33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Machine Learning Classification of MRI Features of Alzheimer's Disease and Mild Cognitive Impairment Subjects to Reduce the Sample Size in Clinical Trials

Javier Escudero [*], *Member, IEEE*, John P. Zajicek, Emmanuel Ifeachor, *Member, IEEE*, and the Alzheimer's Disease Neuroimaging Initiative [#]

Abstract—**There is a need for objective tools to help clinicians to diagnose Alzheimer's Disease (AD) early and accurately and to conduct Clinical Trials (CTs) with fewer patients. Magnetic Resonance Imaging (MRI) is a promising AD biomarker but no single MRI feature is optimal for all disease stages. Machine Learning classification can address these challenges. In this study, we have investigated the classification of MRI features from AD, Mild Cognitive Impairment (MCI), and control subjects from ADNI with four techniques. The highest accuracy rates for the classification of controls against ADs and MCIs were 89.2% and 72.7%, respectively. Moreover, we used the classifiers to select AD and MCI subjects who are most likely to decline for inclusion in hypothetical CTs. Using the hippocampal volume as an outcome measure, we found that the required group sizes for the CTs were reduced from 197 to 117 AD patients and from 366 to 215 MCI subjects.**

I. Introduction

ALZHEIMER'S DISEASE (AD) poses a huge burden in modern societies [1]. In 2006, there were 26.6 million patients with AD worldwide and the prevalence is expected to grow fourfold by 2050 [1]. AD starts several years well before the criteria for clinical diagnosis are met and the diagnosis itself can only been confirmed by autopsy [2], [3]. Moreover, the diagnostic accuracy is relatively low and it depends on the setting where the evaluation takes place [2]. This hinders the AD patients' treatment and management [2].

Disease-modifying drugs for AD are expected to appear in the future but current treatments are only symptomatic [2], [4], [5]. In order for any disease-modifying drug to be most effective, the disease must be detected early. In the case of AD, this may imply to diagnose it at the Mild Cognitive Impairment (MCI) stage, when subjects have memory problems but they do not suffer from dementia [2], [4], [5]. Thus, there is a need for more accurate diagnostic tools that will allow an early detection of AD to maximize the benefits of future therapies.

On the other hand, it is challenging to conduct Clinical Trials (CTs) in AD. Promising drugs in laboratory tests are often ineffective when tested on humans [2], [6]. Due to the slow progression of AD, CTs may run over long periods of time. This increases costs and drop-out rates [4], [5]. The heterogeneity of MCI [5] and the large variability in the clinical scales used as outcome measures in the CTs also lead to increases in the required sample sizes [4]. Moreover, most clinical scales of neurodegenerative diseases do not satisfy the criteria for rigorous linear measurements and it is difficult to know which variable they measure [6]. Thus, an important task in AD research is to develop new approaches to decrease the sample size needed in the CTs of AD and MCI so that fewer patients need to be recruited.

Magnetic Resonance Imaging (MRI) is a promising biomarker to run CTs more efficiently and to help in AD diagnosis [2]–[4], [7]. It measures cerebral atrophy and it is already included in the differential diagnosis of AD from other dementias [7]. MRI technology is widely available and suitable for multi-center studies [4], [5]. Hippocampal and temporal lobe atrophy correlate with clinical decline in AD. In later stages of the disease, the atrophy extends through all the neocortex. The cortical thickness could also predict the progression from MCI to AD [4], [7].

Nonetheless, a few barriers still prevent the inclusion of biomarkers in CTs and in the diagnosis of AD. There is no fully established predictive relationship between any biomarker and the clinical outcomes yet [4] and no single biomarker is optimal to monitor all disease stages [3]. For instance, medial temporal atrophy cannot serve on its own as an absolute criterion to diagnose AD at the MCI stage [7].

Classification with Machine Learning can help to integrate biomarkers and discover data patterns useful for diagnosis and monitoring of disease [8]–[11]. Thus, a combination of biomarkers might help to reflect the evolution of AD better

Manuscript received March 25th, 2011. This work was supported in part by the UK National Institute for Health Research (NIHR) under its Programme Grants for Applied Research scheme (RP-PG-0707-10124; Clinical Trial Methods in Neurodegenerative Diseases). The Foundation for the National Institutes of Health (www.fnih.org) coordinates the private sector participation of the $60 million ADNI public-private partnership that was begun by the National Institute on Aging (NIA) and supported by the National Institutes of Health. *Asterisk indicates corresponding author.*

J. Escudero [*] and E. Ifeachor are with the Signal Processing and Multimedia Communications Research Group, School of Computing and Mathematics, University of Plymouth, Plymouth, PL4 8AA, UK (phone: +44 1752 586295; javier.escudero@ieee.org, e.ifeachor@plymouth.ac.uk).

