Alzheimer's Disease Prediction Using ML

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfilment of the requirements to award the degree of

Bachelor of Technology

In

Computer Science and Engineering

Submitted by

Candidate Name

Sathvika Kurmala – AP21110010071

Srujitha Devineni – AP21110010086

Tanya Kavuru - AP21110010113

Vijaya Vyshnavi Muvvala – AP21110010119

Under the Guidance of

Dr.Murali Krishna Enduri
SRM University–AP
Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

[Nov, 2023]

Certificate

Date: 6-Nov-23

This is to certify that the work present in this Project entitled "**Alzheimer's Disease Prediction Using ML**" has been carried out by **srujitha Devineni** under my/our supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology/Master of Technology in **School of Engineering and Sciences**.

E. Murali Krish.

Supervisor

(Signature)

Dr. Murali Krishna Enduri

Designation - Assistant Professor,

Affiliation - SRM University, AP.

Acknowledgement

The project that we have completed under your guidance is an excellent opportunity for our education and career development. We consider ourselves as extremely fortunate people to having the opportunity to interact with the professionals who guided us throughout this semester

We thank our Professor Murali Krishna Enduri sir, who had provided insights and expertise whenever required that greatly assisted the course project. Additionally, we want to express our gratitude to sir for sharing their pearls of knowledge with us during the course project.

We perceive this opportunity as one of the milestones in the progression of our careers. In order to achieve our intended career objectives and succeed in our careers, we will try to utilize the acquired skills and information as effectively as possible and to develop them in our careers.

Table of Contents

Abstract:

Alzheimer's disease is a neurogenerative disorder that produces a particular global healthcare challenge. Early diagnosis is required for effective treatment. This research explores the potential of machine learning and deep learning techniques for predicting Alzheimer's disease. A dataset containing various MRI scans, neuroimaging, and features was used to train and contemplate machine learning models.

The dataset we have comprehended includes structural MRI scans, cognitive test scores, and genetic markers of individuals, both with and without Alzheimer's disease. Numerous engineering and selection techniques were applied to enhance the performance of the model. Various classification algorithms which include SVM, random forest, and deep neural networks, were used in the implementation that predict Alzheimer's disease. These models were strictly evaluated using measures such as accuracy, precision, recall, F1-Score, and area under the receiver operating characteristic curve. The results indicate that ML models can effectively predict Alzheimer's disease based on a combination of neuroimaging features.

This demonstrates the potential of ML in aiding early Alzheimer's disease diagnosis, which is important for personalized treatment. Future work may involve refining and validating these models on larger and more diverse datasets, as well as exploring the integration of multi-modal data sources for even more robust predictions.

Statement of Contributions

Person I (Srujitha): Srujitha was crucial in creating the model and organising the paper as a whole. Her assistance in developing a unified framework made a big difference in the content's flow and organisation. Srujitha was responsible for outlining the research's structure, making sure it made sense, and laying the groundwork for the paper.

Person II (Tanya): Tanya was crucial in gathering data and helped with the first draught of the manuscript. she also made use of her ability to create visual aids, such as pictures and diagrams, to improve the content's overall presentation and readability.

Person III (Vyshnavi): Vyshnavi was in charge of the literature review; she carefully gathered relevant research papers and combined the data to produce an in-depth summary. Her work was essential in providing the paper with a solid academic foundation.

Person IV (Sathvika): Sathvika added coherence and clarity to the subject matter, which helped with paper writing. Additionally, She was instrumental in the creation of diagrams that brought ideas to life and successfully conveyed difficult concepts.

Abbreviations

CNN	Convolutional Neural Networks
ML	Machine Learning
SVM	Support Vector Machine
KNN	K-Nearest Neighbor

List of Tables

List of Figures

1. Introduction

Alzheimer's is a devastating neurodegenerative disorder, which has emerged as a global health concern as the population of the world continues. [13] This disorder is the most common cause of dementia among older people. It is characterized by a progressive decline in cognitive function, resulting in memory loss and a range of behavioural and psychological symptoms. The burden of this disorder extends through individual patients which affects their families, caregivers, and healthcare systems. [7] Due to the absence of a cure and the limited effectiveness of available treatments, early diagnosis, intervention, and detection have become more important in handling the disease and improving the quality of a patient's life. The early stages of Alzheimer's disease are often subtle which makes it more challenging to diagnose accurately. Cognitive tests and medical imaging play crucial roles in diagnosis but may lack the specificity required for early detection. This is where we use machine learning, a branch of artificial intelligence. Machine learning techniques are capable of handling complex and multidimensional data, they offer the potential to augment and, in some cases, transfigure Alzheimer's disease diagnosis. The sub-sections of this research will dive into the methodology, dataset, feature engineering, and various machine learning algorithms used in this study. Results and discussions will be present in the performance of these models, along with insights into the importance of specific features.

a. Usage of the Report

The main purpose of this project is to elaborate the details about the project we had done, it has several potential uses and applications, which benefits various stakeholders, those include healthcare professionals, researchers, and the public. This report precisely focuses on Early detection and diagnosis, Support, Research and Development, Public Awareness and Education, and also to improve caregiver support.

