context, and decide upon the diagnosis of each of the analyzed subjects. 177 different sets of images were analyzed, which were randomly and equally sampled from AD, CN, and MCI patients (59 each). Since the point of fixation was measured in pixels, nearest neighbour interpolation was used to convert from a point in a slice image to a brain voxel. At the end of this process, a brain volume was obtained for each of the analyzed subjects, with the value of each voxel representing the total amount of time the physician gazed at it.

The assumption that the physician was extracting information from only one image voxel, when fixating a particular point, is too restrictive, since the human eye has maximal acuity within a neighbourhood around that point. To approximate this behaviour, slice-by-slice Gaussian filtering of the volume described earlier was performed. Probability mass functions were generated to model the discrete probability distribution of choosing any particular voxel. This allowed for random sampling of a chosen number of voxels, according to the probability defined by those distributions, this way selecting an arbitrary number of voxels to be fed to the subsequent classification stages.

2.2. Alternative feature selection methods

Three alternative feature selection strategies were implemented, for comparison purposes. The first was also based on medical information, in the form of previously defined ROIs. The same expert physician whose visual observation was monitored was asked to identify the regions which matter the most in his observations of functional brain images. Seven distinct regions, with some of them having symmetrical counterparts were thus pointed out: (1) Lateral temporal - right and left; (2) Mesial temporal - right and left; (3) Inferior frontal gyrus/Orbitofrontal; (4) Inferior anterior cingulate; (5) Superior anterior cingulate; (6) Dorsolateral parietal - right and left; (6) Posterior cingulate and precuneus. These regions were manually segmented in the original image volume reference space, on axial planes. A model of the probability of choosing a particular voxel was built, with the same procedure used for the eye tracking-based alternative (see section 2.1).

The other two selection methods are fully automatic filter strategies: PBCC and MI ranking of features. Ranking is a commonly used mechanism of feature selection because of its simplicity, scalability and good empirical results [8]. In the following, a general problem of pattern recognition shall be considered, where each pattern is represented by a set of features and viewed as point $\mathbf{x}$ in an n-dimensional feature space and is associated with an output variable y. The i^{th} component of $\mathbf{x}$ (i^{th} feature) can be interpreted as a realization of X_i, a random variable with an unknown underlying distribution. Similarly, let Y represent the random variable of which the outcome y is a realization.

The PBCC is a special case of the Pearson correlation coefficient, in which one of the variables is dichotomous and the other is continuous. If one takes into account the dichotomous nature of Y (assuming values 1 or 0), the Pearson correlation coefficient formula can be simplified to:

$$R_{pb}(i) = \frac{M_1 - M_0}{s_{X_i}} \sqrt{\frac{n_1 n_0}{m^2}}, \tag{1}$$

where M_1 and M_0 are the mean values of the variable X_i for the patterns in group 1 and 0, respectively, n_1 and n_0 are the number of patterns in each of the groups, m is the total number of patterns and s_{X_i} is an estimate of the standard deviation of X_i. In the context of feature selection, any of the extreme values this coefficient may assume (-1 and 1) would hint for high relevance of the feature for final prediction. Herein, the ranking coefficient is therefore taken as the absolute value of PBCC.

The MI of two random variables is a concept arising from information theory. In the present case, we consider the mutual information between the random variables X_i and Y, for $i = 1, \ldots, n$, which is defined as:

$$MI(i) = \sum_x \sum_y P(X_i = x, Y = y) \log \frac{P(X_i = x, Y = y)}{P(X_i = x)P(Y = y)}. \tag{2}$$

The probability density functions in (2) were estimated through the use of histograms. Unlike PBCC, which measures linear dependence, MI can assess any kind of relationship between variables, because it is directly based on the probability density functions.

