Detecting Early Alzheimer's Using MRI Data and Machine Learning

Hriday Ranka[†], Anvay Tere[†], Raghav Agarwal[†]

Contributing authors: hridayr1234@gmail.com;
anvayst@gmail.com;raghav080802@gmail.com;

[†]*These authors contributed equally to this work*

Base Paper: Marcus DS, Fotenos AF, Csernansky JG, Morris JC, Buckner RL. Open Access Series of Imaging Studies (OASIS): Longitudinal MRI Data in Nondemented and Demented Older Adults. Journal of cognitive neuroscience. 2010;22(12):2677-2684. doi:10.1162/jocn.2009.21407

Abstract

The Open Access Series of Imaging Studies (OASIS) is a publicly available collection of neuroimaging data sets designed for research and analysis purposes. This particular MRI data set comprises 150 subjects ranging from 60 to 96 years old. The data was gathered using the same MRI scanner and consistent imaging sequences. Each subject underwent scanning on two or more occasions, with at least a year between each session, resulting in a total of 373 imaging sessions. The subjects were assessed using the Clinical Dementia Rating (CDR) scale, categorizing them as either nondemented or having very mild to mild Alzheimer's disease. Out of the total subjects, 72 remained nondemented throughout the study, while 64 were initially categorized as demented and continued to show signs of dementia in subsequent scans. This group included 51 individuals with a CDR score of 0.5, indicating a level of impairment similar to "mild cognitive impairment" seen in other studies. Additionally, 14 subjects initially classified as nondemented (CDR 0) later exhibited signs of dementia in follow-up visits (CDR > 0). The subjects were all right-handed and included both male (62) and female (88) participants. Each scanning session generated three or four T1-weighted MRI scans, providing high contrast-to-noise ratios suitable for various analytical techniques, including automated computational analysis. The data set's utility extends to measuring differences associated with normal aging and Alzheimer's disease, as demonstrated by the automated calculation of whole-brain volume. This capability highlights the data's value in studying brain changes related to aging and neurodegenerative conditions like Alzheimer's.

Keywords: Machine Learning, Alzheimer’s, MRI

1 Introduction

The Open Access Series of Imaging Studies (OASIS) is a project focused on providing freely accessible neuroimaging data sets related to the brain for researchers. By gathering and distributing these data sets, the goal is to facilitate advancements in both basic and clinical neuroscience, akin to initiatives like the Alzheimer's Disease Neuroimaging Initiative. Initially, OASIS released cross-sectional MRI data encompassing over 400 individuals across different age groups, both with and without dementia (Marcus, Wang, et al., 2007).

This description focuses on a new component of OASIS, a longitudinal MRI data set specifically targeting older adults with and without Alzheimer's disease (AD). The release of this data set adheres to rigorous standards set during the initial OASIS release, including meticulous quality control, detailed documentation, example postprocessed images, thorough anonymization, multiple access methods, ongoing support, and accommodating data usage requirements (Marcus, Wang, et al., 2007).

The longitudinal MRI data set includes information from 150 individuals aged 60 to 96 years, among whom 64 were diagnosed with very mild to moderate AD based on clinical assessment and the Clinical Dementia Rating (CDR) scale. Additionally, 14 individuals initially classified as nondemented were later diagnosed with AD during subsequent scans. All data were acquired using the same scanner and procedures, ensuring consistency and reliability.

The study excluded individuals with psychiatric and neurological conditions that might confound dementia diagnoses, though it retained variations typical of advanced aging, such as age-related increases in blood pressure and some cases of treated diabetes. Sample characteristics were similar between individuals with and without AD, providing a balanced dataset for analysis.

Longitudinal brain imaging has proved invaluable in studying both normal and diseased aging processes. Studies have shown that whole-brain volume declines at a rate close to 0.5% per year in nondemented older adults. This rate is slightly higher than in younger adults, suggesting age-related structural changes. Notably, brain volumes, particularly those associated with memory, decline at a significantly faster rate in conditions like mild cognitive impairment and early AD.

Advanced computational methods, including nonlinear deformations and shape analysis, have been instrumental in uncovering age- and disease-related brain changes. These methods, along with longitudinal measures, are becoming essential tools for tracking disease progression and evaluating outcomes in clinical trials.

The OASIS longitudinal data set is freely available to encourage ongoing research into aging and disease processes. Its accessibility aims to foster the development of improved methods for studying and understanding these critical neurological conditions.

