ISSN 1870-4069

Towards a Proposal of Personalized Medical Decision Support Systems: Analysis of Gene Expression Levels of Diabetes Mellitus, Inflammation and Oxidative Stress in Alzheimer's Disease

Sonia Lilia Mestizo Gutiérrez[1] , Nicandro Cruz Ramírez[2], Gonzalo Emiliano[2], Aranda Abreu[3]

[1] Facultad de Ciencias Químicas, Universidad Veracruzana, Xalapa, Veracruz, Mexico
[2] Centro de Investigación en Inteligencia Artificial, Universidad Veracruzana, Xalapa, Veracruz, Mexico
[3] Centro de Investigaciones Cerebrales, Universidad Veracruzana, Xalapa, Veracruz, Mexico
{smestizo,ncruz,garanda}@uv.mx

Abstract. The increased incidence of Alzheimer's disease (AD) and diabetes mellitus (DM) are emerging as major public health problems worldwide. Both sufferings share pathophysiological characteristics and have no cure. Inflammation of the central and peripheral nervous system has been shown to be the link between DM and AD. Oxidative stress is also associated with AD and DM. The increasing complexity of the problems and the continuous growth of information creates the need for the use of Decision Suport System (DSS) driven by the use of new technologies such as big data and machine learning. In this context, the objective of this work is to use decision trees and Bayesian networks as mechanisms of classification of AD gene expression levels, DM, inflammation and oxidative stress, MMSE (Mini-Mental State Examination) score and the number of neurofibrillary tangles to classify 31 individuals (9 healthy controls and 22 AD patients in three different stages of disease) that could be key in the development of AD. Our results allowed us to generate classification models of different states of AD severity, according to the MMSE and we found that the level of expression of the ADIPOQ gene could play an important role in the onset of AD. Our predictive model can contribute knowledge that could be incorporated into a personalized medical DSS in the future.

Keywords: medical decision support system, decision trees, Bayesian networks, Alzheimer disease, diabetes mellitus.

1 Introduction

Alzheimer's disease (AD) is the most common cause of dementia. It is a slowly advancing neurodegenerative disorder with cognitive impairment, progressive memory loss, and behavioral disorders. Despite major research efforts, there is still no cure, but new research is underway to determine the cause of the disease and detect changes in the

brain before the first symptoms appear. Worldwide, there are 50 million people with dementia [1]. The incidence of AD is increasing and constitutes a public health challenge in our society characterized by the increase of elderly people. There are two proteins involved in the development of AD: the beta amyloid protein (Aβ) that accumulates abnormally in the brain to form extracellular neuritic amyloid plaques and the tau protein that produces the formation of intracellular neurofibrillary tangles. Both alterations increase the levels of inflammation, oxidative stress and lead to the death of neurons. In the last 20 to 30 years, scientists have discussed which protein plays the most important role in the development of the disease. The certainty of the diagnosis of AD is approximately 85% and is only confirmed by post mortem examination. AD is multifactorial in nature and is considered a complex condition resulting from an interaction of environmental and genetic factors. The main risk factor is advanced age, however other potential risk factors have been found such as sex, diabetes mellitus, headaches, lifestyle, hypertension, obesity, dyslipidemia, metabolic syndrome, cerebrovascular disease, smoking, physical inactivity, depression and low levels of education [2]. There are reported findings from genetic studies that have pointed to APP metabolism, immune response, inflammation, lipid metabolism and intracellular trafficking/endocytosis that open the door for exploration of new pathways for genetic testing, prevention and treatment [3].

Diabetes Mellitus (DM) is a chronic disease characterized by a high concentration of glucose in the blood because the body does not produce insulin or does not use it properly. Globally, it is estimated that there are 425 million diabetics and it is estimated that by 2045 it will increase to 629 million [4] making it one of the major health challenges of this century. Several studies converge on the implication of inflammation as a key factor in the relationship of DM with AD [5, 6, 7].

The initial relationship between AD and DM was established in the Rotterdam study where it is revealed that Diabetes Mellitus type 2 (DM2) doubles the risk of patients to develop AD, while patients with Diabetes Mellitus type 1 (DM1) who receive insulin treatment quadruple the risk [8]. Several studies [9, 10, 11] propose the existence of a relationship between AD and DM and some authors have called it "type 3 diabetes" [12, 13, 14].

