International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Supervised Machine Learning Algorithms for Early Diagnosis of Alzheimer's Disease

Prasad B S, Akhilaa

Abstract: ***Alzheimer's is a neurodegenerative disease which can eventually leads to dementia. Mostly occurring in elderly people over the age of 65, it is hard to detect and diagnose correctly. Most common symptoms include memory loss and slow deterioration of cognitive functions. Given that these symptoms are seen often in old people, this hinders the detection of Alzheimer's disease (AD). Alzheimer's is currently incurable, but detection of the disease during its early stage is often beneficial to the patient, since there are treatments which can considerably improve the quality of life of the patient. However this can only be done if the patient has been diagnosed at a stage before any permanent brain damage has been done. Most of the current methods for detecting and diagnosing AD are not good enough. It is the need of the hour to develop better and early diagnostic tools. With the improvements in the field of machine learning, we now have the tools needed to drastically improve detection of Alzheimer's. We examine various machine learning methods and algorithms to find a method which can boost the chances of detecting the disease. We will use the following algorithms: Decision Tree, SVM, Random Forest and Adaboost. The dataset being used is the longitudinal MRI data available included in the OASIS dataset. We will use the aforementioned algorithms on the dataset and compare the accuracies achieved to find an optimal.***

Keywords: ***Adaboost, Alzheimer's disease, Decision Tree, Machine learning, Random forest, Supervised learning.***

I. INTRODUCTION

Alzheimer's is a neurodegenerative disease which can severely affect cognitive functions and social skills of the affected person. Alzheimer's usually occurs in older people, usually around the age of 65 and above. Since some of the symptoms during the early stages of the disease also occur in people with age, many cases of Alzheimer's are overlooked and incorrectly diagnosed. This is extremely problematic for those affected, since if not treated early Alzheimer's can cause significant brain damage. However, even during later stages, there are cases of wrong diagnosis.

Manuscript published on 30 September 2019
* Correspondence Author
Prasad B S,* Department of ISE, CMR Institute of Technology, Bangalore, India. Email: prasad.t@cmrit.ac.in
Akhilaa, Department of ISE, CMR Institute of Technology, Bangalore, India. Email: prasad.t@cmrit.ac.in

Even though Alzheimer's currently has no cure, it can be treated to an extent. However such treatment is only effective if it is administered before any significant brain damage has occurred. Hence there is an urgent need for better diagnostic tools which can accurately detect the probability of a patient prone to Alzheimer.

We try to find a method of diagnosing Alzheimer's disease in patients using machine learning technologies. We will be applying four classification algorithms and will look in to their result to find the most suitable one. The decision tree, Support Vector Matrix (SVM), Random Forest and Adaboost. Using brain MRI data, we will use each of these algorithms in an effort to classify test subjects based on whether they have Alzheimer's or not. The accuracy of these algorithms will then be compared, along with their pros and cons in order to choose a method which can be used as a diagnosis tool with high accuracy.

II. RELATED WORK

Previous attempts have been made to make use of machine learning to detect and diagnose Alzheimer's, with most of the attempts using MRI data. Some of the work will be mentioned in this section.

In a published paper titled, "Twin SVM based Classification of Alzheimer's Disease Using Complex Dual-Tree Wavelet Principal Coefficients and LDA" [3], the authors propose a method of detecting those with Alzheimer's using a combination of multiple methods. The method proposed worked using trans-axial images of the brain MRI, using images received from the Alzheimer's disease Neuroimaging Initiative (ADNI).

Another paper, "Detections of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrian and machine learning" [4], proposes to use a CAD system based on eigenbrain and machine learning to effectively diagnose Alzheimer's disease, it also effectively identifies the brain regions which are affected by the disease. This method generates an eigenbrain for each subject using the 3D volumetric data. It then uses SVM to make a prediction of the subjects affected by Alzheimer's disease.

While we have described only two papers, there are quite a few other papers published with the same goal. The two above papers work directly with raw MRI data and images. In this paper we work using certain biomarkers obtained from MRI images and data, along with a few socioeconomic factors in relation to the subjects as well.

III. DATA SET

The dataset we will be using is the MRI data taken from Open Access Series of Imaging Studies (OASIS). It is a openly available dataset which can be used to train a machine learning model to diagnose patients for Alzheimer's disease. Most well-known machine learning methods in this field look to use raw MRI data and images. However, we will be looking for only certain biomarkers along with socioeconomic markers. The OASIS database consists of both longitudinal and cross-sectional MRI data. For the purposes of this paper, we will be using only the longitudinal MRI data.

