Simulating Functional Brain Images in Alzheimer's Disease

Abstract—**The Small Sample Size problem is a common issue in neuroimaging, where the number of features frequently surpass the number of samples, making generalization a difficult task and leading to inconsistent results. To overcome this problem, we have developed a new algorithm intended to generate a new set of functional images with similar characteristics to an original one, in our case, a subset of the PET ADNI database. This way, a bigger dataset can be used to evaluate our methodology, providing better generalizable results. The procedure involves modeling the original database with PCA in order to project the images to the eigenbrain space, and then modelling the coordinate set of the original images in the eigenbrain space using KDE. Once these parameters are obtained, new coordinates can be generated using the density functions of the original database and a random number generator. The generated dataset is then evaluated to see if the new dataset features match the original ones.**

Index Terms—**Positron Emission Tomography, Alzheimer's Disease, Kernel Density Estimation, Support Vector Machines, Principal Component Analysis, Simulated Brains**

I. Introduction

The rise of neuroimaging in the last years has provided physicians and radiologist with the ability to study the brain with unprecedented ease. This led to a new biological perspective in the study of neurodegenerative diseases, allowing the characterization of different anatomical and functional patterns associated with them. Disorders such as Alzheimer's Disease (AD) or Parkinsonism have been widely studied by means of both structural and functional imaging.

From the very beginning, the main use of these images has mainly consisted of visual and semi-quantitative analysis, depending on manual demarcation of regions of interests. Furthermore, medical imaging yields massive data that sometimes is underused or directly dismissed. Therefore, a great effort has been made in the implementation of objective systems able to perform these analysis in a semi-automatic way. A whole new set of statistical tools and algorithms has lead to what has been called the Computer Aided Diagnosis (CAD) paradigm.

There is a fundamental drawback in the evaluation of new CAD algorithms: the small sample size problem [1]. This problem has been exhaustively studied in a number of papers, and partial solutions, such as the use of better generalizable classifiers [2], [3], [4], [5] and feature extraction methods [6], [7], [8] have been proposed. Nevertheless, this work focus on another approach: instead of optimizing the computation of different parameters to make the system better generalizable, we propose the simulation of new functional brain imaging samples to enlarge our training and test data.

Our system is based on the modelling of existing, widely-used, functional imaging datasets, although it can be expanded to other modalities. It uses an strategy based on a Principal Component Analysis (PCA) modelling of a existing dataset (in our case, the widely-known ADNI), converting the original dataset to a eigenbrain based space like in [8]. Afterwards, the statistical distribution of the coordinates in the eigenbrain space is estimated and new coordinates can be used to generate new images with similar properties.

Our strategy is intended to serve as a computational aid to reduce bias in the evaluation of new CAD methods. However, apart from this main use, there exist the inviting possibility to use the simulated datasets as a training tool in neurology.

This paper is organized as follows. First, in Section II the methodology and the different steps are detailed. Next, in Section III the simulated dataset is evaluated and the results are discussed. Finally, in Section IV, some conclusions are drawn, and future work is proposed.

II. Methodology

A. Principal Component Analysis

In order to generate a new set of functional images it is necessary to extract some common features that will be used as sources of the generated images. To do so, we will follow a procedure similar to the one used in [8]. There, the whole dataset is projected to a new space defined by "eigenbrains", the set of eigenvectors that are produced after applying Principal Component Analysis (PCA) to the dataset. Thereafter, each individual image can be considered a set of coordinates in the eigenbrain space (also known as component scores in the common PCA language), and the linear combination of the eigenbrains using these coordinates as weights can be used to reconstruct the original image.

PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations $\mathbf{X}$ of possibly correlated variables (in this work, the data matrix $K \times N$ containing K images of N voxels) in a set $\mathbf{W}$ of M linearly uncorrelated variables that we call eigenbrains (also known as principal components or mixing matrix) that, being orthogonal, define the eigenbrain space. The set of coordinates of the original dataset in the eigenbrain space $\mathbf{S}$ (size $K \times M$) can be estimated as follows:

$$\mathbf{S} = \mathbf{X}\mathbf{W} \tag{1}$$

An efficient algorithm to compute PCA is using Singular Value Decomposition (SVD). This decomposes the data matrix as follows:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^* \tag{2}$$

where $\mathbf{U}$ is a $K \times K$ orthogonal matrix, $\mathbf{\Sigma}$ is a $K \times M$ nonnegative real diagonal matrix, and the $M \times M$ unitary matrix $\mathbf{V}^*$ denotes the conjugate transpose of the $M \times M$ unitary matrix $\mathbf{V}$. With this decomposition, Eq. 1 becomes:

$$\mathbf{S} = \mathbf{X}\mathbf{W} = \mathbf{U}\mathbf{\Sigma} \tag{3}$$

Since the conjugate transpose of a unitary matrix is its inverse, theset of eigenbrains $\mathbf{W}$ can be estimated as:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{W}^{-1} \rightarrow \mathbf{W} = \mathbf{V} \tag{4}$$

B. Kernel Density Estimation via Diffusion

There exist a number of Probability Density Function (PDF) estimation methods, of which the histogram is the most widely known. Non-parametric methods such as Kernel Density Estimation (KDE) are gaining popularity for their capabilities and ease of use. KDE uses kernels, that is, nonnegative functions of zero mean that integrates to one, to estimate the PDF.

