Predicting Early Alzheimers Disease using Classification Models

Ambika Huluse Kapanaiah
Masters of Data Science
University of canberra
Canberra, Australia
u3227622@uni.canberra.edu.au

Shourya Teja Marneni
Masters of Data Science
University of canberra
Canberra, Australia
u3225636@uni.canberra.edu.au

***Abstract*— Alzheimer's is a neurological disorder, which causes brain to shrink and brain-cells to die. Most of the Alzheimer's positive cases will result in dementia and is a fatal one. We can find lots of research papers and analysis done on detecting Alzheimer's using several machine learning/deep learning techniques. In this paper, we are going to detect and predict early Alzheimer's by using Machine learning algorithms such as Logistic regression, Support Vector Machine, and Random Forest. Finally, we are choosing the best model after employing the K-Fold cross validation method and evaluating each model's performance using Accuracy, Precision, Recall and F1 score.**

Keywords—Logistic Regression, Support vector Machine, Random Forest, Accuracy, Precision, Recall, F1-Score, confusion matrix, etc.,

I. INTRODUCTION

Alzheimer's disease is one of the most common brain diseases which worsens over time. It is hard and not so evident to notice the symptoms in the early stages of the disease.

Mild memory loss is the early symptom of this disease and the person with dementia will have a continuous decline in thinking and social skills that affects the ability of the brain to function independently. Over the time the symptoms become predominant as there will be degenerative changes in the brain which causes damage to neurons in the brain that controls the cognitive functions of an individual such as: thinking, learning, walking, planning, etc. This stage is known as dementia. Based on the research, among approximately 50 million people worldwide with dementia, 75% are estimated to have Alzheimer's disease. Medication may temporarily improve or have impact on slow progression of symptoms. Nevertheless, there is no such treatment that cures Alzheimer's disease.

There are several stages in this disease, Mild Cognitive Impairment (MCI) is the prodromal stage. The symptoms may develop in the intermediate stage namely progress Mild Cognitive Impairment (pMCI) and it may also develop at stable Cognitive Impairment (sMCI). Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET) and Diffusion Tensor Imaging (DTI) are the tools used for neuroimaging Alzheimer's. Early diagnosis will concededly decrease the risk of further deterioration.

According to the data acquired after conducting a survey of 150 people in Europe [1], 82% found credence in the importance of early detection of AD and 69% of the people agreed and knows personally those people with this disorder and also believe that early detection would have helped the patient(Fig1). These factors show how essential it is to not only diagnose AD but also to have an early prediction of this disease so that, respective actions in betterment of the patients are taken before it is too late.

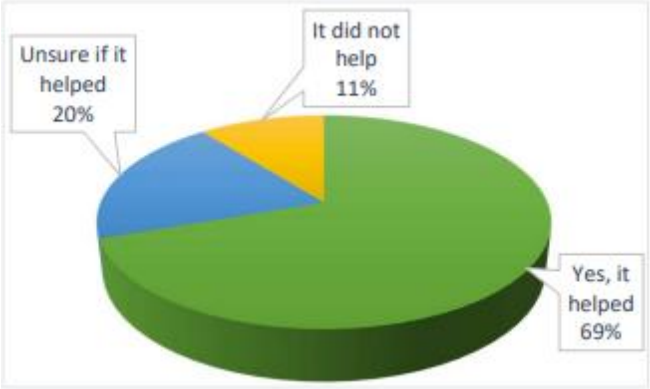


Fig.1 The percentage corresponds to the number of people who believe early detection of AD helped their family/friends according to the survey

II. LITERATURE & REVIEW

We have referred the following research papers and the best model chosen from each author based on the evaluation metrics.

1. *Smart Home Base Prediction of Symptoms of Alzheimer's Disease Using Machine Learning and Contextual Approach* [2]: In this contextual method researchers have come up with random forest model which measures the performance of activities of patients in terms of mobility and mood traits. The Author has compared the SVM and Random Forest models in predicting the symptoms. And Random Forest has given the highest Accuracy of 94.1%, Precision 95.6%, recall of 95.4% and F1 score of 95.2% and choosen as best among both.

2. *Early-Stage Alzheimer's Disease Diagnosis Method* [3]: For the accurate prediction of Alzheimer's disease this research paper has employed Logistic Regression, Decision Tree and SVM. Logistic regression model predicted the results with the Accuracy of 98.12%, Specificity of 95% and Sensitivity of 90% and was the best among others.

