Classification of Alzheimer's Disease State Using Grouped Demographic and MRI Data

Jason Hooker | December 7, 2021
Brown DSI
https://github.com/jchooker18/mri-alzheimers-classification

1 Introduction

Alzheimer's disease is the leading cause of dementia, which refers to the loss of memory and other cognitive abilities to the degree that they affect daily life. The strongest known risk factor for the development of Alzheimer's is older age - people aged 65 or older account for the majority of Alzheimer's cases [1]. Ongoing research examines the effect of the disease on the brain and seeks to identify cases of Alzheimer's through brain imaging processes. This project attempts to develop a tool that will be able to classify a previously unseen patient's level of dementia based on demographic information and data collected from MRI scans.

The data used for this project comes from the Open Access Series of Imaging Studies (OASIS) Brains project [2]. The dataset used is "OASIS-2: Longitudinal MRI Data in Nondemented and Demented Older Adults." 150 subjects between the ages of 60 and 96 participated in the study, with each being scanned on at least two separate occasions spanning at least one year, generating a total of 373 data points in the set. Both men and women participated in the study, and all subjects were right-handed. For each imaging session, each subject was assigned a Clinical Dementia Rating (CDR) of either 0 (Normal), 0.5 (Very Mild Dementia), 1 (Mild Dementia), or 2 (Moderate Dementia). The tool created in this project will predict a CDR classification of 0, 0.5, or 1 - patients with a CDR level of 2 will be excluded, as only three patients reached this level over the course of the study. A short description of each of the 15 features in the dataset is provided in Table 1.

The OASIS Brains project datasets have been fairly widely analyzed by Kaggle users. Most, however, have either used a different dataset or have addressed slightly different machine learning questions. From among the top analyses on Kaggle, user Def Me(X) approaches the dementia prediction question without accounting for the group structure introduced into the dataset by having multiple MRI scanning sessions per subject. This spreads subjects across train-validation-test sets, likely producing accuracy (~70%) that may not be replicated on previously unseen subjects [3]. The team of Choi et. al. account for the group structure by removing all but the first MRI scanning session for each subject, thereby reducing the size of their dataset from 373 points to 150 (several accuracy results reported from 75-84%) [4]. User Sheryas P J appends the OASIS-2 and OASIS-1 datasets and does not account for the group structure, potentially inflating accuracy (~80%) [5]. Overall, the goal for this project is to utilize the inherent group structure in the chosen dataset to produce a tool that more would more reliably predict the CDR level for previously unseen patients upon deployment.

Subject ID	Unique identifier for subject
MRI ID	Unique identifier for scanning session
Group	Nondemented, demented, or converted (began nondemented, became demented)
Visit	Number of subject's scanning session (1-5)
MR Delay	Days since subject's first scanning session
M/F	Gender (male or female)
Hand	Handedness
Age	Age in years
EDUC	Years of education
SES	Socioeconomic status (1: highest, 5: lowest)
MMSE	Mini-Mental State Examination score (range is from 0 = *worst* to 30 = *best*) (Folstein, Folstein, & McHugh, 1975)
CDR	Clinical Dementia Rating (0 = *no dementia*, 0.5 = *very mild AD*, 1 = *mild AD*, 2 = *moderate AD*) (Morris, 1993)
ASF	Atlas scaling factor (unitless). Computed scaling factor that transforms native-space brain and skull to the atlas target (Buckner et al., 2004)
eTIV	Estimated total intracranial volume (cm^3) (Buckner et al., 2004)
nWBV	Normalized whole-brain volume, expressed as percent of all voxels in the atlas-masked image that are labeled as gray or white matter by the automated tissue segmentation process (Fotenos et al., 2005)

Table 1 Feature Descriptions [6]

2 Exploratory Data Analysis

This section presents a selection of the figures created during exploratory data analysis.

