Detecting Early Alzheimer's

Team Members:
Himanshu Chandel, Manish Talekar, Sujhan Das, Aishwarya Kulkarni

Introduction

Alzheimer's disease is unfortunately a very frequent and common disease. Alzheimer's disease (AD) is the most frequent cause of dementia and is a neurodegenerative disorder with unknown causes and pathophysiology that mostly affects older people. Selective memory impairment is the initial clinical manifestation of Alzheimer's disease, and while therapies exist to alleviate some symptoms, there is presently no cure. Patients with probable Alzheimer's disease are evaluated with magnetic resonance imaging (MRI). Localized and generalized shrinking of brain tissue can be seen on MRI scans. According to certain research, MRI features may be able to anticipate the rate of decline in Alzheimer's disease and may help guide therapy in the future. However, physicians and researchers will need to apply machine learning techniques to reliably forecast a patient's progression from mild cognitive impairment to dementia to get to that point.

It is estimated that over 5 million people in the United States currently suffer from the stage of Alzheimer's disease known as dementia, which equates to more than one out of every nine people over the age of 65. According to studies, the population of 5 million people in the United States is predicted to triple in the next 20-30 years. So early diagnosis allows for prompt access to medications and medical attention. Through this project, we propose creating a sound model that can assist clinicians in doing so and predicting early Alzheimer's disease. This will help clinical experts to identify treatments and therapies beforehand.

There are several benefits of detecting Alzheimer's beforehand. The medications are often more effective in early Alzheimer's. This is because their effectiveness is quite limited and often seems to result in maintaining the person's current functioning, and thus, slowing the disease process, rather than reversing the symptoms. This also helps the patient buy more time to plan for medical and financial decisions.

Dataset Description

The dataset is generated by the **Open Access Series of Imaging Studies (OASIS**) that makes neuroimaging datasets available to the scientific community. We had chosen the OASIS-3 dataset which is their latest release. For the access of the dataset, we signed up and waited a week to get access to the OASIS-3 project dataset. The OASIS-3 project has longitudinal neuroimaging, clinical, cognitive, and biomarker dataset for normal aging and Alzheimer's Disease.

In OASIS-3, retrospective data has been compiled for more than 1000 participants that were collected across several ongoing projects over the course of 30 years. The Participants or Subjects of these projects includes 609 cognitively normal adults and 489 individuals at various stages of cognitive decline ranging in age from 42 to 95 years of age. For the anonymity of all participants, they were assigned a new random identifier and all dates were removed and normalized to reflect days from entry into study. The dataset contains over 2000 MR sessions. Many of the MR sessions are accompanied by volumetric segmentation files produced through ‘Freesurfer’ processing. The parent datasets are ADRC, ‘FreeSurfers’ MR Sessions contributed to the determining columns of our dataset. All data is available via www.oasis-brains.org.

Age upon admission, height, weight, and CDR evaluations are all included in the OASIS datatype "ADRC Clinical Data". Clinicians did a diagnostic impression intake and interview as part of the evaluation, which resulted in a coded dementia diagnosis that was recorded in the OASIS datatype "ADRC Clinical Data" dx1-dx5. "Cognitively normal," "AD dementia," "vascular dementia," and contributory variables such as vitamin deficiency, alcoholism, and mood disorders are all diagnoses for this variable. The columns MMSE, CDR, APOE, dx1 were taken from this dataset.

In OASIS-3, there are a total of 2168 MR sessions with a range of scan types to assess structure, vascular integrity, and functional networks. Pictures were used for volumetric segmentation using FreeSurfer, and 1912 segmentations were approved for inclusion in OASIS-3 after passing quality inspection. MR session contributed to our main dataset by contributing the Age, Education, Race and Gender(M/F) columns.

The whole brain, whole cerebral cortex, cortical white matter, and subcortical gray volumes included with the OASIS-3 release were calculated using FreeSurfer segmentations. Cortical reconstruction and volumetric segmentation of T1-weighted images were performed on all MR imaging sessions using the FreeSurfer image analysis package, which is described and publicly available online (http://surfer.nmr.mgh.harvard.edu/). The IntraCranialVol, CortexVol, TotalGrayVol, CorticalWhiteMatterVol and TOTAL_HIPPOCAMPUS_VOLUME columns were obtained from this dataset.

