INTERNATIONAL JOURNAL OF RESEARCH IN MEDICAL SCIENCES & TECHNOLOGY

e-ISSN:2455-5134; p-ISSN: 2455-9059

Leveraging Data Science Linked Tools and Techniques in the Efficacious Detection and Diagnosis of Alzheimer's Disease

Bahisht Samar
Amity University, Noida

Paper Received: 27th May 2022; **Paper Accepted:** 20th June 2022;
Paper Published: 29th June 2022

DOI: http://doi.org/10.37648/ijrmst.v13i01.017

How to cite the article:

Bahisht Samar, Leveraging Data Science Linked Tools and Techniques in the Efficacious Detection and Diagnosis of Alzheimer's Disease, IJRMST, January-June 2022, Vol 13, 175-186, DOI: http://doi.org/10.37648/ijrmst.v13i01.017

e-ISSN : 2455-5134
p-ISSN : 2455-9059

International Journal of Research
in
Medical Sciences & Technology

Open Access, Peer Reviewed Medical Journal

WE INVITE ARTICLES FROM MEDICAL, DENTAL, AYURVEDA, PHYSIOTHERAPY & HOMEOPATHIC FACULTIES

Published by : IJRMST PUBLICATION, GHAZIABAD
9555269393
info@ijrmst.com www.ijrmst.com

ABSTRACT

This paper investigates data for 9 common Alzheimer's Disease risk factors, from three different categories; Medical History, Lifestyle, and Demography. The dataset used consists of 185 normal control, 177 early mild cognitive impairment, 161 late mild cognitive impairment and 127 Alzheimer's Disease subjects. The initial experiment had training results of 0.92 sensitivity, 0.935 specificity and 0.771 precision. However, during the test stage the final output was 0.741 sensitivity, 0.515 specificity and 0.286 precision. The results of this experiment did not give a clear classification or definite predictive value. Involving more variables and underlying data could provide a better outcome. This paper is a part of a long-term study that focuses on the classification and ranking the importance of Alzheimer's Disease risk factors using Machine Learning predictive models and classifications techniques.

INTRODUCTION

Researchers from different fields such as biology, physiology, neurology, computer science and others have been exploring this, ultimately fatal disease, for decades. Although, there have been no major breakthroughs and scientists are still unsure of what is the actual cause of Alzheimer's Disease (AD) or have any cure for it, there is a valuable amount of knowledge and information that has been gained on the disease.

Like any disease, it is important that we know its risk factors and avoid them. Since scientists are still unsure of the actual cause of AD, however, there has been a lot of research to establish the risk factors for AD. General Health Practitioners (GP) would usually rely on diagnosing AD through its symptoms and several standards and procedures. However, AD shares most of its symptoms with other types of dementia; therefore, GPs can sometimes give a wrong diagnosis. The existence of these two protein "Plaques" and "Tangles" is what indicates and confirms the existence of AD [1]. False and inaccurate diagnoses are common when it comes to early diagnosis of dementia. This is because GPs rely on manual evaluation and mental examinations before they turn to brain imaging. It is difficult to manually diagnose AD, or any other types of dementia at an early stage before most of its symptoms are noticeable. Therefore, it is important to use computer analysis to analyse as much patient’s data as possible for a better evaluation and more accurate diagnosis. First, it is important to express the complexity of AD progression, hence, why its risk factors fall into multiple categories from biological risk factors to behavioural risk factors. The main categories of AD risk factors are age, genetics, medical history, lifestyle, and characteristics / demography. Health services providers and major research institutions around the world have provided a list of risk factors and declared some of them as high-risk factors, which poses potential development indication of AD. However, these risk factors do not mean that they are the real reasons behind the development of AD. This is because the pathology of AD progresses through different channels.

With AD, it is important to understand the behaviour or its risk factors and their interrelationship. This study can be viewed in three different phases. Phase one will provide a classified list of AD risk factors and their relevancy to pathology of AD with the use of machine learning tools, phase two will produce a similar outcome but based on manual evaluation from the existing research. The third phase will use the importance of the risk factors from both phases one and two to quest for predictive patterns of AD. The merge of the outcome from both phases one and two will boost the accuracy of the predictive patterns.