2.3. SVM classifier and performance estimation

SVMs have proven to be able to provide good results in classifications performed in almost empty spaces [11]. The underlying concept was first introduced for tasks of binary classification of linearly separable data: for a given input space, the SVM algorithm tries to find a separating hyperplane that not only separates the instances belonging to each class, but also maximizes the margin to any of the data points. The interested reader can find the complete derivations in [12]. The search for the hyperplane with the aforementioned properties can be formulated as a well-known optimization problem:

$$\text{Minimize:} \quad \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{l}\xi_i \tag{3}$$

$$\text{Subject to:} \quad y_i(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad \forall i \tag{4}$$

where vector $\mathbf{w}$ and the constant b parametrize the separating hyperplane, $\phi(\mathbf{x}_i)$ is the image of the input vector $\mathbf{x}_i$ in an arbitrary dimension-space where the data (hopefully) become linearly separable (or separated with less errors), ξ_i are positive slack variables that allow for some errors to be made and C is a constant defined by the user that allows for the tuning of the misclassification cost. Input patterns are commonly mapped to higher dimensional spaces through the use of kernels, which are operators that define the inner product of the features' images in that new space. Herein, linear and RBF kernels were employed, with the latter introducing a new hyperparameter to the classification problem (γ). There are several publicly available packages that numerically solve the optimization problem defined in (3) and (4). Herein LIBSVM was used, a library developed by C. Chang and C. Lin [13] and available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Table 1. Clinical and neuropsychological profile of the groups studied. Values are presented according to $(mean \pm standard\ deviation)$. MMSE - Mini-Mental State Examination; CDR - Clinical Dementia Rating.

Group	**AD**	**MCI**	**CN**
Age	78.3 ± 6.6	77.7 ± 6.9	77.38 ± 4.9
Sex (%M)	58	68	64
MMSE	19.6 ± 5.0	25.7 ± 3.0	29.2 ± 0.9
CDR	0.5/higher	0.5	0

In order to obtain an estimate of the generalization performance of the several classifiers implemented, and simultaneously find the optimal parameters (C for the linear kernel, C and γ for RBF), a nested cross validation (CV) strategy was implemented, which is an extension of the simple CV to situations where parameters optimization is required (and therefore disjoint train, validation and test sets are required). This method provides an unbiased and conservative estimate of the classifier's generalization performance.

2.4. ADNI image data

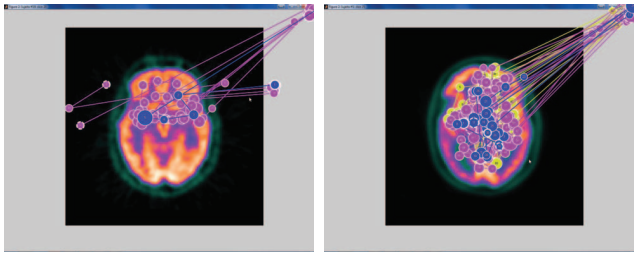
The data used to test the several implemented classifiers were taken from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The main goal of ADNI has been to test whether MRI, PET, other biological markers and clinical and neuropsychological assessments may be combined to measure the progression of MCI and early AD. For the present study, 59 CN, 59 MCI and 59 AD subjects with available FDG-PET images were selected. Table 1 summarizes the description of these groups of patients. Every FDG-PET volume had been subject to a preprocessing procedure, with format, orientation and resolution uniformization purposes. At the end of this process, all volumes were $128 \times 128 \times 60$ voxels wide.

2.5. Experimental design

To study the performance of the different selection methods, each of them was combined with the two forms of SVM kernel (linear and RBF). In addition to the two medical and two automatic criteria, a random selection strategy was also performed. Further, for each classifier configuration (selection method and kernel type), several alternatives of the number of selected features were tested: 50, 100, 250, 500, 1000, 2500, 5000, 10000 and 25000. For the automatic and random alternatives, a previous segmentation of the brain area in the image volumes was carried out by simple thresholding, in order to discard extra-cranial voxels. 5-fold nested CV was implemented to perform parameters optimization and obtain generalization performance estimates for every classifier configuration, in AD vs. CN, MCI vs. CN and AD vs. MCI classifications.