Column Description

IntraCranialVol:

The intracranial volume is the total volume of the brain, meninges, and cerebrospinal fluid within the cranium. Some of the research studies we looked at suggested that a larger brain volume provides a stronger cerebral reserve against the effects of Alzheimer's disease (AD), retaining cognitive performance in the presence of neurodegeneration and so delaying symptom onset. In men, the total intracranial volume was 1469 +/- 102 cm3, while in women it was 1289 +/- 111 cm3.

CortexVol:

Cortical volume represents the amount and size of neurons, dendritic processes, and glial cells. Alzheimer's disease typically destroys neurons and their connections in parts of the brain involved in memory, including the cortex and hippocampus.

TotalGrayVol:

Grey matter and white matter are two types of tissue that make up the brain's central nervous system. Alzheimer's is a grey matter disease whereas white matter plays a crucial role in the progression of the disease. Most of the brain's neuronal cell bodies are found in grey matter. Regions of the brain involved in muscular control and sensory perception are found in grey matter.

CorticalWhiteMatterVol:

White matter is present in the brain's deeper tissues (subcortical). It comprises nerve fibers (axons), which are nerve cell extensions (neurons). White matter has an impact on learning and brain functioning by altering action potential distribution, acting as a relay, and coordinating communication across different brain regions.

TOTAL_HIPPOCAMPUS_VOLUME:

The hippocampal formation is a complex brain structure located deep within the temporal lobe. It plays a crucial function in memory and learning. T the hippocampus is regarded to be primarily involved in storing long-term memories and making those memories resistant to forgetting. It is also suggested to play a function in navigation and spatial processing. It is one of the big factors use to determine the phase of Alzheimer.

Mmse:

The MMSE (Mini Mental State Examination) is a 30-point exam that assesses one's capacity to think (or "cognitive disability"). The MMSE is a step toward a diagnosis if someone has reason to believe they are developing Alzheimer's disease or another dementia. Researchers studying Alzheimer's disease utilize the test to determine a person's level or stage of dementia. It is the most extensively used test for determining whether someone has dementia. The following are the results of the test: time and place orientation (knowing where you are or day of the week) – memory (short-term and recall); ability to pay attention and solve difficulties (like spelling a simple word backwards); linguistics (identifying common objects by name).

Apoe:

The *APOE* gene is involved in making a protein that helps carry cholesterol and other types of fat in the bloodstream. Everyone has two copies of the APOE gene: people with E2/E2 have the lowest overall risk for Alzheimer's and those with E4/E4 have the highest risk.

Cdr:

The CDR (Clinical Dementia Rating Scale) is calculated using a semi-structured interview with the subject and the caregiver (informant) as well as the clinician's clinical assessment. Memory, orientation, judgment, and problem solving, community affairs, home and hobby performance, and personal care are some of the cognitive and behavioral domains that are tested. The CDR is based on a scale of 0–3: no dementia (CDR = 0), questionable dementia (CDR = 0.5), MCI (CDR = 1), moderate cognitive impairment (CDR = 2), and severe cognitive impairment (CDR = 3).

Group:

The patients belonging to the dataset are divided into the following categories: Normal Cognition, Moderate Dementia, Mild Dementia and Severe Dementia according to their cognitive status.

MMSE	Group
25-30 points	Normal Cognition
21-24 points	Mild Dementia
10-20 points	Moderate Dementia
9 points or lower	Severe Dementia

Data Preparation

Data preprocessing and preparation included checking the data, removing the missing and null values in dataset. We also worked on classifying the data into categories like converting Male/Female to 0/1. Correlations were computed and feature selection for models was done accordingly. Also, the OASIS has very friendly data selection website such as there was option to edit column, join columns, select columns that was very useful.

Imputation**:** We found that there are 3 columns (Apoe, Age, Education) which have missing values and hence we replaced them by using the most frequent values of each column.