b. Overview or Background

Early diagnosis of this disorder is challenging but is crucial for slowing down the progression of the disorder and to improve the quality of life for the individuals who are affected. [8]ML based prediction models offers to detect the disease in its early stages. We gathered large number of datasets which consists of medical records, MRI preprocessed data, MRI scans, genetic information and other information of the patient which is relevant. These datasets serve for training predictive models. ML models require careful selection of features that are relevant to Alzheimer's disease prediction.

c. Problem Statement

The main objective of this research is to develop an accurate and reliable ML model so that it can predict the risk of Alzheimer's disease or study its presence in every individual. This model will use many data sources, which includes, medical records, MRI scans, survey data, neuroimaging data, genetic information, and required data types. The reason for early detection is it leads to more effective interventions, improved patient care, and a better health care system. Development of reliable predictive models can impact the lives of individuals who are at risk of or currently living with Alzheimer's disease.

d. Motivation

The need to build a predictive model of Alzheimer’s disease prediction using Machine Learning arises from the urgent need to address the challenges associated with this neurogenerative disorder. This research study is important because as millions of individuals are affected by this disorder, this disease is expected to rise, placing an increase of burden in healthcare systems and in society as well. This disorder currently has no cure expect of few medications.

[12] Early prediction can help the disorder to slow down its progression, and also improves the quality of life for affected individuals, and probably we can delay the onset of severity of this disorder. With the present health records, neuroimaging technologies, and genetic testing, there is huge amount of data tat is available for research. ML offers a way to evaluate this data for predictive modeling.

[9] Predictive models can help in the development of personalized treatments for individuals who are victims of Alzheimer’s disease. This type of care can lead to more effective interventions and best outcomes. By identifying and predicting the risk outcomes as early as possible can reduce the long-term cost of medical facilities.

As the field of ML and AI is evolving day by day and is becoming advanced, every researcher is motivated by the opportunity to apply cutting-edge technology to a crucial medical problem, which leads to breakthroughs in diagnosis and treatment. Precisely the main motivation is to improve the quality of individuals life which is at risk of or affected by Alzheimer’s disease. Early detection and prediction can lead to a greater level of independence and a better standard of living for affected patients and their families.

In Alzheimer's disease prediction using machine learning, several common knowledge patterns and observations have emerged from research and clinical practice. These patterns provide insights into the disease and guide the development of predictive models. Some of the common knowledge patterns observed include:

- Progressive Cognitive Decline
- Age as a Significant Risk Factor
- Neuroimaging Abnormalities
- Longitudinal Data
- Clinical and Medical History
- Ethical Considerations

1. Literature analysis

Our research adventure started with an intensive evaluation of academic papers, aiming to pinpoint an most useful research consciousness for our Alzheimer's prediction study. Our primary hobby lay in the application of numerous machine mastering algorithms and picture processing techniques, seeking a challenge that could be intellectually stimulating and nearly possible. After thorough attention, we decided to middle our research on predicting Alzheimer's chance using those techniques.

The research papers we examined provided precious insights that enriched our understanding of the diverse strategies hired in Alzheimer's prediction thru system gaining knowledge of and photograph processing techniques. These readings illuminated the strengths and weaknesses of different approaches, presenting important perspectives at the hurdles faced whilst utilizing these methods for Alzheimer's threat evaluation. They additionally guided us in formulating our studies problem: the improvement of a new, extra accurate, and efficient approach for Alzheimer's hazard prediction, surpassing the capabilities of current methods.

Exploration of Relevant Literature: In the following segment, we will delve into the content material of four particular papers that we extensively studied to deepen our comprehension of the demanding situations related to Alzheimer's chance prediction the usage of machine learning algorithms and image processing techniques.

In [1] they used various ML techniques like Random Forest, SVM, XG Boost, Max Voting, and Decision tree classifiers for predicting AD vs normal subjects. For this, they used a dataset consisting of 150 patients' MRI data. They used various data mining techniques to pre-process data for adjusting missing values and extracting features. After analysis of the data they finally took parameters like EDUC, MMSE, SES, CDR, Etiv, nWBv, and ASF into consideration.

They measured the performance using accuracy, precision, recall, and F1 score. Out of all the used ML techniques Random Forest and XG boost received higher accuracies and precision. XG Boost and voting received higher recall values and F1 scores.

In [2] they tried to find Alzheimer's disease at an early stage. This stage is known as the MCT stage which results in memory loss. By predicting the disease at this stage it helps to decrease the risk to life. For this, they used Machine learning algorithms like SVM and Decision Trees.

For predicting the disease they took parameters like MMSE score, no. of visits, and education of patients into consideration out of all other parameters available in the dataset. By using the SVM algorithm they attained an accuracy of 85% and by using Decision trees they attained an accuracy of 83%.

In [3] they created a novel method employing 3D deep convolutional neural networks to effectively differentiate mild Alzheimer's from cognitively normal and cognitively impaired people. For this, they used structural MRI and data sets were from ADNI and NACC. Based on the sizes and thickness of formerly identified brain regions that are known to be involved in the course of disease, we have created a reference model for comparison.

[14] When classifying individuals with moderate Alzheimer's dementia and mild cognitive impairment from those with cognitively normal subjects, the deep learning model produced an AUC of 85.12. This indicates that the model is accurate. It obtains an AUC of 62.45 in the more difficult task of identifying MCI. In comparison to the reference model built that is volume/thickness model, which requires the volumes and thickness to be retrieved beforehand, it is also much faster.