The existence of large volumes of biomedical data provides a great opportunity for better understanding, prediction and decision making of conditions. Microarrays are a powerful technique for the measurement of gene expression data that allow the comparison of the relative abundance of messenger RNA generated in different biological tests. The analysis of microarrays is a challenge due to its high dimensionality and complexity so machine learning techniques have been used with satisfactory results. Our work aims to use supervised learning techniques (decision trees and Bayesian networks) to classify gene expression levels of Alzheimer's disease, diabetes mellitus, inflammation and oxidative stress from a public database of 31 individuals, MMSE scores and number of NFT (neurofibrillary tangles) in order to contribute to a better understanding of AD and provide knowledge for the development of earlier and more accurate diagnosis, as well as the development of more appropriate treatments leading to future personalized treatments for incorporation into a personalized medical DSS.

In the next sections we present the state of the art, the methodology that we followed for building classifiers, the results and finally the conclusions and future work.

2 State of Art

Several works have used automatic learning techniques (neural networks, support vector machines, bagging, boosting, information gain, random forests, genetic algorithms) for the analysis of the levels of expression of AD [15, 16, 17]. Some work using Bayesian nets has also been carried out [18, 19].

Recently the use of machine learning for the classification of gene datasets has increased. In one study decision trees were used to classify a gene dataset of AD. Classification models were generated according to Mini-Mental State Examination (MMSE) scores to identify expression levels of different proteins that could determine the involvement of genes involved in various pathways of AD pathogenesis. The results showed that the MMSE score and relevance association score are the most significant attributes for gene classification. In the functional gene classification analysis, they reported that APOE, PSEN1, GRN, ACE, BCHE, PRNP, IL1A are strongly related to AD [20]. Machine learning techniques (decision trees, quantitative association and hierarchical cluster) have been used to identify potential genes for the prognosis of AD through the use of different biological sources (microarrays, PubMed, GO and PPI network). The results reported a set of significant genes (down/up) related to AD [21]. Park and colleagues formulated a new random forest-based method that allows the classification of gene-gene interaction of gene expression profiles. The proposed method was evaluated using AD data with remarkable accuracy, the result of gene-gene interaction could be used for the construction of a genetic network to explain underlying mechanisms of AD [22].

In a recent study, decision trees were used to report a genetic risk profile derived from a set of candidate genes related to cognitive performance selected a priori in order to explore the combined effect of these genes on cognitive impairment rates during the preclinical stage of AD. The results support the hypothesis that the combination of genes associated with cognitive performance makes it possible to identify groups with accelerated rates of cognitive impairment [23].

3 Methodology

The development of this study was divided into two main phases. During the first phase we made the selection of the microarrays database, the analysis of properties of the data and the preprocessing techniques were applied and in the second phase the genes of interest for our study were selected and the techniques of decision trees and Bayesian networks were applied to obtain the knowledge models that represent the patterns of behavior in the levels of genetic expression of Alzheimer's disease. Finally, we evaluated and interpreted the results obtained.

3.1 Database Selection

In this work we made use of the microarray database GDS810 obtained from the *National Center for Biotechnology* (NCBI) *Gene Expression Omnibus* (GEO) database ([HG-U133A] *Affymetrix Human Genome* U133A Array) [24]. Expression levels of 23,283 genes from 31 individuals were obtained from the CA1 region of the hippocampus and correspond to 9 control patients, 7 with incipient Alzheimer's disease, 8 with moderate Alzheimer's disease and 7 with severe Alzheimer's disease. The dataset includes MMSE scores and number of NFT [25].

3.2 Property Analysis and Preprocessing

For the analysis of the properties of the data we proceeded to explore, clean and adjust the data. We removed clones and pseudogenes from the database. Regarding the preprocessing of Affymetrix microarray data, RMA (Robust Multi-Array Average), GCRMA (GeneChip Robust Multi-Array Average), MAS5 (MicroArray Suite 5.0) and Expresso (Gautier, et al., 2004) were used in the normalization phase using the Affy R [26] Bioconductor package.

3.3 Gene Selection

Our interest focused on the expression values of genes related to AD: APP, APOE, BACE1, NCSTN, PSEN1, PSEN2, MAPT and INPP5D, MEF2C, HLA-DRB5/DRB1, NME8, ZCWPW1, PTK2B, CELF1, SORL1, FERMT2, SLC24A4, CASS4 [27], as well as DM genes: HLA-DQB1, TCF7L2, ACE, PPARG, HLA-DQA1, APOE, ADIPOQ and inflammation: TNF, IL6, IL1B, IL10, TLR4, IL1RN, LTA, IL1A, CD14, PTGS2, CRP reported by Genotator [28], and oxidative stress-related genes: ANXA6, ARAF, CBX7, DHX16, EBP, FGF13, HIF1A, TNIP1 or NAF1, NDUFS1, NFE2, POLD1, RAB15, SGK2, SMAD5, STAT5B, UBA7, WNT2B [29]. We added MMSE score and number of NFT.