METHODS

Currently there has been little progress made in relation to developing a complete early diagnose approach of AD by using Intelligence Data Analytic and Machine Learning techniques for early prediction. The current work into prediction of AD relates to research using lifelogging technology to monitor memory decline or to diagnose the disease at a very late stage. According to the Alzheimer's Association there are no current working methods to diagnose AD at a very early stage and the “current diagnosis of Alzheimer's relies largely on documenting mental decline.”[2] The methods used to diagnose AD are cognitive tests such as the Mini Mental Score Examination test and in some cases a brain scan is required. Unfortunately, these methods detect AD at a very late stage when all of the symptoms appear.[3] However, this research will focus on developing an evolving framework to effectively diagnosis and predict AD at a very early stage using the data collected for AD patients. The framework will continuously use large sets of related data to AD patient collected from multiple sources. The collected data will feed the framework with the input required to deploy computational modelling and machine learning techniques to predict and diagnose AD. The research will depend on collecting data from multiple existing datasets such as ADNI, and it will include collected data that related to DNA, dietary, medical history, lifestyle and any other related data linked to risk factors of AD. Away from biological methods, this study will conclude with a working method for early diagnosis of AD. The proposed research aims to investigate AD with effective use of Machine Learning (ML) techniques to predict AD at a very early stage using its predictive risk factors. The overall target is to look for possible solutions by using data science, machine learning and artificial intelligence, in order to develop an intelligent data analytic framework to predict and visualise data for early diagnosis and prediction for people with AD. The research aims to develop a framework consisting of the following stages; construction of a baseline dataset, deploying Machine Learning techniques on the dataset, increasing the dataset variables and deploying Deep Learning techniques to model high-level abstractions in the data as the datasets increases [4][5][6][7][8], tracking the changes and developing a weighting formula, then finally produce a computational method that will provide an early diagnostic tool of AD and help general participant with early decision-making.

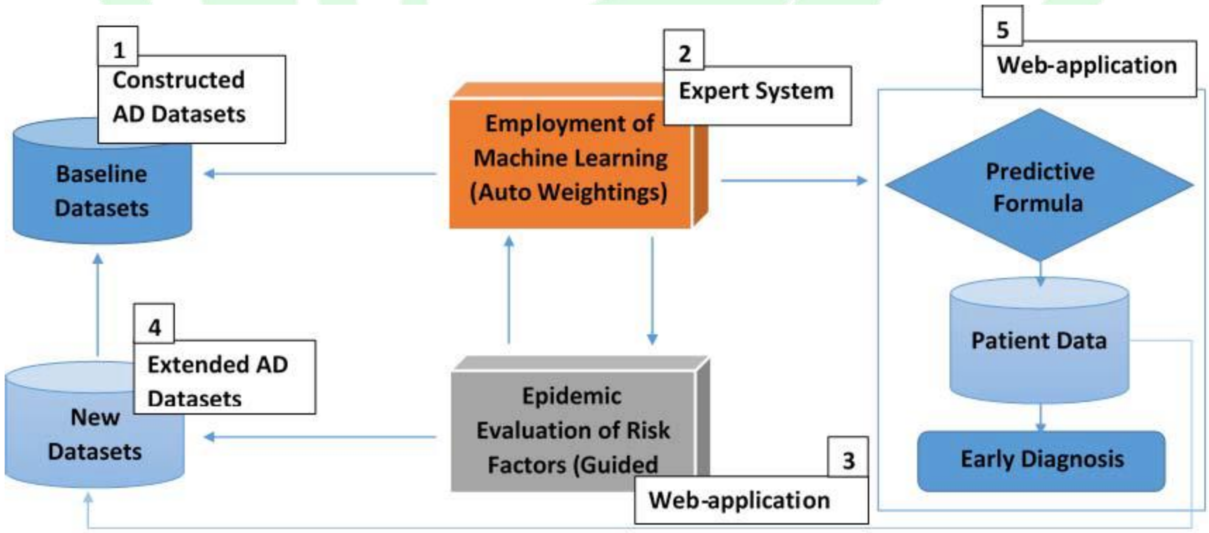


Fig.1. Diagram of proposed work

Elaboration:

- Subsets of data related to major risk factors of AD will be extracted from multiple existing (published) databases and constructed into a Machine Learning ready dataset.
- Continuous learning technique will be employed to analyse the constructed dataset and provide feedback on the importance of the variables and AD risk factors. This part of the framework and the employment of ML techniques will be done using MatLab, R studio and other integration services.
- A web-based sub-system will be developed to calculate guided weighting for each risk factor. This system will rely on validated discrete knowledge manually inputted by either system admin, or professionals through crowdsourcing. This will be used to influence the weighting used by ML techniques, as well as decision making when adding new variables to the dataset.
- The feedback from both ML techniques and epidemic evaluation will be used to determine which new data needs to be collected and what variables should be added to the baseline dataset.[9]
- Whilst constantly learning from the datasets, the predictive formula will continuously be update to provide as accurate prediction as possible. The prediction formula will feed into a live system in which live patient data will be stored. This part of the framework will keep track of patient records and trigger warnings when establish.[10]

FEATURE SELECTION

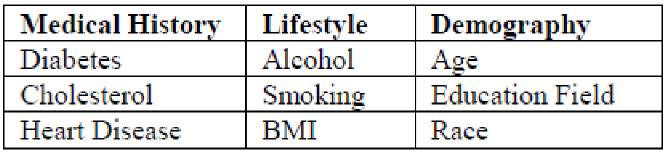
Feature selection widely used technique in research on big data; researchers explore domains with hundreds to tens of thousands of variables or features. Therefore, many feature selection techniques used to address these challenges in order to select relevant data and to remove irrelevant, redundant, and noisy information from the data [11]. There are many features selection and these methods are categorized in three different classes based on how the selection algorithm and the model building are combined. The three classes of feature selection methods are; filter method, wrapper method and embedded method. A demonstration of how feature selection used to help researchers; Feature selection technique was employed in the work of Dimitrios Ververidis from the VTT Technical Research Centre of Finland. Titled: “Feature selection and time regression software: Application on predicting AD progress” [12].

In this work the data features will be extracted and selected in accordance to the work carried out on AD risk factors. The aim of this work is use AD risk factors as a mean to predict AD at an early stage before it progresses to severe stage. With this type of research Neural Network Classifiers will be used

to identify the predictive AD risk factors. There are several risk factors for AD which will be used in this work. The work will use the ADNI datasets in the initial experiments, then it will use other datasets to expand the hunt for early prediction of AD.

ADNI collected data for subjects who are Normal Control (NC), Mild Cognitive Impairment (MCI), AD (AD), Significant Memory Concern (SMC), Early Mild Cognitive Impairment (EMCI) and Late Mild Cognitive Impairment (LMCI) patients. Initially, before data cleaning ADNI 2 had 1171 subjects; 343 NC, 204 AD, 230 EMCI, 159 LMCI, 131 MCI and 104 SMC. In the ADNI datasets there are over 63 tables containing rich amount of data for all of the patient participated in the study. Each table has a large number of variables ranging between 11 to over 71. However, since this report is focused on early prediction of AD or pre-dementia stage, the experiments is carried out on data that relates to the behavioural markers and the variables selected are matching the features discussed in the risk factors section of this report. The early prediction of AD will be attempted through a series of experiments on the ADNI datasets. First experiment will be carried out on ADNI 2 baseline data. The variables in this first experiment will be selected in accordance to AD risk factors as categorized by current research; medical history, lifestyle and demography. Three risk factors have been taken from each category. TABLE 1 lists the risk factors use as variables in this experiment.

TABLE 1: AD Risk Factors used in the experiment



Medical History	Lifestyle	Demography
Diabetes	Alcohol	Age
Cholesterol	Smoking	Education Field
Heart Disease	BMI	Race

To start the experiments on the ADNI 2 dataset, the first step was to clean the data, removing rows with missing values, normalizing the values and converting strings to numeric values. However, after the performance of data cleaning the distance between the data volume for each class has changed and resulted for the data to be imbalanced. The graph below (Fig.2) shows an explore of the data after the deletion of subjects with missing data. This resulted in a very small number of subjects who are classed as SMC compare to the rest of the classes, which, makes the dataset imbalanced.