The KDE estimates f from a number of independent and identically distributed samples $(x_1, x_2, \ldots x_n)$, in the following manner:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big), \tag{5}$$

where $h > 0$ is the bandwidth, a smoothing parameter. This parameter is of fundamental importance as it defines the tradeoff between the bias of the estimator and its variance, and it often requires to be set *a priori*.

Many algorithms attempt to overcome this parameter estimation using, for example, a simple rule of thumb [9] or performing a preliminary normal model of the data [10], which invalidates the original motivation for applying a nonparametric method. Conversely, the Kernel Density Estimation by Diffusion[11] used in this article uses a data-driven automatic estimation of the bandwidth, which unlike most methods, does not rely on arbitrary normal reference rules.

C. Empirical Random Number Generation

After estimating the empirical PDF of the coordinates, we aim to generate a new set that match the distribution of the originals. To do so, we will use a random number generator that provides uniformly distributed random numbers in the interval $[0, 1]$. Then, the final values are interpolated from the empirical Cumulative Distribution Function (CDF), using the generated random numbers as query points.

D. Database and Preprocessing

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org).

		N	Age
Normal	M	62	75.52 ± 4.70
	F	39	76.71 ± 5.05
	Tot.	101	75.98 ± 4.85
AD	M	57	76.67 ± 7.23
	F	38	74.29 ± 7.53
	Tot.	95	75.72 ± 7.40

TABLE I
DEMOGRAPHICS OF THE DATASET.

In this work, the ^{18}F-FDG PET images, used to estimate the metabolic activity of the brain, are used to generate and validate the generated images. 95 PET images from AD affected subjects and 101 images from Normal Controls (NC) have been used to construct the original set from which the simulation parameters will be obtained (see more demographic details in Table I).

Both the original and the simulated dataset will be intensity normalized to adjust the intensity range and allow further processing. We will use a normalization to the maximum strategy, that has been successfully applied in a number of works involving functional imaging [12], [13], [14]. In this strategy, the image is normalized to the maximum intensity, computed using the average of the 3% higher intensity voxels.

III. RESULTS AND DISCUSSION

A. Validation

We have already described how the images were obtained. To evaluate if the newly generated database matches the original one, we perform a classification analysis using feature selection by means of t-Test, within a stratified 10-fold cross validation. The final aim is to obtain values of accuracy (acc), sensitivity (sens) and specificity (spec) and their standard deviation for both the original and the simulated database, in order to compare them and test that the simulated database keeps similar discrimination ability while reducing the variance of the estimates.

The classification analysis is performed using a Voxels As Features (VAF) approach [2], that is considered a good estimator of visual analysis. The classifier used is a Support Vector Machine (SVM) with linear kernel as implemented in `LIBSVM` [15]. Selection of parameter C is performed using an inner 5-fold cross-validation on the training subset.

B. Classification Results

We have performed a VAF analysis of the original dataset (containing 95 AD-affected subjects and 101 normal controls) to estimate the performance and standard deviation of the different parameters. Later, we have repeated the exact same procedure with a simulated database containing 500 AD and 500 NC subjects. The accuracy, sensitivity and specificity obtained in both cases are shown in Table II.

	acc (±SD)	sens (±SD)	spec (±SD)
Original	0.857 ± 0.094	0.833 ± 0.162	0.881 ± 0.104
Simulated	0.853 ± 0.035	0.836 ± 0.053	0.870 ± 0.073

TABLE II
CLASSIFICATION RESULTS FOR THE ORIGINAL DATASET AND THE SIMULATED $N = 1000$ DATASET.

C. Discussion

We have proposed an algorithm that generates a series of functional brain images as coordinates in the eigenbrain space. After the reconstruction, these new images closely resemble the original ones, as can be seen in Fig. 1. There, original NOR and AD subjects are compared to two simulated NOR and AD subjects.