In [4] The paper discussed the software of system mastering fashions to expect dementia in patients using longitudinal Magnetic Resonance Imaging (MRI) statistics from the OASIS venture. Various device mastering fashions, such as SVM, Logistic Regression, Decision-Tree, and Random-Forest, were employed to make predictions. Before great-tuning, those fashions performed exclusive ranges of accuracy, with SVM displaying the highest trying out accuracy of seventy-four%. After satisfactory-tuning, SVM outperformed other fashions, accomplishing 92% accuracy. However, some overfitting troubles had been mentioned inside the Decision Tree and Random Forest models.

They also supplied insights into facts preprocessing, consisting of filling missing values and function scaling, and using assessment metrics like accuracy, remember, AUC, and the confusion matrix for version overall performance evaluation. It highlights that SVM is the satisfactory-appearing model across all metrics and recommends using a bigger dataset and additional machine studying fashions, inclusive of AdaBoost, KNN, Majority Voting, and Bagging, for further development. The gadget goals to offer the public with a tool for assessing the chance of dementia in person patients via inputting MRI information.

In precis, the paper centered on the improvement and assessment of gadget studying fashions to expect dementia the use of MRI records. It demonstrates that SVM is the most accurate model, with ninety-two% accuracy after first-rate-tuning, and emphasizes the ability of device studying in early dementia detection and public use.

As a conclusion, the introduction has shed light on the critical issues that Alzheimer's disease presents and emphasized the importance of early detection and intervention. Using cutting-edge technologies becomes essential as the number of cases of this neurodegenerative illness rises worldwide. The approach used in this study will be covered in detail in the sections that follow, detailing the tedious steps involved in feature engineering, dataset collection, and the use of various machine learning algorithms. The methodology section clarifies every aspect of our strategy for handling the difficulties of Alzheimer's disease prediction and acts as a link between the theoretical foundation established here and the real-world application of predictive modelling.

2. Methodology

A major global health concern is Alzheimer's disease, for which early detection is essential to successful treatment [10]. The objective of this research is to improve the accuracy and robustness of Alzheimer's disease identification by applying a dual-modality strategy that makes use of both numerical and MRI image datasets. Several classification techniques were used on the numerical dataset to find patterns indicating of Alzheimer's disease. Cross-validation was then used to carefully evaluate the models' reliability. The study also included an MRI image dataset, from which relevant characteristics were extracted using image processing techniques. This pre-processed MRI dataset was then subjected to deep learning algorithms, opening up a new method for Alzheimer's diagnosis. In the upcoming sections, we'll explain in detail how we selected our data, the Algorithms we used, and the steps we took to ensure the study was conducted ethically and fairly.

- **First let's discuss the methodology used for numerical dataset.**

2.1.)Data Selection, Pre-processing and EDA (Exploratory data analysis):

1) Data Selection

We obtained the numerical dataset from Kaggle, specifically the "oasis_longitudinal.csv" file, with the label "MRI and Alzheimer's." This dataset offers an extensive range of personal data, including important features like demographic information, cognitive scores, and complex brain properties.
As seen in the figure below, the dataset has a crucial "Group" column that divides people into groups like "Demented," "Nondemented," and "Converted." This column is a vital part of our research because it helps categorize cases as either Alzheimer's or not. The dataset also includes a variety of other useful columns, such as gender, age, and education.

	Subject ID	MRI ID	Group	Visit	MR Delay	M/F	Hand	Age	EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
0	OAS2_0001	OAS2_0001_MR1	Nondemented	1	0	M	R	87	14	2.0	27.0	0.0	1987	0.696	0.883
1	OAS2_0001	OAS2_0001_MR2	Nondemented	2	457	M	R	88	14	2.0	30.0	0.0	2004	0.681	0.876
2	OAS2_0002	OAS2_0002_MR1	Demented	1	0	M	R	75	12	NaN	23.0	0.5	1678	0.736	1.046
3	OAS2_0002	OAS2_0002_MR2	Demented	2	560	M	R	76	12	NaN	28.0	0.5	1738	0.713	1.010
4	OAS2_0002	OAS2_0002_MR3	Demented	3	1895	M	R	80	12	NaN	22.0	0.5	1698	0.701	1.034
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
368	OAS2_0185	OAS2_0185_MR2	Demented	2	842	M	R	82	16	1.0	28.0	0.5	1693	0.694	1.037
369	OAS2_0185	OAS2_0185_MR3	Demented	3	2297	M	R	86	16	1.0	26.0	0.5	1688	0.675	1.040
370	OAS2_0186	OAS2_0186_MR1	Nondemented	1	0	F	R	61	13	2.0	30.0	0.0	1319	0.801	1.331
371	OAS2_0186	OAS2_0186_MR2	Nondemented	2	763	F	R	63	13	2.0	30.0	0.0	1327	0.796	1.323
372	OAS2_0186	OAS2_0186_MR3	Nondemented	3	1608	F	R	65	13	2.0	30.0	0.0	1333	0.801	1.317

373 rows × 15 columns

Fig 1: Numeric Dataset

2.1.2) Data Preprocessing:

In order to get the dataset ready for further analysis, we carefully preprocessed it:

Handling missing values:

After loading and exploring the dataset, namely the "oasis_longitudinal.csv" file , we noticed that some important columns—namely, "SES" (Socioeconomic Status) and "MMSE" (Mini-Mental State Examination)—had missing values. We chose a calculated approach to deal with these missing values in order to improve the robustness of our models. i.e to replace the missing values in these columns with the corresponding medians. This method of imputation helps in keeping the integrity of the dataset by decreasing the effect of missing data on future analyses.