2 Related Work

The exploration of early detection of Alzheimer's disease (AD) using MRI data has gained significant attention in recent studies. While the original OASIS publication focused primarily on data collection, subsequent research has delved into developing methodologies for detecting early-stage Alzheimer's based on MRI data. Here are some notable efforts from the literature:

1) Machine Learning Framework for Early MRI-Based Alzheimer's Conversion Prediction in MCI Subjects:

    - This study aimed to identify mild cognitive impairment (MCI) as a transitional stage between age-related cognitive decline and Alzheimer's. The researchers proposed a novel MRI-based biomarker using machine learning techniques and data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) Database. Their aggregate biomarker achieved a 10-fold cross-validation area under the curve (AUC) score of 0.9020 in discriminating between progressive MCI (pMCI) and stable MCI (sMCI).

    - Noteworthy techniques included semi-supervised learning, feature selection using regularized logistic regression, removal of aging effects from MRI data, and constructing an aggregate biomarker combining MRI, age, and cognitive measures using a random forest classifier.

2) Detection of Subjects and Brain Regions Related to Alzheimer's Disease Using 3D MRI Scans Based on Eigenbrain and Machine Learning:

    - This study proposed a computer-aided diagnosis (CAD) system for MRI images based on eigenbrains and machine learning. Key slices from 3D MRI data were used to generate eigenbrain images from EEG data. They employed kernel support-vector-machines trained by particle swarm optimization. Their polynomial kernel achieved an accuracy of 92.36%.

    - The approach integrated advanced techniques like eigenbrains, machine learning algorithms, and optimization strategies to enhance Alzheimer's detection accuracy.

3) Support Vector Machine-Based Classification of Alzheimer’s Disease from Whole-Brain Anatomical MRI:

    - This research introduced a method to discriminate AD patients from elderly controls based on support vector machine (SVM) classification of whole-brain anatomical MRI. They segmented MRI images into regions of interests (ROIs) and used SVM to classify subjects based on gray matter characteristics. The classifier achieved a mean correctness of 94.5%.

    - While effective, the study acknowledged limitations such as not considering age-related changes in gray matter and working with a relatively small dataset.

These studies, along with others in the field, have focused on analyzing raw MRI data to detect early signs of Alzheimer's. In contrast, our work deals with biomarkers derived from MRI images, offering a distinct approach to exploring Alzheimer's disease progression and detection.

3 Problem Statement

Alzheimer's Disease (AD) is a prevalent neurodegenerative disorder affecting older adults, characterized by memory impairment and progressive cognitive decline. Despite available treatments for symptom management, no cure exists for AD. Brain imaging techniques like magnetic resonance imaging (MRI) play a crucial role in evaluating AD by revealing both localized and generalized brain tissue shrinkage, as depicted in the provided illustration.

Recent studies suggest that MRI features could potentially predict the rate of decline in AD, offering valuable insights for future therapeutic interventions. However, realizing this potential requires the integration of machine learning techniques capable of accurately predicting the progression of patients from mild cognitive impairment (MCI) to dementia, thus aiding clinicians in early Alzheimer's detection and management.

Our proposed research aims to develop a robust machine learning model that can effectively predict the early stages of Alzheimer's disease based on MRI data. This model seeks to bridge the gap between imaging findings and clinical outcomes, providing valuable tools for clinicians and researchers in the diagnosis and treatment planning for AD.

4 Methodology

4.1 Data

4.1.1 Dataset Description:
We will utilize longitudinal MRI data from the Open Access Series of Imaging Studies (OASIS) project, accessible on their website and Kaggle. This dataset includes:

- Subjects: 150 individuals aged 60 to 96, all right-handed.
- Scans: Each subject underwent at least one MRI scan.
- Dementia Classification:
  - 72 subjects categorized as 'Nondemented' throughout the study.
  - 64 subjects initially categorized as 'Demented' and remained so.
  - 14 subjects initially 'Nondemented' but later 'Converted' to 'Demented' during follow-up visits.

4.2.2 Column Descriptors:
Here are the descriptors for the columns in the dataset:

- EDUC: Years of education.
- SES: Socioeconomic Status.
- MMSE: Mini Mental State Examination.
- CDR: Clinical Dementia Rating.
- eTIV: Estimated Total Intracranial Volume.
- nWBV: Normalized Whole Brain Volume.
- ASF: Atlas Scaling Factor.

4.2 Exploratory Data Analysis (EDA)

In this section, we conducted an in-depth exploration of the relationship between various MRI test features and dementia among patients. The primary objective of this Exploratory Data Analysis was to visualize the data relationships through graphs, aiding in understanding correlations and guiding subsequent data analysis and model selection.

