Journal of Computer Allied Intelligence
ISSN: 2584-2676
Vol.02 (01), Feb 2024, pp.1-7

Research Article

Machine Learning in Predicting Alzheimer's Disease: Exploring Applications and Advancements

Rekha Gangula[1,*] and Dayakar Thalla[2]

[1] Department of CSE(DS) Vaagdevi Engineering College, Bollikunta, Warangal, Telangana,506005, India.
[2]Assistant professor, Department Of CSE, Bollikunta, Warangal, Telangana,506005, India.
[*]Corresponding Author: Dr.Rekha Gangula. Email: gangularekha@gmail.com

Received: 25/01/2024; Accepted: 20/02/2024
DOI: https://doi.org/10.69996/jcai.2024001

Abstract: Alzheimer's disease (AD), a predominant form of dementia that accounts for 60 to 70 percent of cases in the elderly population. AD significantly affects daily functioning, memory, cognition, and behaviour, presenting a substantial global health challenge with approximately 50 million dementia cases worldwide and an annual incidence of 10 million new cases. Using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort, we conduct a systematic analysis of several machine learning models to predict genetic variance linked to Late-Onset Alzheimer's Disease (LOAD). Our experimental results demonstrate that the most effective models achieve an impressive 72 percent area under the Receiver Operating Characteristic (ROC) curve in the evaluation of LOAD genetic risk. This highlights the promise of machine learning models as valuable tools for assessing the genetic susceptibility to LOAD. Furthermore, our exploration into the strategic selection of learning models unveils the potential for identifying novel genetic markers linked to the disease. This improves our ability to anticipate outcomes and advances our understanding of the fundamental processes behind Alzheimer's disease. The findings presented herein contribute to the evolving field of precision medicine by offering insights into the application of machine learning in understanding and predicting the genetic factors associated with LOAD.

Keywords: Machine learning; alzheimer's disease; late-onset alzheimer's disease; alzheimer’s disease neuroimaging initiative; ROC.

1 Introduction

To date, analyses have been conducted mainly by specialists, MRIs, post-emission tomography, functional MRIs, and diphusion tensor imaging are some examples of which requiring a high level of expertise. MRIs need a high level of imagery. Alzheimer's disease is the most prevalent type of dementia that affects those over 65. It is characterised by a progressive decline in cognitive and memory abilities. A timely and early diagnosis of AD and its prodromal stage is required; is critical to slow the progression of dementia (MCI). A accurate diagnosis is possible in this respect from brain imagery, and a rigorous diagnostic system assisted by examining neuroimaging information may provide a more detailed and reliable approach and could potentially improve precision in the diagnosis. Traditional research techniques for neuroimaging biomarkers are focused on mass univariant statistics to evaluate neuropsychiatric conditions assuming various regions function independently. However, because of our current knowledge of brain functioning, this statement is not acceptable.

Adipose Disease (AD) was a neurodegenerative disorder which progressively damages the function of the brain. It is typical to lose cognitive abilities like speech, manners, memory, and thought processes. The illness leads to dementia and ultimately death. AD accounts for 60 to 80 percent of dementia cases in those 65 years of age and older [1]. Age is not the sole risk factor for AD; it has been demonstrated that some hereditary factors increase the chance of Early Onset AD (EOAD) at a young age (<60). Age is AD's sole risk factor. Other than the age variations, EOAD is clinically viewed quite much as late-coming AD (LOAD) is seen and in certain clinical and pathological aspects many aspects of the disorder correlate again with the standard. The genetic variants in APP, PSEN1 and PSEN2 are characterised by maternal heritage EOAD, which are associated with amyloids, but account for only 5% of total AD.