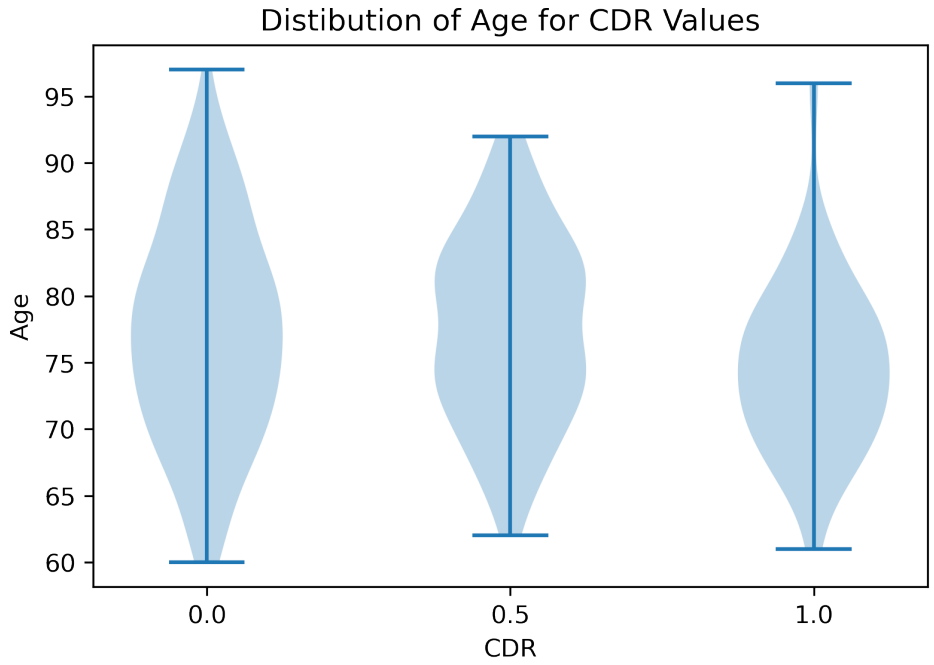

Figure 1 This figure shows the kernel density plot for the age of subjects at each visit. Surprisingly, given age is the greatest known risk factor for dementia, the mean ages across each level are quite similar. If anything, higher levels of CDR appear to be more concentrated among younger subjects than those who were older at the time of scanning. This is likely due to the relatively small number of subjects in the study that produced this dataset. This does, however, essentially control for age and allow the model created in this project to depend on other features more heavily.

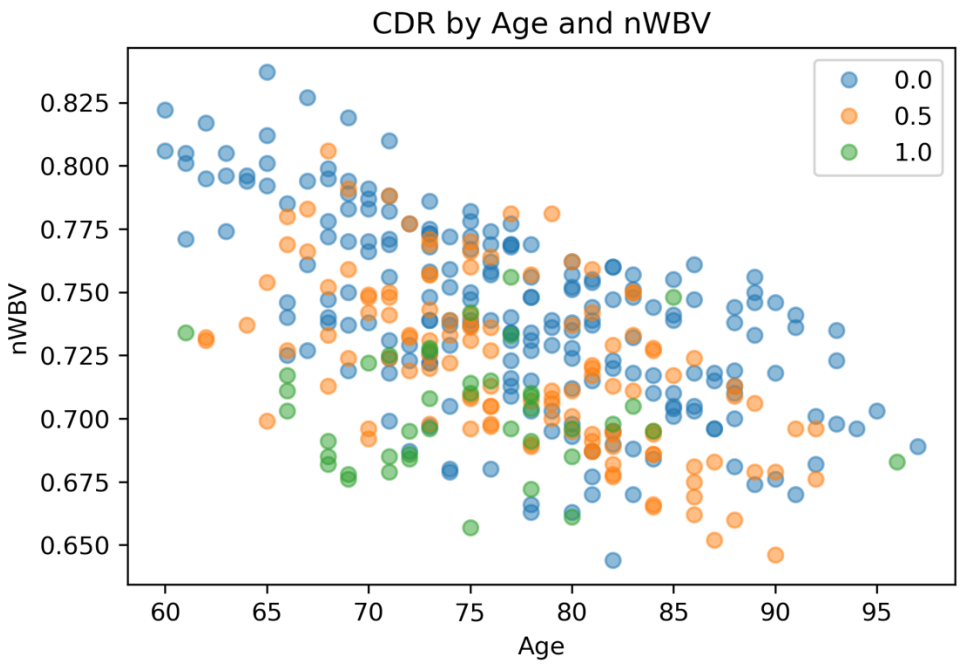

Figure 2 This figure displays normalized whole-brain volume (nWBV) as a function of age, with the color of each point representing CDR classification for that scanning session. There may be a negative correlation between nWBV and age, meaning brain volume may typically decrease with age. The three CDR levels are fairly equally represented across all ages, though there may be a trend of 0 CDR points being toward the top of the clustering, 0.5 CDR points being toward the middle, and 1.0 CDR points being near the bottom. This does not hold for every point, but suggests a correlation between decreasing nWBV and increasing CDR levels.