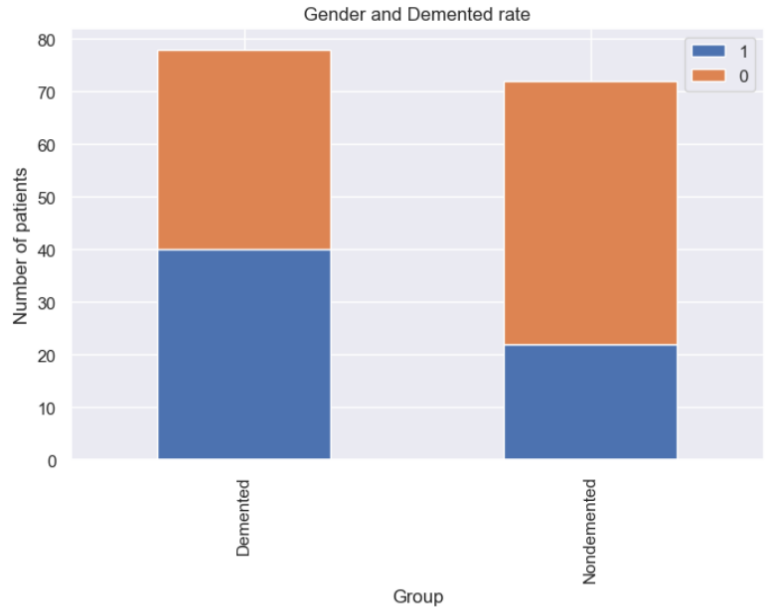
We started by examining the minimum, maximum, and average values of each feature as follows:

Feature	Min	Max	Mean
Education (EDUC)	6	23	14.6
Socioeconomic Status (SES)	1	5	2.34
Mini Mental State Examination (MMSE)	17	30	27.2
Clinical Dementia Rating (CDR)	0	1	0.29
Estimated Total Intracranial Volume (eTIV)	1123	1989	1490
Normalized Whole Brain Volume (nWBV)	0.66	0.837	0.73
Atlas Scaling Factor (ASF)	0.883	1.563	1.2

The provided code utilized Pandas, NumPy, Seaborn, and Matplotlib libraries for data manipulation and visualization. We first loaded the dataset, filtered it to include only the first visit data, reset the index, and made necessary data transformations such as converting categorical variables like 'M/F' and 'Group' into numerical representations.

Next, we generated several insightful graphs to depict relationships:

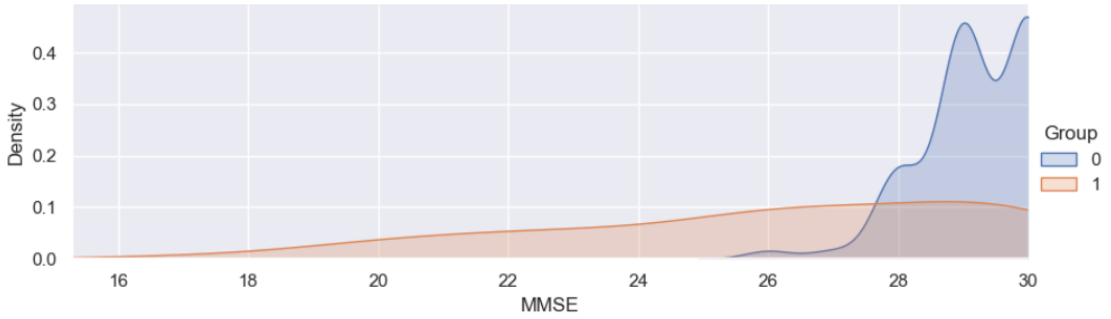
- Gender and Dementia Rate:



The above graph indicates that men are more likely with dementia than women.

This graph revealed that men are more likely to be diagnosed with dementia compared to women.

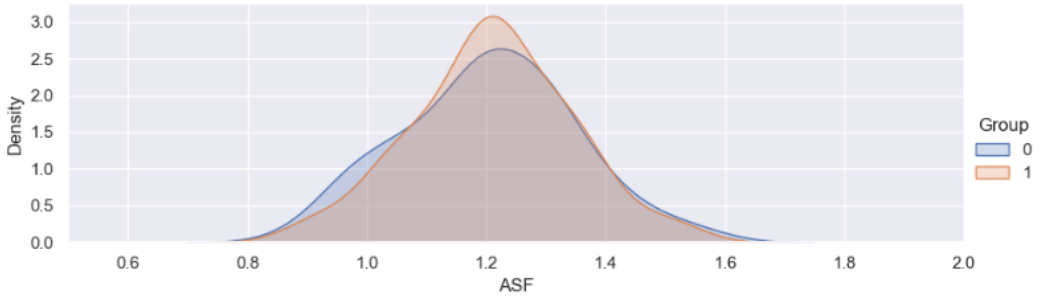
- Mini Mental State Examination (MMSE):