3. RESULTS

Two examples of the results obtained from the eye tracking experiment are displayed in Fig. 1. The dots represent points of fixation in the image, and their radius is proportional to the fixation duration. One may immediately realize that not all slices were observed



(a) **(b)**

Fig. 1. Visualization of eye tracking measurements for two example slices: (a) slice 20 and (b) slice 21, when considering the data for all analyzed subjects.

for the same amount of time, neither were they for the same number of times. Useless fixations points, like the ones on the upper right corner of both images, were discarded.

Fig. 2 displays the best performing configuration for each classification problem and for each feature selection method. The number of features and SVM kernel that maximized the result are indicated.

Generally, it can be said that no pronounced differences were found between the different selection methods studied, although there seem to be some tendencies in the results which may be scrutinized. For AD vs. CN, acc. was close to 90% when a selection method was used, and down to 85% when random sampling of features was performed instead. The best average acc., sens. and spec. were achieved by the eye tracking-driven procedure, at 5000 selected features, scoring 91.4($\pm$4.3)%, 90.0($\pm$7.2)% and 92.8($\pm$3.2)% respectively. The medically-driven features reached results close to their best for small numbers of features (at 250 features, they provided marks that were close to their best).

In classifications where the MCI state was involved, the performance of the classifiers is clearly poorer. For MCI vs. CN, none of the selection methods delivered results substantially better than random choice. Automatic methods seemed to benefit from larger feature spaces, what suggests that the criteria for information selection couldn't identify the valuable features, possibly because there was high variability in the populations. In AD vs. MCI classification, the medically driven selection approaches didn't perform substantially better than random sampling. Contrarily, the automatic selection shows a somewhat superior performance. Unlike what was found in the previous classification task, there is evidence that the automatic methods were capable of identifying meaningful features that produced good results for relatively low dimensional spaces. Further, it seems that the model of the physician's visual behaviour selects voxels which are not the most useful for this particular classification task (results improved for larger feature spaces, where the effect of selection fades out).

The lower performance observed for classifications where MCI was present may be related to the fact that, in this experiment, the expert physician was asked to perform a multi-class classification (AD vs. MCI vs. CN). It is plausible to assume that the regions that hold the most valuable information for AD vs. CN comparison are not among the most relevant for the other classification tasks.

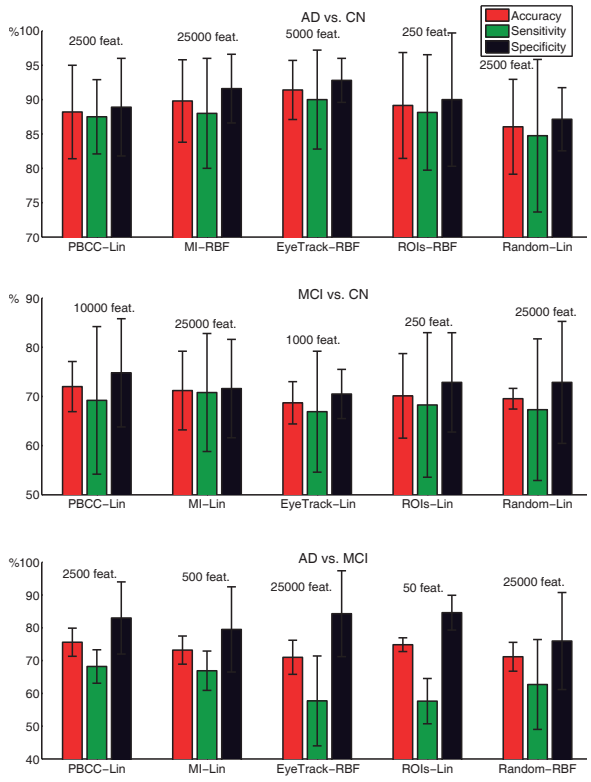


Fig. 2. Performance of the best classifier for each of the feature selection methods, in (a) AD vs. CN; (b) MCI vs. CN; (c) AD vs. MCI classification.