3. *A Novel Modelling Technique for Early Recognition and Classification of Alzheimer's disease* [4]: From the experimental results, the classification accuracy obtained when random forest classifier used and is 98.42% and the classification accuracy obtained when Tree Bagger classifier used is 98.17%. The random forest classifier gave more accuracy rate which is been chosen as the best model.

III. METHODOLOGY

In this paper, we have chosen the Longitudinal MRI Data in Nondemented, and Demented Older Adults dataset provided by OASIS. By applying the following Pattern Recognition and Machine Learning algorithms on the dataset, we will be able to predict the early stage of Alzheimer's disease. Here in Alzheimer's data set, our aim is to find the probability of the case as Demented or Non-demented based on the predictors available in the dataset. We have chosen the following three classification methods.

A. Logistic Regression

Logistic regression is a supervised classification machine learning algorithm, used to predict the probability of an outcome. It is basically used for binary classification problems with positive and negative classes (y= {+, -}) or multi-class classification (y= {0,1,2, 3..}).
The logistic regression function can be mathematically expressed as follows.

$$p = \frac{1}{1 + e^{-(b0+b1*x)}}$$

It is a classification model rather than a regression model despite its name. Logistic Regression is more effective and simpler for linear and binary classification problems. This model achieves very good performance with linearly separable classes. This is a default algorithm which is using in the classification industry as base line.

There are Multiple types of Logistic regression:

- Binary Logistic Regression,
- Multinomial Logistic Regression,
- Ordinal Logistic Regression

If we have a model with outputs 0 and 1,

Hypothesis => Z = WX+B,

$h^0(X)$ = sigmoid(Z) then

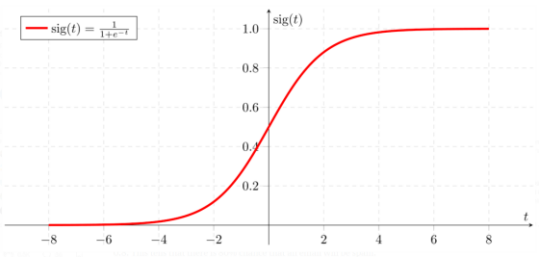


Fig 3. Logistic Regression

If Z moves towards positive infinity the y(predicted) will be assigned 1 and when Z moves towards negative infinity y(predicted) will be assigned 0.

B. Support Vector Machine

SVMs is a supervised machine learning algorithm which can be used for both regression and classification problems. The observations will be separated by a hyperplane in the space to classify the samples into different classes.

Separation of samples is defined based on the maximum margin from each class. In the Fig.4, C is at the maximum distance from both c + b and c - b support vectors.

It is mathematically represented as follows.

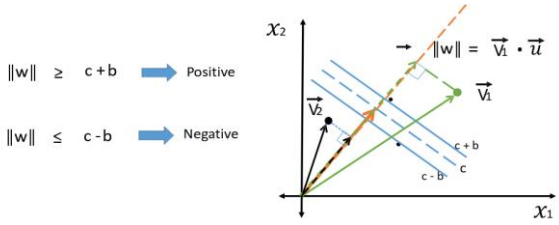


Fig 4. Support Vector Machine

Here are the SVM kernels: Linear, RBF (radial bias function), Sigmoid, and Polynomial functions as in Fig.5.

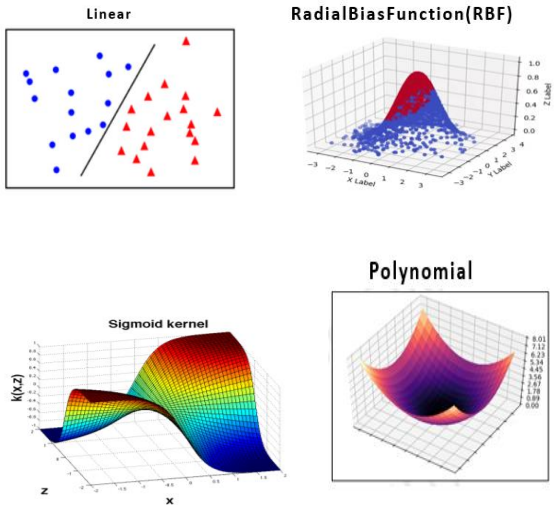


Fig 5. SVM Kernel (a) Linear (b)Radial Bias Function (RBF) (c) Sigmoid (d) Polynomial

C. Random Forest

Random forest works based on ensemble learning method wherein it generates group of base learning algorithms and then combine the results to get the higher accuracy. Each of the base learners can use different parameters, sequence training sets and so on. There are 2 types of ensemble-learning one is bagging; in which various models are built in parallel. All the models vote to give the final prediction. And another one is boosting; in which the decision trees are executed sequentially by correcting the incorrect observation

from the previous model prediction and improvising the learning in every sequence.