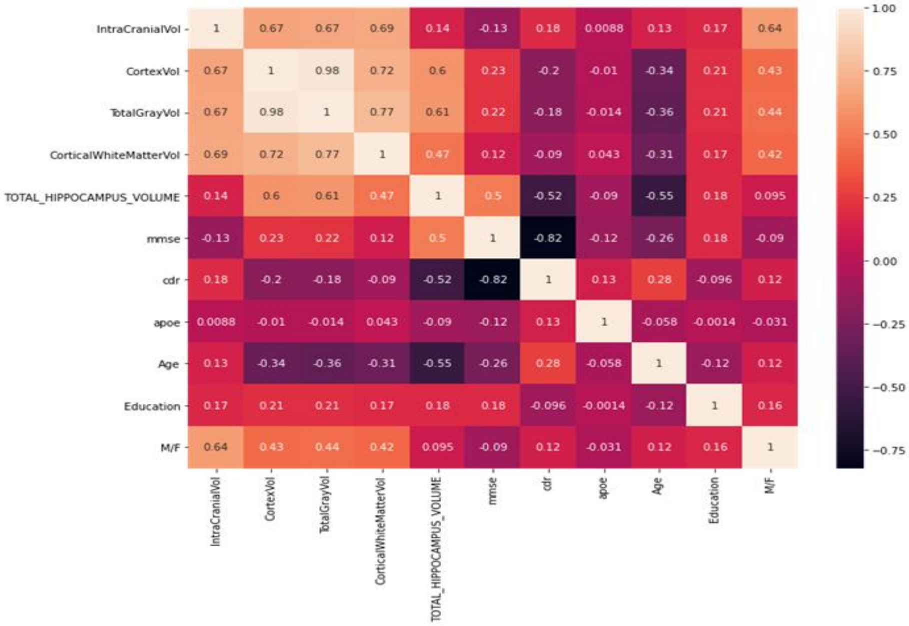
Problem Statement

The MMSE has a maximum score of 30 points. The scores are grouped as follows:
Focusing on the score we have explored using 4-class classification models keeping it as our dependent variable and a good criterion to determine insightful relationships between the columns.

MMSE	Group
25-30 points	Normal Cognition
21-24 points	Mild Dementia
10-20 points	Moderate Dementia
9 points or lower	Severe Dementia

Exploratory Data Analysis

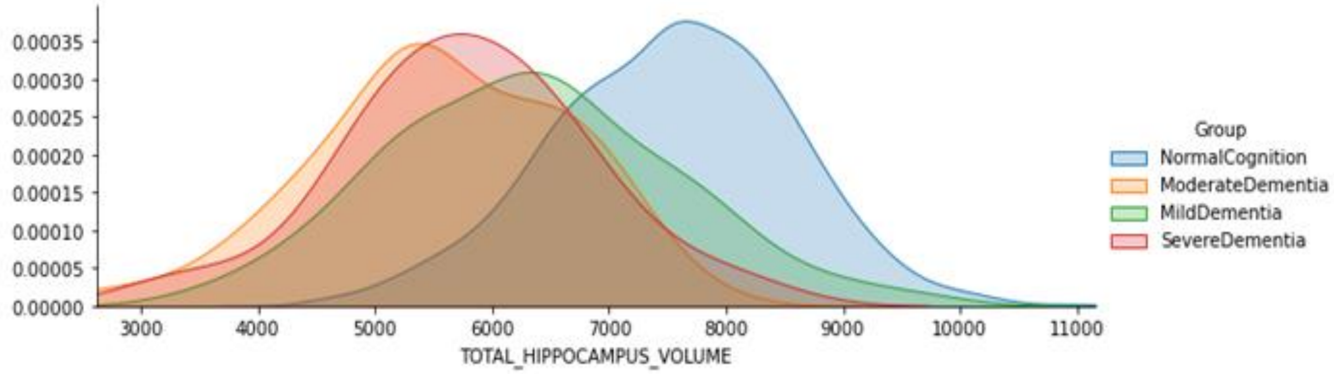
The Correlation heatmap was extremely useful in determining which columns are strongly correlated-



Now, picking up the most correlated columns with MMSE are Total Hippocampus Volume and Cdr. The column Age, Cortex Volume and total gray volume are also showing some average correlation with MMSE.