Recently, Machine Learning (ML) techniques that take account of interrelationships among areas are a computer-assisted research technology that is attractive and important and have been used widely for automated diagnosis with analysis of neuropsychiatric disorders. Though several methods for the automatic prediction of neurological disorders have been used, two main investigation directions include the vector-based and profound-learning (DL) model diagnostic support machine (SVM). Comprehensive studies have been published in this regard concerning medical imaging, using machine learning techniques. SVM-based, automated neuropsychiatric diagnostic models7–9 often employ hand-crafted features because they cannot remove adaptive functionality [9]. The functional connectivity and patterns (FC) indicative of brain region correlation is a common feature of current SVM-based diagnostic models. Individual FC patterns are derived for couples of segmented brain patches marked by an automated anatomic labelling. While efficient, SVM was criticised for the fact that it lacks raw data and for the expert use of design methods for extracting information.

2 Related Works

2.1 Using machine learning to diagnose Alzheimer's disease through neuroimaging

AD was the primary cause of dementia. Though life expectancy is higher globally, the number of instances in low- and middle-income countries has recently climbed quickly, particularly in developing nations with ageing populations. In 2015 more than 46 million people had dementia worldwide, and it is estimated that this figure will nearly double every 20 years. Therefore, several countries and international organisations have prioritised studies on AD.

The processes underlying the disease begin long before the first clinical signs become apparent. The preclinical or presymptomatic level, in the absence of any symptoms, when pathological changes occur (Dubois et al., 2016; Sperling et al., 2011). And there are mild cognitive impairments, often recollective moments, but they are not discouraged. This step is known as mild cognitive impairment (Dubois and Albert, 2004; Albert et al., 2011). Next, along with the worsening memory problems there begins to be an issue with the vocabulary, management and motor functions, leaving the patient unable to carry out regular tasks. This process is referred to as dementia.

AD is a complex pathology that coexists with numerous. The first was accumulation in extracellular space of amyloid beta (Aβ) proteins that cause amyloid plaques in the brain. Up to 20 years before the diagnosis this accumulation can begin. The development of neurofibrillary tangles of tau proteins that interact with one another in the neuron and lead to death is another phenomenon observed. This results in the killing of neurons, which normally follow a certain pattern, and subsequent functional loss associated with the affected areas, resulting in brain atrophy. Although some types of family exist, AD is largely a sporadic illness. The most common genetic factor, in sporadic cases, is the presence of allel 4 ApoE4 (Mahley, Weisgraber, and Huang, 2006).

2.2 Interest of ML for identification of AD

A significant task is an early and precise detection of AD patients. An attractive way to achieve this is to establish ML approaches which can use various data types to early detection of patients with AD. As mentioned earlier, a subject passes through various stages before AD is created. Different ML problems can be formulated depending on the stage being considered.

A first question is to separate AD dementia patients from normal cognitive subjects. AD dementia (CN). This is a classification task which is hereinafter referred to as AD vs CN. This task had been dedicated to previous ML studies in AD diagnosis. In general, good classification performance in most studies with accuracy between 85% and 95% was achieved for this mission (Rathore et al., 2017). In brain imaging and cognitive testing, variations between a CN subject and an AD patient can easily be identified so that functional applications are minimal. However, they may be helpful in improving diagnostic trust.

3. Materials and Methods

3.1 Data set

Data from the Alzheimers Disease NeuroimagingInitiative (ADNI) data base have been collected for the preparation of this paper (adni.loni.usc.edu). The partnership between the public and commercial sectors Michael W. Winner, M.D., the principal investigator, developed ADNI. Testing the viability of combining clinical and neuropsychological evaluations (AD), post-emission tomography (PET) and other biological indicators, and the development of mild cognitive impairment (MCI) was the main objective of ADNI.

Individuals with the ADNI dataset can also have several different types of diagnosis, including: normal (CN) cognitive disability, precocious and late (eMCI) and Alzheimer's disease (AD). ADNI 1 consisted of two hundred CN, four hundred MCI and 200 AD males. This has been expanded by ADNI GO to 200 people with eMCIs and 500 ADNI rollers 1. The rollovers were finally incorporated by ADNI 2 by 150 CN, 150 eMCI, 150 lMCI and another 200 AD.