J.P. Zajicek is Head of the Clinical Neurology Research Group, Peninsula College of Medicine and Dentistry, University of Plymouth, Derriford, Plymouth, PL6 8BX, UK.

[#] Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (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. ADNI investigators include (complete listing available at http://www.loni.ucla.edu/ADNI/Collaboration/ADNI_Authorship_list.pdf).

[3], [7]. Machine Learning classification methods can also be useful for computer-assisted decision support [9]–[11]. This approach has been applied to classify MRI scans in AD and MCI [7], [8], [10]. Although Support Vector Machines (SVMs) are the most widely used classifiers, alternative techniques are available. However, these have not been fully tested yet [10]. In addition to helping in diagnosis, classification with Machine Learning can boost the power of CTs so that smaller sample sizes (fewer patients) are needed to prove the effects of a new treatment [8]. This relies on the hypothesis that the subjects who are more likely to decline according to a classifier will suffer greater clinical decline, thus showing the effects of the drug better [7], [8].

Thus, the aim of this study is twofold. Firstly, we test the diagnostic accuracy of four classifiers [9] applied to MRI features [12] of AD and MCI subjects. Secondly and more importantly, we investigate whether those classifiers can reduce the required sample sizes in CTs of AD and MCI by selecting the subjects that are most likely to decline. We present here preliminary results addressing these two issues. We build on previous findings [8] and focus our tests on automatically extracted MRI features to compare four classifiers: Logistic Regression (LR), SVM, Radial Basis Function (RBF), and C4.5 tree learner [9], [13].

II. Methods

A. ADNI Database

Data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of CTs. The initial goal of ADNI was to recruit about 200 cognitively normal (CN) older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years, and 200 people with early AD to be followed for 2 years. For up-to-date information see www.adni-info.org.

B. Data Retrieval

The ADNI clinical and numeric summary data used in this study are as of February 7th, 2011. The database was queried for basic demographic (age, gender, and years of education), clinical (number of ApoE ε4 alleles, ADAS-Cog, and MMSE scores), and MRI (cross-sectional FreeSurfer, v.4.3 [12]) data from CN, MCI, and AD subjects at baseline. This query reported 813 records. The FreeSurfer measures did not fully pass the overall quality control in 239 cases. Thus, 574 subjects were available for analysis. However, there is a larger proportion of males in the ADNI MCI subgroup [4].

TABLE I
Demographic Data of the ADNI Subjects Included in this Study

	CN (*N*=180)	MCI (*N*=222)	AD (*N*=122)
Gender (%)	50.6 / 49.4	51.4 / 48.6	52.5 / 47.5
Age	75.92 ± 4.89	74.45 ± 7.38	74.94 ± 7.78
Years of education	15.97 ± 2.96	15.50 ± 3.03	14.78 ± 3.11
ApoE ε4 (%)	73.9 / 25.0 / 1.1	42.3 / 44.6 / 13.1	31.1 / 48.4 / 20.5
ADAS-Cog	6.08 ± 2.86	11.94 ± 4.58	18.98 ± 6.35
MMSE	29.10 ± 1.00	26.93 ± 1.82	23.11 ± 2.09

Data are given as mean ± standard deviation (SD), except for the gender and ApoE ε4 distributions, where the relative frequencies of male / females and number of subjects with 0 / 1 / 2 ApoE ε4 alleles are given.

To avoid artifactual influences of the gender in the classification, 50 male MCI subjects were randomly removed from the sample [8]. Thus, the analyses were performed on 524 subjects, whose basic data are detailed in Table I.

FreeSurfer v.4.3 was used to compute features from T_1 MRI scans acquired at 1.5T [12]. This involved motion correction, affine transformation, intensity normalization, and removal of non-brain tissues. A non-linear wrapping of a brain atlas was applied to the subject's scan to perform an atlas-based tissue segmentation of the subcortical structures. Then, cortical parcelation was done. Finally, volumetric, surface and cortical thickness summaries of the brain were computed. No manual editing was used, but the segmentation was quality-controlled. The numeric summaries are available at the ADNI website [14]. The reader interested in segmentation algorithms is referred, for instance, to [15]. We considered all variables computed with FreeSurfer to take advantage of the potentially complementary information of diverse MRI features [7]. This resulted in 328 variables per subject. Volumetric values were normalized to the intra-cranial volume [7]. To assess the reduction in the sample sizes of CTs, the hippocampal volumes of all 96 AD and 185 MCI subjects with available MRI follow-ups were retrieved.

