2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT) | 979-8-3503-1212-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/IConSCEPT57958.2023.10170234

# Early Alzheimer's Detection Using Random Forest Algorithm

Pranjlee Kolte
*Alumni, Department of Electronics and Communication Engineering*
*Shri Ramdeobaba College of Engineering and Management*
*Nagpur, India*
koltepg@rknec.edu

Aditya Shrivastava
*Alumni, Department of Electronics and Communication Engineering*
*Shri Ramdeobaba College of Engineering and Management*
*Nagpur, India*
*shrivastavaar_1@rknec.edu*

Himanshu Choudhary
*Alumni, Department of Electronics and Communication Engineering*
*Shri Ramdeobaba College of Engineering and Management*
*Nagpur, India*

Nandani Rabra
*Alumni, Department of Electronics and Communication Engineering*
*Shri Ramdeobaba College of Engineering and Management*
*Nagpur, India*
rabrans@rknec.edu

Anushka Khadatkar
*Alumni, Department of Electronics and Communication Engineering*
*Shri Ramdeobaba College of Engineering and Management*
*Nagpur, India*
khadatkarar@rknec.edu

Divya Shrivastava
*Faculty, Department of Electronics and Communication Engineering*
*Shri Ramdeobaba College of Engineering and Management*
*Nagpur, India*
shrivastavadd1@rknec.edu

***Abstract*–Alzheimer's disease (AD) is a progressive neurological ailment causing damage to brain cells. Beginning with mild symptoms that usually goes unnoticed, the disorder gets worse as it progresses hindering the general abilities of person. Early AD symptoms being ordinarily simple, detection occurs only on disease progression to an advance irreversible stage. Early detection of AD is thus critical to reduce the adverse effects of the disease. Earlier detection can prove promising for the development of specific treatment strategies that improve or slow AD progression. Machine Learning (ML) approach has become increasingly useful in the detection of Alzheimer's disease in recent years. In this paper, early detection of Alzheimer's disease using different machine learning algorithms for predictive categorization of patients is presented. The study suggests that random forest algorithm offers best performance for early prediction of Alzheimer's disease with an accuracy of 93.69%. A GUI for users to enter parameters for early detection and display the categorized result for random forest algorithm is also designed.**

***Keywords: Alzheimer's Disease, Early Detection, Machine Learning, Random Forest Algorithm***

## I. INTRODUCTION

Dementia is a general term referred for sever health disorder that hinders the ability to remember, think, or make decisions affecting millions of people around the world. Alzheimer's Disease (AD), most common type of dementia, is a degenerative neurological disorder that progressively damages brain cells. It affects the part of the brain responsible for thought, memory and language thereby affecting the people suffering from AD in numerous ways. The disorder begins with mild conditions like loss of memory, confusion and difficulty in carrying out a conversation or to respond. The severity increases eventually with disease progression to the extent of inability to recognize familiar people, places or to perform routine tasks. AD is a mental blow that cannot be reversed with the condition of patient worsening as the ailment advances. It can even cause death in elder people. Research suggests that most individuals with the disease are 65 years or older. After the age of 65 years, the risk of AD doubles every five years. Though age is not the only factor affecting AD, risk increases with increasing age. Advancing age being the biggest risk factor for AD combined with steadily increasing life expectancy; AD is becoming more and more prevalent, emerging as a global concern [1]. Over the years, clinical research on AD led to the effective treatment strategies only intended for slowing down or preventing neuronal death in AD [2] [3]. Recognizing, categorizing and differentiating the mildly common symptoms of AD at an early stage can drastically affect the treatment and understanding of the disease. Early risk prediction of AD calculated using the clinical history of the risk factors, medical tests and non medical parameters can prove to be very beneficial for deciding such strategies and preventing brain tissue damage or at least slow down disease's progression rate. This requires highly accurate detection of AD at an early stage for improving treatment efficacy. Although, there is no specific scheme that satisfies for the early detection, evidence proves that combination of neuroimaging, cerebrospinal fluid (CSF), Mini-Mental State Examination (MMSE), blood biomarkers and patient's history provide can also give important introspect and contribute for accurate and earlier diagnosis of AD [2].