The dataset consists of the data of 150 subjects between the ages of 60-96. 64 subjects were already classified as demented, and 72 as non-demented. 14 subjects are part of the converted category, which describes those who were initially classified as non-demented before being found as having dementia at a later time.

The datasets consists of 15 columns. Some of the more important ones are described here:

- EDUC: The years of education the subject has undergone.
- SES: The socio-economic status of the subject. Ranges from 1-5, with 5 being the highest.
- MMSE: Mini Mental State Examination which is utilized to measure the cognitive ability of the patients. The scores calculated based on tests of memory, language, spatial skills, etc.
- CDR: Clinical Dementia Rating. Consists of scoring based on various problem based and social problems. Scales is from 0-3.
- eTIV: Estimated Total Intracranial Volume. Used for volumetric analysis of the brain.
- nWBV: Normalize Whole Brain Volume. Used for diagnosing neuropsychiatric disorders.
- ASF: Atlas Scaling Factor. Normalisation used for head size correction

IV. DATA VISUALIZATION AND PRE PROCESSING

We attempted to analyze the data present in the database in order to try and find the relationships between the data with regard to Alzheimer's and dementia. The idea was to plot the data graphically in a bid to observe any possible trends present in the data. The data showed that dementia tends to occur more often in males than in females.

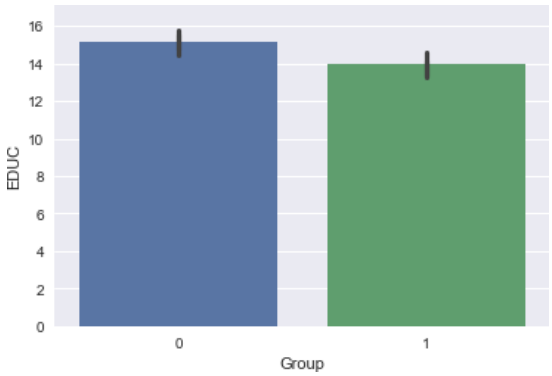

Fig 1. Plot of Years of education and Demented Rate

With regards to age, dementia seemed to be more concentrated in the range of 70-80 years. Socio-economic factors such as education clearly indicated that the rate dementia was in less educated. Also, brain volume in non-demented patients found to be higher when compared to their demented counterparts.

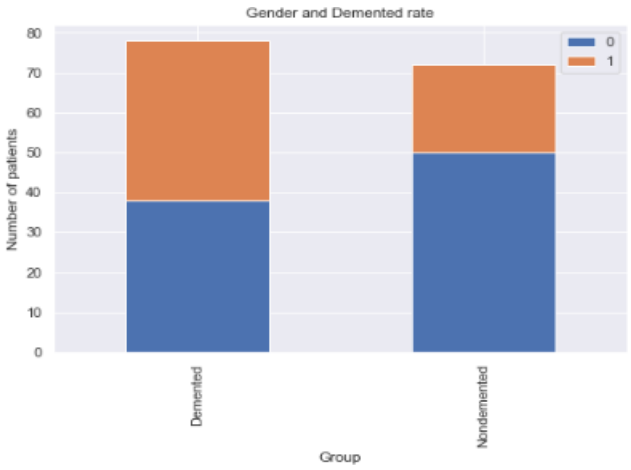

Fig 2.Plot of Gender and Demented Rate

Analyzing important attributes of the dataset such as Atlas Scaling Factor (ASF) Fig.3, Estimated Total Intracranial Volume (eTIV) Fig.4 and Normalized Whole Brain Volume (nWBV) Fig.5 by plotting them against Mini-Mental State Examination (MMSE) taken as the x-axis with reference to demented and nondemented group of datasets.

