American Journal of Biomedical Engineering 2020, 10(1): 1-8
DOI: 10.5923/j.ajbe.20201001.01

Investigation of Genome-Wide Association SNPs and Alzheimer's Disease

Mark Nnh Mikhail[1], Ahmed Y. Sayed[2], Mai S. Mabrouk[3,*], Ayman M. Eldeib[1]

[1]Systems and Biomedical Engineering Department, Faculty of Engineering, Cairo University, Cairo, Egypt
[2]Department of Engineering Mathematics and Physics, Faculty of Engineering, El- Mataria, Helwan University, Cairo, Egypt
[3]Biomedical Engineering, Misr University for Science and Technology, 6th of October, Giza, Egypt

Abstract The aim of this work is to measure the influence of genome-wide association study single nucleotide polymorphisms (SNPs) in Alzheimer's disease (AD). Data mining methods were tested. Data used were obtained from ADNI database. Subjects were 214 normal controls (NCs), 364 subjects with mild cognitive impairment (MCI), and 179 subjects with early AD. Linear regression (LR), random forests (RF) and multifactor dimensionality reduction (MDR) models were used. The results demonstrate the effectiveness of using RF and MDR. The MDR model produced the best sensitivity in all comparisons using only 3 SNPs. Regarding specificity, LR resulted in the best specificity in two comparisons (NC vs. MCI and MCI vs. AD), while MDR produced the best specificity in comparison of NC vs. AD. Several significant polymorphisms associated with MCI and AD were identified. RF and MDR are alternatives to existing methods for detecting genetic interactions.

Keywords Alzheimer's disease, Single nucleotide polymorphisms, Linear regression, Random forest, Multifactor dimensionality reduction

1. Introduction

It is typically believed that genes and biomarkers involved in age-related diseases, such as coronary artery disease, cerebrovascular disease, and Alzheimer's disease (AD), play a vital role in human ageing [1].

AD is a complex neurodegenerative disorder that affects up to eighty-one million persons worldwide [2]. AD is usually divided into two types: (i) cases with strong familial clustering, which often show Mendelian disease transmission mechanism and generally exhibit an early (65 years) or very early (50 years) age of onset (collectively referred to as EOAD) and (ii) cases of later-onset age (LOAD) (typically well beyond 65 years), showing no obvious familial aggregation. A strong genetic basis is known for AD, with heritability estimates of approximately 80% [3]. To identify the genes involved in the common LOAD, efforts have focused on conducting genome-wide association studies (GWAS) [4].

Whole-exome sequencing and GWAS are recommended to identify the risk gene variants for LOAD [5,6]. Their main role is to identify rare coding variants that were recently recognized as a risk for LOAD [6-9]. With respect to common variants for AD, since 2009, five large GWAS and one meta-analysis have identified more than twenty loci significantly associated with LOAD. According to their potential role in the process causing AD, these genes were classified into three groups: (1) lipid metabolism: APOE, CLU, ABCA7 and SORL1; (2) immune response: CR1, CD33, MS4A, EPHA1, ABCA7, CLU, HLA-DRB5/DRB1 and INPP5D; and (3) endocytosis: BIN1, PICALM, EPHA1, RIN3, CD2AP, SORL1, MEF2C and MADD [5,10-14]. However, in most cases, the identified single nucleotide polymorphisms (SNPs) have small to moderate effect sizes, and the proportion of heritability explained is quite modest.

The aim of this work is to measure the influence of GWAS SNPs on gene expression in AD [15]. Due to limitations of the linear model and other parametric statistical models, machine learning and data mining methods will be tested for the same data, mainly random forests (RF) and multifactor dimensionality reduction (MDR), as explained by Moore et al. (2010) [16].

2. Methods

2.1. Experimental Setup

The data used in the present study were downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The ADNI is a consortium of universities and medical centers that was established to develop standardized

* Corresponding author:
msm_eng@yahoo.com (Mai S. Mabrouk)
Published online at http://journal.sapub.org/ajbe

imaging techniques and biomarker procedures in normal subjects, subjects with MCI, and subjects with early AD [17].

The subjects for the study were classified as normal controls (NC), subjects with MCI, or subjects with early AD. The criteria for classification of the subjects were described in McKhann et al. (1984) [18].