C. Classification With Machine Learning

Only baseline data (normalized to the [0,1] range considering only the training data [9]) are used to develop the classifiers. In addition to the MRI features, age, gender, years of education, and number of ApoE ε4 alleles are presented to the classifiers [8]. The ADAS-Cog and MMSE values are not used for classification to avoid circular inference [8]. The algorithms are applied with the software Weka (version 3.6.3), which provides a well-known set of feature selection and classification methodologies [9], [13].

1) Feature selection

Including irrelevant variables into a classifier can lead to overfitting and hinder the interpretation of the model. Hence, a feature selection stage is implemented to omit irrelevant information [9]–[11], [16]. A filter method, which is independent of any classifier, is applied [9], [11]. This consists of a forward selection (Weka's *BestFirst*) to look for combinations of features with high individual predictive value of the diagnostic class and low inter-correlation (Weka's *CfsSubsetEval*) [9]. Although univariate filtering might degrade the performance of multivariate classifiers,

CfsSubsetEval takes into account the degree of redundancy among variables [9].

2) Classification Algorithms

The selection of the best classifier for diagnosis is an open problem. Hence, we compare four different classifiers: LR, SVM, RBF, and C4.5.

LR is a classical technique that estimates the class probabilities by applying the *logit* transformation to a linear regression model [10], [16]. The *Logistic* function in Weka implements the LR [9].

SVMs can implement non-linear decision boundaries by transforming the input data to a new space. A straight hyperplane in the transformed space corresponds to a non-linear boundary in the original space. Then, the SVM optimizes the decision boundary that offers the greatest margin between classes. The transformation can be done with several kernels. We consider a polynomial kernel whose degree can range between 1 and 2, as preliminary analyses showed higher accuracy with this type of kernel than with a Gaussian one. Additionally, we vary the complexity parameter of the SVMs from $\log_{10}(C)$ equal to –4 to +4. This classifier is applied with the SVM Sequential Minimal Optimization (*SMO*) algorithm in Weka [9], [10].

RBF is an artificial neural network with input, hidden, and output layers. The nodes in the hidden layer represent a particular point in the data space. Their output values depend on the distance between this point and the data instances. A Gaussian activation function transforms these distances into non-linear similarity measures, which are combined linearly at the output. The centers and widths of the Gaussian functions are estimated with *k*-means clustering. We consider values for the number and minimum standard deviation of the clusters between 1 and 4, and 0.05 and 0.55, respectively. The *RBFNetwork* function in Weka applies this classifier [9].

The C4.5 learner uses a tree-like structure to arrange a set of decisions into a hierarchy. The nodes in the tree structure involve testing a particular variable and the leaves provide the classification that applies to all instances that reach them. Features to be tested on each node are selected according to their information gain. A pruned C4.5 tree learner, implemented in the Weka's *J48* algorithm, is tested. We optimize the confidence level for pruning and minimum number of instances per leaf between the values of 0.05 and 0.70, and 1 to 10, respectively [9], [10].

D. Diagnostic Accuracy and its Impact on CTs

In this study, we deal with two research questions. The first one has to do with the fact that the diagnoses of AD and MCI are difficult to ascertain [2]. Thus, we test the accuracy of classifiers for CN vs. AD and CN vs. MCI separation. The second, and most important, issue is related to the large number of patients needed in CTs of AD and MCI [6], [8]. To test if machine learning helps to run more efficient CTs, we use the previously trained classifiers to select the patients most likely to decline and investigate if the sample sizes of hypothetical CTs in AD and MCI decrease.

Ten different stratified full runs of a ten-fold cross-validation are used to evaluate the classification performance of the classifiers [9]. In each fold of cross-validation, another ten-fold cross-validation was performed on the training samples to optimize the tunable parameters of each model via a grid-search procedure [9]. The classifiers were compared on the basis of their accuracy and area under the ROC curve (AUC) values [9], [16].