Standardisation of groups:
As we can see in the below figure,there is a crucial step in which the categorization label "Converted" is combined with "Demented." resulting in a consistent and uniform classification. This kind of strategic standardisation is crucial because from the clinical perspective the individuals who transition from "Converted" to "Demented" are essentially experiencing the cognitive decline caused by Alzheimer's disease .

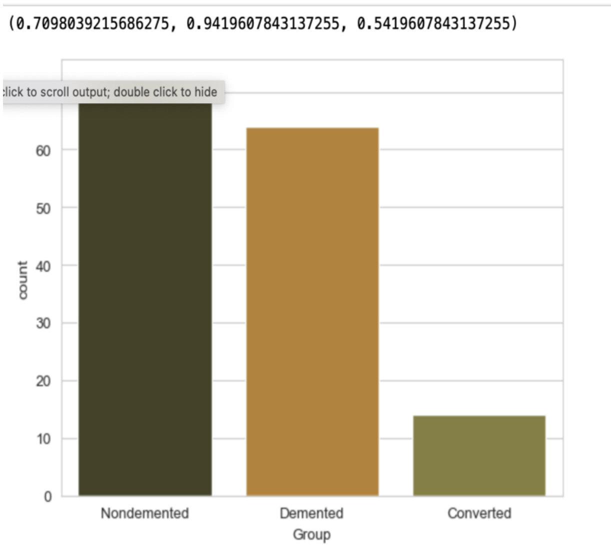

Fig 2.1: Group attributes before conversion

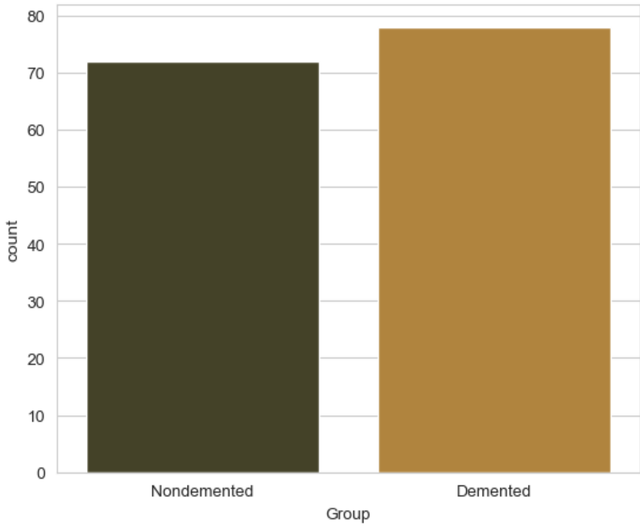

Fig 2.2: Group attributes after conversion

2.1.3.) EDA (Exploratory data analysis):

Gender Analysis:
To shed light on how men and women are represented in the dataset, we carefully looked at the frequency of demented cases in relation to gender.
The results provided fascinating new information about the distribution of dementia cases by gender:
Demented Cases(people with alzehmer’s):
Men (M): 40 instances
Women (W): 38 instances
Cases without dementia (people without alzehmer’s):
Male (M): 22 cases
Women (W): 50 instances
This gender analysis highlights an important finding: men are more likely than women to experience mental illness. These gender-specific patterns offer important contextual information that may impact how results are interpreted and advance a more sophisticated investigation of Alzheimer's disease within various demographic groups.

```
             M   F
Demented     40  38
Nondemented  22  50

Text(0.5, 1.0, 'Gender and Demented rate')
```

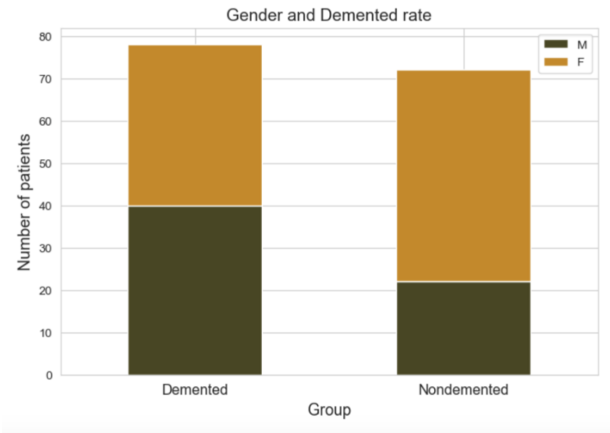

Fig 3: Gender vs Demented rate Bar graph

Education and Dementia:

In this study, we examined the educational histories of those who were categorized as having dementia with those who were not, An interesting pattern was found by the analysis:
Patients with dementia typically had lower levels of education.
Patients who were not demented: Usually had greater educational attainment.
This comparative analysis is visually represented in the figure below, which also shows the differences in education between patients with and without dementia.