2.2. Dataset

In the present study, data were downloaded from the ADNI web site (ADNI 1 data) in November 2016. In this work, only ADNI1 subjects with all detailed clinical information and genotype data were included. Whole-genome sequencing (WGS) data of 819 individuals were obtained from the ADNI database (http://adni.loni.usc.edu/). An initial quality control (QC)-based filtering step was performed using PLINK [19] and applied to the selected datasets. The following QC procedures were carried out in order, according to the steps suggested by Shi et al. (2012) [1].

- SNPs with genotyping rate less than 0.95 (--geno 0.05) were excluded from further analysis.
- SNPs with a minor allele frequency (MAF) less than 0.01 (--maf 0.01) were excluded from further analysis.
- A list of SNPs with MAF between 0.01 and 0.05 was generated (--freq). Within this short list, SNPs with a genotyping rate less than 0.99 (--geno 0.01) were excluded (--exclude) from further analysis.
- SNPs with a Hardy-Weinberg Equilibrium p value less than 0.001 (--hwe 0.001 --hwe-all) were excluded from further analysis, irrespective of status (AD cases or controls).
- Individuals with a genotyping rate less than 0.95 (--mind 0.05) were excluded from further analysis.

Table 1. Demographic characteristics of the participant groups

Characteristic	Study Group			P value			
	NC N=214	MCI N=364	AD N=179	Overall	NC vs. MCI	NC vs. AD	MCI vs. AD
Age, mean ± SD, y	75.67 ±4.91	74.74 ±7.32	74.44 ±7.33	0.230			
Education, mean ± SD, y	16.07 ±2.80	15.68 ±3.04	14.65 ±3.17	<0.001	0.384	<0.001	0.001
Gender,%				0.003	0.005	0.895	0.005
- Males	53.7	65.7	53.1				
- Females	46.3	34.3	46.9				
Marital status,%				0.008	0.001	0.149	0.254
- Married	70.8	80.5	80.4				
- Widowed	15.9	11.5	10.6				
- Divorced	6.5	6.6	5.0				
- Never Married	7.0	1.4	3.9				
Ethnicity, %				0.366			
- Not Hisp/Latino	98.6	96.7	96.6				
- Hisp/Latino	0.9	3.0	2.2				
- Unknown	0.5	0.3	1.1				

Table 2. Gene regions and SNPs used in the current study

Gene	Location	SNP ID
ABCA7	19p13.3	rs3764650
BIN1	2q14.3	rs744373, rs7561528
CD2AP	6p12.3	rs9296559
CD33	19q13.41	rs3865444
CLU	8p21.1	rs11136000, rs7012010
CR1	1q32.2	rs6701713, rs3818361, rs1408077
EPHA1	7q35	rs11771145, rs11767557
EXOC3L2	19q13.32	rs597668
FERMT2	14q22.1	rs17125944
MS4A6A/ MS4A4E	11q12.1	rs670139, rs610932
NME8	7p14.1	rs2718058
PICALM	11q14.2	rs3851179, rs541458, rs543293, rs677909
SLC24A4 RIN3	14q32.12	rs10498633
SORL1	11q24.1	rs2070045, rs661057
APOE-ε4	19	Number of Copies

This yielded a total of 757 subjects, including 179 LOAD patients, 364 MCI patients and 214 NC.

SNPs belonging to the top AD candidate genes listed on the AlzGene database [20], together with those listed by Lambert et al. (2013) [13] and Nettiksimmons et al. (2016) [21], were selected for use in this study if they were present in the ADNI database. Table 1 summarizes the demographic characteristics of the participant groups, while table 2 summarizes the top candidate genes used in this study and their identification numbers, all of which have been proposed to play some role in AD.

2.3. Linear Regression (LR) Model

Binary logistic regression analysis was performed under an additive model that included age and sex as covariates to test for associations between each SNP allele and LOAD risk. The data were divided according to age into groups in which each contains an approximately equal number of samples. All significant SNPs were then put into a stepwise multivariable regression model to evaluate the association of SNPs and LOAD susceptibility. All statistical analyses were performed using IBM SPSS version 24.0. The adjusted p value of ≤ 0.05 was defined as statistically significant.