3.4 Decision tree and Bayesian network

The performance of our classifiers is based on precision (number of correct classifications divided by the size of the test set), sensitivity (number of AD patients correctly identified) and specificity (correct identification of patients without AD).

For the analysis of gene expression levels using decision trees [30, 31] and Bayesian networks [32,33] the clones and pseudogenes were removed from the database. The data were analyzed using a WEKA software utilizing decision tree J48 classification algorithm and Bayesian network (Naive Bayes algorithm) with 10-fold cross-validation. We used a decision tree because they provide models that are easy to interpret and understand thanks to their ability to select and classify attributes according to their relevance [34].

In the generation of the Bayesian network we use the CAIM (Class-attribute Interdependence Maximization) [35] and MDL (Minimum Description Length) [36] methods provided by WEKA (Waikato Environment for Knowledge Analysis) [37, 38].

4 Results

In Fig.1, the decision tree obtained from the levels of genetic expression of AD, DM, inflammation, oxidative stress, MMSE score and the number of neurofibrillary tangles is presented with an accuracy of 87.09%, sensitivity of 90.90% and specificity of 77.77%.

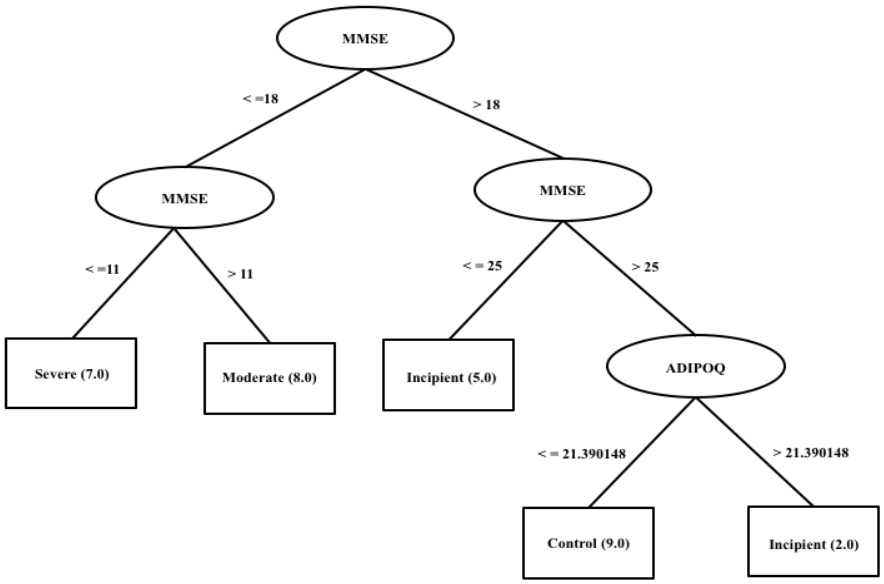

Fig. 1. Decision tree of the main genes related to AD, DM, inflammation, oxidative stress, MMSE and number of neurofibrillary tangles.

As we can see, the most informative variable is the MMSE. The J48 algorithm provides MMSE score cut-off values for each stage of the disease: normal > 25, incipient 19-25, moderate 18-12 and severe < 12, which are similar to those used in clinical practice to classify an individual's cognitive status. The importance of our model is that it allows us to identify individuals at an early stage of AD when the MMSE score is above 25 points and the level of expression of ADIPOQ (Adiponectin, C1Q and Collagen Domain Containing) is greater than 21.390148, an individual is classified as AD incipient. The ADIPOQ gene is only expressed in adipose tissue. Obesity has been reported to be a significant risk factor for the development of metabolic syndrome and other degenerative diseases. One study found that serum adiponectin level correlated positively ($r=0.683$, $P<0.001$) with MMSE score in patients with AD [39]. Our work corroborates the results obtained in this study, however it is more explanatory since it allows us to identify AD at an early stage. Another advantage is that our model is transparent and understandable for human experts who are not machine learning specialists.

The model generated by Naive Bayes with the discretization technique CAIM correctly classified 29 of the 31 samples: 7/7 of severe AD, 8/8 of moderate AD, 6/7 of incipient AD and 8/9 of healthy control. Model accuracy was 93.54%, sensitivity 95.45% and specificity 88.88%. In the model obtained by Naive Bayes with the MDL discretization technique, an accuracy of 90.32%, sensitivity of 90.90% and specificity of 88.88% were obtained. From our results, the best classification model was obtained using Naive Bayes with the CAIM discretization technique.

