Prediction of Alzheimer stages for given MRI scans

Aaradhya Verma (2022004) Aarya Khandelwal (2022007)

I. Problem Statement

Alzheimer's disease is a progressive neuro-degenerative disorder that affects millions of people worldwide, leading to cognitive decline and ultimately severe dementia.
Early detection and accurate classification of Alzheimer's disease stages are crucial for timely intervention and treatment. Magnetic Resonance Imaging (MRI) scans offer detailed structural insights into the brain, presenting a valuable resource for diagnosing Alzheimer's disease.
In this study, our objective is to elucidate the data processing methodologies employed on the original dataset and to assess various machine learning models for their efficacy in classifying Alzheimer's disease stages based on MRI scans.

II. Dataset

It comprises images of MRI (Magnetic Resonance Imaging) scans , each resized to 128×128 pixels. It contains approximately 5200 MRI images, with patients classified into four distinct classes:

- Class 1: Mild Demented (896 images)
- Class 2: Moderate Demented (64 images)
- Class 3: Non-Demented (2000 images)
- Class 4: Very Mild Demented (2240 images)

Dataset: https://www.kaggle.com/datasets/sachinkumar413/alzheimer-mri-dataset/data.

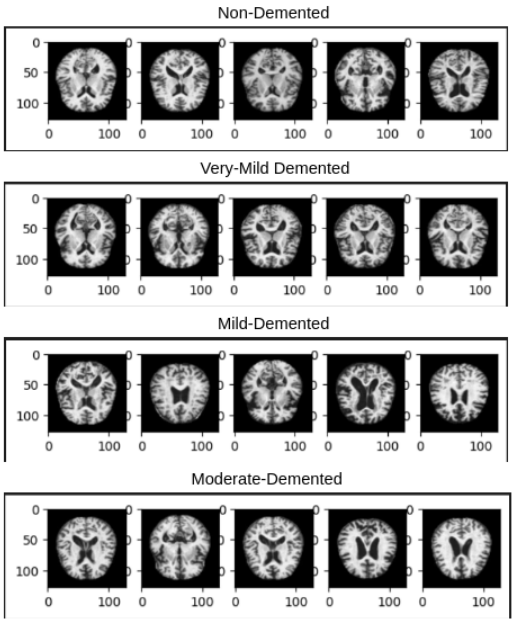

Fig. 1. Samples of every class

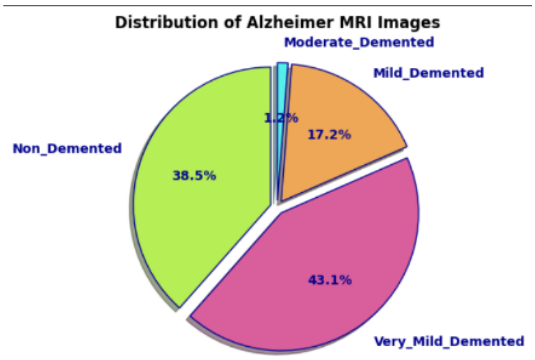

Fig. 2. Data distribution

III. Data Pre-processing

- **Image Loading:** The dataset was divided into four categories based on disease stage. All of the images were gray-scaled to reduce data-complexity and ensure uniformity in pixel values across all images.
- **Data Splitting:** The dataset was split into training, validation, and test sets. The training set constituted 80% of the data, while the validation and test sets comprised 10% each.
- **Data Augmentation:** To enhance the diversity of the training dataset and improve model generalization, data augmentation techniques were applied. Brightness was adjusted and Gaussian noise was addition to create augmented versions of the original images.

Fig. 3. Data Augmentation

- **Principal Component Analysis (PCA):** PCA was employed to reduce the dimensionality of the feature space, focusing solely on dimensions that have a significant impact on the data. The optimal number of principal components was determined based on the cumulative explained variance, with the objective of retaining the maximum possible variance in the data.