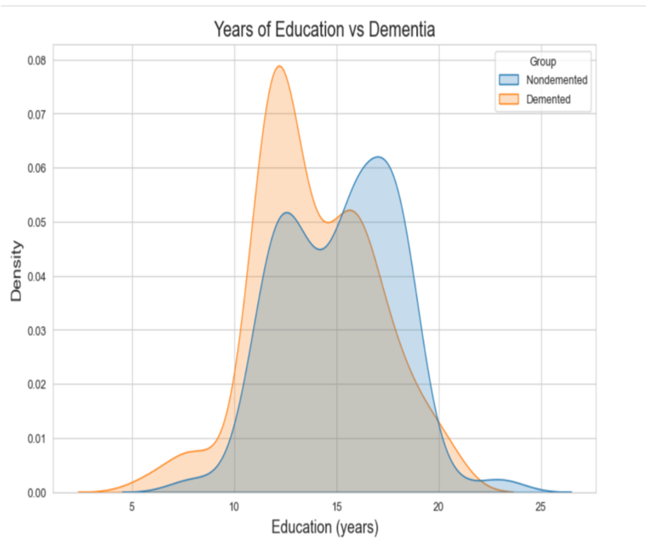

Fig 4: Years of Education vs Dementia

Correlation Examination:

We used a heatmap to perform a correlation analysis in order to obtain understanding of the relationships between various numerical features. The "oasis_longitudinal.csv" dataset was used to create the correlation map, which shows the correlation coefficients between the variables visually. The correlation value, which ranges from -1 to 1, is annotated in each heatmap cell to show the direction and strength of the correlation.

The heatmap helps us understand how different factors may be related by highlighting possible correlations between different features. This analysis helps us better understand the dataset overall and directs our feature selection process.

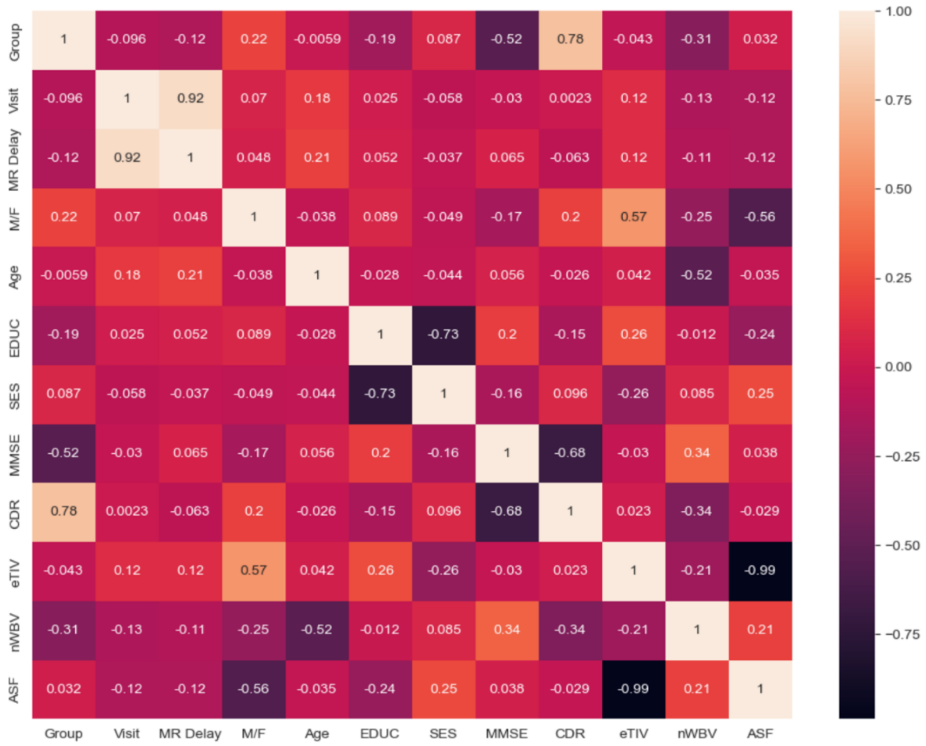

Fig 5: Correlation Matrix (Heat Map)

We wrap up our data groundwork by navigating missing values and identifying gender and education patterns, with men exhibiting higher rates of dementia and lower education levels associated with dementia. Now for the really important part: classification analysis, which is a crucial stage in our study into Alzheimer's.

2.2.) Implementation of classification algorithms:

In this phase, we employed various classification algorithms on the pre-processed numerical dataset using the "oasis_longitudinal.csv" file. The main goal was to predict the existence of Alzheimer's disease, denoted by the "Group" column, by employing attributes like age, gender, education, socioeconomic status (SES), MMSE scores, and brain properties.

Used Algorithms:

2.2.1) Logistic Regression

[15] Logistic regression is a useful tool for classification tasks because it models the probability of a binary outcome. The[6] logistic function is used by the algorithm to determine how the independent variables(i.e x) and the likelihood of a specific result(i.e y) relate to one another. It seeks to identify the logistic curve that best fits the relationship and best describes it.