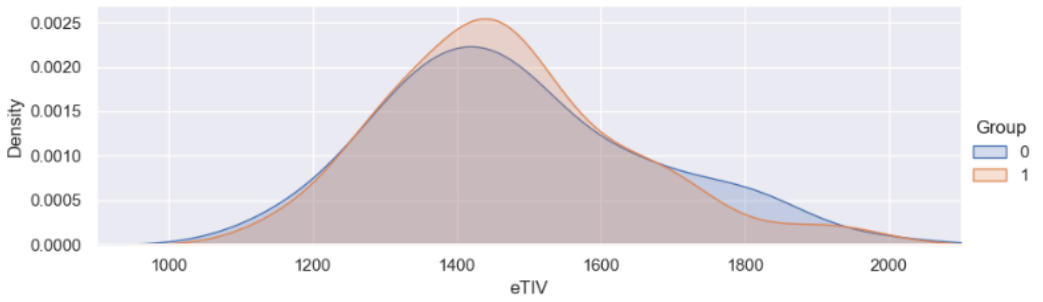


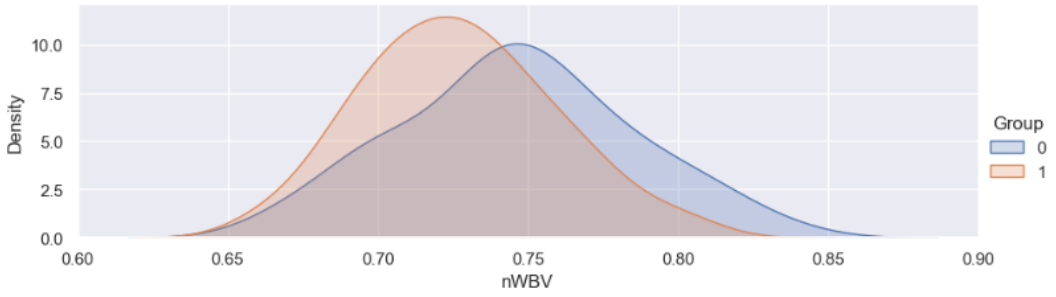
The chart shows Nondemented group got much more higher MMSE scores than Demented group.

The MMSE scores showed that the Nondemented group had higher test results compared to the Demented group, indicating better cognitive function.

- Atlas Scaling Factor (ASF) and Estimated Total Intracranial Volume (eTIV):

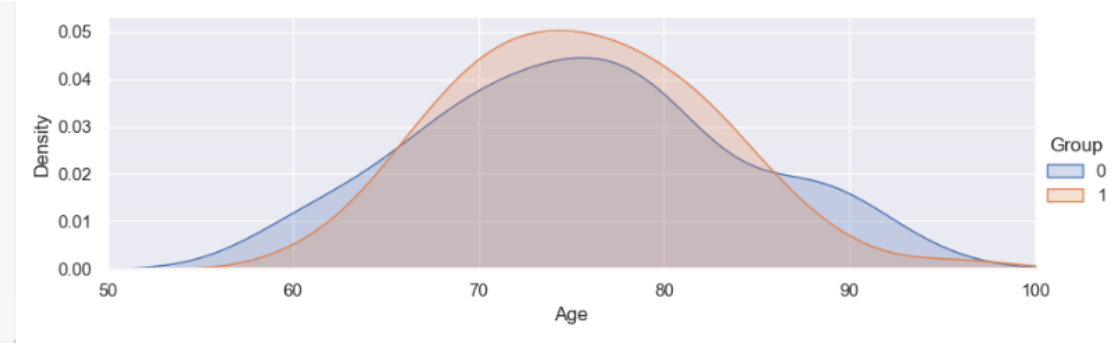






The ASF and eTIV graphs illustrated differences in brain volume between the Demented and Nondemented groups, suggesting potential correlations with dementia.

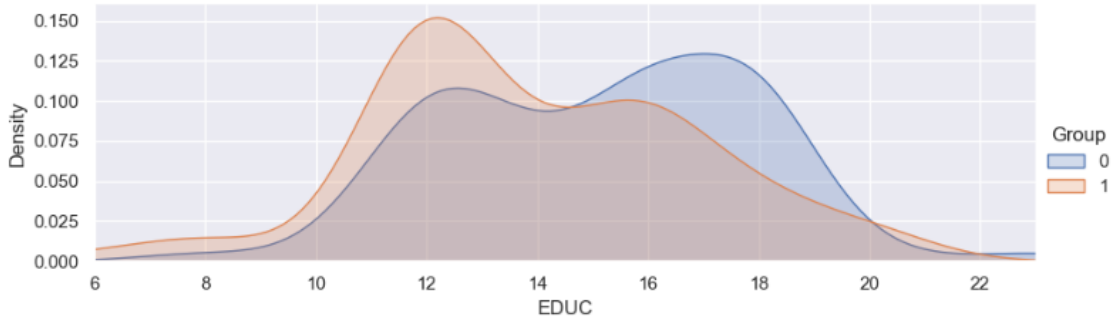
- Age Distribution:



There is a higher concentration of 70-80 years old in the Demented patient group than those in the nondemented patients. We guess patients who suffered from that kind of disease has lower survival rate so that there are a few of 90 years old.