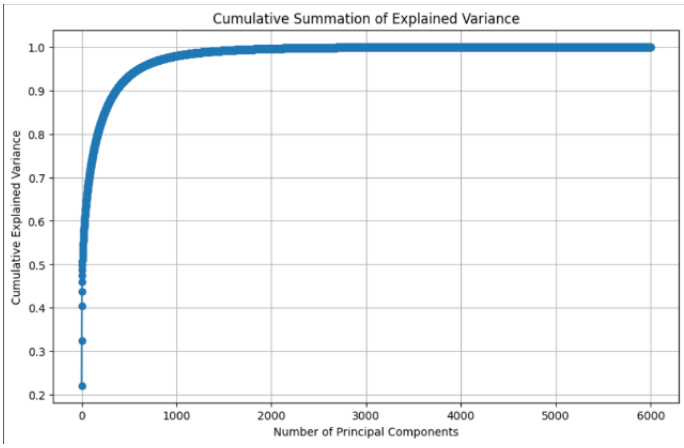

Fig. 4. PCA: dimension-data-explanation plot

IV. Models Applied

- **Decision Tree Classification:** It is a versatile supervised learning algorithm used for both classification and regression tasks. It works by partitioning the data into subsets based on the features' values, creating a tree-like structure of decisions. At each node of the tree, the algorithm selects the feature that best splits the data, optimizing some criterion (like Gini impurity or entropy) to maximize information gain or minimize impurity. The resulting tree can be used to make predictions by following the path from the root to a leaf node, where each leaf corresponds to a class label (in classification) or a numerical value (in regression).

- **Multiclass Logistic Regression:** It is a statistical method used for classification which extends the principles of binary logistic regression to handle multiple classes by employing a set of linear functions, one for each class, combined with the soft-max function to convert raw predictions into probabilities. The class with the highest probability is then assigned as the predicted class for a given input.

- **Random Forest Classification:** Random Forest is an ensemble learning method that utilizes multiple decision trees to make predictions. It operates by constructing a multitude of decision trees during training, where each tree is trained on a random subset of the training data and a random subset of the features. During prediction, each tree in the forest independently predicts the outcome, and the final prediction is determined by a majority vote (for classification) or an average (for regression) of the predictions made by individual trees. Random Forest helps to reduce overfitting compared to a single decision tree by combining the predictions from multiple trees, thereby improving the model's generalization ability.

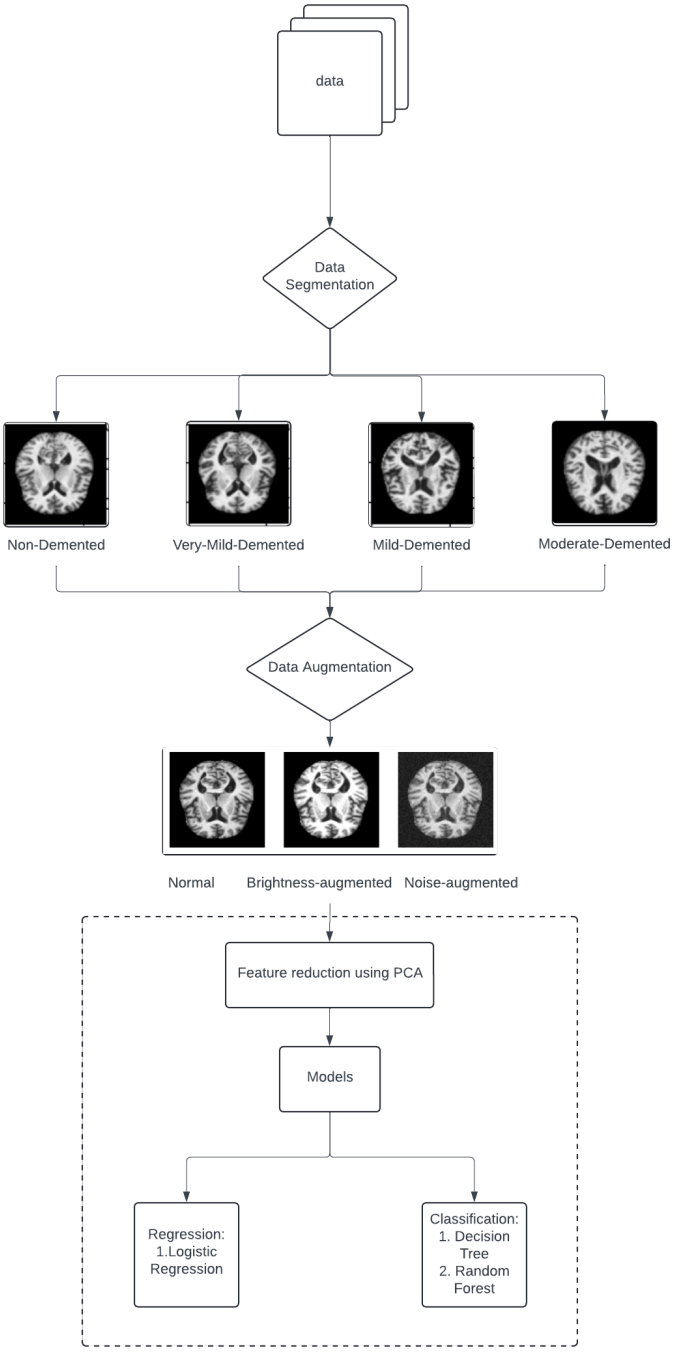

Fig. 5. Project Schema

V. Results

- **Decision Tree:** It performed with an accuracy of **66.15%** on *validation dataset* and **66.92%** on *test dataset*.