Researchers numerous attempts for early AD detection following various techniques have bared fruits, sometimes with realized objective and in some cases with a scope of improvement. In recent years, following the exploration of Machine Learning (ML) to solve complex computations and problem solving, ML for early AD detection has gathered attention of the researchers [4]. ML has showcased tremendous potential to cope with the need of medicinal field for early prognosis, diagnosis and suggestive treatment of various diseases [5] [6]. Exploring various algorithms has also been a part of study of the researchers. Random Forest (RF) algorithm, one of the recently talked ML algorithms, has been successfully explored and applied in numerous applications. RF has

been explored on neuroimaging data for the highly accurate prediction of Alzheimer's disease [7]. RF algorithm has been increasing explored for various applications in recent years and has shown a good deal of usefulness [8].

In this paper, various ML algorithms are explored for early AD detection. The proposed work is organized around dataset preparation, training and testing various ML algorithms for early detection of AD; of which RF emerges with highest efficient algorithm for the application.

## II. Related Work

Frequently use of ML for pattern recognition has led to its use in MRI for detection of AD. Study of ML techniques for AD detection has been explored by various researchers [9]. A good amount of research for early AD detection has been carried using different statistics and techniques including Speech analysis, EEG analysis, MRI Imaging etc. Data collection process for these methods is difficult and not easily accessible to people. As a result, a quick and easy machine learning-based system capable of detecting the existence of dementia using a person's clinical and demographic record might be useful in delivering rapid diagnosis. The data from MRI scans, as well as demographic evaluation such as the subject's education level, the Mini-Mental State Exam (MMSE) scores, the subject's socio economic status, and other factors, are used to develop ML models that predict the existence of dementia in the individual in this study [10]. ML is frequently used for image based automated pattern recognition. Classic ML algorithms, such as SVM and linear judgment analysis, have been successfully deployed for detecting AD at early stages using MRI images [11]. Research and reviews related to AD detection and improve patient's life quality has also been done extensively. Researchers have also been exploring the multimodal measurement data that can support AD detection [12] [13] [14]. Neural network based approach for feature extraction from MRI image has also been proposed for AD detection [15]. ML is not restricted to AD prognosis but can also provide support for AD treatment by caregivers [16]. Numerous research papers are targeted at reviewing and comparing the results of various researchers. AD recognition accuracy rates from previous works are shown in Table I.

TABLE I. Accuracy of Models

| Algorithm | Accuracy (%) |
|---|---|
| Logistic Regression | 61.65 |
| Naïve Bayes | 84.76 |
| Support Vector Machine | 87.89 |
| Random Forrest | 89.69 |
| Decission Tree | 81.78 |

## III. Proposed Work

### A. Dataset

Open Access Series of Imaging Studies (OASIS) is a series of open source neuroimaging data sets of brain facilitating and supporting research in the field of neuroscience. The dataset consists of four sets viz. OASIS-1, OAIS-2, OASIS-3 and OASIS-4. OASIS-2: Longitudinal MRI Data is the dataset used for training and testing the network for the work presented in this paper [17]. It consists of MR session data over a long-term for a subject group of 150 demented and non-demented persons between the ages of 60 and 96 with a total of 373 MRI sessions. Multiple MRI sessions were taken at an interval of at least one year on two distinct visits of the individuals. In case of single MRI session, 3 or 4 separate T1-weighted MRI scans were collected for each individual. The participants, including men and women, were all right-handed. Throughout the MRI sessions, 72 of the individuals were classified as non-demented. During the initial visit, 64 of the participants were classified as demented, and this classification stayed the same for successive sessions. Sessions of 51 individuals showed Alzheimer's disease, ranging from mild to moderate. Another 14 individuals were first identified as non-demented but were later classified as demented on subsequent visits. OASIS-2 Longitudinal MRI dataset consists of 15 features recorded at every MRI scanning session as described in table II through V.