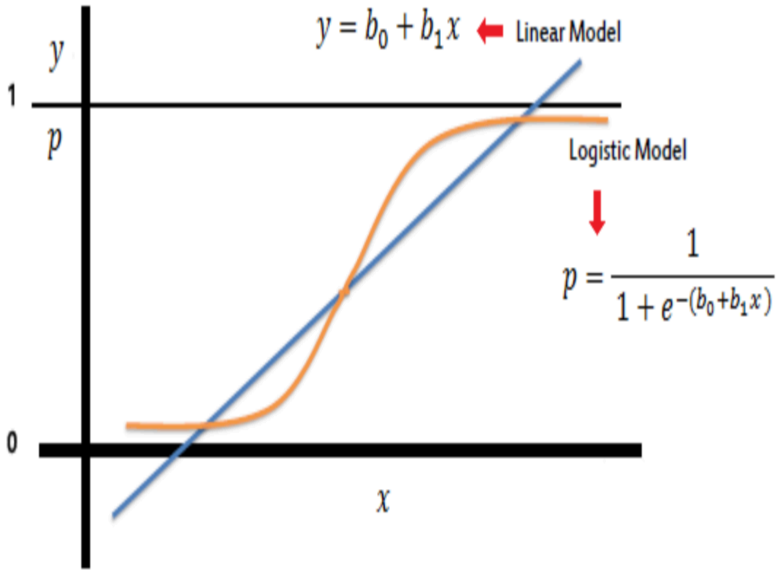

Fig 6: Logistic Regression

With its default hyperparameters, our model used the Logistic Regression algorithm. The regularisation parameter and solver method are typically set to default values. The solver method controls how the algorithm handles optimisation, while regularisation aids in preventing overfitting ,Logistic Regression accuracy is 64% in our dataset. Notable characteristics that enhanced the model's capacity to forecast binary outcomes were Age, MMSE, and eTIV.

2.2.2) Navie Bayes (Gaussian)

The Gaussian Naive Bayes Classifier functions similarly to an experienced investigator. Its methodology treats each clue or feature as an independent piece of a puzzle, assuming that they are independent of one another. The "Gaussian" component suggests that it interprets these hints as following a consistent pattern, much like how numbers fit into a bell the graph. Even with its seemingly simple methodology, this approach works surprisingly well, especially when the puzzle pieces don't depend on each other too much, In order to reflect the idea that the features have a common pattern, our model adopted the Gaussian Naive Bayes variant. Given the numeric nature of the features and the plausible assumption of normalcy, this selection is in line with our dataset. The Gaussian Naive Bayes model, configured with default hyperparameters, achieved an accuracy of 73.33%. Crucial features such as Age, MMSE, and eTIV significantly contributed to the model's predictive capability.

2.2.3) Navie Bayes (Bernouli)

Based on the Bayes theorem, Bernoulli Naive Bayes is a probabilistic classification algorithm. Working well with binary or binarized features, it assumes independence between features. Our model is appropriate for our Alzheimer's prediction task because it determines the likelihood of being a member of a specific class based on the presence or absence of features. We made use of the scikit-learn BernoulliNB implementation. This particular Naive Bayes variant is made for binary features. The model makes predictions based on the presence or absence of characteristics associated with an increased risk of Alzheimer's disease.

Bernoulli Naive Bayes obtained a 56% accuracy rate using the default hyperparameters.

2.2.4) k-nearest Neighbors (KNN)

K-Nearest Neighbours groups instances according to the majority class of their neighbours. It works on the principle that instances that are similar are probably a part of the same class. To

identify the class of a new data point in the context of Alzheimer's prediction, the algorithm takes into account the properties of the 'k' nearest data points.
The KNeighbors Classifier algorithm was utilised by our model, with the parameter 'n_neighbors' set to 3. This parameter establishes how many neighbours are taken into account when classifying something. Selecting three neighbours achieves a balance between accuracy and model complexity.
KNN in our dataset, achievied an accuracy of 73.33%. Noteworthy features such as Age, MMSE, and eTIV contributed to the model's ability to classify instances based on their proximity in the feature space. The choice of three neighbors optimized the balance between model simplicity and predictive accuracy.

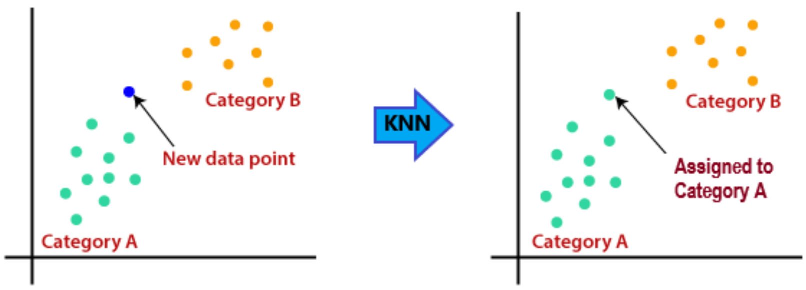

Fig 7: KNN

2.2.5)Decision Tree Classifier

Consider approaching decision-making as a sequence of yes-or-no questions that guide us, step by step, towards the correct response. In a similar way, Decision Tree Classifier operates. It creates a decision tree-like structure by dividing the dataset into groups based on the evaluation of various features and their values. Every branch points the model in the direction of a prediction by representing a decision based on a feature.