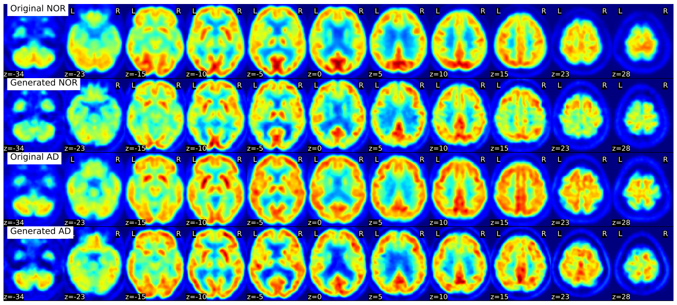

Fig. 1. Comparison between simulated and original images from AD and NOR classes.

Some of the most widely documented patterns of AD progression are present in both original and simulated subjects. Both original and simulated NOR subjects present a higher contrast between frontal and occipital lobe than the original or generated AD subjects. A decrease in FDG uptake in the posterior cingulate gyri and precunei can be observed as well [2], [8]. Note that all images are normalized to the maximum, therefore the decrease in uptake can be seen as a smaller contrast between these areas and the rest of the brain. The contribution of the low-frequency eigenbrains (see Fig. 2) highlights some of these effects: cingulate gyri and precunei differences are modeled in eigenbrain 1, and the differences in glucose metabolism between the frontal and the occipital lobe can be perceived in eigenbrain 3. Asymmetry is also influenced by other eigenbrains, such as number 7.

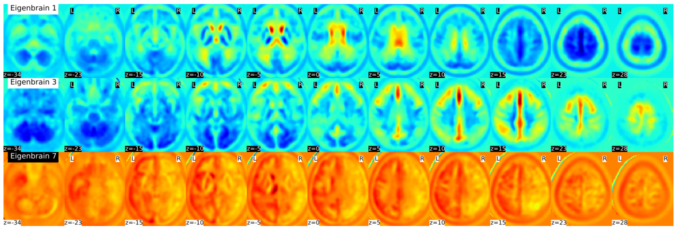

Fig. 2. Illustration of some of the most relevant eigenbrains.

We have also checked if the discriminative information found in the original database is kept in the simulated by applying a two-sample t-test to both datasets. Voxelwise p-values were obtained after applying the Bonferroni correction. This yielded the significant regions shown in Figure 3. Most significant areas are located at the Temporoparietal region (see z=-10, 20k 30 and 40) and the Precunei (z=20 and 30). Other relevant areas, such as the Hippocampus (z=0) or Cyngulate Gyri (z=0,10,20) are highlighted as well.

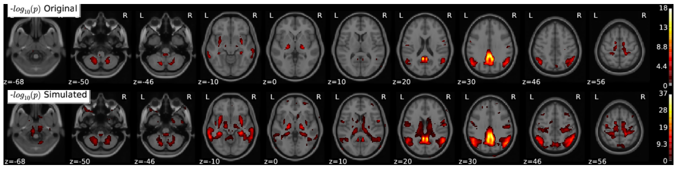

Fig. 3. Significant regions obtained after performing a corrected t-Test on the original and the simulated dataset.

Regarding the classification analysis, we have shown in Figure 4 a visual comparison of the data found in Table II. Both data confirm that the generated dataset provides the same discriminative power as the original, whereas reducing the variance of the performance estimates. This could be of great help in functional neuroimaging studies where the small sample size problem frequently appears.

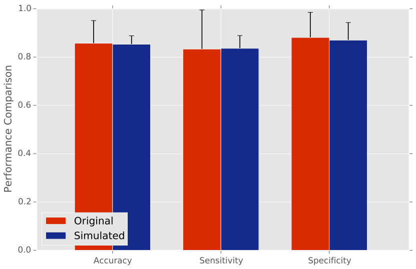

Fig. 4. Performance comparison between the original and the simulated datasets under the VAF paradigm.

Conceptually this approach is different to other brain image simulators such as [16]. Most simulation software is focused on MR imaging, which provides a higher resolution. Our approach would not be recommendable in this task, given the inherent properties of the eigenbrains, which probably could not model the anatomical detail of such images. However, it has shown its ability in modeling lower resolution images such as the FDG-PET database used here, and could be expanded to other modalities and tracers such as PiB-PET or SPECT-HMPAO, also used in AD or SPECT-DATSCAN, used in Parkinson's Disease.

IV. Conclusions and Future Work

In this work, a new algorithm to simulate functional brain images is presented. It is intended to generate large-samples dataset to overcome the small sample size problem, very frequent in neuroimaging. To do so, an original dataset is modelled used Principal Component Analysis, and the projections of the subjects in the eigenbrain space are obtained.

Them, new coordinates in the eigenbrain space are generated and reconstructed so that we simulate a new dataset that keeps the characteristics of the original.