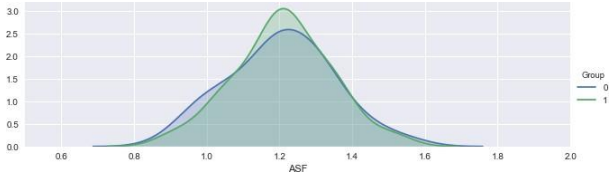

Fig 3. ASF Vs MMSE

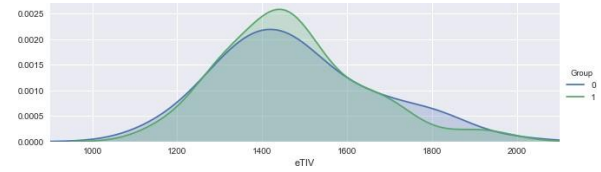

Fig 4. eTIV Vs MMSE

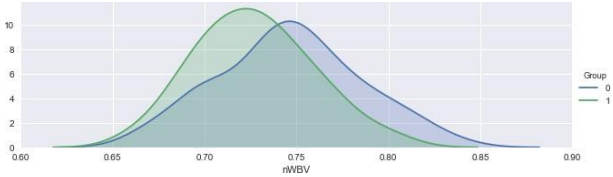

Fig 5. nWBV Vs MMSE

The dataset we used had a few columns with missing values, which required pre-processing. A common method to handle this is to remove the row with the missing value. However as there are many missing values this method can cause bias in the dataset. This problem can be reduced by using a method called imputation. Imputation involves replacing missing values with substituted value.

In this case we substituted the missing values with the median of the values of the particular column.

V. SUPERVISED MACHINE LEARNING MODELS

A. Decision Tree

Decision tree algorithm is one of the well-known supervised learning algorithm which is frequently used for classification and regression. Decision trees works using rules which are learnt by using training data. The best attribute is placed as the root of the tree. The remaining attributes is split into subsets based on the values of the root attribute. This is done iteratively to all the subsets till we get leaf nodes. There are two ways of choosing the best attribute: information gain and gini index. We have used gini index in our work. The gini index is the measurement of how often a randomly chosen element will be wrongly identified. Hence an attribute with a lower gini index is preferred. When used on our dataset, the Mini Mental State Examination (MMSE) attribute had the lowest gini index and was chosen as the root of the tree. Using the decision tree model on the dataset gave an accuracy of approximately 82%.

B. Random Forest

Random forest algorithm is works on similar to that of decision tree algorithm. It creates multiple decision trees, each different from the other. It then merges all the trees and obtains the result. As the name suggests, Random Forest adds an element on randomness when building the decision trees. Unlike the original decision tree algorithm, Random Forest does not look for the best feature when splitting a node. It instead takes a random subset of features for each tree and finds the best feature in that. It is also possible to take random attributes for splitting features instead of looking for the best possible one. The features in the dataset are given a value based on their feature importance- a measure of how much a particular feature contributes to the classification across the entire forest of decision trees. Random Forest randomly selects observations and features across the various trees and attempts to merge them together. Since Random Forest uses subsets of features, it results in smaller trees, thus reducing the risk of overfitting the data. However the computation of Random Forest is slower due to the presence of multiple decision trees. Using the Random Forest model gave us accuracies around 84%.

C. Adaboost

Adaboost, also known as Adaptive Boost is a popular boosting algorithm. Boosting algorithms are used in classification problems to create a strong classifier by using multiple weak classifiers. It works by using training data to build a model, then creating a second one which attempts to correct any errors in the first model. Adaboost is used with weak learners. Weak learners are models having accuracy slightly higher than random chance. Adaboost sets weights on classifiers and samples. This is done in such a way as to force classifiers to concentrate on observations that are difficult to correctly classify. We managed to achieve accuracies of around 85% using our dataset.

D. Support Vector Machine (SVM)

Support Vector Machine or SVM for short is a classification algorithm which uses a hyperplane in n-dimensional space to classify the features. Here 'n' refers to the number of features. The dimensions of the hyperplane changes with the number of features. SVM uses kernel functions to change dimensions of data from lower to higher dimensions. This helps when dealing with datasets in which the data is not easily separable. SVM chooses the right hyperplane by seeing the most accurate hyperplane classifier and then measuring the margin length from the SVM points. To construct the optimal hyperplane, SVM uses an iterative training function which helps reduce the error. The SVM model had an accuracy of around 81% when used on our dataset.

VI. LIMITATIONS

While using machine learning shows a good degree of accuracy, it also has some disadvantages.

A. While the algorithms provide accuracies in the 80s, it is not perfect. For a disease like Alzheimer's where the early stages are critical, higher accuracies are required for correct diagnosis.
B. The availability of limited data reduces the accuracy of our comparisons.
C. Since the exact biological factors for the disease are not known, it reduces the accuracy of the result.

VII. RESULTS AND CONCLUSION

Alzheimer's is a hard disease to correctly detect and diagnose. Conventional clinical methods are mediocre at best and ineffective at worst. Machine learning provides a solution to this. While not perfect, it provides a drastic increase in accuracy, making it a useful tool.