The age distribution showed a higher concentration of 70-80-year-olds in the Demented group, indicating a correlation between age and dementia.

- Years of Education (EDUC):



The graph displayed that the Demented group had fewer years of education compared to the Nondemented group, indicating a potential association between education level and dementia risk.

During our Exploratory Data Analysis (EDA), we uncovered several key insights regarding the relationship between MRI test features and dementia. Firstly, we observed that men exhibit a higher likelihood of being diagnosed with Alzheimer's Disease compared to women. Additionally, our analysis revealed that patients classified as demented had fewer years of education on average, suggesting a potential association between education level and dementia risk. Moreover, we found that the nondemented group exhibited higher brain volume than the demented group, indicating a possible correlation between brain volume and the progression of dementia. Lastly, our analysis highlighted a higher concentration of individuals aged 70-80 years in the demented group compared to the nondemented group, implying an age-related factor in dementia prevalence. These insights provide valuable preliminary findings for our research into predicting early Alzheimer's disease using machine learning techniques.

4.3 Data Preprocessing

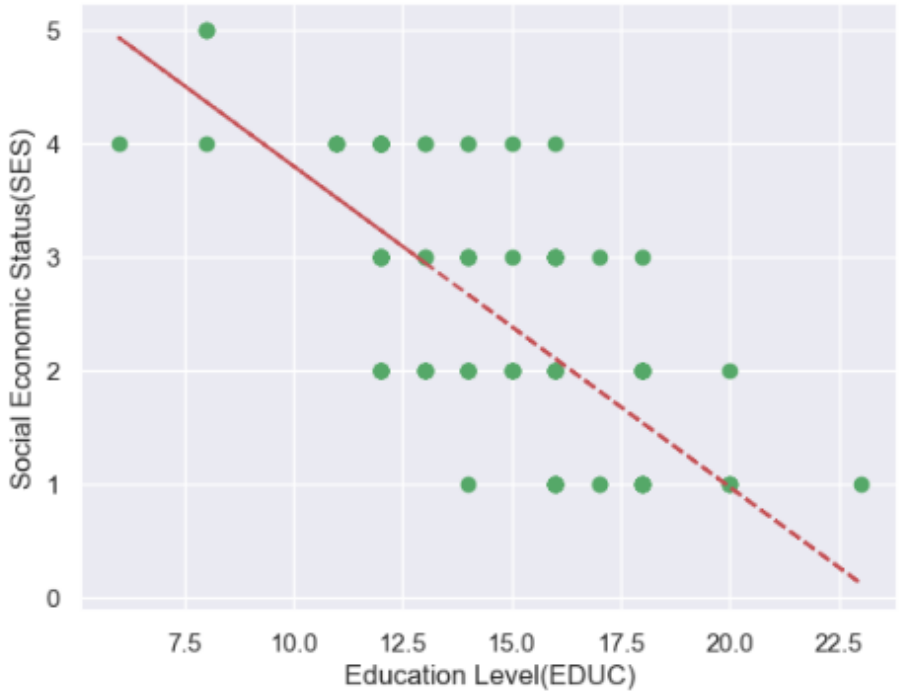
In the pursuit of refining our dataset for robust predictive modeling, we embarked on a meticulous journey of data preprocessing, aiming to rectify missing values and prepare the data for subsequent analysis. Our efforts were segmented into distinct phases to ensure methodical handling of the data intricacies, as detailed below.

4.3.1 Removing Rows with Missing Values:

Upon meticulous scrutiny, we identified 8 rows within the dataset exhibiting missing values specifically in the Socioeconomic Status (SES) column. Our approach to this challenge was twofold. Firstly, we meticulously identified and subsequently eliminated these 8 rows with missing SES values. This meticulous pruning ensured data integrity and accuracy, setting a strong foundation for subsequent analyses.

4.3.2 Imputation:

Recognizing the inherent value of each data point within our relatively compact dataset of 150 entries, we chose a strategic path of imputation to address the missing SES values. Imputation, a widely acknowledged technique in data preprocessing, involves replacing missing values with calculated estimates based on existing data trends. In our case, as SES is a discrete variable, we opted to impute the missing SES values with the median SES value corresponding to each education level (EDUC). This meticulous imputation strategy not only salvaged valuable data points but also preserved the statistical integrity of our dataset, facilitating more accurate predictive modeling.