4. CONCLUSIONS

The present work dealt with innovative approaches to the CAD of AD, tackling feature selection by capturing the medical analysis expertise directly from the physician's visual behaviour. The eye tracking-driven selection yielded the best result for AD vs. CN classification, but for classification tasks where the MCI state was present, the method didn't outperform the alternatives. Nevertheless, the results encourage further work, through the inclusion of visual data from more physicians and the refining of the model of the visual observation, that may include more complex forms of information processing in the neighbourhood of the gaze point and extract meaning out of the sequence of areas observed.

Our method performed better than most expert physicians [1] in AD vs. CN classification (maintaining the potential for clinical use), and than most automatic classification implementations working with FDG-PET images only, found in the literature [9, 3].

5. ACKNOWLEDGEMENTS

This work was supported by Fundação para a Ciência e Tecnologia (FCT/MCTES) through the ADIAR project (PTDC/SAU-ENB/114606/2009). 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. Refer to `http://www.loni.ucla.edu/ADNI/Collaboration/ADNI_Authorship_list.pdf` for a complete listing of ADNI investigators.

References

[1] D. H. S. Silverman, "Brain ^{18}F-FDG PET in the diagnosis of neurodegenerative dementias: Comparison with perfusion SPECT and with clinical evaluations lacking nuclear imaging," *Journal of Nuclear Medicine*, vol. 45, no. 4, pp. 594–607, Apr. 2004.

[2] S. Klöppel, C. M. Stonnington, J. Barnes, F. Chen, C. Chu, C. D. Good, I. Mader, L. A. Mitchell, A. C. Patel, C. C. Roberts, N. C. Fox, C. R. Jack, J. Ashburner, and R. S. J. Frackowiak, "Accuracy of dementia diagnosis - a direct comparison between radiologists and a computerized method," *Brain*, vol. 131, no. 11, pp. 2969–2974, 2008.

[3] M. Silveira and J. Marques, "Boosting Alzheimer disease diagnosis using PET images," *Pattern Recognition, International Conference on*, pp. 2556–2559, 2010.

[4] Y. Fan, S. M. Resnick, X. Wu, and C. Davatzikos, "Structural and functional biomarkers of prodromal Alzheimer's disease: A high-dimensional pattern classification study," *NeuroImage*, vol. 41, no. 2, pp. 277 – 285, 2008.

[5] C. Davatzikos, Y. Fan, X. Wu, D. Shen, and S. M. Resnick, "Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging," *Neurobiology of aging*, vol. 29, no. 4, pp. 514 – 523, 2006.

[6] J. Stoeckel and G. Fung, "SVM feature selection for classification of SPECT images of Alzheimer's disease using spatial information," in *ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining*, Washington, DC, USA, 2005, pp. 410–417, IEEE Computer Society.

[7] C. Hinrichs, V. Singh, L. Mukherjee, G. Xu, M. K. Chung, and S. C. Johnson, "Spatially augmented lpboosting for AD classification with evaluations on the ADNI dataset," *NeuroImage*, vol. 48, no. 1, pp. 138 – 149, 2009.

[8] I. Guyon, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[9] I. A. Illán, J. M. Górriz, J. Ramírez, D. Salas-Gonzalez, M. M. López, F. Segovia, R. Chaves, M. Gómez-Rio, and C. G. Puntonet, "18F-FDG PET imaging analysis for computer aided Alzheimer's diagnosis," *Information Sciences*, vol. 181, no. 4, pp. 903 – 916, 2011.

[10] A. T. Duchowski, *Eye Tracking Methodology: Theory and Practice*, Springer, 2007.

[11] Robert P.W. Duin, "Classifiers in almost empty spaces," in *In 15th International Conference on Pattern Recognition*, 2000, pp. 1–7.

[12] C. Cortes and V. Vapnik, "Support-vector networks," in *Machine Learning*, 1995, pp. 273–297.

[13] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.