5 Conclusions

In the development of this work we evaluated classifiers with decision tree techniques and Bayesian networks of AD gene expression levels, DM, inflammation and oxidative stress, MMSE score and the number of neurofibrillary tangles. In the decision tree, the MMSE score was the most important attribute, however we found that the level of ADIPOQ expression can play a crucial role in distinguishing between a normal cognitive state and incipient EA when the MMSE score is considered normal. In summary, we successfully modeled different states of AD with accuracies of 87.09% (decision tree), 93.54% (Naive Bayes with CAIM) and 90.32% (Naive Bayes with MDL) and showed that the level of expression of ADIPOQ has potential to be considered in the early diagnosis of AD so our results could contribute with knowledge for a future implementation of a personalized medical DSS. The development of this work demonstrates that the use of machine learning techniques, provide favorable results for the early diagnosis of AD. The models obtained can become the knowledge base of a personalized medical DSS. A limitation of our is the sample size, so as future work is proposed the use of artificial instance generators to improve performance.

Acknowledgements. This work was supported by support for the reincorporation of former fellow PROMEP DSA/103.5/16/10415/EXB-552, 47856 project.

References

1. Alzheimer's Disease International: The state of the art of dementia research: New frontiers Homepage, https://www.alz.co.uk/research/WorldAlzheimerReport2018.pdf?2, last accessed 2018/09/15.
2. Kivipelto, M., Mangialasche, F., Ngandu, T.: Lifestyle interventions to prevent cognitive impairment, dementia and Alzheimer disease. Nature reviews. Neurology, doi: 10.1038/s41582-018-0070-3 (2018)
3. Reitz, C., Mayeux, R.: Alzheimer disease: epidemiology, diagnostic criteria, risk factors and biomarkers. Biochemical pharmacology 88(4):640–651 (2014)
4. International Diabetes Federation, Homepage, https://www.idf.org/, last accesed 2018/09/15.
5. Lue, L., Andrade, C., Sabbagh, M., Walker, D.: Is there inflammatory synergy in type II diabetes mellitus and Alzheimer's disease? J Alzheimers Dis. 1:1–9 (2012)
6. De Felice F., Ferreira S.: Inflammation, defective insulin signaling, and mitochondrial dysfunction as common molecular denominators connecting type 2 diabetes to Alzheimer´s disease. Diabetes 63:2262–72 (2014)

7. Jiang, C., Li, G., Huang, P., Liu, Z., Zhao, B.: The Gut Microbiota and Alzheimer's Disease. J Alzheimers Dis. 58(1):1–15 (2017)
8. Ott, A., Stolk, R., Hofman, A., van, H., Grobbee, D., Breteler, M.: Association of diabetes mellitus and dementia: The Rotterdam Study. Diabetologia 39:1392–1397 (1996)
9. Pasquier, F., Boulogne, A., Leys, D., Fontaine, P.: Diabetes mellitus and dementia. Diabetes & Metabolism 5(32):403–414 (2006)
10. Arab, L., Sadeghi, R., Walker, D., Lue, L., Sabbagh, M.: Consequences of Aberrant Insulin Regulation in the Brain: Can Treating Diabetes Be Effective for Alzheimer's Disease. Neuropharmacol 9(4):693–705 (2011)
11. Adeghate, E., Donáth, T., Adem, A.: Alzheimer disease and diabetes mellitus: do they have anything in common? Curr Alzheimer Res. 10(6) (2013)
12. Steen, E., Terry, B., Rivera, E., Cannon, J., Neely, T., Tavares, R., Xu, X., Wands, J., de la Monte, S.: Impaired insulin and insulin-like growth factor expression and signaling mechanisms in Alzheimer's disease—is this type 3 diabetes? J Alzheimers Dis. 7(1):63–80 (2005)
13. Kroner, Z.: The relationship between Alzheimer's disease and diabetes: Type 3 diabetes? Altern Med Rev. 14(4):373–9 (2009)
14. de la Monte, S.: Wands, J. Alzheimer's disease is type 3 diabetes-evidence reviewed. J Diabetes Sci Technol. 2(6):1101–13 (2008)
15. Walker, P., Smith, B., Liu, Q., Famili, A., Valdés, J., Liu, Z., Lach, B.: Data mining of gene expression changes in Alzheimer brain. Artif. Intell. Med. 31(2):137–154 (2004)
16. Scheubert, L., Lustrek, M., Schmidt, R., Repsilber, D., Fuellen, G.: Tissue-based Alzheimer gene expression markers comparison of multiple machine learning approaches and investigation of redundancy in small biomarker sets. BMC Bioinformatics 13(266) (2012)
17. Jain, M., Dua, P., Dua, S., Lukiw, W.J.: Data Adaptive Rule-based Classification System for Alzheimer Classification. J Comput Sci Syst Biol 6:291–297 (2013)
18. Armañanzas, R., Larrañaga, P., Bielza, C. Ensemble transcript interaction networks: a case study on Alzheimer's disease. Comput Methods Programs Biomed. 108(1):442–50 (2012)
19. Zhang, B., Gaiteri, C., Bodea, L.G., Wang, Z., McElwee J., Podtelezhnikov, Emilsson, V.: Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. Cell. 153(3):707–20 (2013)
20. Kumar, A., Singh, T.R.: A New Decision Tree to Solve the Puzzle of Alzheimer's Disease Pathogenesis Through Standard Diagnosis Scoring System. Interdisciplinary Sciences: Computational Life Sciences 9(1):107–115 (2017)
21. Martínez-Ballesteros, M., García-Heredia, J. M., Nepomuceno-Chamorro, I. A., Riquelme-Santos, J. C.: Machine learning techniques to discover genes with potential prognosis role in Alzheimer's disease using different biological sources. Information Fusion 36:114–129 (2017)
22. Park. C., Kim, J., Kim, J., Park, S.: Machine learning-based identification of genetic interactions from heterogeneous gene expression profiles. PLoS ONE 13(7):e0201056 (2018)
23. Porter, T., Villemagne, V. L., Savage, G., Milicic, L., Ying Lim, Y., Maruff, P., Laws, S.M.: Cognitive gene risk profile for the prediction of cognitive decline in presymptomatic Alzheimer's disease. Personalized Medicine in Psychiatry 7(8):14–20 (2018)
24. Gene Expression Omnibus, http://www. ncbi.nlm.nih.gov/geo, last accesed 2018/07/21
25. Blalock, E.M., Geddes, J.W., Chen, K.C., Porter, N.M., Markesbery, W.R., Landfield, P. W.: Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proc. Nat. Acad. Sci. U.S. A. 101:2173–2178 (2004)
26. Gautier, L., Cope, L., Bolstad, B.M., Irizarry, R.A. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20:307–315 (2004)