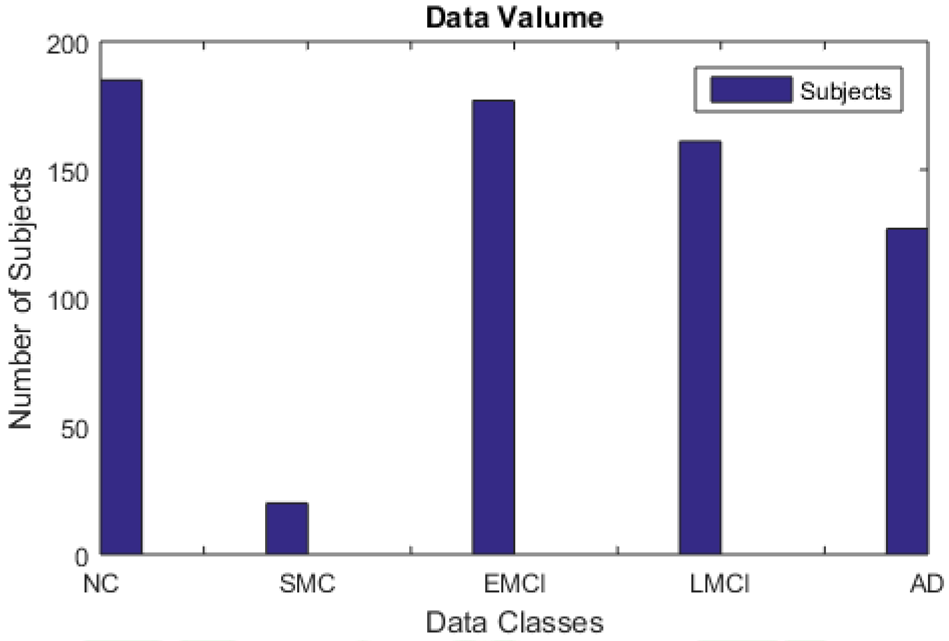


Fig.2 – ADNI Data Volume

DATA ANALYSIS

Running the experiment on an imbalanced data will not give a correct accuracy as it will mislead the artificial agents to give insufficient results. To make the data more balance, all subjects with SMC class were removed from the dataset. The following graphs and tables show an overview of the final dataset that will be used in the following experiments.