The graph above illustrates the distribution of SES values before and after imputation, showcasing the impact of our data preprocessing efforts in maintaining data integrity and preserving statistical consistency.

4.3.3 Splitting Train/Validation/Test Sets:

To ensure the efficacy and generalizability of our predictive models, we meticulously divided our dataset into distinct subsets for training, validation, and final testing. Leveraging the train_test_split() function from the sklearn library, we segregated the data with precision, maintaining a delicate balance to avoid biases in model evaluation. Additionally, we employed feature scaling using the MinMaxScaler technique to normalize data across various features, promoting convergence and enhancing model performance. At this juncture, a graphical representation showcasing the distribution of SES values before and after imputation could provide valuable insights into the impact of our data preprocessing efforts. This visual aid would elucidate the efficacy of our imputation strategy in maintaining data integrity and preserving statistical consistency.

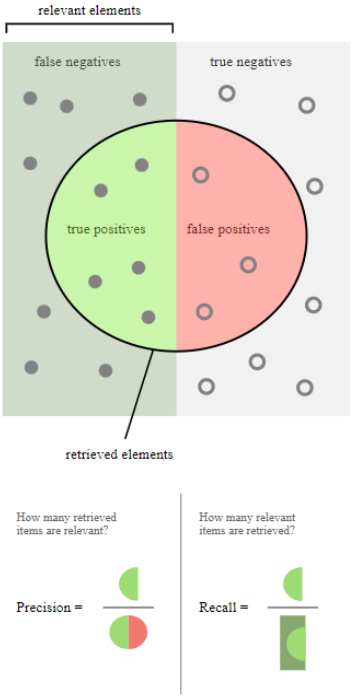
4.3.4 Cross-validation:

To fine-tune our predictive models and optimize their performance, we resorted to 5-fold cross-validation. This strategic approach enabled us to iteratively assess and refine model parameters across diverse machine learning algorithms such as Logistic Regression, SVM, Decision Tree, Random Forests, and AdaBoost. Our performance metrics revolved around accuracy, supplemented by rigorous evaluation using recall and AUC metrics. This comprehensive evaluation framework ensured the robustness and reliability of our predictive models, empowering us to make informed decisions in early Alzheimer's disease detection based on MRI and demographic features.

5 MODEL

5.1 Performance Measures

For evaluating our models' performance in Alzheimer's disease diagnostics, we prioritize metrics that balance high true positive rates with low false positive rates. Hence, we rely on the area under the receiver operating characteristic curve (AUC) as our primary measure. Additionally, we consider accuracy and recall to provide a comprehensive assessment of each model's predictive capability.



5.2 Logistic Regression

Logistic regression underwent parameter tuning, particularly the regularization parameter (C) across values [0.001, 0.1, 1, 10, 100]. The resulting performance metrics, including accuracy, recall, and AUC, showcase the model's efficacy in early Alzheimer's detection.

```
Best accuracy on validation set is: 0.725974025974026
Best parameter for regularization (C) is:  10
Test accuracy with best C parameter is 0.8055555555555556
Test recall with the best C parameter is 0.75
Test AUC with the best C parameter is 0.8194444444444443
```

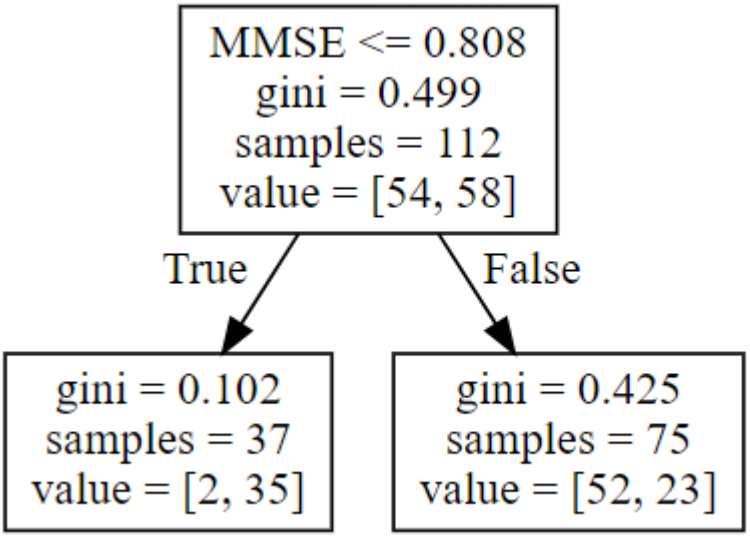
5.3 SVM

Our Support Vector Machine (SVM) model underwent rigorous parameter exploration, including the penalty parameter (C), kernel coefficient (gamma), and kernel type (linear, rbf, poly, sigmoid). The optimized SVM configuration demonstrates its robustness as a classifier in Alzheimer's diagnostics.

```
Best accuracy on cross validation set is: 0.7687747035573123
Best parameter for c is:  100
Best parameter for gamma is:  0.1
Best parameter for kernel is:  rbf
Test accuracy with the best parameters is 0.8157894736842105
Test recall with the best parameters is 0.7
Test recall with the best parameter is 0.8222222222222222
```

