Title: Predicting Alzheimer's Disease using Machine Learning approaches.

M. Pinheiro[1] PhD, S. Neto[1] PhD, V. Enes[1,2] MSc, G. Moura[1] PhD, M. Santos[1] PhD

[1] Institute of Biomedicine—iBiMED, Department of Medical Sciences, University of Aveiro, 3810-193 Aveiro, Portugal
[2] Center for Research & Development in Mathematics and Applications (CIDMA), University of Aveiro, 3810-193 Aveiro, Portugal

Keywords: Alzheimer's Disease, Machine Learning, ADNI
Corresponding author: M.Pinheiro, monsanto@ua.pt

Alzheimer's disease (AD) is the most common form of dementia, a general term for memory loss and other cognitive abilities serious enough to interfere with daily life.
Due to the difficulty of identifying AD in early stages we want to test different models to help predict patients with AD and determine the accuracy of conventional machine learning algorithms and neural networks to evaluate each data set and several combinations between them.

Data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) public database (http://adni.loni.usc.edu/) (1). The ADNI was launched in 2003 as a public partnership by several organizations, including the National Institute on Ageing (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies, and non-profit organizations.

Several studies were made with ADNI data, base only with Magnetic resonance imaging (MRI) data, achieving very high values of precision 96.1 (2) or ~98% of accuracy (3).
For the purpose of this study, genotype, gene expression and MRI data were obtained from the ADNI-GO/2 database. Only individuals with a condition in last time point of Normal (control) or AD (case) were considered. We exclude all Mild Cognitive Impairment (MCI).

SNP genotype data. Since genetic risk factors can help scientists to focus on relevant biological pathways and form effective hypothesis, identifying risk genetic markers associated with brain imaging, can help understanding the underlying biological mechanisms. We downloaded the ADNI-GO/2 genotyping data, performed quality control and population stratification using the approach described at (4).
In this study we used only samples with ancestral relation with Western European Ancestry (CEU), to limit potential effects of population stratification.
Parsing genotype data, the additive model was chosen because it has reasonable power to detect both additive and dominant effects, codifying each SNP with values 0, 1 or 2. The SNPs with less than 5% variation or non variation at all, across all samples, were removed from the study.

Imaging data. The MRI data used in this study were previous pre-processed by Dementia Research Centre, UCL Institute of Neurology Brain, applying the method of Boundary Shift Integral (5) in the Brain and Ventricular regions. The data were also obtained from the ADNI database (FOXLABBSI_02_07_19.csv). The data have several time points for each participant and because of the absent of some time points in the samples the calculation of a slope between points were performed.

Gene expression. Gene expression profiling from blood samples of ADNI participants was contributed by Bristol-Myers Squibb (BMS) and performed at the BMS laboratories for 811 ADNI participants from the ADNI cohort performed by an Affymetrix array. The file with data is "ADNI_Gene_Expression_Profile.csv".

First approach was to reduce the number of features, identifying which of them can contribute most to the classification of the phenotype. Using Weka (6),with the Information Gain algorithm based on entropy, we calculate the contribution of each feature to the identification of the phenotype. By choosing the highers in the rank we can apply several types of machine learning approaches to identify the best model.

We used Support Vector Machine (SVM), Random Forest, eXtreme Gradient Boost and Neural Networks using Scikit-learn machine learning package (7) that gave a possibility to identify the best model of an algorithm that best adapts to the data based on grid search, with a cross validation of 5 fold, and optimizing for Area Under The Curve (AUC) and Accuracy. We run each method 100 times to calculate an average of accuracy and identify which parameters in the grid search had the best result.

Table 1: Results for each dataset and algorithms. The number of features are in front of each dataset. The columns Train and Test has the number of cases an control for each dataset. The precision mean are in the bottom followed by standard deviation.

		GWAS (86)	GWAS (31)	Image (5)	GWAS (31) + Image (5)	Expre. (13)	Expre. (13) + GWAS (31)	Espre. (13) + GWAS (31) + Image (5)
Train	Control	155	155	155	155	82	82	82
	Case	275	275	275	275	39	39	39
Test	Control	18	18	18	18	4	4	4
	Case	20	20	20	20	6	6	6
Algorithms	SVM	52 (0)	54 (0)	58 (0.2)	28 (0)	36 (0)	78 (0.1)	70 (0)
	RF	36 (0.2)	64 (0.1)	60 (0.03)	67 (0.03)	71 (0.1)	71 (0.1)	78 (0.1)
	XGBoost	52 (0)	64 (0)	61 (0.04)	64 (0)	60 (0)	72 (0.1)	72 (0)
	NN	42 (0.2)	47(0.2)	36 (0.2)	53 (0.2)	32 (0.1)	36 (0.2)	37 (0.2)

There was not a distinct winner but we can consider that RF had the best overall performance. The NN had the worst results with a very low values of precision. There is an increasing of precision when gene expression was added but this can be an artfact because the number of individuals descreased significatly.

Based on the results we can achieved a maximum of 78% of precision in two/three combinations of data and two distinct algorithms. The higher values of precision were obtained when we combined several types of data, like imagining with genotype and gene expression.

For future work we can resort to stacking or meta-ensembling the 1st level predictive models to generate a 2nd level model which tends to outperform all of them, but is very difficult to achieve a precision of 96% of Jha D. et al. (2) with this type of data.

Acknowledgements

This work was supported by the Integrated Programme of SR&TD “pAGE – Protein aggregation across the lifespan” (reference CENTRO-01-0145-FEDER-000003), co-founded by Centro 2020 program, Portugal 2020, European Union, through the European Regional Development Fund.

MEDPERSYST project (POCI-01-0145-FEDER-016428-PAC) “Redes sinapticas e abordagens compreensivas de medicina personalizada em doenças neurocomportamentais ao longo da vida (SAICTPAC/0010/2015)
MEDISIS : Promoção do envolvimento dos atores chave da região centro para a transferência de conhecimento e criação de novas parcerias nas áreas da Medicina de Sistemas, de acordo com a definição das unidades de investigação do DiscoveriesCTR; project (CENTRO-01-0246-FEDER-000018; CENTRO2020).

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). 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.

References

1. Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack C, Jagust W, et al. The Alzheimer’s disease neuroimaging initiative. Neuroimaging Clinics of North America. 2005.
2. Jha D, Alam S, Pyun J-Y, Lee KH, Kwon G-R. Alzheimer’s Disease Detection Using Extreme Learning Machine, Complex Dual Tree Wavelet Principal Coefficients and Linear Discriminant Analysis. J Med Imaging Heal Informatics. 2018;
3. Khagi B, Kwon GR, Lama R. Comparative analysis of Alzheimer’s disease classification by CDR level using CNN, feature selection, and machine-learning techniques. Int J Imaging Syst Technol. 2019;
4. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nat Protoc. 2010;
5. Freeborough PA, Fox NC. The boundary shift integral: An accurate and robust measure of cerebral volume changes from registered repeat MRI. IEEE Trans Med Imaging. 1997;
6. Reutemann P, Hall M, Frank E, Witten IH, Holmes G, Pfahringer B. The WEKA data mining software. ACM SIGKDD Explor Newsl. 2009;
7. Klikauer T. Scikit-learn: Machine Learning in Python. TripleC. 2016.