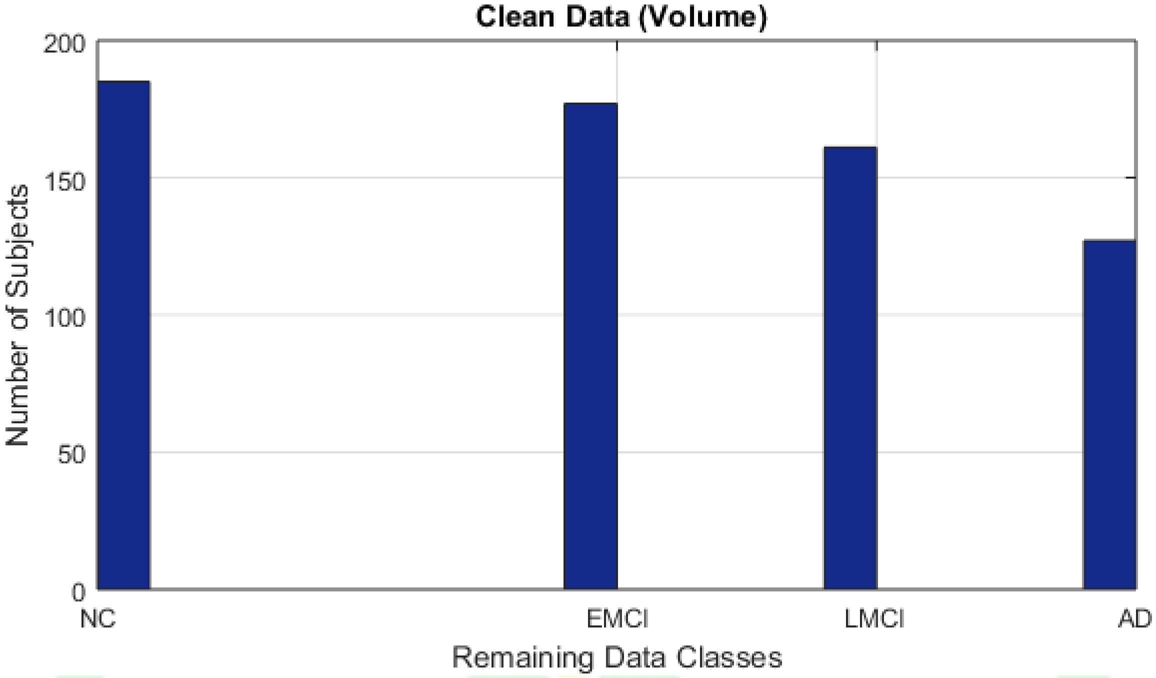


Fig.3 – Explore of Data Using MatLab Graphs

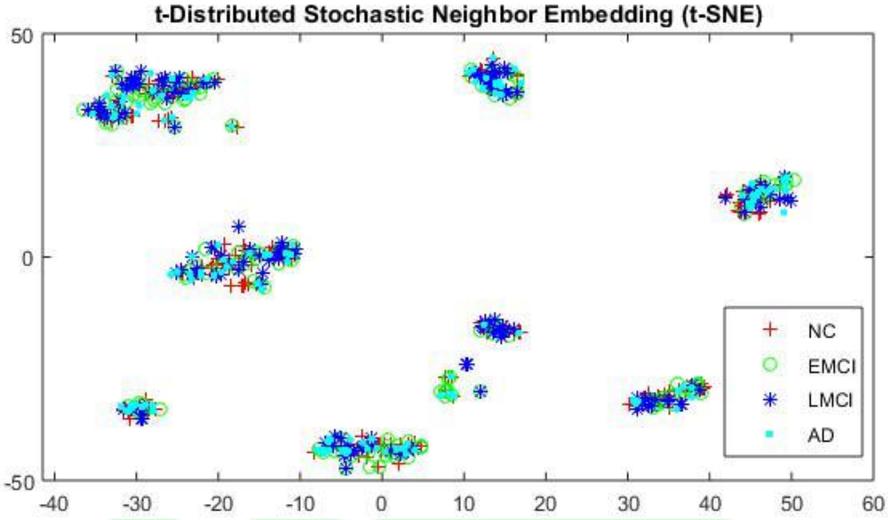


Fig.4 - Explore of Data Using t-SNE on MatLab

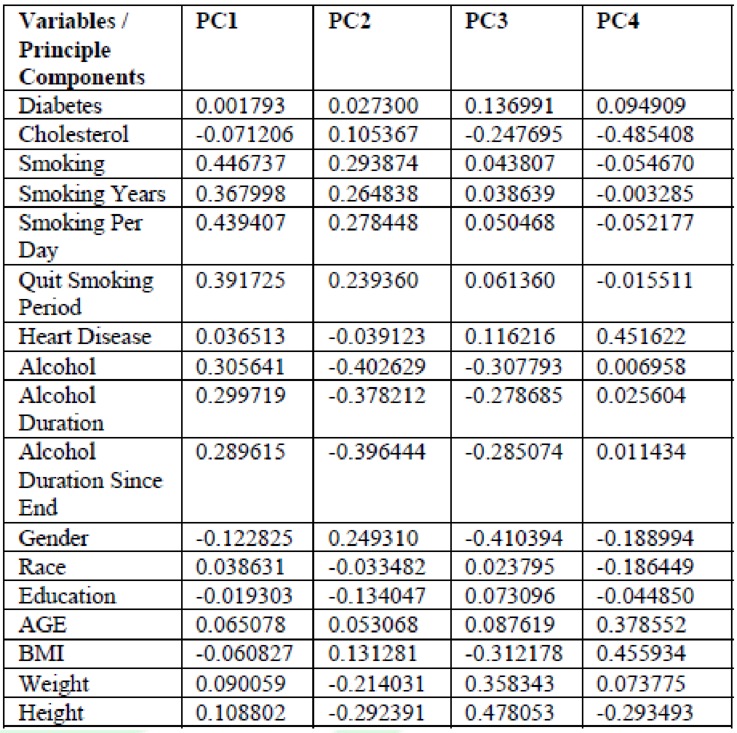
During this phase the number of subjected was reduced by almost 50%, now with a total remaining number of 650 subjects to be studied. TABLE 2 and the graph in Fig.3 give an overview of the dataset for each of the four remaining classes (Labels).