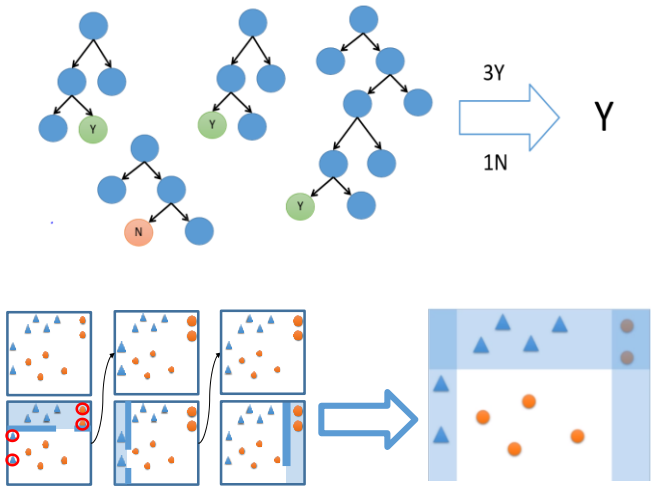


Fig 6. Random Forest

D. Evaluation Metrics:

Evaluating the performance of a machine learning model is very important. There are different evaluation metrics such as Accuracy, precision, recall etc. Using these metrics, we will be able to improve the model's overall prediction power before we deploy it out for production for unseen data.
Here, we are using a dataset where we predict whether a subject is demented or non demented. Using the classification metrics such as Accuracy, Precision and False Negative Rates. We are going to evaluate the three classification models Logistic Regression, SVM and Random Forest. Using Confusion Matrix, we will be calculating these metrics. We concentrate on a model which has least False Negative Rates Where model predicted a True labelled as False i.e., it is the count of Demented subjects which are predicted as Non demented by the model.

	Predicted Class		
Actual Class		**P**	**N**
	P	TP	FP
	N	FN	TN

Tab 1. Confusion Matrix

Confusion matrix: The matrix summarizes the performance of a classification algorithms. It shows the number of correct and number of incorrect predictions with counts predicted by the algorithms by each class.
True Positive: Predicted values correctly predicted as actual positive.
True Negative: Predicted values correctly predicted as an actual negative
False Positive: Count of negative values predicted as positive.
False Negative: Count of Positive values predicted as negative.

Accuracy: It is the ratio of total classes correctly predicted by total observation.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

Precision: It is the ratio of correctly predicted positive outcomes by total number of positive outcomes predicted.

$$Precision = \frac{TP}{TP + FP}$$

Recall: It is the fraction of correctly predicted positive outcomes by total number of actual positive outcomes from the model.

$$Recall = \frac{TP}{TP + FN}$$

F1 Score: It is the combination of precision and Recall. It is the harmonic mean of precision and Recall. If we have an uneven class distributed data, we will concentrate on F1 score rather than accuracy.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

IV Experiment

We have followed the below pipeline in this project.

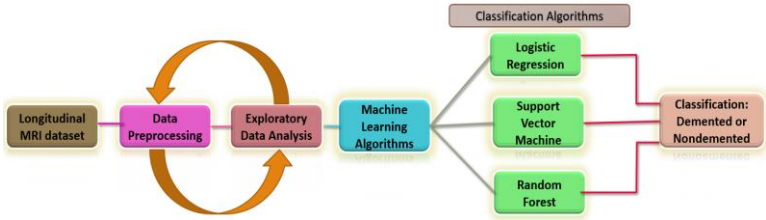


Fig 7. Pipeline of the project

A. Dataset Description:

The Open Access Series of Image Studies (OASIS) is a data resource by the Washington University Alzheimer's Disease Research Centre, Dr. Randy Buckner at the Howard Hughes Medical Institute (HHMI) at Harvard University, the Neuroinformatic Research Group (NRG) at Washington University School of Medicine, and the Biomedical Informatics Research Network (BIRN) made the MRI datasets of brain available for free to be used by scientific community to facilitate future discoveries in basic and clinical neuroscience. In this Project we will be using longitudinal MRI Data.