We have the complete genome sequence samples available for 812 people in the ADNI database (WGS). These WGS were sampled on a 2.5 M Illumina Chipset with one nuclear polymorphism of 2379,855 These WGS were sampled (SNPs). Owing to the peculiar existence of the MCI and the uncertainty that the patient may be moved to Alzheimer's disease grades are purely binary and only those with CN or AD diagnoses are the examples for binary classification.

The outcomes of the Alzheimer's International Genomics Project have also aided in the classification of the traits, guided algorithm learning, and artificially increased data set worth. The IGAP is a broad two-stage research focussing on the studies in individuals with an ancestry in Europe of genome-wide associations (GWAS). In Stage 1, four of the previously published GWAS data sets (EADI-European Alzheimer's Initiative, ADGC Cohorts for Heart and Ageing Research, in the CHA Genomic Epidemic Consortium) comprised 17,008 Alzheimer's cases and 37,154 monitors. IGAP used data on 7,055,881 single nuclear polymorphism (SNPs) to analyse these datasets. Step 2 involved génotyping 11,632 SNPs and examining their interactions in a separate group of 11,312 controls and 8,572 Alzheimer's patients. The integration of the results from the first and second stages was carried out last.

3.2 Tools and Software

The Program for the reading and convert of the WGS Variant Call Format data was also used for the quality control pipeline as well as in a lightweight Binary Pedigree Files (BED) PLINK format. Tensorflow[2] on the GPU backend and Keras[2] for the deep learning framework have been used in the implementation of the Python 3.5 code to access the binary pedigree files of Python from the Python library.

3.3 Validation Methodology

The principal analysis variable is the area of the receiver operations, the field of curve (ROCAUC). 5-Fold Cross-Validations are used for statistically significant findings in the data collection, and they are not overfitting for validation. This is achieved in both the direct and complex genetic disorder models. The testing of the IGAP train samples and the non-IGAP test samples is carried out without cross-validation. directly. For the process of data enhancement, we perform training on the various subgroups and then explicitly assess 138 unrelated ADNI individuals, or in the entire subgroup of 471 individuals.

4. Results and Discussion

In the first investigation, several significant snps were directly analysed using the ADNI data set. In order of significance for a number of SNP, the previously collected SNPs were used as inputs in the clump file. The ADNI data set is divided by 5 times Cross Validation in tests and training. Figure 1 shows the value of the resulting value for a machine learning system obtained as the maximum ROC AUC score of 0.66 with around 20 SNPs using the random forest method.

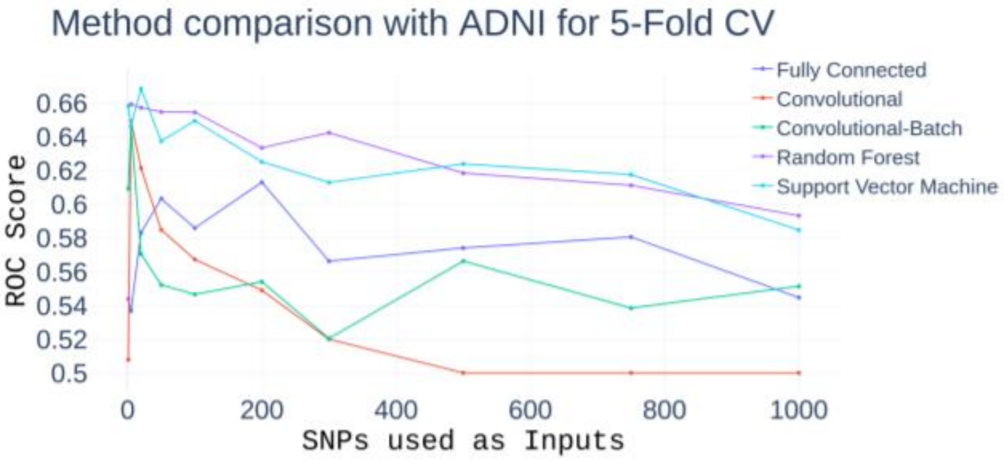

Figure 1: Comparison of the process with the Complete ADNI Dataset ROC AUC Output Analysis