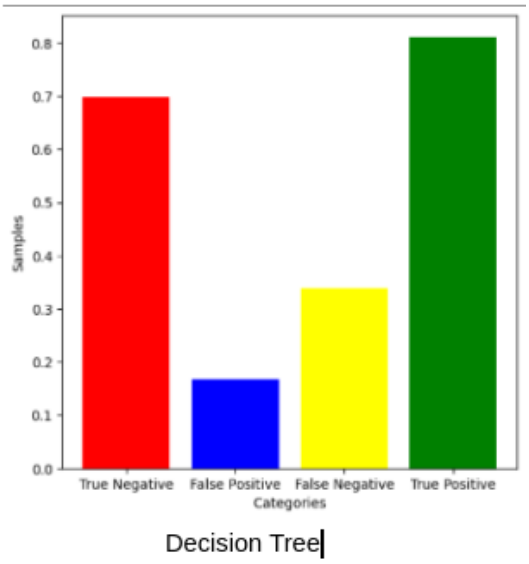

Fig. 6. Decision Tree

Data	Accuracies (%)
Total	66.923
Non-Demented	68.817
Very-Mild-Demented	73.755
Mild-Demented	53.773
Moderate-Demented	28.57

- **Logistic Regression:** The algorithm performed with an accuracy of **85%** on *validation dataset* and **82.307%** on *test dataset*. And **Confusion Matrix** is used as metric to visualize the true labels against the predicted labels.

Data	Accuracies (%)
Total	82.307
Non-Demented	84.44
Very-Mild-Demented	78.733
Mild-Demented	87.7735
Moderate-Demented	100.00

- **Random Forest Classifier:** The algorithm performed with an accuracy of **96.7307%** on *test dataset*.

Data	Accuracies (%)
Total	96.730
Non-Demented	97.849
Very-Mild-Demented	99.547
Mild-Demented	90.566
Moderate-Demented	71.428

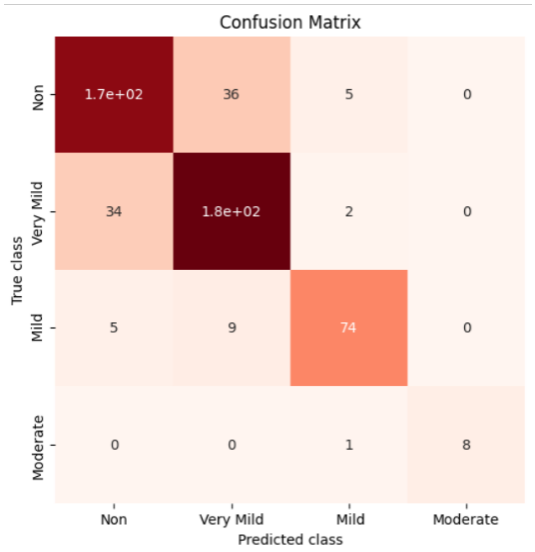

Fig. 7. Logistic Regression: Confusion Matrix

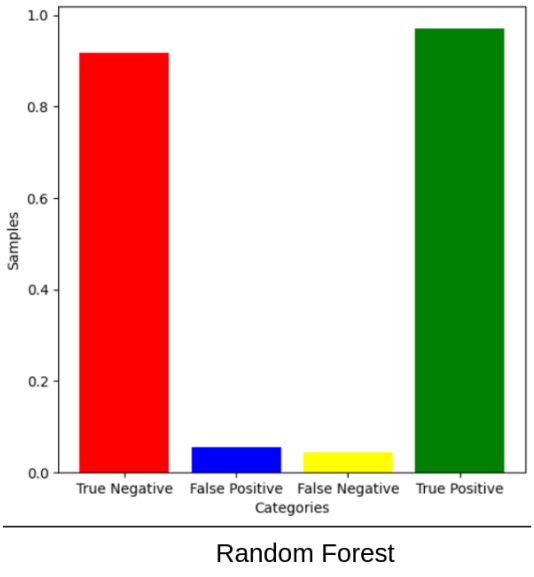

Fig. 8. Random Forest