There are 150 subjects and 373 MRI Sessions at various stages of cognitive decline ranging in age from 60 to 98 years. There are 14 features which includes Subject ID, MRI ID, Group (Demented/ Non-Demented/ Converted), Visit (Number of MRI Scan), MRI Delay(the number of days difference between 2 scans), Male/Female, Hand (Right/Left Handed), Age, Education, Socio-Economic status (SES), Mini mental

state examination score (MMSE), Clinical dementia rating (CDR), Estimated total intracranial volume (eTIV), Normalized whole brain Volume (nWBV) and Atlas Scaling Factor(ASF).

B. Data preprocessing:

The Alzheimer's dataset has 375 observation and 15 features among which 10 are numeric and remaining are categorical variables. There are patients with uneven number of visits i.e., few patients have visited 5 times, few 3 and some only once. To avoid complexity in visits we have picked patients with visit=1 only. For the patients who got confirmation of dementia after multiple visits, the status has been changed to "Converted" and is changed for the visit=1 also. So, there won't be any data loss by filtering the data with patients with visit=1. We have dropped unwanted features such as Subject ID, MRI ID, Visit, MR Delay and Hand from the dataset. We changed the status of the patients with 'Converted' status to 'Demented' as we are considering the patients with visit 1 only. So, we have observation with labels of "Demented" and "Non demented". Using Label Encoder, we have converted the features 'M/F', 'Group' and 'Hand' to discrete values such as in 'Group', observation who are 'Demented' are now 1 and 'non-Demented' are now 0.

There are 8 missing values in SES feature, so we created two datasets. Among the two one is with missing values removal, making the dataset 142 x 13 in shape and in the second one, we have imputed the missing values with mean of the features which make no changes to the shape of the main data frame (the dataset with shape 150 x 13).

C. Exploratory Data Analysis

We performed Exploratory Data analysis to maximize insight into a data set, uncover the underlying structure, and extract important variables. We came to know that:

- More men are demented than women (Fig 8).
- Maximum Demented patients are from 70-80 years of age group.
- Patients with a smaller number of years of education are more demented.
- Patients with higher mini mental examination score (MMSE) are Nondemented.
- Clinical dementia ratings are given properly For Demented it is 0.5 and above and for Nondemented it is 0. (Fig 9)
- Normalized whole brain Volume (nWBV) looks more shrunk in demented patients. (Fig 10)

The Exploratory data analysis of the features 'SEX', 'Age', 'EDUC', 'MMSE', 'CDR', 'eTIV', 'nWBV', 'ASF', and 'SES' have shown strong to medium relationship with the target variable Group (Demented/Non-Demented).

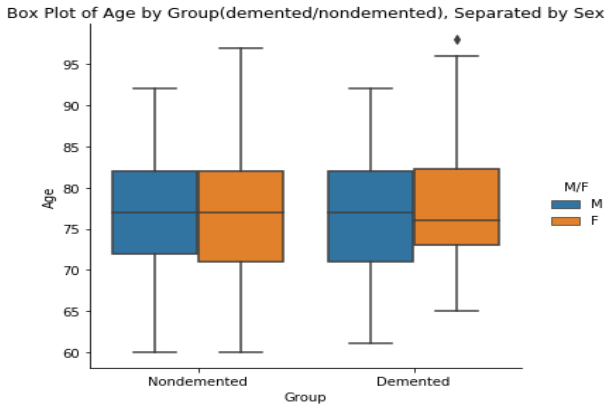


Fig 8. Box plot of age and group based on gender

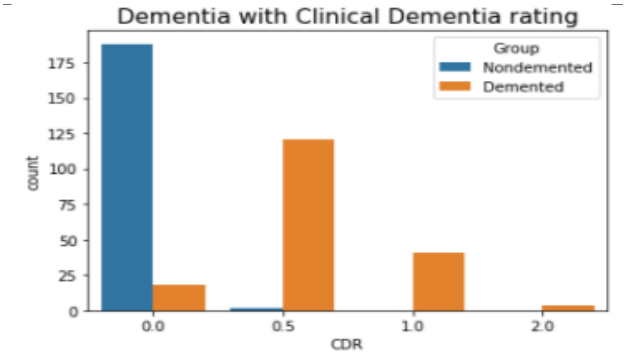


Fig 9. Bar plot for Clinical Dementia Rating (CDR)