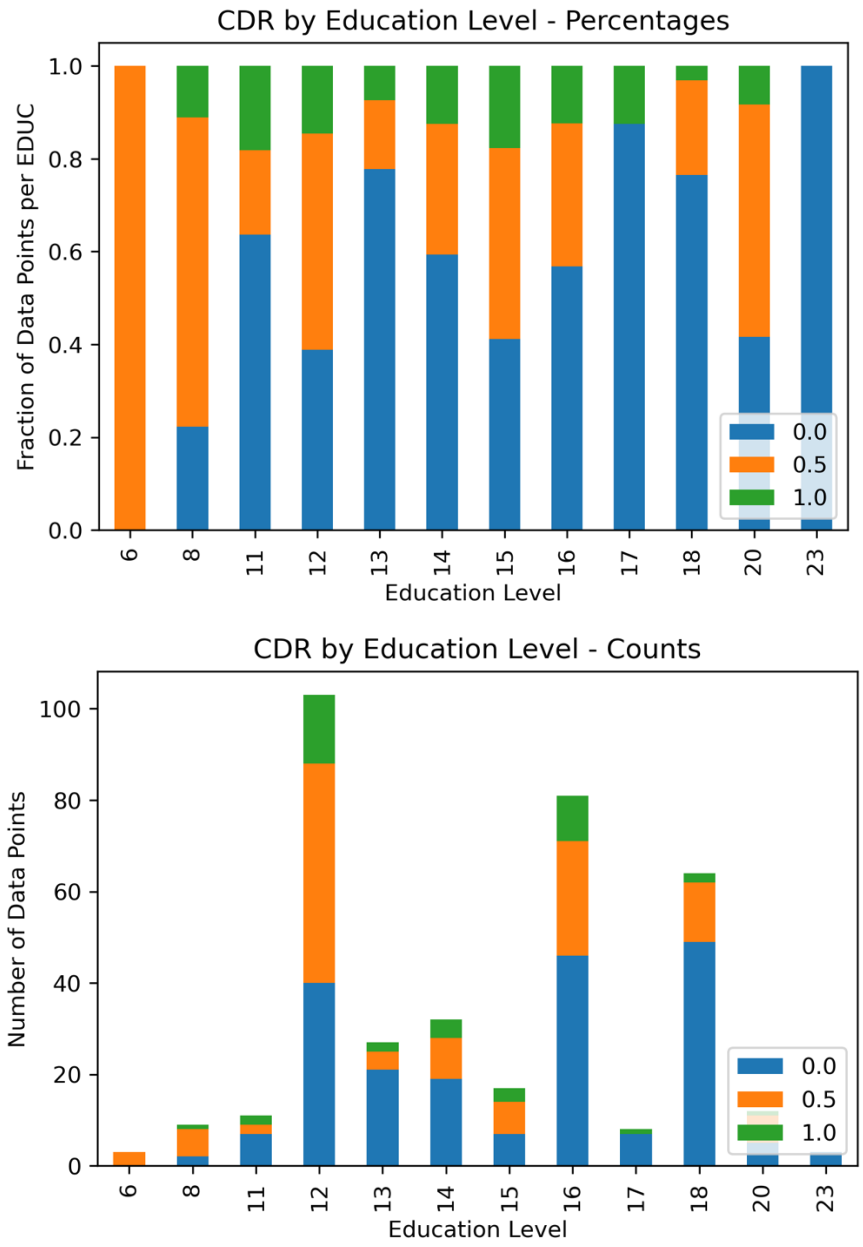

Figure 3 This figure displays the relationship between years of education and CDR (percentages and raw counts per education level). The top chart shows what may be a relationship between higher levels of education and lower levels of CDR – it is important, however, to keep in mind the number of data points present for each education level. The bottom chart shows that most points fall within 12, 16, or 18 years of education. Low CDR appears to increase and high CDR decreases across those three levels.