A prediction score was developed using any significant SNP from the previous multivariable regression. This score was used for receiver operator characteristics (ROC) curve analysis with the calculation of the highest Youden Index to calculate best cut-off values that differentiate between the three diagnoses and for the evaluation of this prediction by sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and accuracy.

2.4. Random Forests (RF) Model

Twenty percent of the dataset was randomly selected and used as the training set, and consequently, the remaining 80% served as the learning set. An RF model is a collection of individual decision tree classifiers, where each tree in the forest has been trained using a bootstrap sample of subjects from the data, and each attribute in the tree is chosen from among a random subset of attributes. Individual trees are constructed as follows from data having N samples and M attributes:

1. Choose a training set by selecting 20% of samples.
2. At each node in the tree, m attributes are randomly selected from the entire set of M attributes in the data (the magnitude of m is constant throughout the forest building).
3. Choose the best split at that node from among the m attributes.
4. Iterate the second and third steps until the tree is fully grown [16].

RF was performed using RapidMiner Studio software, version 8. The ReliefF algorithm was used for attribute selection.

The RF model using regression resulted in a regression tree. This regression tree was used to calculate a prediction for diagnosis. ROC analysis with the calculation of the highest Youden Index was used to calculate best cut-off values for this prediction that differentiate between the three diagnoses.

2.5. Multifactor Dimensionality Reduction (MDR) Model

MDR is a computational strategy for detecting and characterizing non-linear patterns of gene–gene interactions in genetic association studies. MDR was developed as a genetic model-free non-parametric machine learning strategy for identifying combinations of genetic and environmental factors that are predictive of a discrete clinical end point [22].

MDR was performed using Multifactor Dimensionality Reduction Open Source software, version 3.0.2.

2.6. Implementation

An I7 PC was used together with the following software:

- PLINK used for QC and was downloaded from the following site: http://pngu.mgh.harvard.edu/purcell/plink/.
- IBM SPSS version 24.0
- RapidMiner Studio software, version 8 (RapidMiner, Inc. Boston, MA, USA), downloaded from https://rapidminer.com.
- Multifactor Dimensionality Reduction Open Source software, version 3.0.2, produced by The Computational Genetics Laboratory at Dartmouth Medical School, Hanover, New Hampshire, USA The software is available for download at http://www.multifactordimensionalityreduction.org/.

3. Results

3.1. Linear Regression (LR) Model

The results of the multiple regression analysis are presented in table 3. The results indicate that the following SNPs are significant: APOE-ε4, CD33 (rs3865444), CR1 (rs1408077), BIN1 (rs7561528) and EPHA1 (rs11771145). The prediction score resulting from the stepwise LR model was calculated using the following equation:

Score = 0.823 + (APOE-ε4*0.322) + (BIN1_ rs7561528-AA*0.143) + (EPHA1_ rs11771145-AG*0.013) – (CD33_ rs3865444-TG*0.032) – (CR1_ rs1408077-GG*0.105)

Table 4 shows the prediction accuracy results, sensitivity, and specificity for the prediction score. The highest accuracy occurred in differentiation between NC and AD (accuracy = 70.2%, sensitivity = 69.8% and specificity = 70.64%), followed by differentiation between MCI and AD but with limited sensitivity (accuracy = 64.6%, sensitivity = 38.5% and specificity = 78.0%). The smallest accuracy occurred in the differentiation between NC and MCI (accuracy = 61.7%,

sensitivity = 57.4% and specificity = 70.6%).

3.2. Random Forests (RF) Model

A regression tree of significant predictor SNPs was constructed, and it contained the following SNPs: APOE-ε4, CR1 (rs3818361), PICALM (rs543293), EPHA1 (rs11767557), NM8 (rs2718058), PICALM (rs3851179) and CLU (rs11136000). The regression tree is presented in Fig. 1. The prediction score increased from the NC to the MCI and to the AD groups. ROC analysis was used to detect the best cut-off values to differentiate between the three diagnoses, and 0.856 was found to differentiate between NC and MCI, 0.866 to differentiate between NC and AD, and 0.864 to differentiate between MCI and AD. All were statistically significant. However, the results are highly significant for the differentiation of NC vs. MCI and NC vs. AD ($p<0.001$) and weakly significant ($p=0.014$) for the differentiation of MCI vs. AD.