The next step was to attempt to simulate a scenario and use a larger data set to determine how the methods function. Figure 3 illustrates how using more data samples results in a substantially more accurate categorization, as demonstrated by the simulation. More intriguingly, the increase in sample sizes makes it more advantageous to employ more SNPs as inputs. Thus it is advantageous to use the highest ranking SNPs for limited data sets but the findings are optimised by integrating more SNPs into large data sets and a more accurate classification is achievable. Figures 4,5 and 6 display this. Figure 7 and 8 can be evaluated to assess the major improvement in output by using two separate subsets (500 and 10,000 respectively).

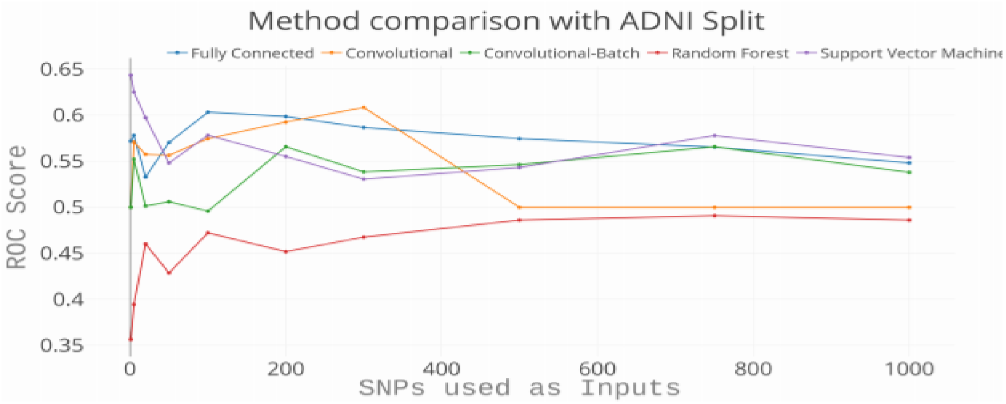

Figure 2: Method compared with Split ADNI data set Output analysis of the ROC AUC scores

Detailed analysis of ML methods is illustrated in Figures 3, 4 and 5. The balance of the error, the ROCAUC, the precision and specificity and sensitivity are shown as bar graphs for both classifiers and filter combinations. Those plots show that the lowest output is the motor support vector machine (SVM) with the lowest maximum redundancy (mRMR) filter. However, when utilising the Methods Collection, the LASSO approach yielded the greatest results among all ML methods, with a ROC AUC of 0.719. In addition, the ML Methodology has been improved.

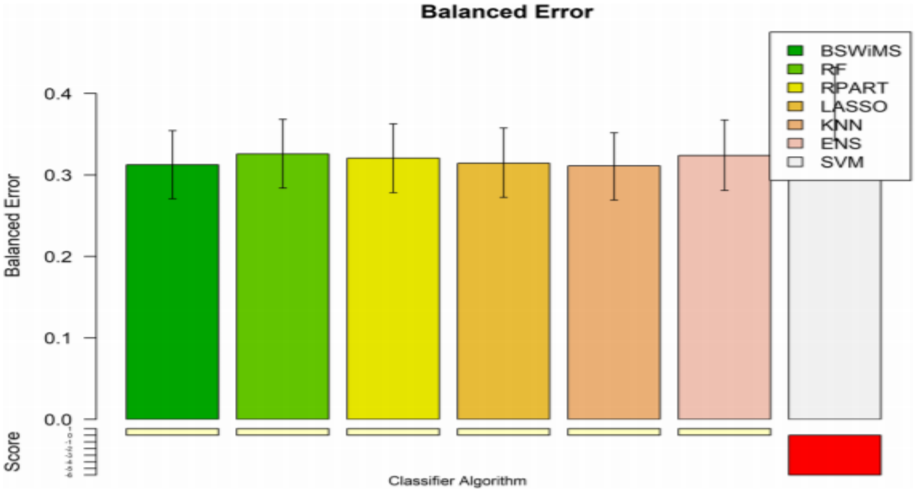

Figure 3: Balancing FRESA.CAD Classification Error Comparison of the balanced error obtained with the ADNI Discovery Dataset for Cross-Validation and use of the top 2500 SNPs as feedback by the various classification techniques of the FRESA.CAD Benchmarking.