TABLE II. OASIS-2 Subject Characteristic Data

| Group | Subjects | Age (years) | Sex (M/F) | Hand (L/R) |
|---|---|---|---|---|
| Demented | 143 | 76.09±7.03 | 83/60 | 0/134 |
| Converted | 41 | 79.43±7.18 | 16/25 | 0/41 |
| Non-demented | 189 | 76.80±7.84 | 61/128 | 0/189 |

TABLE III. OASIS-2 Subject Demographic Data

| Group | Education (years) | SES (1/2/3/4/5) | MMSE |
|---|---|---|---|
| Demented | 13.6±2.8 | 25/30/27/45/6 | 24.3±4.58 |
| Converted | 15.51±2.56 | 24/9/6/2/0 | 28.41±2.41 |
| Non-demented | 15.22±2.71 | 40/72/43/33/2 | 29.22±0.93 |

TABLE IV. Clinical Data

| Group | CDR (0/0.5/1/2) | eTIV ($mm^3$) | nWBV (mg) | ASF |
|---|---|---|---|---|
| Demented | 0/99/41/3 | 1477.81± 163.89 | 0.717± 0.032 | 1.2027± 0.132 |
| Converted | 18/21/2/0 | 1576.36± 158.28 | 0.723± 0.03 | 1.2± 0.121 |

### B. Feature Selection

Feature selection highly impacts the performance of the ML model as training and testing on redundant or irrelevant dataset features may affect the performance of the model. Correlation based feature selection, a type of filter based feature selection method, is used for removing less relevant features as a preprocessing step before testing various algorithms for AD early detection owing to its lower computational complexity and independence from algorithms to be used [18]. Selecting the features to train and test the ML model is done with the help of a

correlation matrix summarizing the relation between different input variables is shown in figure 1. The matrix representing the feature relation numerically with both positive and negative values can be interpreted with the help of magnitude and sign. Higher the magnitude, stronger is the correlation between features. Positive magnitude represents regular correlation and negative an inverse correlation. Using the correlation matrix, number of input features from the dataset has been reduced to get the optimum features most significant to the work.

TABLE V. LIST OF FEATURES WITH RELATIVE VALUE IN OASIS-2 DATASET

| Feature | Description | Relative Value |
|---|---|---|
| MRI Id | Unique MRI Identification Number (An individual can have more than one MRI) | Variable depending on number of scans |
| Patient Id | Unique Patient Identifying Number (for every individual) | Variable depending on number of individuals |
| Age | Patient's Age (at the time of MR scan) | Years |
| M/F | Gender of Patient | Male or Female |
| Hand | Patient's Significant Hand | Left or Right |
| EDUC | Patient's Education Background | Years |
| SES | Socio Economic Status of Patient | Lower, Lower-Middle, Middle, Middle-Upper, Upper |
| MMSE | Mini Mental State Examination Score | 0 – 30<br>0 – more likely to be demented<br>30 – least likely to be demented |
| CDR | Clinical Dementia Rating | 0 – 3<br>0 – normal<br>0.5 – questionable/very mild dementia<br>1 – mild dementia<br>2– moderate dementia<br>3– severe dementia |
| eTIV | Estimated Total Intracranial Volume | 1488±176.13<br>cm3 |
| nWBV | Normalized Whole Brain Volume | 0.730±0.037<br>mg |
| ASF | Atlas Scaling Factor | 1.195±0.138 |
| Delay | Interval between MRI sessions | Days |
| Visit | Visit's ordinal number for the MR session at the testing facility | Days |
| Groups | Type of dementia | Classification is three levels deep<br>Dementia – Person suffers from severe dementia<br>Converted – Following the first examination, the individual has converted to a severe dementia condition<br>Non-Demented – Person does not have dementia |