Table 5 shows the prediction accuracy results for the prediction score. The highest accuracy occurred in the differentiation between NC and AD (accuracy = 69.4%, sensitivity = 67.1% and specificity = 71.4%), followed by the differentiation between NC and MCI (accuracy = 61.6%, sensitivity = 58.3% and specificity = 67.7%). The lowest accuracy occurred in the differentiation between MCI and AD (accuracy = 52.6%, sensitivity = 67.1% and specificity = 45.5%).

Table 3. Results of multiple regression

Gene	SNP	Allele	OR (95% C.I.)	Adjusted P value
A- Comparison of NC and MCI				
APOE-ε4	Number of Copies	0	Ref.	**<0.001**
		1	2.709 (1.812-4.050)	**<0.001**
		2	10.997 (3.813-31.713)	**<0.001**
CD33	rs3865444	TT	Ref.	**0.035**
		GG	1.819 (0.905-3.655)	**0.093**
		TG	2.453 (1.204-4.998)	**0.013**
B- Comparison of NC and AD				
APOE-ε4	Number of Copies	0	Ref.	**<0.001**
		1	4.479 (2.745-7.307)	**<0.001**
		2	23.870 (7.917-71.981)	**<0.001**
CR1	rs1408077	TT	Ref.	**0.034**
		GG	0.148 (0.029-0.752)	**0.021**
		TG	0.216 (0.041-1.125)	**0.069**
BIN1	rs7561528	GG	Ref.	**0.051**
		AA	2.273 (1.083-4.771)	**0.030**
		AG	1.588 (0.954-2.645)	**0.075**
C- Comparison of MCI and AD				
APOE-ε4	Number of Copies	0	Ref.	**0.029**
		1	1.575 (1.033-2.400)	**0.035**
		2	1.983 (1.123-3.501)	**0.018**
EPHA1	rs11771145	GG	Ref.	**0.038**
		AA	0.572 (0.290-1.129)	**0.107**
		AG	0.616 (0.412-0.921)	**0.018**

Table 4. Prediction accuracy results for LR of the SNPs

	Sensitivity 95% C.I.	Specificity 95% C.I.	PPV 95% C.I.	NPV 95% C.I.	Accuracy 95% C.I.
NC vs. MCI	57.4% 52.2-62.6	70.6% 64.0-76.6	76.8% 71.1-80.4	49.3% 44.1-57.1	61.7% 57.7-65.7
NC vs. AD	69.8% 62.5-76.5	70.6% 64.0-76.6	66.5% 59.5-73.6	73.7% 66.9-79.2	70.2% 65.4-74.7
MCI vs. AD	38.5% 31.4-46.1	78.0% 73.4-82.2	46.3% 40.1-54.0	72.1% 65.3-77.0	64.6% 60.6-68.8