We have tested the algorithm with a FDG-PET dataset from the Alzheimer's Disease Neurogimaging Initiative. The simulated dataset obtained similar discrimination ability while reducing the variance of the performance estimates, as it could be expected. By visual analysis, the new simulated images resemble the original FDG-PET images, and preserve the typical Alzheimer's Disease patterns. The algorithm could be used to generate large sample datasets in the evaluation of Computed Aided Diagnosis systems, making the performance estimates more generalizable and consistent. It could also have a educational application in the training of visual abilities in physicians.

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] S. Raudys and A. Jain, "Small sample size effects in statistical pattern recognition: recommendations for practitioners," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 3, pp. 252–264, 1991.

[2] J. Stoeckel, N. Ayache, G. Malandain, P. M. Koulibaly, K. P. Ebmeier, and J. Darcourt, "Automatic Classification of SPECT Images of Alzheimer's Disease Patients and Control Subjects," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI*, ser. Lecture Notes in Computer Science, vol. 3217. Springer, 2004, pp. 654–662.

[3] G. Fung and J. Stoeckel, "SVM feature selection for classification of SPECT images of Alzheimer's disease using spatial information," *Knowledge and Information Systems*, vol. 11, no. 2, pp. 243–258, 2007.

[4] F. J. Martínez-Murcia, J. M. Górriz, J. Ramírez, C. G. Puntonet, and D. Salas-González, "Computer Aided Diagnosis tool for Alzheimer's Disease based on Mann-Whitney-Wilcoxon U-Test," *Expert Systems with Applications*, vol. 39, no. 10, pp. 9676–9685, Aug. 2012.

[5] A. Ortiz, J. M. Górriz, J. Ramírez, and F. J. Martínez-Murcia, "Lvq-SVM based CAD tool applied to structural MRI for the diagnosis of the alzheimer's disease," *Pattern Recognition Letters*, vol. 34, no. 14, pp. 1725–1733, Oct. 2013.

[6] P. Markiewicz, J. Matthews, J. Declerck, and K. Herholz, "Robustness of multivariate image analysis assessed by resampling techniques and applied to FDG-PET scans of patients with Alzheimer's disease," *Neuroimage*, vol. 46, pp. 472–485, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/B6WNP-4VFK7X3-3/2/e7833cb1d62f98e28326352e45981d00

[7] L. Khedher, J. Ramírez, J. Górriz, A. Brahim, and F. Segovia, "Early diagnosis of alzheimers disease based on partial least squares, principal component analysis and support vector machine using segmented MRI images," *Neurocomputing*, vol. 151, p. 139150, Mar. 2015.

[8] 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, "^{18}F-FDG PET imaging analysis for computer aided Alzheimer's diagnosis," *Information Sciences*, vol. 181, pp. 903–916, Feb. 2011.

[9] A. W. Bowman and A. Azzalini, *Applied smoothing techniques for data analysis*. Clarendon Press, 2004.

[10] M. Jones, J. S. Marron, and S. Sheather, "Progress in data-based bandwidth selection for kernel density estimation," *Computational Statistics*, vol. 11, no. 3, pp. 337–381, 1996.

[11] Z. Botev, J. Grotowski, D. Kroese *et al.*, "Kernel density estimation via diffusion," *The Annals of Statistics*, vol. 38, no. 5, pp. 2916–2957, 2010.

[12] P. Saxena, D. G. Pavel, J. C. Quintana, and B. Horwitz, "An Automatic Threshold-Based Scaling Method for Enhancing the Usefulness of Tc-HMPAO SPECT in the Diagnosis of Alzheimer's Disease," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI*, ser. Lecture Notes in Computer Science, vol. 1496. Springer, 1998, pp. 623–630.

[13] D. Salas-Gonzalez, J. M. Górriz, J. Ramírez, I. A. Illán, M. López, F. Segovia, R. Chaves, P. Padilla, and C. G. Puntonet, "Feature selection using factor analysis for Alzheimer's diagnosis using F-FDG PET images," *Medical Physics*, vol. 37, no. 11, pp. 6084–95, Nov. 2010.

[14] M. López, J. Górriz, J. Ramírez, D. Salas-Gonzalez, R. Chaves, and M. Gómez-Río, "SVM with bounds of confidence and pls for quantifying the effects of acupuncture on migraine patients," in *Hybrid Artificial Intelligent Systems*. Springer, 2011, pp. 132–139.

[15] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 127, Apr. 2011.

[16] R. K.-S. Kwan, A. C. Evans, and G. B. Pike, "An extensible MRI simulator for post-processing evaluation," in *Visualization in Biomedical Computing*. Springer, 1996, pp. 135–140.