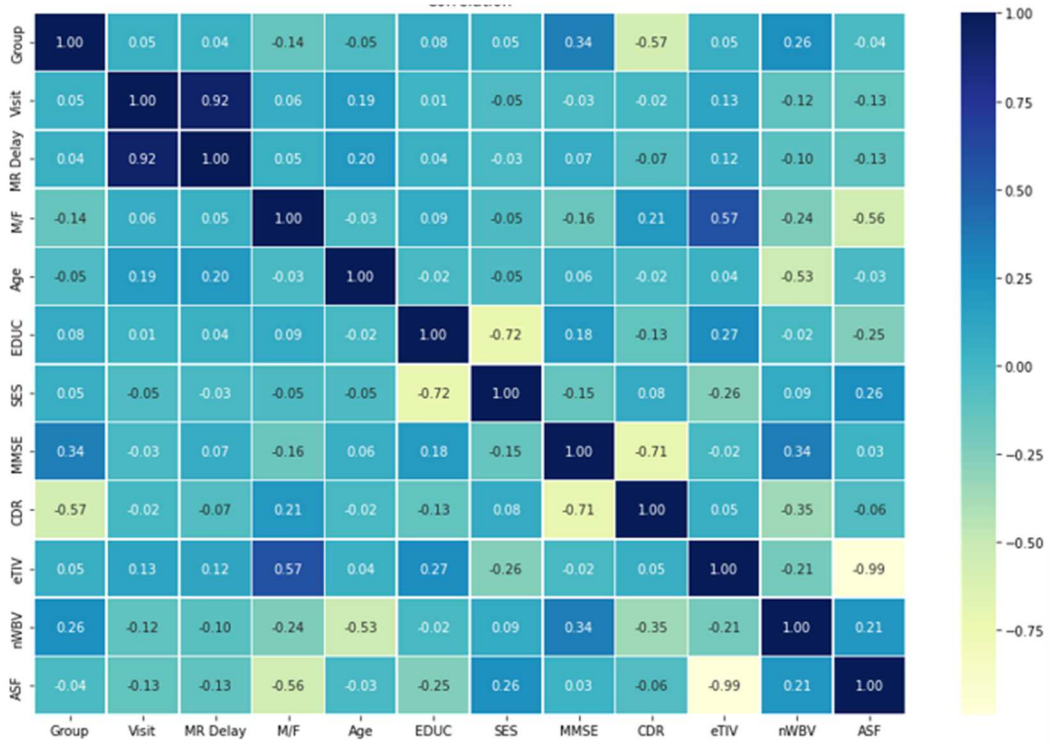

Fig. 1 Correlation Matrix for OASIC-2

Dataset has been further optimized by removing the outliers. This gave the possibly best concise dataset to work on various ML algorithms. The final features considered to train and test the algorithms are as mentioned below.

- Age
- Sex
- Education Background
- Socio Economic Status
- Mini Mental State Examination Score
- Normalized Whole Brain Volume
- Clinical Dementia Rating
- Atlas Scaling Factor
- Estimated Total Intracranial Volume
- Groups

Label encoding of above features with categorical values in string format have been done. Finally, the dataset has been split in train test set with a ration of 3:7 and five models have been evaluated.

### C. *Machine Learning Algorithm and Evaluation*

The refined dataset is then used for designing various ML algorithms. A ratio of 7:3 has been used for training and testing the five ML models. Logistic Regression, Naive Bias, Support Vector, Decision Tree and Random Forest ML algorithms have been designed utilizing 30% of the dataset and the testing of the algorithms have been done using the remaining 70%. Confusion matrix indicating the accuracy of algorithms is shown figure 2.

### D. *Model Deployment*

Based on the performance of the models, the best predictive model has been considered for the developing an application. The application has two main functionalities viz. taking feature values as input from the user and displaying the prediction. The input features are taken through an HTML form, converted into a NumPy array and fed to the Alzheimer detection RF model. The model predicts and provides the output feature. It is then displayed on the application.