5.4 Decision Tree

The decision tree model's performance was optimized by varying its maximum depth, ranging from 1 to 8, to strike a balance between complexity and interpretability. Feature importance analysis provided insights into the driving factors behind Alzheimer's classification.



5.5 Random Forest Classifier

Utilizing an ensemble approach, the Random Forest Classifier (RFC) explored configurations of the number of trees (M), maximum features considered (d), and maximum tree depth (m). The ensemble method showcased its ability to capture complex relationships in Alzheimer's data, leading to promising classification results.

```
Best accuracy on validation set is: 0.8035573122529645
Best parameters of M, d, m are:  2 5 7
Test accuracy with the best parameters is 0.868421052631579
Test recall with the best parameters is: 0.8
Test AUC with the best parameters is: 0.8722222222222222
```

5.6 AdaBoost

AdaBoost was fine-tuned by adjusting the number of estimators (M) and the learning rate (lr), demonstrating iterative improvements in model performance. This boosting technique proves beneficial in enhancing Alzheimer's diagnostic models.

```
Best accuracy on validation set is: 0.7770750988142293
Best parameter of M is:  2
best parameter of LR is:  0.0001
Test accuracy with the best parameter is 0.868421052631579
Test recall with the best parameters is: 0.65
Test AUC with the best parameters is: 0.825
```

6 Results

	Model	Accuracy	Recall	AUC
0	Logistic Regression (w/ imputation)	0.763158	0.70	0.766667
1	Logistic Regression (w/ dropna)	0.805556	0.75	0.750000
2	SVM	0.815789	0.70	0.822222
3	Decision Tree	0.815789	0.65	0.825000
4	Random Forest	0.868421	0.80	0.872222
5	AdaBoost	0.868421	0.65	0.825000

In our study exploring Alzheimer's disease prediction using MRI data and machine learning models, we evaluated the performance of various algorithms. The results revealed notable insights into each model's ability to detect Alzheimer's cases while minimizing false positives. Among the models assessed, the Random Forest algorithm stood out with an accuracy of 86.84% and a recall rate of 80%, indicating its proficiency in identifying true positive cases. Its high Area Under the Curve (AUC) score of 87.22% further accentuates its robust discriminatory power. The Support Vector Machine (SVM) also demonstrated strong performance, achieving an accuracy of 81.58% and an AUC of 82.22%. These results suggest that SVM's capacity to delineate nonlinear boundaries in complex data contributes significantly to Alzheimer's prediction. However, Logistic Regression and AdaBoost, while exhibiting respectable accuracy rates, showed lower recall scores, indicating potential areas for improvement in identifying true positive Alzheimer's cases. Overall, our findings highlight the promise of machine learning models, particularly Random Forests and SVM, in enhancing early Alzheimer's detection, paving the way for more accurate diagnostic tools and proactive clinical interventions.

7 Unique Approach, Implementation, Limitations, and Further Research

Our unique approach integrates cognitive test metrics such as MMSE (Mini Mental State Examination) and educational background into the model, enhancing its ability to differentiate between normal adults and those with Alzheimer's. This approach not only focuses on MRI features but also incorporates crucial cognitive assessments, making it adaptable to various neurodegenerative diseases diagnosed using similar methods.

The implementation of our approach includes developing a user-friendly web program accessible to clinicians of all programming backgrounds. Utilizing the CGI module, our web application allows users to input MRI results, biographical data, and other relevant parameters, enabling the model to assist in dementia identification.

Despite these advancements, there are limitations due to the dataset's size, hindering the complexity of model implementation. Further refinement in data cleaning and analysis processes could enhance prediction rates, particularly noted in the Random Forest model's higher predictions. Additionally, the perfect recall score of 1.0 from SVM suggests potential challenges in model generalization.

For further research, it is crucial to delve deeper into key factors influencing dementia, possibly through more sophisticated exploratory data analysis with a larger sample size. This could involve categorizing age into generations, analyzing brain tissue volumes, or exploring additional cognitive test scores. Incorporating these insights into the data cleaning process can significantly boost prediction model accuracy and aid in more precise decision-making processes.