3 Methods

The dataset is not independent and identically distributed (IID), as it has both a group structure and a time-series component. For this project, the time-series component was removed by excluding the MRI ID, Visit, and MR Delay features. The group structure is based on each subject undergoing at least two imaging sessions, so the data points for each subject must be present in only one of the training, validation, or test sets to prevent data leakage, given the attempt to classify a *previously unseen* patient's CDR. The Subject ID and Group features were removed so as not to introduce bias into the model by utilizing features that will be unavailable to it when deployed, and the Hand feature was excluded because it contains the same value for all data points. Two rows contained missing values for MMSE, and those rows were dropped for preprocessing.

The data was split by utilizing GroupShuffleSplit to assign 20% of subjects to the test set, then performing four-fold cross-validation using GroupKFolds to partition the rest of the data into four different pairs of training and validation sets. Given the small size of the dataset, utilizing cross-validation helped to estimate the uncertainty due to the random splitting of points into training and validation sets. These methods yield a 60-20-20 train-val-test split, and checks were performed to ensure that each value of CDR was adequately represented in each of the sets, given that a relatively low percentage (11%) of the data points have a CDR of 1. The OrdinalEncoder was applied to EDUC, as education level is easily ranked. The M/F and SES features were preprocessed using the OneHotEncoder - gender cannot be ranked, and there were missing values for SES, which can be treated as their own category. The Age, MMSE, and nWBV features are continuous and reasonably bounded and therefore used the MinMaxScaler, and the remaining continuous features eTIV and ASF used the StandardScaler. Lastly, the target variable was encoded to give labels 0, 1, and 2. After preprocessing, the dataset contained 14 features.

Seven different machine learning models were trained on the preprocessed data: logistic regressions with L1, L2, and elastic net regularizations; random forest classifier; support vector machine classifier; and XGBoost classifier. The hyperparameters for each model presented in table 2 were tuned using a grid search brute-force approach, finding the parameter combination yielding the best validation score for each model. Accuracy was the evaluation metric, as we care about correctly predicting each CDR class evenly. The test score was calculated on the holdout set for each model using the best parameters, and this whole process (including splitting and preprocessing the data) was repeated using 20 different random states so as to measure uncertainties to randomness in splitting and non-deterministic methods used in modeling.

Model	Hyperparameters & Values Tried
Log (L1)	**C:** 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2, 1e3, 1e4
Log (L2)	**C:** 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2, 1e3, 1e4
Log (EN)	**C:** 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2, 1e3, 1e4; **l1_ratio:** 0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 0.99
RF	**max_depth:** 1, 3, 10, 30, 100; **max_features:** 0.5, 0.75, 1.0
SVC	**gamma:** 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2, 1e3; **C:** 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2, 1e3
KNN	**n_neighbors:** 1, 3, 10, 30, 100; **weights:** uniform, distance
XGBoost	**max_depth:** 1, 2, 3, 10, 30, 100*

Table 2 Hyperparameters tuned for each model
* Other hyperparameters were explored but did not increase accuracy, and so were left out to decrease computational complexity

Based on the results of our model comparisons (discussed in next section), the final model was a random forest model with max_depth = 3 and max_features = 0.75. To calculate final accuracy and baseline scores and measure feature importances, this model was trained on new data splits using GroupShuffleSplit to allocate 80% of data for training and 20% for testing in each of 100 different random states. Averages were taken across all random states to determine final scores and feature importances.

4 Results

Figure 4 shows the mean accuracy test score for each model over the initial 20 random states. The initial baseline score was 0.54, corresponding to accuracy if the most common CDR, 0, were predicted for each test case. All scores were above baseline, and the random forest models had the highest average score at approximately 0.72. The standard deviations of the test scores, representing uncertainty due to pipeline process randomness, are represented by the black bars in the figure.