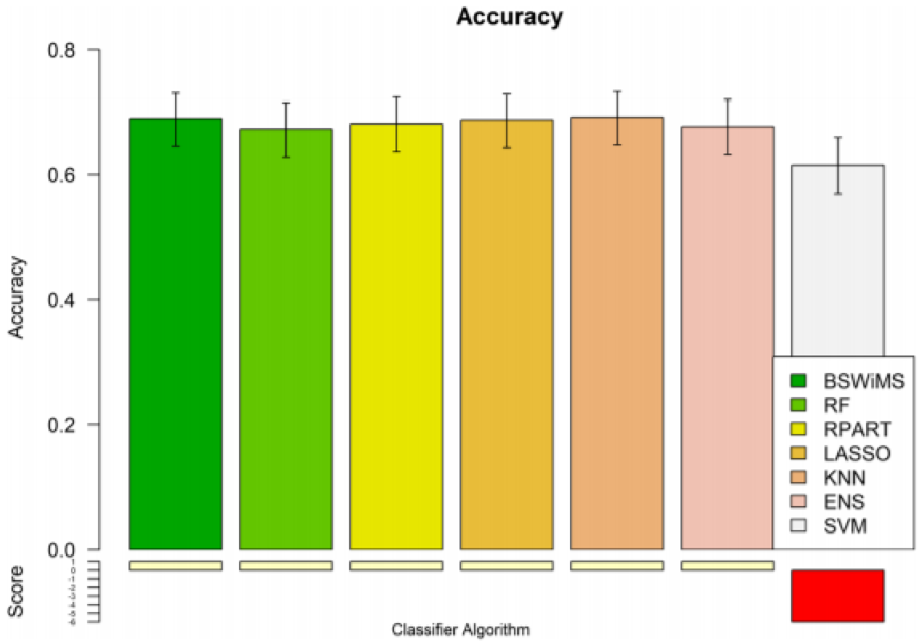

Figure 4: Comparison of the accuracy attained with the top 2 500 SNPs as an input and ADNI-discovery data for cross-validation using FRESA.CAD benchmarking techniques Comparison of classifying systems' accuracy. Comparing and contrasting Comparing and contrasting