To explore the dataset suitably, different data analysis toolboxes on MatLab, MiniTab 16 were deployed, this includes t-

Distributed Stochastic Neighbour Embedding (t- SNE), Principle Component Analysis (PCA), Independent Component Analysis (ICA), and Square Prediction Error (SPE). The t-SNE is a Machine Learning algorithm commonly used for dimensionality reduction in data visualization. t-SNE was apply on the dataset, giving the results showing in Fig.4, it shows mixed clusters detached apart. Ideally the perfect results that we had hoped for is that each cluster will contain a majority of one class. However, as shown in Fig.4; the clusters have almost equal mixture from all classes. Which, means that the algorithm struggled to differentiate between the categories of the subjects. Though this algorithm also shows a large distance between the clusters which means on a dimensional level it managed to differentiate between the variables (risk factors).

Other useful methods used to explore and visualize this dataset is the Principal Component Analysis (PCA) and Independent Component Analysis (ICA), both Fig.5 illustrate the use of this technique on Matlab and MiniTab 16. PCA is used to emphasize the variation, dimensions reduction and bring out the strongest patterns in the dataset. ICA is a method for separating a multivariate signal into additive subcomponents.

TABLE 2 –PCA Coefficient for Each Variable on MiniTab 16



Variables / Principle Components	PC1	PC2	PC3	PC4
Diabetes	0.001793	0.027300	0.136991	0.094909
Cholesterol	-0.071206	0.105367	-0.247695	-0.485408
Smoking	0.446737	0.293874	0.043807	-0.054670
Smoking Years	0.367998	0.264838	0.038639	-0.003285
Smoking Per Day	0.439407	0.278448	0.050468	-0.052177
Quit Smoking Period	0.391725	0.239360	0.061360	-0.015511
Heart Disease	0.036513	-0.039123	0.116216	0.451622
Alcohol	0.305641	-0.402629	-0.307793	0.006958
Alcohol Duration	0.299719	-0.378212	-0.278685	0.025604
Alcohol Duration Since End	0.289615	-0.396444	-0.285074	0.011434
Gender	-0.122825	0.249310	-0.410394	-0.188994
Race	0.038631	-0.033482	0.023795	-0.186449
Education	-0.019303	-0.134047	0.073096	-0.044850
AGE	0.065078	0.053068	0.087619	0.378552
BMI	-0.060827	0.131281	-0.312178	0.455934
Weight	0.090059	-0.214031	0.358343	0.073775
Height	0.108802	-0.292391	0.478053	-0.293493

Furthermore, the data was explored using the Square Prediction Error (SPE) plot to measure the quality of a predictor. The graphs and coefficients result in TABLE 3 show an apparent indication that the type of study will be a nonlinear regression.