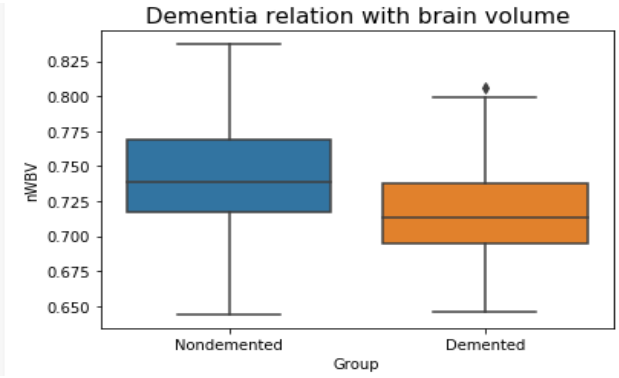


Fig 10. Box on Group and Normalized whole brain volume

V Results

We conducted experiments to check the classification performance metrics of various classifiers such as logistical regression, Support vector machine and Random Forest. The dataset has been divided in 75:25 ratio as training and testing dataset respectively. We used the methods like Grid Search cv and Randomized search CV (SVM) to find best hyperparameter and tuned the model with best one. We also have used 5-fold cross validation to choose the best parameter. The performance evaluation metrics are observed later to check the performance of classifiers.

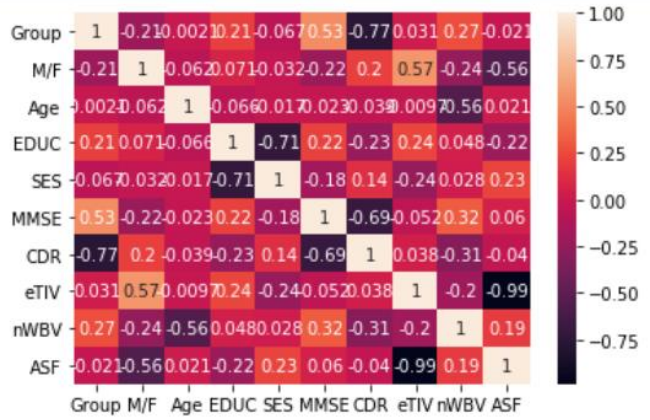


Fig 11. Correlation Heatmap.

Comparison of Results Between Logistic regression, Support Vector Machine and Random Forest Algorithms.

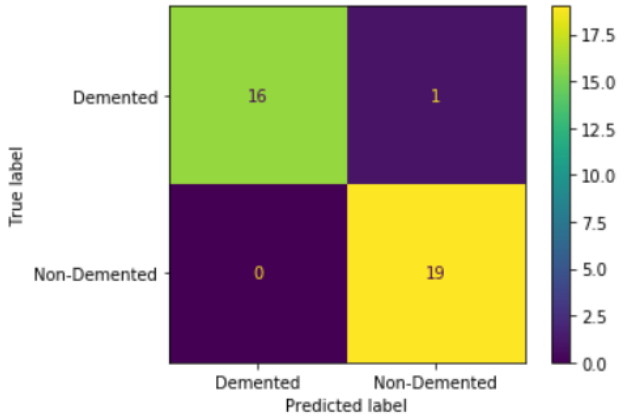
Below are the Best Parameters for the dataset where missing values are replaced with mean of that feature

Classifier	*Best Parameters*	*Best Score*
Logistic	*C = 10*	*0.9019*
SVM	*gamma = 0.01, C = 100, degree = 1, kernel = 'poly'*	*0.8928*
Random Forest	*n_estimator = 10, max_feature = 4, max_depth =6*	*0.9019*

The performance of each model is stated below.

Classifier	*Accuracy*	*Precision*	*Recall*	*F1 - Score*
Logistic	*0.8947*	*0.897*	*0.8947*	*0.8947*
SVM	*0.9473*	*0.952*	*0.9473*	*0.9473*
Random Forest	*0.8947*	*0.9*	*0.8888*	*0.9473*

For this Dataset the best model is Support vector machine. We used Randomized Search CV with 5-fold cross validation and got the best parameters as C = 100, gamma = 0.01, Degree = 1, kernel = 'poly' with the best score of 89.28%. The accuracy of the model is 94.73%, precision 95.2% and recall of 94.73%.



There are 0 False negatives in this model i.e., the model has predicted 0 Demented observation as non-Demented and predicted 1 non-demented observation as Demented making false positives 1.