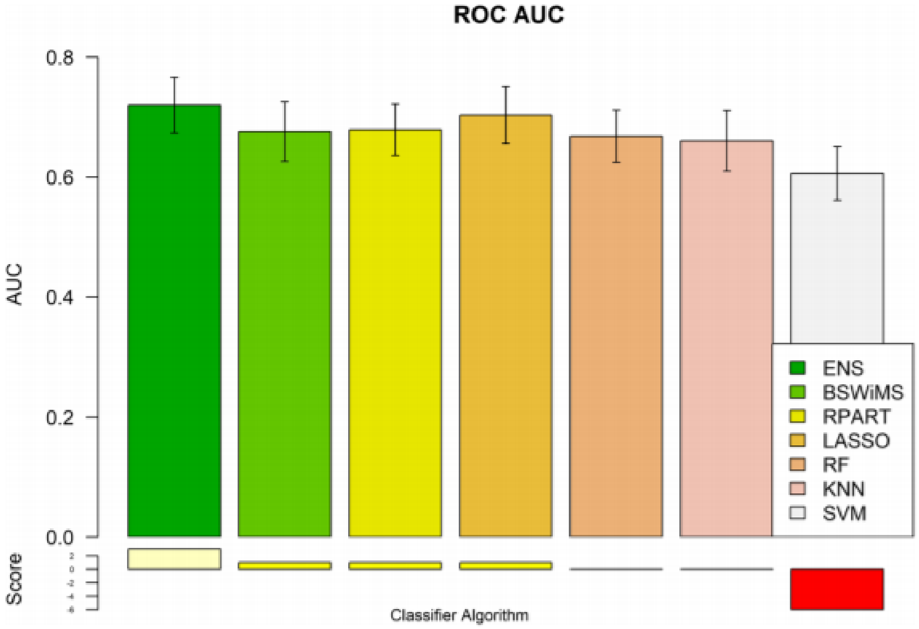

Figure 5: Comparison of the ROC AUC Score obtained by the different FRESA.CAO classification methods with the cross-validation ADNI-detection dataset and with the top 2500 SNPs as input. ROC AUC FRESA.CAD Benchmarking

4. Conclusion

We draw a range of conclusions in this paper. Final learning is possible even with the endemic neuroimaging issues, when training data are small and the sample dimensionality very high without the use of handcrafted features. In addition, we have developed an explanatory visualization map to block expert information intervention; this protocol is to serve as a tool to distinguish biomarkers relevant to the condition and other neuropsychiatric disorders. The test results on the ADNI data showed that our model provides good performance and efficiency compared to modern models. But the paper has many limitations: First, since there is only a small number of subjects for training and testing to facilitate end-to-end learning, there is only a modest increase in output relative to previous traditional models. We assume, however, that this methodology demonstrates a greater capacity for the generalisation of learning models than handmade methods where a greater amount of data will be available in the future. This study proposes that machine learning approaches based on a large number of genetic variants should be used to predict the late onset of Alzheimer's disease. Experiments' findings demonstrate that the suggested model is committed to providing insightful forecasts for LOAD clinical diagnosis.

Acknowledgement: Not Applicable.
Funding Statement*:* The author(s) received no specific funding for this study.
Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

[1] A. Mart´ın, A. Ashish, B. Paul Barham, B.Eugene, C. Zhifeng, C. Craig *et al.*, “TensorFlow: Largescale machine learning on heterogeneous systems,” 2015. Software available from tensorflow.org.

[2] A. Athanasios, D.Vasileios N. Mantzavinos, H. Greig, A. Mohammad *et al*., “Bayesian model for the prediction and early diagnosis of alzheimer’s disease,” *Frontiers in Aging Neuroscience*, vol.9, no.77, 2017.

[3] M. Panpalli Ates, Y. Karaman, S. Guntekin, and M.A. Ergun, “Analysis of genetics and risk factors of alzheimers disease*,” Neuroscience*, vol. 325, pp 124 – 131, 2016.

[4] Ch. Franc, Keras. https://keras.io, 2015.

[5] D. Carole and M. Maria Glymour, “Prediction to prevention in alzheimer’s disease and dementia,” *The Lancet Neurology*, vol.17, no. 5, pp. 388– 389, 2018.

[6] E. Evan, J. Eichler, F. Jonathan , G. Greg, K. Augustine M. Suzanne, *et al.*, “Missing heritability and strategies for finding the underlying causes of complex disease,” *Nature Reviews Genetics*, vol.11, no. 446, pp. 06, 2010.

[7] E. Evan, F. Eichler, F. Jonathan, G. Greg, K. Augustine M. Suzanne *et al*., “Missing heritability and strategies for finding the underlying causes of complex disease,” *Nature Reviews Genetics*, vol.11, no. 446, pp.06 ,2010.

[8] P. Valentina, S. Maryam, P. Richard W. Julie H. John *et al*., “Polygenic score prediction captures nearly all common genetic risk for alzheimer’s disease,” *Neurobiology of Aging*, vol, no.49, pp.214.e7–214.e11, 2017.

[9] G. Rekha, Ch. Sudha, R. Shashi and N. Muddam., “A Conceptual Framework For Understanding The Role Of Machine Learning In Artificial Intelligence,” http://sersc.org/journals/index.php/IJAST/article/view/6531

[10] E. Andre, K. Brett, A. Roberto, A. Novoa, M. Susan *et al*., “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no.115, pp.01-03. 2017.