Pioneering Precision: Machine Learning Prognosticates Early Stage Alzheimer's Disease

Nimitha Gopinath

School of Engineering Science, Simon Fraser University, Burnaby, BC

Abstract

Alzheimer's disease (AD) is a prevalent neurodegenerative disorder that primarily affects older adults, leading to dementia. MRI analysis faces challenges such as complexity, subjectivity, and interrater variability. This study employs machine learning models and MRI analysis to enhance early prediction and diagnostic accuracy for Alzheimer's disease (AD).

The Random Forest model achieves outstanding performance with 86.84% accuracy, effectively differentiating between healthy individuals and those with AD by incorporating factors like MMSE and education level. This project contributes to the understanding and management of AD.

Dataset Description

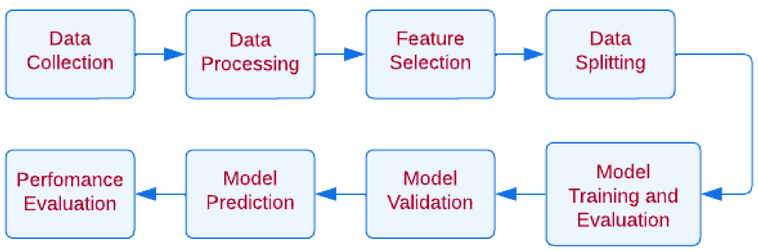
The dataset used consists of MRI-related data from the Open Access Series of Imaging Studies (OASIS) project, available on their website and Kaggle. The dataset includes subjects categorized as follows throughout the study:

- 72 subjects were initially grouped as **'Nondemented.'**
- 64 subjects were initially categorized as **'Demented'**"Demented' and remained so.
- 14 subjects were initially categorized as **'Nondemented.'** but later classified as **'Demented.'** ('Converted' category).

This dataset can be effectively used to train various machine learning algorithms.

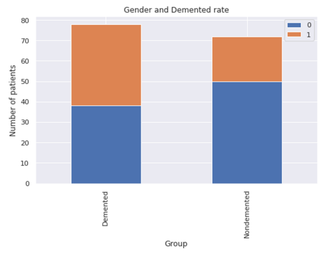
Methodology

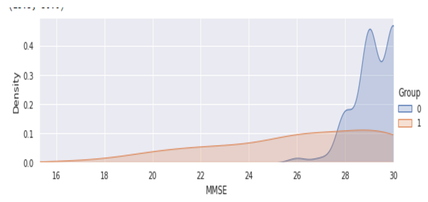
A systematic study explored early Alzheimer's detection using MRI data and crucial experimental steps.



Explanatory Data Analysis

During Exploratory Data Analysis (EDA), we identified gender disparities in dementia incidence, variations in MMSE scores, and differences in brain volume ratios between the Demented and Non-demented groups, providing valuable insights into the dataset.





1. **Males** are more susceptible to demented conditions, including Alzheimer's Disease, than females.
2. **Demented patients** have fewer years of education.
3. **The Nondemented group** has a higher brain volume than the Demented group.
4. A notable concentration of individuals aged **70-80 years** was observed in the Demented group, suggesting an age-related factor in dementia occurrence.

Data Processing and Evaluation

In the data processing stage, one of the fundamental aspects is data cleaning, a critical step aimed at ensuring the integrity, consistency, and quality of the dataset.

1. **Rows** with missing values in the SES column were removed to ensure dataset completeness and reduce bias.
2. **Median Imputation** addressed missing values in the SES column.
3. The dataset was randomly divided into **train, validation, and test sets**.
4. **5-fold cross-validation** determined optimal parameters for machine learning models: **Decision Tree, Random Forest, SVM, logistic regression, and AdaBoost**.

Perfomance Evaluation of Machine Learning Models

The analysis of machine learning models for Alzheimer's disease prediction yields insightful results, showcasing the performance of various models:

No.	Model	Accuracy	Recall	AUC
0	Logistic Regression (w/imputation)	0.789474	0.70	0.794444
1	Logistic Regression (w/dropna)	0.8055556	0.75	0.750000
2	SVM	0.815789	0.70	0.82222
3	Decision Tree	0.815789	0.65	0.825000
4	Random Forest	0.868421	0.80	0.87222
5	AdaBoost	0.868621	0.65	0.825000

Table 1. Model Performance.

The **Random Forest model outperformed other models**, accurately identifying cases of Alzheimer's disease with high precision.

Limitations and Future Scope

The **limitations** of this study:

- Undefined test value ranges require precise identification of influential variables.
- Larger datasets enhance generalizability of findings.
- Unexplored variables may impact the study's focus on specific features.
- Assumptions and limitations of evaluated models affect their applicability.

In terms of **future scope**:

- Investigate additional factors (genetic markers, lifestyle) influencing Alzheimer's prediction.
- Explore advanced machine learning techniques and ensemble models for improved accuracy.
- Conduct longitudinal studies to track disease progression and evaluate predictive models.
- Collaborate with healthcare professionals, incorporate clinical data for enhanced relevance.

Conclusion

This project utilizes advanced machine learning models and the OASIS dataset to predict Alzheimer's disease. The Random Forest model achieves exceptional performance with 86.84% accuracy, 80% recall, and 87.22% AUC. Incorporating essential features like MMSE enhances the model's capability to differentiate between healthy individuals and those with Alzheimer's. Early prediction and intervention are crucial for mitigating symptom severity. This project contributes to the field of neurodegenerative diseases, demonstrating the effectiveness of machine learning models in diagnosis and prognosis.

References

[1] Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J. *Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects*. NeuroImage, 104, 398-412, 2015.

[2] 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, 13, 2019.