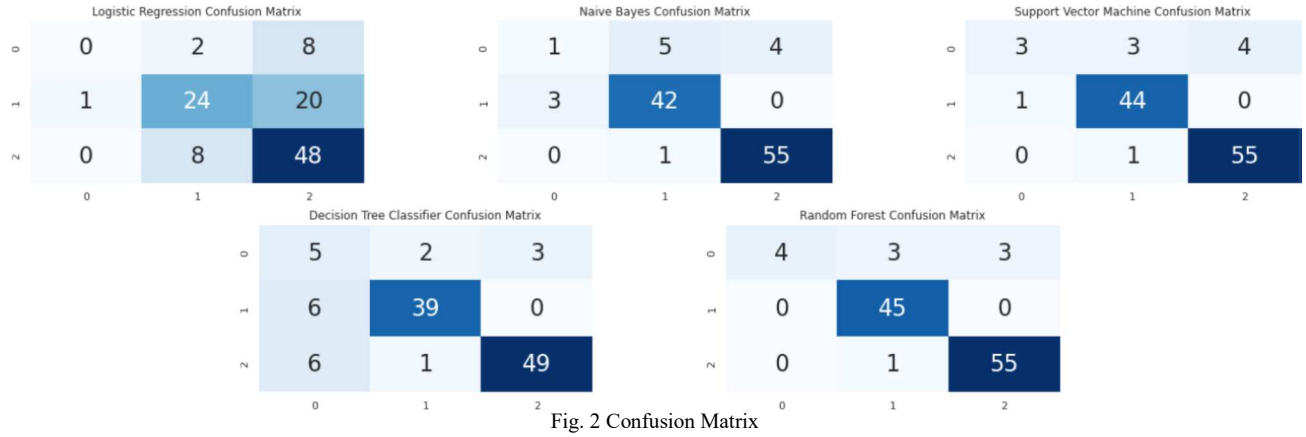


Fig. 2 Confusion Matrix

## IV. RESULT

The work presented in this study explores the application of ML in the field of medical research. The study contributes to the field of neuro degenerative diseases by showcasing the potential application useful to the medical professionals who want to get a quick diagnosis of the patient. Five ML algorithms, Logistic Regression, Naive Bayesian, Support Vector Machine, Random Forest Classifier and Decision Tree have been trained and tested on the optimized OASIS-2 dataset with 30% of dataset for training and 70% for testing. The accuracy rate of the five models is presented in Table VI. A comparison of the accuracy rate for the five algorithms is shown in figure 3. The models are remarkably accurate with the decreased features when compared to other studies. Increasing the features can improve the accuracy. It can be seen that RF provides the highest accuracy followed by SVM, NB, DY and LR.

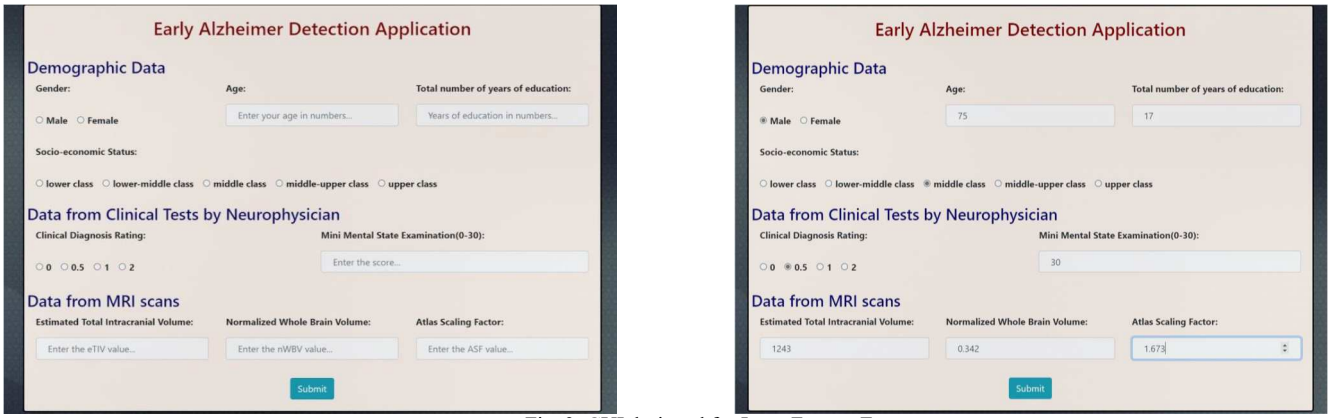


Fig. 3. GUI designed for Input Feature Entry

## V. CONCLUSION

Alzheimer's has emerged globally as disease adversely affecting elderly. The incurable disease can only be slowed down by timely medical intervention and clinical therapy. ML has been employed to detect AD accurately by various researchers. In this paper, comparative studies of ML algorithms have been successfully applied for early detection of AD with RF emerging as highly affective algorithm. The output feature have three classes for describing the individual as Converted, Demented and Non-Demented based on the input features extracted from the correlation matrix. The five models have been evaluated based on this correlation matrix. Random Forest prevailed as the best performing model for AD detection on all the output features with an accuracy of 93.69%. This is the highest accuracy appearing in comparison with all ML models appearing in the other relevant studies. RF has deployment for the GUI application developed using the flask web framework in python for backend and HTML, CSS, Bootstrap for frontend.

TABLE VI. ACCURACY OF MODELS

| Algorithm | Accuracy (%) |
|---|---|
| Logistic Regression | 64.86 |
| Naïve Bayes | 88.29 |
| Support Vector Machine | 91.89 |
| Random Forest | 93.69 |
| Decision Tree | 83.78 |

## VI. FUTURE SCOPE

The proposed work is an effort to evaluate the performance of some basic ML algorithms for early AD detection. However, the variables outside the dataset affecting AD are not considered in this work. This restricts the performance of the models. These variables like family history, accidents, etc. can further enhance the accuracy.