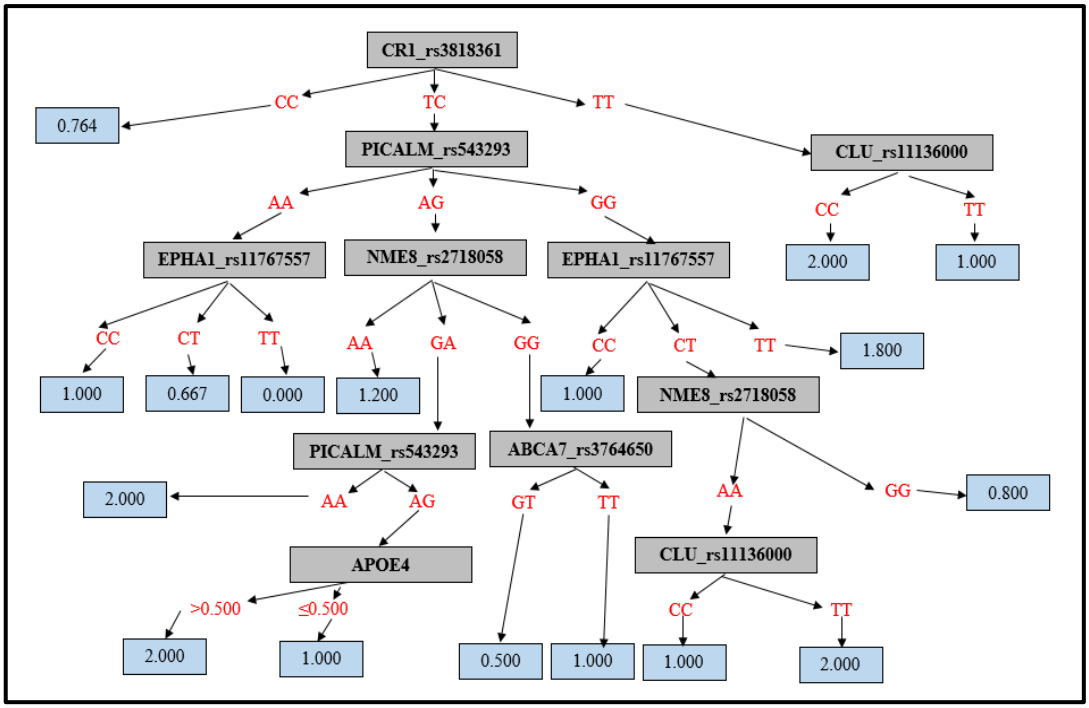

Figure 1. Random Forests Model Regression Tree

Table 5. Prediction accuracy results for RF of the SNPs

	Sensitivity 95% C.I.	Specificity 95% C.I.	PPV 95% C.I.	NPV 95% C.I.	Accuracy 95% C.I.
NC vs. MCI	58.3% 52.4-63.9	67.7% 59.9-74.9	76.9% 72.3-80.9	46.8% 42.5-51.1	61.6% 56.9-66.1
NC vs. AD	67.1% 58.9-74.7	71.4% 63.8-78.3	68.1% 61.9-73.6	70.6% 65.1-75.5	69.4% 63.9-74.5
MCI vs. AD	67.1% 58.9-74.7	45.5% 39.7-51.3	37.7% 34.2-41.4	73.8% 68.4-78.5	52.6% 47.8-57.3

Table 6. Prediction accuracy results for MDR model of the SNPs

Gene	SNP	OR 95% C.I.	Sensitivity 95% C.I.	Specificity 95% C.I.	PPV 95% C.I.	NPV 95% C.I.	Accuracy 95% C.I.
NC vs. MCI							
CD2AP + MS4A4E + APOE-ε4	rs9296559 rs670139 Copies	4.01 2.80-5.74	65.1% 60.0-70.0	68.2% 61.5-74.4	77.7% 73.9-81.1	53.5% 49.3-57.6	66.3% 62.3-70.1
NC vs. AD							
CR1+ EPHA1+ APOE-ε4	rs1408077 rs11771145 Copies	7.31 4.67-11.43	72.6% 65.5-79.0	73.4% 66.9-79.2	69.5% 64.2-74.4	76.2% 71.4-80.5	73.0% 68.4-77.4
MCI vs. AD							
CD2AP+ EPHA1+ SORL1	rs9296559 rs11771145 rs661057	3.72 4.29-5.57	76.5% 69.6-82.5	53.3% 48.0-58.5	44.6% 42.3-48.0	82.2% 77.7-86.0	61.0% 56.7-65.1

3.3. Multifactor Dimensionality Reduction (MDR) Model

MDR did not result in a prediction score (as in linear regression and RF). However, it resulted in model using (if --- then). Table 6 shows the prediction accuracy results, sensitivity, and specificity for the MDR model of the SNPs.

For the comparison of NC vs. MCI, MDR used CR2AP rs9296559, MS4A4E rs670139 and the APOE-ε4 gene, which resulted in an odds ratio (OR) of 4.01 and an accuracy of 66.3% (sensitivity = 65.1% and specificity = 68.2%). For

the comparison of NC vs. AD, MDR used CR1 rs1408077, EPHA1 rs11771145 and the APOE-ε4 gene, which resulted in an OR of 7.31 and accuracy of 73.0% (sensitivity = 72.6% and specificity = 73.4%). Comparison of MCI and AD used CD2AP rs9296559, EPHA1 rs11771145 and SORL1 rs661057 and resulted in an OR of 3.72 and an accuracy of 61.0% (sensitivity = 76.5% and specificity = 53.3%).