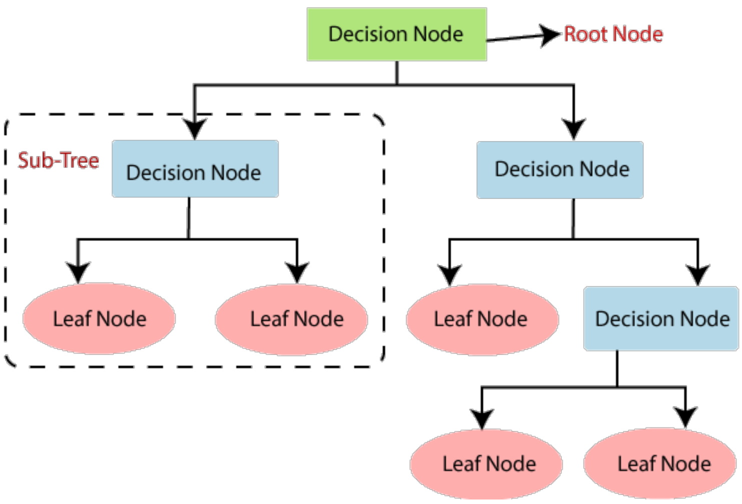

Fig 8: Decision Tree

The Decision Tree Classifier using the 'entropy' criterion was adopted by our model. This criterion gauges the disorder or impurity of nodes, assisting the model in making judgement calls that result in more precise predictions. In addition, we limited the tree's maximum depth to five to avoid overfitting and overcomplication with our particular dataset.

Configured with 'entropy' and a maximum depth of 5, our Decision Tree Classifier achieved an accuracy of 70.67%. Crucial features like Age, MMSE, and eTIV played a vital role in the decision-making process.

2.2.6)Random Forest Classifier

Combining predictions from several decision trees is how Random Forest, a reliable ensemble learning technique, works. Every tree is a distinct entity that presents its own view of the data. The combined effect of these separate contributions generates the final prediction, which is more accurate and resistant to overfitting.

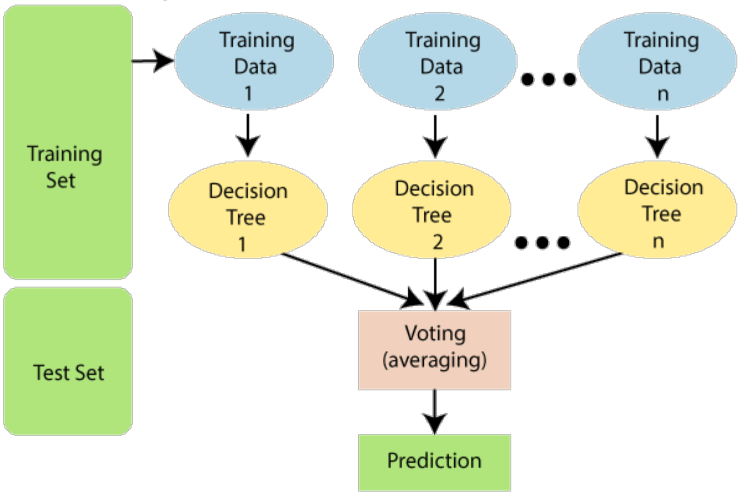

Fig 9: Random Forest Classifier

We used a Random Forest Classifier with the following parameters in our Alzheimer's prediction model: n_estimators=200, Criterion='gini,' Max Depth=8, and Max Features= 0.5. This customised setup allowed for a 78.67% accuracy rate.
Age, the Mini-Mental State Examination , and the estimated Total Intracranial Volume were the three key predictors that our model highlighted. These characteristics are crucial for predicting when Alzheimer's disease will appear.

2.2.7) support vector Machine(SVM)

[16] Support Vector Machines (SVM) operate by finding the optimal hyperplane that best separates different classes in a dataset. In our case, the 'linear' kernel signifies a linear decision boundary, which is suitable for our dataset. SVM identifies support vectors, which are data points crucial for determining the position and orientation of the hyperplane. The algorithm aims to maximize the margin between classes, enhancing its robustness in classifying new instances.

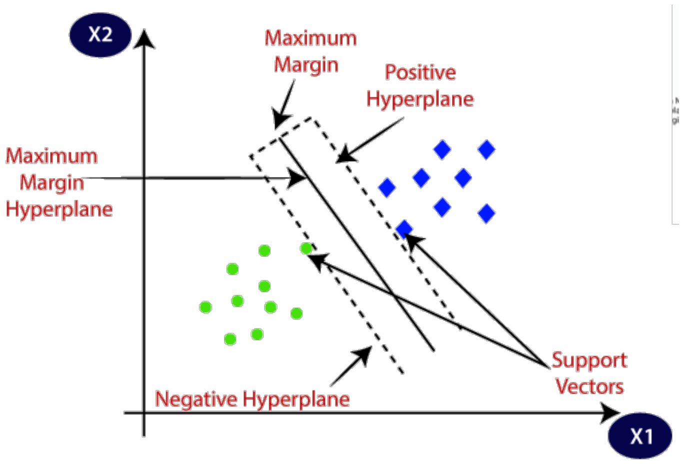

Fig 10: SVM

Additionally, [5] the regularization parameter (C=0.1) ensures a balance between fitting the training data well and preventing overfitting, we employed a linear SVM, characterized by a linear decision boundary. This type of SVM is effective when the relationship between features and classes is approximately linear, as observed in our dataset.
The SVM with a linear kernel and C=0.1 demonstrated performance of 78.67%. The selected features - Age, MMSE, and eTIV - played a crucial role in determining the optimal hyperplane.