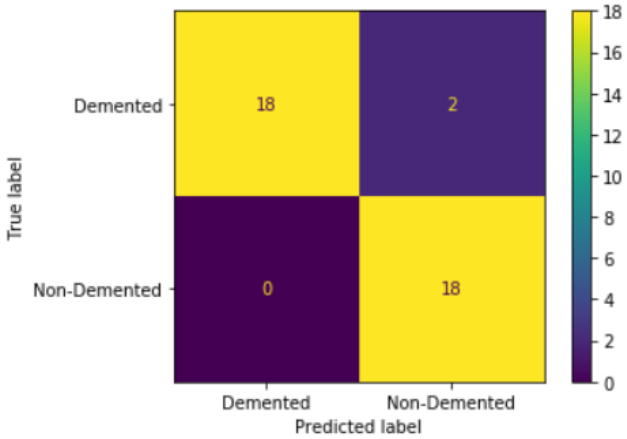
Below are the Best Parameters for the dataset where we have removed the rows of missing values.

Classifier	*Best Parameters*	*Best Score*
Logistic	*C = 5*	*0.848*
SVM	*gamma = 0.01, C = 100, degree = 1, kernel = 'poly'*	*0.8928*
Random Forest	*n_estimator = 4, max_feature = 4, max_depth =1*	*0.8861*

The performance of each model is stated below.

Classifier	*Accuracy*	*Precision*	*Recall*	*F1 - Score*
Logistic	*0.9166*	*0.9182*	*0.9166*	*0.9167*
SVM	*0.9722*	*0.9136*	*0.9722*	*0.9721*
Random Forest	*0.8888*	*0.9411*	*0.8421*	*0.0.8888*

For this Dataset the best model is Support vector machine. We used Randomized Search CV with 5-fold cross validation and got the best parameters as C = 100, gamma = 0.01, Degree = 1, kernel = 'poly' with the best score of 89.28%. The accuracy of the model is 97.22%, precision 91.36% and recall of 97.22%.



There are 0 False negatives in this model i.e., the model has predicted 0 Demented observation as non-Demented and predicted 2 non-demented as demented by the model making False positive 2.

VI Future Work

The current dataset has only the status of dementia of the patient as Demented/Nondemented. And the size of the dataset is less. In the future if we can access the MRI scan data and acquire huge dataset, it would be possible to predict Alzheimer's better than the one which we have employed in this paper. If we able to get MRI scan images, we would be

able to implement deep learning methods such as Convolutional neural networks. This would be very useful in suggesting early medication and avoid further damage to neuron cells.

VII Conclusion

The research work which we have conducted here is on predicting the early Alzheimer's using different machine learning classifiers. The performance metrics have been checked for accuracy and Low False negative rate while predicting early detection of Alzheimer's disease. Three classifiers such as logistic regression, support vector machine and Random forest have been used. Among the above three classifier SVM gave the best performance; 94.73% in terms of accuracy when the missing values are imputed and 97.22% accuracy when missing values are removed. Therefore, we suggest that SVM is the best classifier for the prediction of AD using the Longitudinal MRI dataset.

REFERENCES

[1] Shetty, Ayush, Dikshi Mehta, Pranali Rane, and Shruti N. Dodani. "*Detection and Prediction of Alzheimer's disease using Deep learning: A review*." In 2021 4th Biennial International Conference on Nascent Technologies in Engineering (ICNTE), pp. 1-7. IEEE, 2021.

[2] Alberdi, A., Weakley, A., Schmitter-Edgecombe, M., Cook, D.J., Aztiria, A., Basarab, A. and Barrenechea, M., 2018 "*Smart home-based prediction of multidomain symptoms related to Alzheimer's disease"* IEEE journal of biomedical and health informatics, 22(6), pp.1720-1731.

[3] Memon, M.H., Li, J., Haq, A.U. and Memon, M.H., 2019, December "*Early-stage Alzheimer's Disease diagnosis method*" In 2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing (pp. 222-225). IEEE.

[4] Dinu, A.J. and Manju, R., 2021, May" A *Novel Modelling Technique for Early Recognition and Classification of Alzheimer's disease*" In 2021 3rd International Conference on Signal Processing and Communication (ICPSC) (pp. 21-25). IEEE.

[5] https://www.nia.nih.gov/health/alzheimers-disease-fact-sheet

[6] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[7] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[8] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[9] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

[10] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html