Among all these the Hippocampus Volume is one of the factors that helps to detect the early Alzheimer.

When we check out how the distinct groups have the distribution of hippocampus volume, we reconfirmed that the people having high MMSE values also have more hippocampus volume hence low chance of AD.



Modelling

We started training our model with different machine learning techniques to find the best model which can be trained on our training dataset giving highest accuracy on the test dataset. The team's primary intention was to explore how machine learning can make a difference in the clinical environment. For that we have developed several algorithms which perform at a good accuracy level. Below are the different types of models we ran,

Model Name	Accuracy (%)
Deep Learning Model (SGD)	91.30%
Change of Solver optimizer in MLP	89.30%
MLP	89.10%
Random Forest Classifier	88.46%
SVM	88.46%
Logistic Regression	88.07%
MLP with early stopping	87.80%
Decision Tree	87.47%
AdaBoost	87.07%

After running above models, we found the SGD as the best model with highest accuracy on the test dataset.

1) Logistic Regression:

Logistic Regression is a supervised learning classification algorithm used to predict the probability of a target variable. Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classifications problems such as Alzheimer detection and many other.

The Y-of the Logistic Regression

During our Alzheimer detection experiment We have used *"Group"* as a target variable which is a multi-class variable having following classes:

- NormalCognition
- MildDementia
- ModerateDementia
- SevereDementia

Hence, we have used One-vs-rest (OvR) method which is a heuristic method for using binary classification algorithms for multi-class classification problems.

The X-of the Logistic Regression

- IntraCranialVol
- CortexVol
- TotalGrayVol
- CorticalWhiteMatterVol
- TOTAL_HIPPOCAMPUS_VOLUME
- cdr
- apoe
- Age
- Education
- M/F

Results using Logistic Regression: The accuracy of the model is 88.07% (Coefficient of determination) which says that 90.45 % accurately the observed outcomes are replicated by model.

2) Random Forest Classifier:

This works by using a collection of multiple random decision trees and it's much less sensitive to the training data and hence a change in training data creates low variance. Here, we use multiple trees and so it is called *Forest.* Random forest helps reduce overfitting in decision trees and helps to improve accuracy. This technique has a capability to focus both on observations and variables of a training data for developing individual decision trees and take maximum voting for classification.

Training Model

Here we are building new data sets randomly from our original data by keeping the same number of rows. In our experiment we are creating random trees with estimators, random feature selections and then feeding to our RandomForestClassifier, values for number of trees, depth of trees and number of features are obtained from loops. After the model is built, we are performing k-fold cross validation in our experiment its 5-fold cross validation. After this we are computing mean cross-validation accuracy and storing better scores.

These steps are repetitively performed in each iteration of for loop.

Rebuilding model

In this step we rebuild a model on the combined training and validation set then we predict model accuracy based on best accuracy on validation set, best parameters and Test accuracy with best paraments

Results:

Best accuracy on validation set is: 92%
Best parameters of M, d, m is: 12 ,5, 8
Test accuracy with the best parameters is: 88.46%
This shows that our model can predict with 88 % accuracy given patients data.

3) Support Vector Machine (SVM):

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The algorithm creates a line or a hyperplane which separates the data into classes. SVM works well when there is a clear margin of separation between classes.

Training Model

In our model, we used a for loop to iterate over many values to get the best value for our parameters: **C**, **k,** and gamma. A range of values (0.001, 0.01, 0.1, 1, 10, 100, 100) were tried and tested for each of the parameters. The C value controls the penalty of misclassification. A large value of C would result in a higher penalty for misclassification and a smaller value of C would result in a smaller penalty of misclassification. With a larger value of C, a smaller margin will be accepted if the decision function is better at classifying all training points correctly. A large value of the C parameter suited our model.

Rebuilding model

The model was rebuilt with the best parameters obtained to get required accuracy and test scores.