Once the classifiers have been trained, they are used to select the 50% and 33% of the MCI and AD subjects who are most likely to decline according to the outcome of these classifiers [8]. The atrophy of the hippocampus after 12 months is considered the outcome measure in a hypothetical CT. The minimum sizes per group required to detect a 25% reduction in the atrophy rate for a two-arm (treatment and placebo) study with 80% power and two-sided test are computed (with and without the selection of subjects) by [8]:

$$n = \frac{2\sigma^2\left(z_{1-\alpha/2} + z_{\text{power}}\right)^2}{\left(0.25\beta\right)^2}, \quad (1)$$

where $\alpha = 0.05$, β and σ are the significant level, mean and SD of the decrease in the outcome measure. z_p is the value of the standard Gaussian distribution for $P[Z < z_p] = p$ [8].

III. Results and Discussion

First of all, feature selection was applied within each of the cross-validation folds to select the most informative variables. The average number (mean ± SD) of selected features was 36.37 ± 4.29 for the classification of CN vs. AD, and 24.41 ± 3.29 for CN vs. MCI. Some features were always selected in both classification tasks: age; number of ApoE ε4 alleles; right and left hippocampal, left entorhinal cortex, and left amygdale volumes; and average cortical thickness of the left middle temporal cortex. For the CN vs. AD classification, some other features extracted from areas around the temporal lobe (e.g., volume and thickness of the left inferior and middle temporal cortex) were also selected in all folds. This automatic feature selection agrees with previous reports that emphasized the relevance of regions in or near the medial temporal lobe in the progression of AD [4], [7], [10]. This supports the feature selection step, even though further validations and comparisons with wrapper methods are required.

Secondly, the classifier parameters were optimized. The average degree of the SVM polynomial kernel was 2.0 in the CN vs. AD task and 1.7 for the MCI. In both cases, the $\log_{10}(C)$ was about 0.3. Regarding the RBF, the data were mapped to a mean number of 2.5 clusters for both AD and MCI with a minimum SD of about 0.5. As for the tree learner, the optimized confidence level for pruning was 0.6 for AD and 0.3 for MCI, with a minimum of 6 and 8 instances per leaf for AD and MCI, respectively.

Thirdly, the classifiers were tested. The accuracy and AUC appear in Table II. These values are similar to the

TABLE II
AVERAGE PERFORMANCE (MEAN ± SD) OF THE CLASSIFIERS

Experiment	Classifier	Accuracy	AUC
CN vs. AD	LR	85.63 ± 5.94	0.919 ± 0.055
	SVM	89.17 ± 5.08	0.884 ± 0.056
	RBF	87.94 ± 5.24	0.874 ± 0.058
	C4.5	83.93 ± 6.17	0.833 ± 0.064
CN vs. MCI	LR	72.51 ± 6.79	0.803 ± 0.067
	SVM	72.65 ± 6.61	0.726 ± 0.067
	RBF	70.92 ± 7.24	0.710 ± 0.088
	C4.5	72.69 ± 7.24	0.725 ± 0.073

TABLE III
MINIMUM SAMPLE SIZES PER ARM TO DETECT A 25% ANNUAL RATE OF CHANGE IN THE HIPPOCAMPAL VOLUME WITH ALL SUBJECTS AND WHEN THOSE MOST LIKELY TO DECLINE ARE PRE-SELECTED WITH CLASSIFIERS

Case	Subset	LR	SVM	RBF	C4.5
AD	No selection	197	197	197	197
	50%	183	196	174	148
	33%	218	117	132	143
MCI	No selection	366	366	366	366
	50%	243	290	420	380
	33%	131	215	467	292

classification rates found in [8] and higher that those of [10] for a smaller dataset of CN and very mild AD subjects. However, comparing classification rates between studies is not straightforward due to the different datasets. The performance of the CN vs. MCI task was lower than for AD patients due to the overlap between CNs and MCIs [8]. The AUC values of the LR were significantly higher ($p < 0.05$) that those of the other classifiers. There were no significant differences among the accuracy values of the algorithms.

Finally, the previously trained classifiers were used to reduce the samples sizes for CTs by selecting the 50% and 33% of the subjects most likely to decline. Table III details the sample sizes for AD and MCI patients in CTs with the hippocampal volume as outcome measure. The sizes calculated without this scheme ("No selection") are also given in Table III. Notable reductions in the number of subjects needed to run the hypothetical CT were achieved in most cases. The results confirm the utility of SVMs in this type of application [8] and highlight the potential of other classifiers. Yet, it was only possible to reduce the required number of patients in all cases with SVM. Thus, SVM might be the best technique to ensure that the selection of subjects does not decrease the power of the CT.