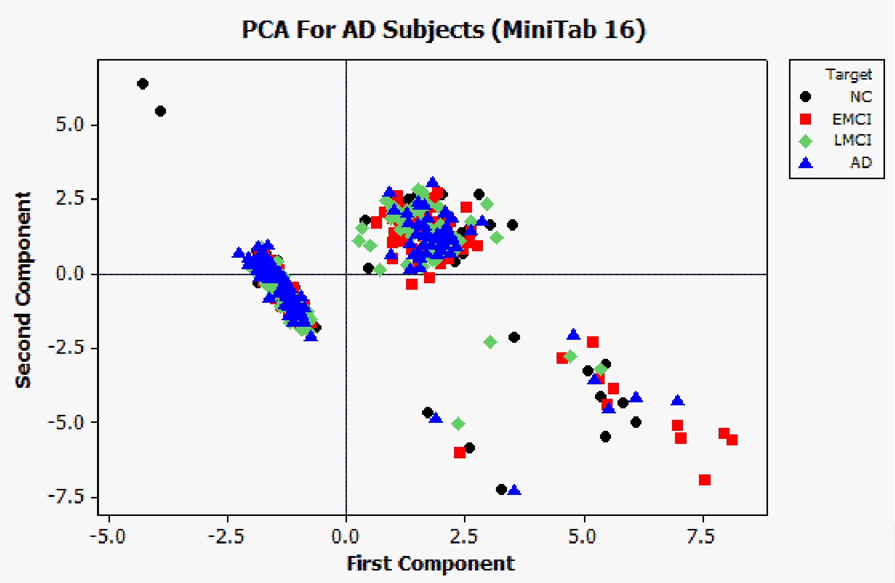


Fig.5 - Explore of Data Using PCA on MiniTab 16

The coefficient data shows an apparent indication that the experiment result of the predictors will not provide a clear outcome. From the data exploratory it is very unclear which variable is the highest predictive factor, which, shows that the classifiers might not give a very clear classification or definite

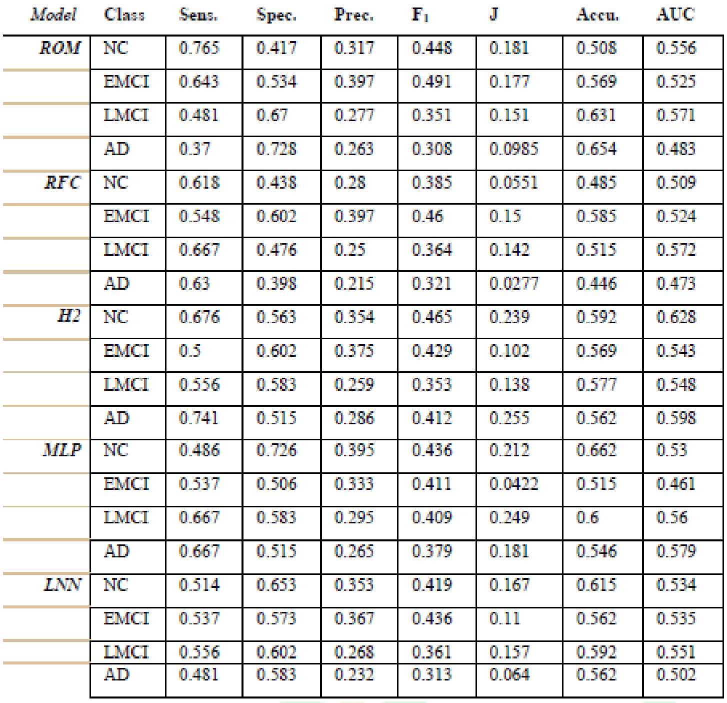
predictive value. In this case future experiments involving more variables and underlying data will provide a better outcome.

RESULTS

During both the training stage and test stage, the five different classifiers were applied consecutively for approximately 30 times for a better accuracy. The contrast between the outcome of the training experiment and the test experiment is very obvious. As expected, the classifiers have performed better

during the training stage. Model H2 performed the best while other classifiers did not have dramatic difference during the test stage. TABLE 3 show the outcome of the test experiments. From a constructive perspective the initial investigation provides a needed foundation to draw a roadmap for future work and it has become apparent that more variable related to AD risk factors are required to improve the accuracy of the classifiers.

TABLE 3 – Testing Overall Results



Model	Class	Sens.	Spec.	Prec.	F_1	J	Accu.	AUC
ROM	NC	0.765	0.417	0.317	0.448	0.181	0.508	0.556
	EMCI	0.643	0.534	0.397	0.491	0.177	0.569	0.525
	LMCI	0.481	0.67	0.277	0.351	0.151	0.631	0.571
	AD	0.37	0.728	0.263	0.308	0.0985	0.654	0.483
RFC	NC	0.618	0.438	0.28	0.385	0.0551	0.485	0.509
	EMCI	0.548	0.602	0.397	0.46	0.15	0.585	0.524
	LMCI	0.667	0.476	0.25	0.364	0.142	0.515	0.572
	AD	0.63	0.398	0.215	0.321	0.0277	0.446	0.473
H2	NC	0.676	0.563	0.354	0.465	0.239	0.592	0.628
	EMCI	0.5	0.602	0.375	0.429	0.102	0.569	0.543
	LMCI	0.556	0.583	0.259	0.353	0.138	0.577	0.548
	AD	0.741	0.515	0.286	0.412	0.255	0.562	0.598
MLP	NC	0.486	0.726	0.395	0.436	0.212	0.662	0.53
	EMCI	0.537	0.506	0.333	0.411	0.0422	0.515	0.461
	LMCI	0.667	0.583	0.295	0.409	0.249	0.6	0.56
	AD	0.667	0.515	0.265	0.379	0.181	0.546	0.579
LNN	NC	0.514	0.653	0.353	0.419	0.167	0.615	0.534
	EMCI	0.537	0.573	0.367	0.436	0.11	0.562	0.535
	LMCI	0.556	0.602	0.268	0.361	0.157	0.592	0.551
	AD	0.481	0.583	0.232	0.313	0.064	0.562	0.502