- The study focuses on clinical and demographic data of an individual. This restricts the features used for prediction of AD. Factors such as family history, genetics and other external factors affecting AD can be further investigated to get a more accurate prediction result.
- Additional functionality like a recommendation system and showing the medical history of the patient can be added.

## VII. REFERENCE


[1] Alzheimer's Association, "2016 Alzheimer's Disease Facts and Figures", Alzheimers & Dementia, April 2016, 12(4):459-509.

[2] S. G. Mueller, M. W. Weiner, L. J Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, L. Beckett, "Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI)", Alzheimers & Dementia, July 2005, 1(1): 55-66.

[3] D. K. Lahiri, M. R. Farlow, N. H. Greig, K. Sambamurti, "Current Drug Targets for Alzheimer's Disease Treatment", Drug Development Research, September 2002, 56: 267–281.

[4] R. Sivakani; Gufran Ahmad Ansari, "Machine Learning Framework for Implementing Alzheimer's Disease", 2020 International Conference on Communication and Signal Processing, July 2020.

[5] P. Singh, N. Singh, K. K. Singh, A. Singh, "Chapter 5 - Diagnosing of disease using machine learning", Machine Learning and Internet of Medical Things in Healthcare, 2021.

[6] S. Mishra, A. Dash, L. Jena, "Use of Deep Learning for Disease Detection and Diagosis", Bio-inspired Neurocomputing, July 2020, pp181-201.

[7] A. Sarica, A. Cerasa, A. Quattrone, "Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review", Frontiers in Aging Neuroscience, 2017.

[8] Md. Zahangir Alam, M. Saifur Rahman, M. Sohel Rahman, "A Random Forest based predictor for medical data classification using feature ranking", Informatics in Medical Unlocker, April 2019.

[9] M. Tanveer, B. Richhariya, R. U. Khan, A. H. Rashid, P. Khanna, M. Prasad, C. T. Lin, "Machine learning techniques for the diagnosis of Alzheimer's disease: A review", ACM Transactions on Multimedia Computing, Communications, and Applications, 2020 Apr 15;16(1s):1-35.

[10] M. F. Folstein S. E. P. R. McHugh, "Mini-mental state. A practical method for grading the cognitive state of patients for the clinician", J Psychiatr res. 1975;12(3):189-98.

[11] Y. Zhang, Z. Dong, P. Phillips, S. Wang, G. Ji, J. Yang, T. F. Yuan, "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, June 2015, 2:9:66.

[12] F. C. Morabito, M. Campolo, C. Ieracitano, J. M. Ebadi, L. Bonanno, A. Bramanti, S. Desalvo, N. Mammone, P. Bramanti, "Deep convolutional neural networks for classification of mild cognitive impaired and Alzheimer's disease patients from scalp EEG recordings", In2016 IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow, September 2016, pp. 1-6.

[13] A. Alberdi, A. Aztiria, A. Basarab, "On the early diagnosis of Alzheimer's Disease from multimodal signals: A survey", Artificial intelligence in medicine, July 2016, 1;71:1-29.

[14] L. Liu, S. Zhao, H. Chen, A. Wang, "A new machine learning method for identifying Alzheimer's disease", Simulation Modelling Practice and Theory, February 2020, 1;99:102023.

[15] D. Jha, Ji-In Kim, Goo-Rak Kwon, "Diagnosis of Alzheimer's Disease Using Dual-Tree Complex Wavelet Transform, PCA, and Feed-Forward Neural Network", PubMed 2017;2017:9060124, June 2017.

[16] Bo Xie, Cui Tao, Juan Li, Robin C Hilsabeck, Alyssa Aguirre, "Artificial Intelligence for Caregivers of Persons with Alzheimer's Disease and Related Dementias: Systematic Literature Review", International Journal of Emerging Technologies and Innovative Research, 2019, 2349:5162.

[17] OASIS: Longitudinal: https://doi.org/10.1162/jocn.2009.21407.

[18] Yang Han; Xing-Ming Zhao, A hybrid sequential feature selection approach for the diagnosis of Alzheimer's Disease. International Joint Conference on Neural Networks, 2016 July.