IV. CONCLUSION

In this preliminary study, we applied four Machine Learning classifiers [9] to MRI features [12] in AD and MCI patients. The ability of the classifiers to select the subjects who are most likely to decline and reduce the sample size in CTs was investigated. The results indicated that the use of SVMs to select the subjects for CTs decreased the sample size per arm (treatment and placebo) from 197 to 117 and from 366 to 215 AD and MCI subjects, respectively. Nonetheless, further analysis, including validation of the feature selection and inclusion of other classification techniques, are needed to corroborate or refute our results.

ACKNOWLEDGMENT

This article presents independent research commissioned by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research scheme (RP-PG-0707-10124). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

REFERENCES

[1] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H. M. Arrighi, "Forecasting the global burden of Alzheimer's disease," *Alzheimers Dement.*, vol. 3, no. 3, pp. 186–191, Jul. 2007.

[2] K. Blennow, M. J. de Leon, and H. Zetterberg, "Alzheimer's Disease," *Lancet*, vol. 368, no. 9533, pp. 387–403, Jul. 2006.

[3] C. R. Jack Jr, D. S. Knopman, W. J. Jagust, L. M. Shaw, P. S. Aisen, M. W. Weiner, *et al.*, "Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade," *Lancet Neurol.*, vol. 9, no. 1, pp. 119–128, Jan. 2010.

[4] J. L. Cummings, "Integrating ADNI results into Alzheimer's disease drug development programs," *Neurobiol. Aging*, vol. 31, no. 8, pp. 1481–1492, Aug. 2010.

[5] M. W. Weiner, P. S. Aisen, C. R. Jack Jr, W. J. Jagust, J. Q. Trojanowski, L. Shaw, *et al.*, "The Alzheimer's Disease Neuroimaging Initiative: Progress report and future plans", *Alzheimers Dement.*, vol. 6, no. 3, pp. 201–211, May 2010.

[6] J. C. Hobart, S. J. Cano, J. P. Zajicek, and A. J. Thompson, "Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations," *Lancet Neurol.*, vol. 6, no. 12, pp. 1094–1105, Dec. 2007.

[7] G. B. Frisoni, N. C. Fox, C. R. Jack, P. Scheltens, and P. M. Thompson, "The clinical use of structural MRI in Alzheimer disease," *Nat. Rev. Neurol.*, vol. 6, no. 2, pp. 67–77, Feb. 2010.

[8] O. Kohannim, X. Hua, D. P. Hibar, S. Lee, Y. Y. Chou,A. W. Toga, *et al.*, "Boosting power for clinical trials using classifiers based on multiple biomarkers," *Neurobiol. Aging*, vol. 31, no. 8, pp. 1429–1442, Aug. 2010.

[9] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, Morgan Kaufmann, 2nd edition, 2005.

[10] R. Chen and E. H. Herskovits, "Machine-learning techniques for building a diagnostic model for very mild dementia," *NeuroImage*, vol. 52, no. 1, pp. 234–244, Aug. 2010.

[11] Y. Huang, H. Zheng, C. Nugent, P. McCullagh, N. Black, K. E. Vowles, and L. McCracken, "Feature Selection and Classification in Supporting Report-based Self Management for People with Chronic Pain," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 1, pp.54–61, Jan. 2011.

[12] B. Fischl, A. van der Kouwe, C. Destrieux, E. Halgren, F. Segonne, D. H. Salat, *et al.*, "Automatically parcellating the human cerebral cortex," *Cereb. Cortex*, vol. 14, no. 1, pp. 11–22, Jan. 2004.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, "The WEKA Data Mining Software: An Update," *ACM SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, Jun. 2009.

[14] "Automated Whole Brain Segmentation Using FreeSurfer," Last access: March 17th, 2011. Available: http://adni.loni.ucla.edu/wp-content/uploads/2010/12/UCSFFreeSurferMethodsSummary.pdf.

[15] A. Mishra, P. W. Fieguth, and D. A. Clausi, "Decoupled Active Contour (DAC) for Boundary Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 310–324, Feb. 2011.

[16] J. Escudero, S. Sanei, D. Jarchi, D. Abásolo, and R. Hornero, "Regional Coherence Evaluation in Mild Cognitive Impairment and Alzheimer's Disease Based on Adaptively Extracted Magnetoencephalogram Rhythms," *Physiol. Meas.*, In press, May 2011.