CONCLUSION

According to the Alzheimer's Association there are no current working methods to diagnose Alzheimer’s Disease at a very early stage and the “current diagnosis of Alzheimer's relies largely on documenting mental decline.”[2] The method used to diagnose Alzheimer’s Disease is by using the Mini Mental Score Examination test and in some cases a brain scan. Unfortunately, these methods detect Alzheimer’s Disease at a very late stage when all of the symptoms appear.[3] However, the more knowledge gain on Alzheimer’s Disease, the closer scientists get to solving its mysterious cause.

Financial support and sponsorship: Nil

Conflict of Interest: None

REFERENCES

1. D. M. Holtzman, J. C. Morris, and A. M. Goate, “Alzheimer’s disease: the challenge of the second century.,” Sci. Transl. Med., vol. 3, no. 77, p. 77sr1, Apr. 2011.
2. “Alzheimer’s & Dementia Testing Advances | Research Center,” Alzheimer’s Association, 2015. [Online]. Available: http://www.alz.org/research/science/earlier_alzheimers_diagnosis.as p#Biomarkers. [Accessed: 17-May-2015].
3. M. Fernández, A. L. Gobartt, and M. Balañá, “Behavioural symptoms in patients with Alzheimer’s disease and their association with cognitive impairment.,” BMC Neurol., vol. 10, no. 1, p. 87, Jan. 2010.

4. P. Yang, M. Hanneghan, J. Qi, Z. Deng, F. Dong, and D. Fan, “Improving the Validity of Lifelogging Physical Activity Measures in an Internet of Things Environment,” in 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, 2015, pp. 2309–2314.
5. J. Qi, P. Yang, M. Hanneghan, and S. Tang, “Multiple density maps information fusion for effectively assessing intensity pattern of lifelogging physical activity,” Neurocomputing, vol. 220, pp. 199– 209, Jan. 2017.
6. H. Schulz and S. Behnke, “Deep Learning,” KI – Künstliche Intelligenz, vol. 26, no. 4, pp. 357–363, Nov. 2012.
7. S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, and D. Feng, “Early diagnosis of Alzheimer’s disease with deep learning,” in 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), 2014, pp. 1015–1018.
8. H.-I. Suk and D. Shen, “Deep Learning-Based Feature Representation for AD/MCI Classification,” Springer Berlin Heidelberg, 2013, pp. 583–590.
9. J. Qi, P. Yang, M. Hanneghan, D. Fan, Z. Deng, and F. Dong, “Ellipse fitting model for improving the effectiveness of lifelogging physical activity measures in an Internet of Things environment,” IET Networks, vol. 5, no. 5, pp. 107–113, Sep. 2016.
10. J. Qi, P. Yang, G. Min, O. Amft, F. Dong, and L. Xu, “Advanced internet of things for personalised healthcare systems: A survey,” Pervasive Mob. Comput., vol. 41, pp. 132–149, Oct. 2017.
11. K. Javed, H. A. Babri, and M. Saeed, “Feature Selection Based on Class-Dependent Densities for High-Dimensional Binary Data,” IEEE Trans. Knowl. Data Eng., vol. 24, no. 3, pp. 465–477, Mar. 2012.
12. D. Ververidis, M. Van Gils, J. Koikkalainen, and J. Lotjonen, “Feature selection and time regression software: Application on predicting Alzheimer’s disease progress.” pp. 1179–1183, 2010.