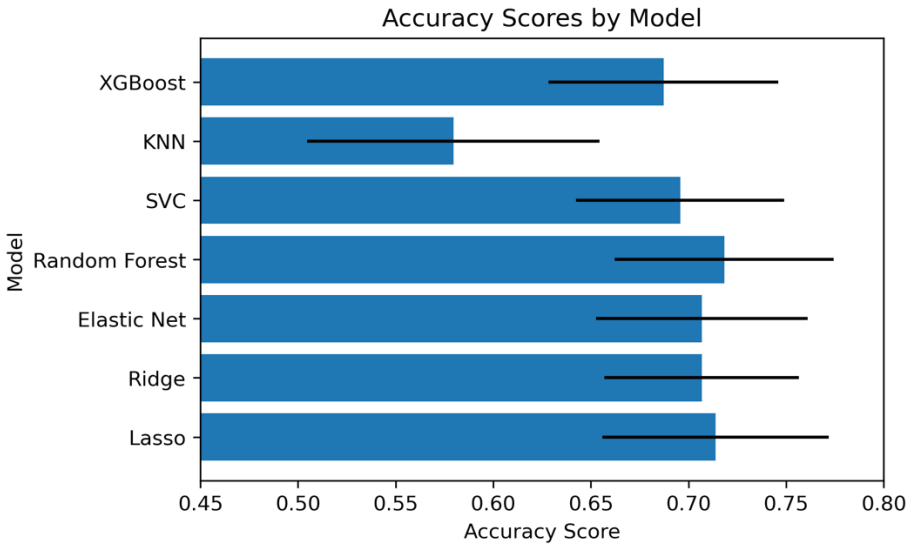

Figure 4 Mean accuracy scores for each model

Using the random forest model with optimal parameters, the mean test score across the 100 secondary random states was 0.720, with a standard deviation of 0.059, while the mean baseline score was 0.544 with standard deviation 0.077. This final accuracy score was 2.98 standard deviations above baseline.

Figure 5 shows mean global feature importances over all random states calculated using mean decrease in impurity (MDI) and permutation of features. Under both methods, MMSE is by far the most important feature. This shows that this nearly 50-year old examination is quite effective in determining dementia levels. After that, nWBV and age are highlighted, and the permutation method gives weight to gender as well. SES appears to have little importance. Figure 6 shows SHAP feature importances for one random state, averaged over all points to show global importance. This method could be used to determine local feature importance if we wanted to know which features contributed to a single person's dementia rating. The relative importance of nWBV is encouraging evidence that MRI-based data may be useful in classifying dementia levels.