4. Disscusions

Prediction is often a primary goal of genomic data analyses. The complexity and high dimensionality of genomic data require flexible and powerful statistical learning tools for effective statistical analysis [23].

The LR model, together with other data mining and machine learning methods (RF and MDR), was tested for the same data, as suggested by Moore et al. (2010) [16].

The results demonstrated the effectiveness of using RF and MDR for identifying AD causal SNPs with acceptable accuracy. The MDR model produced the best sensitivity in all three comparisons (NC vs. MCI, NC vs. AD and MCI vs. AD) using only 3 SNPs for its algorithm. Regarding specificity, LR resulted in the best specificity in two comparisons (NC vs. MCI and MCI vs. AD), while MDR produced the best specificity in the comparison of NC vs. AD. Accordingly, despite the smaller number of predictors in MDR, the classification performance achieved slightly better performance than other methods.

Previous studies concentrated on the use of RF for the analysis of genetic data and found that it is an effective tool for such settings [23]. Wu et al. (2003) [24] compared RF with linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), k-nearest neighbour (KNN) classifier, bagging and boosting classification trees, and support vector machine (SVM) for the separation of early stage ovarian cancer samples from normal tissue samples based on mass spectrometry data. RF performed better than other methods in terms of prediction error rate. Lee et al. (2005) [25] presented a comprehensive comparison of RF to LDA, QDA, logistic regression, partial least square (PLS), KNN, neural network, SVM, and other classification methods using seven microarray gene expression data sets. RF showed the best performance among all tree-based methods. These results suggest that RF is capable of accurate prediction. Carcia-Magarinos et al. (2009) [26] evaluated RF, classification and regression trees (CART), and logistic regression (LR) in 99 simulated scenarios involving different sample sizes, missing data, minor allele frequencies, and other factors. RF was more powerful in detecting true association. CART, RF, and LR yielded similar results in terms of detection of true association; however, CART and RF outperformed LR with regard to classification error. Molinaro et al. (2011) [27] compared RF with Monte Carlo logic regression and MDR in testing SNPs in pro-inflammatory and immuno regulatory genes and the risk of non-Hodgkin lymphoma. RF achieved the best power in these studies.

Sherif et al. (2015) [28] used the ADNI-1 dataset to test for the ability of different Bayesian network structure learning algorithms to detect causal AD SNPs and gene-SNP interactions. They tested four types of classification algorithms: na ïve Bayes (NB), tree augmented Bayes (TAB), Markov blanket (MB), and minimal augmented Markov blanket (MAMB). They demonstrated the effectiveness of using these algorithms for identifying AD causal SNPs with acceptable accuracy. The results indicated that the SNP set detected by MB-based methods has a strong association with AD and achieved better performance than other methods. Abd El Hamid et al. (2016) [29] used correlation-based feature selection (CFS) and chi-square feature selection to find the most important SNPs. The SVM classifier of different kernels has been used on ADNI-1. The results revealed that the SVM-trained model using RBF kernel had a relatively high association with AD and achieved an accuracy of 76.7%.

The good results obtained by MDR models encouraged the development of modified MDR models such as K-Nearest Neighbours MDR (KNN-MDR) [30] and two-step unified model-based MDR (UM-MDR) [31]. However, the high false positive rates, as with the current results, are still a problem. One obvious reason for this finding is multiple testing: the large number of performed tests necessitates that the significance threshold be properly adapted, which is not always easy to do [30].

5. Conclusions

The prediction of complex disease phenotypes from genotype data is an emerging research goal. SNPs and their association with AD can provide insights into the underlying mechanisms and identify SNPs that may serve as targets for therapeutic intervention. In conclusion, several significant SNPs associated with MCI and AD were identified in the APOE ε4, CD33, CR1, BIN1, EPHA1 PICALM, NM8, CLU, CD2AP and SORL1 genes.

The current study showed that RF and MDR are alternatives to other existing methods for detecting genetic interactions, with important advantages. Among the advantages of their use is that they are able to detect interactions between SNPs. Moreover, these methods are non-parametric with no assumed prior distribution, unlike many parametric statistical methods. Nevertheless, parameters (distances, number of neighbours, window definition) are available to allow flexibility in the search strategies, which could help make these methods useful. Finally, using WGS data and the top related genes or adding other modalities, such as PET, MRI, or CSF markers, may improve the prediction accuracy.