Table- I: Accuracy obtained under different algorithms

Algorithm	Accuracy
Decision Tree	81.5%
Random Forest	84%
SVM	81%
Adaboost	82%

As we can see in Table-I, the Random Forest algorithm provides the greatest accuracy of around 84%. The remaining algorithms have similar accuracy. Although Adaboost is a boosting algorithm, it provides an improvement of only 0.5% over the regular algorithm. This is a negligible difference, and is not worth the extra complexity that comes with implementing it over the decision tree algorithm. SVM provides a reasonable accuracy of around 81%. It is however better suited for problems where the data is not easily separable. The decision tree provides a similar accuracy, and has several advantages such as its ease of implementation, robustness and ability to work well with poor quality data. The Random Forest algorithm provides the highest accuracy among the chosen algorithms. Being a more advanced version of the decision tree algorithm, it overcomes some of its problems such as overfitting. Alzheimer's disease at the earliest stage possible. One suggested solution is to attempt applying the Adaboost or another similar boosting algorithm in combination with SVM classification.

If this can be done, it would result in an algorithm with a better accuracy, making it easier to detect and diagnose patients with Alzheimer's disease.

REFERENCES

1. Dessouky MM, Elrashidy MA (2016) Feature Extraction of the Alzheimer's Disease Images Using Different Optimization Algorithms. J Alzheimers Dis Parkinsonism 6:230.
2. Sarica A, Cerasa A, Quattrone A. Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: A systematic review. Frontiers in aging neuroscience. 2017 Oct 6;9:329.
3. Saruar Alam, Goo-Rak Kwon, Ji-In Kim and Chun Su Park, "Twin SVM Based Classification of Alzheimer's Disease Using Complex Dual-Tree Wavelet Principal Coefficients and LDA".
4. Zhang, Yu-Dong & Dong, Zhengchao & Phillips, Preetha & Wang, Shuihua & ji, Genlin & Yang, Jiquan & Yuan, Ti-Fei. (2015). Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning. Frontiers in computational neuroscience. 9. 66. 10.3389/fncom.2015.00066.
5. Nielsen and Didrick, "Tree Boosting with XGBoost- Why Does XGBoost Win Every Machine Learning Competition".
6. Padilla P, Lopez M, Gorriz J, Ramirez J, Salas-Gonzalez D, et al. (2012)NMFSVM Based CAD Tool Applied to Functional Brain Images for the Diagnosis of Alzheimer's Disease. IEEE Trans Med Imaging 31: 207-216.
7. Martinez-Murcia FJ, Gorriz JM, Ramirez J, Puntonet CG, Salas-Gonzalez (2012) Computer Aided Diagnosis tool for Alzheimer's Disease based on Mann–Whitney–Wilcoxon U-Test. Expert Systems with Applications 39: 9676- 9685.
8. Elbeltagi E, Hegazy T, Grierson D (2005) Comparison among five evolutionary based optimization algorithms. Advanced Engineering Informatics 19: 43-53. 6. Xin-She Yang (2014) Nature-Inspired Optimization Algorithms. Elsevier.
9. Holland J (1975) Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor, MI, USA.
10. Tabassum M, Mathew K (2014) A Genetic Algorithm Analysis towards Optimization solutions. International Journal of Digital Information and Wireless Communications (IJDIWC) 4: 124-142G
11. Wimo A, Prince M (2010) World Alzheimer Report 2010: The global economic impact of dementia.

AUTHORS PROFILE

Prasad B S, is currently pursuing his research from Visvesvaraya Technological University in the area of Visual Cryptography and machine learning. His passion is in the field of Cryptography and Network Security, Algorithms and Theory of Computation. He has got more than 10 years of academic experience and 3 years of Industrial experience. He is a life member of Indian Society for Technical Education and Indian Science Congress Association (ISCA). He has served as resource person in many national and international training programs in the field of IoT and Programming skills. He has multiple national and international publications to his credit.

Akhilaa is currently working as Assistant Professor in the Department of Information Science and Engineering, C M R Institute of Technology, Bengaluru. She has 4 years of teaching experience. Her area of research includes Data Analytics, Data Science, Machine Learning and Natural Language Processing. She is Life-time member of Indian Society for Technical Education (ISTE) and Indian Science Congress Association (ISCA). She has published multiple technical e-books under amazon publications. She has developed course materials including lecture notes and laboratory manuals for the students of Visvesvaraya Technological University She also has published multiple technical papers in national and international journals.