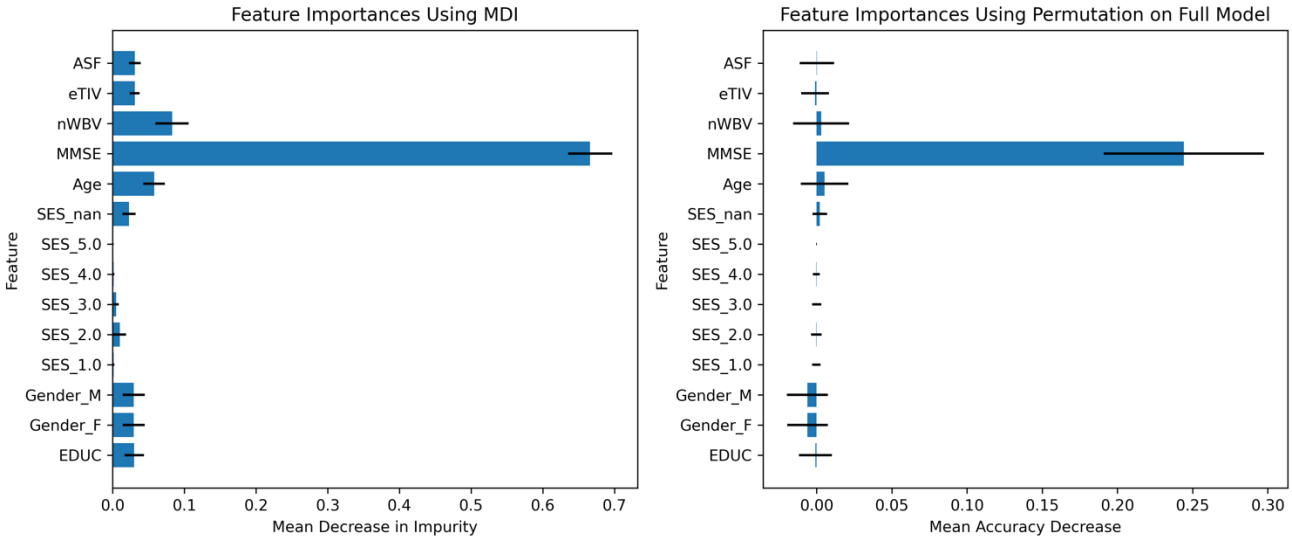

Figure 5 Global feature importances using MDI and feature permutation

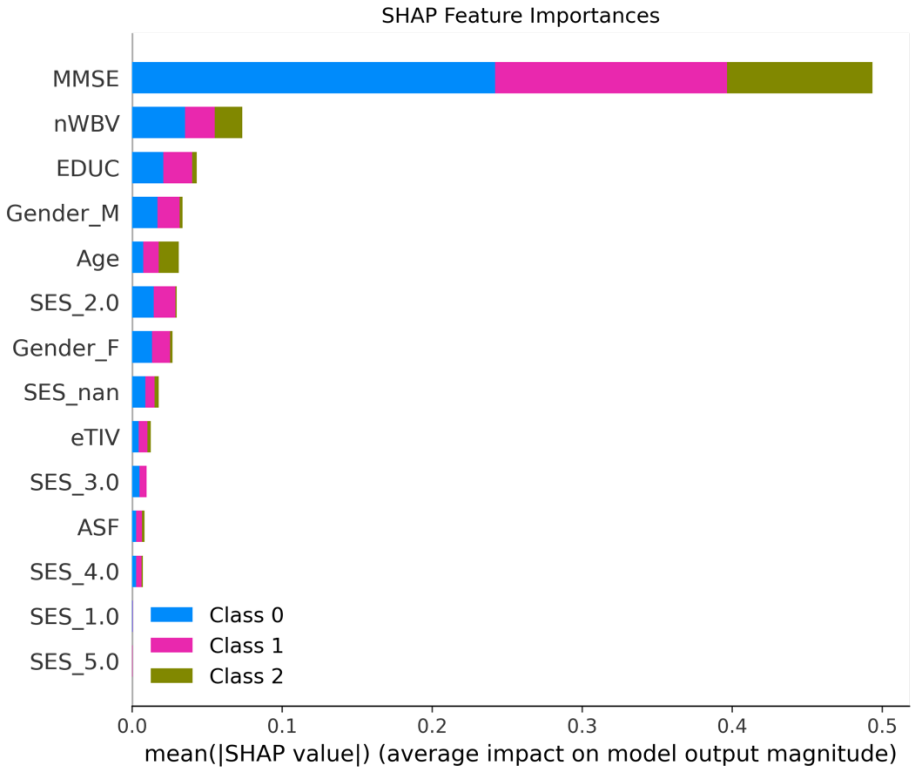

Figure 6 SHAP feature importances

5 Outlook

This model could likely be improved by engineering new features. Collecting more data from the MRI scans would help to assess the ability to predict dementia from brain imaging; dropping the MMSE feature from the model would also force it to rely more on the MRI data. Adding more subjects with moderate/severe dementia to a future study would help to increase accuracy for those levels of CDR. Finally, utilizing the time-series component inherent in the data may also provide increased accuracy and would allow us to address questions regarding dementia development over time.

6 References

[1]: "What Is Alzheimer's?" Alzheimer's Disease and Dementia, Alzheimer's Association, www.alz.org/alzheimers-dementia/what-is-alzheimers.

[2]: "OASIS Brains - Open Access Series of Imaging Studies." OASIS Brains, www.oasis-brains.org.

[3]: Def Me(X). "Dementia Prediction w/ Tree-Based Models." Kaggle, 20 May 2018, www.kaggle.com/ruslankl/dementia-prediction-w-tree-based-models/report.

[4]: Choi, Hyunseok, et al. "DETECTING EARLY ALZHEIMER'S." Kaggle, 6 Mar. 2018, www.kaggle.com/hyunseokc/detecting-early-alzheimer-s#DETECTING-EARLY-ALZHEIMER'S-USING-MRI-DATA-AND-MACHINE-LEARNING.

[5]: Shreyaspj. "Alzheimer's Analysis Using MRI👀👨‍⚕️🩺." Kaggle, 14 Sept. 2021, www.kaggle.com/shreyaspj/alzheimer-s-analysis-using-mri.

[6]: Marcus, Daniel S., et al. "Open Access Series of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults." Journal of Cognitive Neuroscience, vol. 22, no. 12, 2010, pp. 2677–84. Crossref, doi:10.1162/jocn.2009.21407.