27. Lambert, J., Ibrahim-Verbaas, C., Harold, D., Naj, A., Sims, R., Bellenguez, C., DeStefano, A., Amouyel, P.: Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nature Genetics 45:1452–8 (2013)
28. Wall, D., Pivovarov, R., Tong, M., Jung, J., Fusaro, V., DeLuca, T., Tonellato, P.: Genotator: a disease-agnostic tool for genetic annotation of disease. BMC medical genomics 3:50 (2010)
29. Walton, N., Shin, R., Tajinda, K., Heusner, C., Kogan, J., Miyake, S., Chen Q., Tamura, K., Matsumoto, M.: Adult Neurogenesis Transiently Generates Oxidative Stress. PLoS ONE 7(4) (2012)
30. Quinlan, J.: Induction of decision trees. Machine learning 1:81–106 (1986)
31. Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.H., Steinbach, M., Hand, D.J., Steinberg, D.: Top 10 algorithms in data mining. Knowl Inf Syst 14:1–37 (2008)
32. Friedman, N., Geiger D., Goldszmidt, M.: Bayesian networks classifiers. Machine Learning 29:131–163 (1997)
33. Needham, C.J., Bradford, J.R., Bulpitt, A.J., Westhead, D.R.: A primer on learning in Bayesian networks for computational biology. PLoS Comput Biol 3(8) (2007)
34. Geurts, P., Irrthum, A.L.: Wehenkel, Supervised learning with decision tree-based methods in computational and systems biology. Mol. Biosyst. 5(12):1593–1605 (2009)
35. Kurgan, L.A., Cios, K.J.: CAIM Discretization Algorithm. IEEE Transactions on knowledge and data engineering 16:145-153 (2004)
36. Fayyad, U., Irani, K.: Multi-interval discretization of continuous-valued attributes for classification learning. Thirteen. Int. Jt. Conf. Artificial Intell., vol. II, pp. 1022–1027 (1993)
37. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA Data Mining Software: an update. SIGKDD Explorations 11(1):10–18 (2009)
38. Waikato Environment for Knowledge Analysis (WEKA), http://www.cs.waikato.ac.nz/ml/weka/, last accesed 2018/07/10
39. Li, W., Tian, Y., Deng, Y.Y., Feng, X.L., Wang, Y., Feng, H., Hou, D.R.: Correlation between serum adiponectin level and cognitive function in patients with Alzheimer's disease. Nan Fang Yi Ke Da Xue Xue Bao 37(4):542–545 (2016)