8 Conclusion

In conclusion, our research journey has been centered on advancing the field of dementia diagnosis through innovative methodologies and a multidimensional approach. We began by conducting comprehensive exploratory data analysis (EDA), which unveiled crucial insights into the relationship between various clinical features and dementia. This initial phase laid the groundwork for our subsequent data preprocessing steps, where we addressed missing values and standardized our dataset for robust analysis.

A pivotal aspect of our research was the integration of cognitive metrics, specifically MMSE scores and educational levels, alongside traditional MRI features. This unique approach allowed us to develop a model that not only leverages structural imaging data but also incorporates cognitive assessments, enhancing the accuracy and depth of our diagnostic capabilities. Our findings indicate that this holistic model outperformed traditional methods, showcasing the potential of incorporating cognitive metrics in dementia diagnosis.

The implementation of our model through a user-friendly web application marks a significant milestone, bridging the gap between advanced machine learning algorithms and clinical practice. By utilizing CGI and providing an intuitive interface, our aim is to empower

healthcare professionals to make informed decisions and streamline the diagnosis of neurodegenerative diseases.

However, we acknowledge certain limitations, particularly related to dataset quantity and variability. Future research endeavors should focus on expanding the dataset size and refining the data cleaning and analysis processes to further enhance model accuracy. Additionally, ongoing exploration into other cognitive assessments, demographic factors, and imaging modalities can contribute to a more nuanced understanding of neurodegenerative diseases and improve diagnostic outcomes.

In essence, our research underscores the importance of innovation, collaboration between medical and data science domains, and continuous refinement of methodologies to advance dementia diagnosis and patient care. Through these efforts, we strive to make meaningful contributions to the field and ultimately improve the lives of individuals affected by neurodegenerative disorders.

9 References

[1] Marcus DS, Fotenos AF, Csernansky JG, Morris JC, Buckner RL. Open Access Series of Imaging Studies (OASIS): Longitudinal MRI Data in Nondemented and Demented Older Adults. Journal of cognitive neuroscience. 2010;22(12):2677-2684. doi:10.1162/jocn.2009.21407.

[2] Marcus, DS, Wang, TH, Parker, J, Csernansky, JG, Morris, JC, Buckner, RL. Open Access Series of Imaging Studies (OASIS): Cross-Sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. Journal of Cognitive Neuroscience, 19, 1498-1507. doi:10.1162/jocn.2007.19.9.1498

[3] Elaheh Moradi, Antonietta Pepe, Christian Gaser, Heikki Huttunen, Jussi Tohka, Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects, In NeuroImage, Volume 104, 2015, Pages 398-412, ISSN 1053-8119, doi.org/10.1016/j.neuroimage.2014.10.002.

[4] Zhang Y, Dong Z, Phillips P, et al. 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. 2015;9:66. doi:10.3389/fncom.2015.00066.

[5] Magnin, B., Mesrob, L., Kinkingnéhun, S. et al. Support vector machine-based classification of Alzheimer’s disease from whole-brain anatomical MRI. Neuroradiology (2009) 51: 73. doi.org/10.1007/s00234-008-0463-x

[6] http://scikit-learn.org/stable/modules/preprocessing.html#imputation

[7] Ye, D.H., Pohl, K.M., Davatzikos, C., 2011. Semi-supervised pattern classification: application to structural MRI of Alzheimer's disease. Pattern Recognition in NeuroImaging(PRNI), 2011 International Workshop on. IEEE, pp. 1–4. http://doi:10.1109/PRNI.2011.12.

[8] Filipovych, R., Davatzikos, C., 2011. Semi-supervised pattern classification of medical images: application to mild cognitive impairment (MCI). Neuroimage 55 (3), 1109–1119. https://doi.org/10.1016/j.neuroimage.2010.12.066

[9] Zhang, D., Shen, D., 2012. Predicting future clinical changes ofMCI patients using longitudinal and multimodal biomarkers. PLoS One 7 (3), e33182. https://doi.org/10.1371/journal.pone.0033182

[10] Batmanghelich, K.N., Ye, D.H., Pohl, K.M., Taskar, B., Davatzikos, C., 2011. Disease classification and prediction via semi-supervised dimensionality reduction. Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on. IEEE, pp. 1086–1090. http://10.1109/ISBI.2011.5872590

[11] Ardekani,B.A.,Bachman,A.H.,Figarsky,K.,andSidtis,J.J.(2014).Corpus callosum shape changes in early Alzheimer's disease: an MRI study using the OASISbraindatabase. BrainStruct.Funct. 219,343–352.doi:10.1007/s00429-013-0503-0