Abbreviations

AD: Alzheimer's Disease; ADNI: Alzheimer's Disease Neuroimaging Initiative; APOE: Apolipoprotein E; CART: Classification and Regression Trees; CSF: Cerebro-Spinal Fluid; EOAD: Early-Onset Alzheimer's Disease; GWAS: Genome-Wide Association Studies; IBM: International Business Machines Corporation; KNN: K-Nearest Neighbor; KNN-MDR: K-Nearest Neighbors MDR; LDA: Linear Discriminant Analysis; LOAD: Late-Onset Alzheimer's Disease; MAF: Minor Allele Frequency; MAMB: Minimal Augmented Markov Blanket; MB: Markov Blanket; MCI: Mild Cognitive Impairment; MDR: Multifactor Dimensionality Reduction; MRI: Magnetic Resonance Imaging; NB: Naïve Bayes; NC: Normal Controls; NPV: Negative Predictive Value; OR: Odds Ratio; PET: Positron Emission Tomography; PLS: Partial Least Square; PPV: Positive Predictive Value; QC: Quality Control; QDA: Quadratic Discriminant Analysis; RF: Random Forests; ROC: Receiver Operator Characteristics Curve; SNPs: Single Nucleotide Polymorphisms; SPSS: Statistical Package for the Social Sciences; SVM: Support Vector Machine; TAB: True Augmented Bayes; UM-MDR: Unified Model-based MDR; WGS: Whole Genome Sequencing.

ACKNOWLEDGEMENTS

The data used in preparation of this manuscript were obtained from the ADNI database (adni.loni.usc.edu). As such, the investigators within the ADNI study contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this paper.

REFERENCES

[1] H. Shi, O. Belbin, C. Medway, K. Brown, N. Kalsheker, M. Carrasquillo, et al., Genetic variants influencing human aging from late-onset Alzheimer's disease (LOAD) genome-wide association studies (GWAS), Neurobiol Aging 33(8) (2012) 1849.e5-18.

[2] B. Jiao, X. Liu, L. Zhou, M.H. Wang, Y. Zhou, T. Xiao, et al., Polygenic Analysis of Late-Onset Alzheimer's Disease from Mainland China, PLoS One 10(12) (2015) e0144898.

[3] L. Bertram, R.E. Tanzi, Genome-wide association studies in Alzheimer's disease, Hum Mol Genet. 18(R2) (2009) R137-45.

[4] M.I. Kamboh, F.Y. Demirci, X. Wang, R.L. Minster, M.M. Carrasquillo, V.S. Pankratz, et al., Genome-wide association study of Alzheimer's disease, Transl Psychiatry 2 (2013) e117.

[5] J.C. Lambert, S. Heath, G. Even, D. Campion, K. Sleegers, M. Hiltunen, et al., Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease, Nat Genet. 41(10) (2009) 1094-9.

[6] R. Guerreirom, A. Wojtasm, J. Brasm, M. Carrasquillom, E. Rogaevam, E. Majouniem, et al., TREM2 variants in Alzheimer's disease, N Engl J Med. 368(2) (2013) 117-27.

[7] C. Cruchaga,, C.M. Karch,, S.C. Jin, B.A. Benitez, Y. Cai, R. Guerreiro, et al., Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease, Nature 505(7484) (2014) 550-554.

[8] B. Jiao, B. Tang, X. Liu, X. Yan, L. Zhou, Y. Yang, et al., Identification of C9orf72 repeat expansions in patients with amyotrophic lateral sclerosis and frontotemporal dementia in mainland China, Neurobiol Aging 35(4) (2014) 936.e19-22.

[9] M.K. Wetzel-Smith, J. Hunkapiller, T.R. Bhangale, K. Srinivasan, J.A. Maloney, J.K. Atwal, et al., A rare mutation in UNC5C predisposes to late-onset Alzheimer's disease and increases neuronal cell death, Nat Med. 20(12) (2014) 1452-7.

[10] D. Harold, R. Abraham, P. Hollingworth, R. Sims, A. Gerrish, M.L. Hamshere, et al., Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease, Nat Genet. 41(10) (2009) 1088-93.