Results: -

Best accuracy on cross validation set is: 92.4%
Best parameter for c is: 1000
Best parameter for gamma is: 0.1
Best parameter for kernel is: rbf
Test accuracy with the best parameters is 88.4%

4) AdaBoost:

AdaBoost was the first truly successful boosting algorithm created specifically for binary classification. Adaptive Boosting, or AdaBoost, is a prominent boosting strategy that combines several "weak classifiers" into a single "strong classifier." The weak learners in AdaBoost are decision trees with a single split, called decision stumps. AdaBoost works by putting more weight on difficult to classify instances and less on those already handled well. AdaBoost allows us to capture many of these non-linear relationships, which translates into better prediction accuracy on the problem of interest. Most of the research papers we read used this model for predictions which intrigued us and thus led to us exploring how well this model will do compared to others. Interestingly, other models performed much better than this traditional approach.

Training Model

We again iterate over loops to obtain the best values for our model parameters: n_estimators and learning_rate. The maximum number of estimators at which boosting is terminated is an important parameter as it determines the learning procedure. Learning rate is defined as the weight applied to each classifier at each boosting iteration. A higher learning rate increases the contribution of each classifier. There is a trade-off between the learning_rate and n_estimators parameters. A cross validation procedure is performed thereafter. After this we rebuild the model on the combined training and validation set.

Results:

Best accuracy on validation set is: 0.9091
Best parameter of M is: 4
Best parameter of LR is: 0.1
Test accuracy with the best parameter is 0.8707

Conclusion:

UNIQUE APPROACH
The uniqueness of our approach is the fact that we would be including metrices like MMSE, TOTAL_HIPPOCAMPUS_VOLUME, CDR, Education also in our model to train it to differentiate between normal healthy adults and those with Alzheimer's. MMSE is one of the gold standards for determining dementia and hence we think it is an important feature to include. The same fact also makes our approach flexible enough to be applied to other neurodegenerative diseases which are diagnosed using a combination of MRI features and cognitive tests.

Challenges faced:

Data retrieval and processing was a time-consuming obstacle we had to overcome. After finally getting access to the data from the organizations, attaining the data columns that were relevant to our study was also a tedious process. We required domain knowledge about the subject matter which led us to read several research papers to conquer our disadvantage of not belonging to this field or having relevant backgrounds in medical science. This was crucial to help us determine the important columns that we needed to consider from the parent datasets.

Based on the data that we had, we didn’t have enough data for each data point in the MMSE column. Therefore, grouping the data in the column helped us to classify them into 4 groups for better prediction. Another hiccup in our project was determining the y or predictor column. We settled for MMSE rather than CDR because of the nature and quality of the dataset.

Future Scope:

If we can achieve access to datasets containing images of the actual MRI scans, we can treat it as an image classification problem. This will make the detection a smoother process. Deep-learning-based approaches can be proposed for the classification of neuroimaging data related to Alzheimer's disease (AD). One clear criticism of our project is the requirement of the expert use of design techniques to extract informative features from images. One solution to this issue is using end-to-end learning where all steps in the processing pipeline are simultaneously optimized, potentially leading to optimal performance.

Sources:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4265726/

https://www.nia.nih.gov/health/alzheimers-disease-genetics-fact-sheet

https://www.oasis-brains.org/#data

https://towardsdatascience.com/detecting-precursors-of-alzheimers-by-utilizing-deep-learning-a6de0ee0e2d2

https://www.iwh.on.ca/what-researchers-mean-by/cross-sectional-vs-longitudinal-studies

https://chronicdata.cdc.gov/Healthy-Aging/Alzheimer-s-Disease-and-Healthy-Aging-Data/hfr9-rurv

https://www.sciencedirect.com/science/article/pii/S2352873719300393

https://www.medrxiv.org/content/10.1101/2019.12.13.19014902v1

https://aramislab.paris.inria.fr/clinica/docs/public/latest/Converters/OASIS3TOBIDS/

https://www.oasis-brains.org/

https://www.kaggle.com/jboysen/mri-and-alzheimers

https://www.kaggle.com/hyunseokc/detecting-early-alzheimer-s