[11] S. Seshadri, A.L. Fitzpatrick, M.A. Ikram, A.L. DeStefano, V. Gudnason, M. Boada, et al., Genome-wide analysis of genetic loci associated with Alzheimer disease, JAMA 303(18) (2010) 1832-40.

[12] A.C. Naj, G. Jun, G.W. Beecham, L.S. Wang, B.N. Vardarajan, J. Buros, et al., Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease, Nat Genet. 43(5) (2011) 436-41.

[13] J.C. Lambert, C.A. Ibrahim-Verbaas, D. Harold, A.C. Naj, R. Sims, C. Bellenguez, et al., Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease, Nat Genet. 45(12) (2013) 1452-8.

[14] P. Hollingworth, D. Harold, R. Sims, A. Gerrish, J.C. Lambert, M.M. Carrasquillo, et al., Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease, Nat Genet. 43(5) (2011) 429-35.

[15] C.M. Karch, A.T. Jeng, P. Nowotny, J. Cady, C. Cruchaga, A.M. Goate, Expression of novel Alzheimer's disease risk genes in control and Alzheimer's disease brains, PLoS One 7(11) (2012) e50976.

[16] J.H. Moore, F.W. Asselbergs, S.M. Williams, Bioinformatics challenges for genome-wide association studies, Bioinformatics 26(4) (2010) 445-55.

[17] R.C. Petersen, P.S. Aisen, L.A. Beckett, M.C. Donohue, A.C. Gamst, D.J. Harvey, et al., Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization, Neurology 74(3) (2010) 201-9.

[18] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, E.M. Stadlan, Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease, Neurology 34(7) (1984) 939-44.

[19] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, et al., PLINK: a tool set for

whole-genome association and population-based linkage analyses, Am J Hum Genet. 81(3) (2007) 559-75.

[20] AlzGene – Field Synopsis of Genetic Association Studies in AD, Http://www.alzgene.org/ Last accessed on 2nd January 2019.

[21] J. Nettiksimmons, G. Tranah, D.S. Evans, J.S. Yokoyama, K. Yaffe, Gene-based aggregate SNP associations between candidate AD genes and cognitive decline, Age (Dordr). 38(2) (2016) 41.

[22] H.J. Cordell, Detecting gene-gene interactions that underlie human diseases, Nat Rev Genet. 10(6) (2009) 392-404.

[23] X. Chen, H. Ishwaran, Random forests for genomic data analysis, Genomics 99(6) (2012) 323-9.

[24] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, et al., Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data, Bioinformatics 19(13) (2003) 1636-43.

[25] J.W. Lee, J.B. Lee, M. Park, S.H. Song, An extensive comparison of recent classification tools applied to microarray data, Computational Statistics & Data Analysis 48(4) (2005) 869-885.

[26] M. Garcá-Magariños, I. López-de-Ullibarri, R. Cao, A. Salas, Evaluating the ability of tree-based methods and logistic regression for the detection of SNP-SNP interaction, Ann Hum Genet. 73(Pt 3) (2009) 360-9.

[27] A.M. Molinaro, N. Carriero, R. Bjornson, P. Hartge, N. Rothman, N. Chatterjee, Power of data mining methods to detect genetic associations and interactions, Hum Hered. 72(2) (2011) 85-97.

[28] F.F. Sherif, N. Zayed, M. Fakhr, Discovering Alzheimer Genetic Biomarkers Using Bayesian Networks, Adv Bioinformatics 2015 (2015) 639367.

[29] M.M. Abd El Hamid, Y.M.K. Omar, M.S. Mabrouk, Identifying genetic biomarkers associated to Alzheimer's disease using Support Vector Machine, Biomedical Engineering Conference (CIBEC), 8th Cairo International (2016).

[30] S. Abo Alchamlat, F. Farnir, KNN-MDR: a learning approach for improving interactions mapping performances in genome wide association studies, BMC Bioinformatics 18(1) (2017) 184.

[31] W. Yu, S. Lee, T. Park, A unified model based multifactor dimensionality reduction framework for detecting gene-gene interactions, Bioinformatics 32(17) (2016) i605-i610.