The SVM with a linear kernel and C=0.1 demonstrated strong performance, achieving an accuracy of 78.67%. The selected features - Age, MMSE, and eTIV - played a crucial role in determining the optimal hyperplane. This model is well-suited for our dataset, effectively classifying instances and providing a reliable foundation for further analysis.

2.2.8)Ada-Boost classifier

Adaptive Boosting, or AdaBoost, [17] is an ensemble learning method that builds a powerful predictive model by combining several weak learners. The algorithm allows for the correction of errors in subsequent models by assigning more weight to instances that were incorrectly classified in the previous iteration. This process occurs in each iteration. Until a predetermined number of weak learners, or estimators, are combined, this iterative process is carried out.
The AdaBoost Classifier was employed in our model, with the parameter n_estimators=100. In other words, the algorithm trained 100 weak learners, each of which concentrated on fixing the mistakes made by the one before it. The selection of 100 estimators achieves a compromise between retaining computational efficiency and reaching high accuracy.
The AdaBoost Classifier, with its ensemble approach, achieved an accuracy of 80%. The selection of noteworthy features, including Age, MMSE, and eTIV, played a pivotal role in enhancing the model's predictive capabilities. The iterative nature of AdaBoost, emphasizing instances with classification difficulties, makes it a powerful tool for improving model performance over time.

2.2.9)XG-Boost Classifier

XGBoost, also known as eXtreme Gradient Boosting, is an ensemble learning technique that builds decision trees one after the other. Every tree makes up for the mistakes made by the one before it, creating a reliable and accurate model. XGBoost blends the adaptability of gradient boosting with the advantages of boosting algorithms such as AdaBoost. It continuously improves the model's predictions by minimising the loss function through gradient descent optimisation.Our model made use of XGBoost's default hyperparameters, which represent the default setup without any

additional tuning. XGBoost is a dependable option for a range of datasets because its default settings frequently yield excellent performance.
The XGBoost model, with its default hyperparameters, demonstrated exceptional accuracy, achieving a rate of 84%. Noteworthy features, including Age, MMSE, and eTIV, played a vital role in enhancing predictive performance. The ensemble nature of XGBoost ensures that the model is robust, making it a powerful tool for our dataset.

- **Let's now explore the methodology applied to MRI images datasets.**

Apart from analysing numerical data, we extended our methodology to integrate significant insights obtained from imaging data. Our goal was to find complex patterns and relationships in the visual data contained in brain scans by utilising cutting-edge image processing techniques and deep learning. This multifaceted approach broadens the scope of our analysis and creates new opportunities for gaining a comprehensive understanding of Alzheimer's disease.".In the upcoming sections, we'll explain in detail how we selected our data, the Algorithms we used, and the steps we took to ensure the study was conducted ethically and fairly.

2.3) Data Selection and Pre-processing :

2.3.1) Data Selection

We acquired the pre-processed dataset labelled Alzheimer's MRI images from Kaggle. We can observe that this dataset contains the following figures: "Moderate_Demented," "Mild_Demented," "Very_Mild_Demented," and "Non_Demented." Images that are exclusive to their respective categories are gathered in these directories to guarantee a fair representation.'Moderate_Demented' comprises 64 images, 'Mild_Demented' 896 images, 'Very_Mild_Demented' 2240 images, and 'Non_Demented' 3200 images. Our deep learning and classification algorithms are built on this careful selection, which guarantees that they are trained on a wide range of visual data.

Fig 11.1: Very mild demented

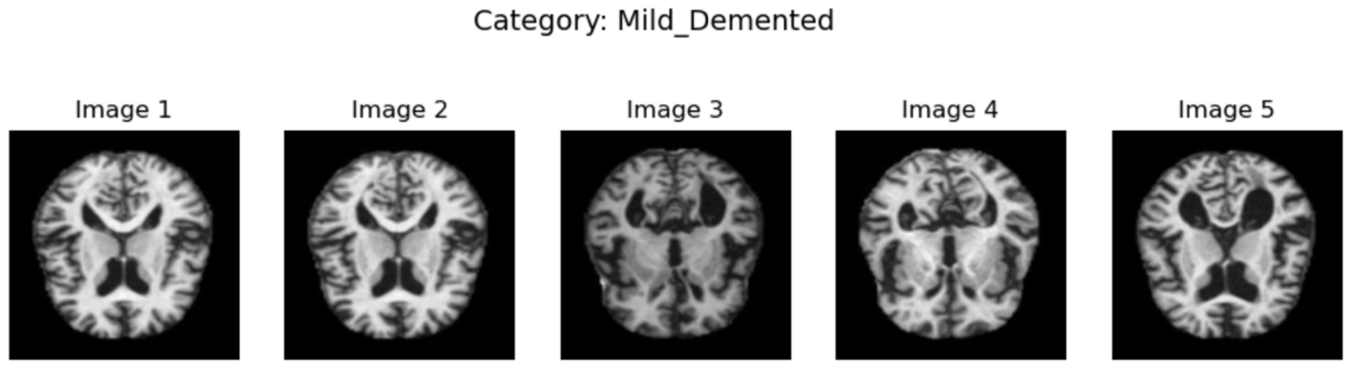

Fig 11.2: Mild demented

Fig 11.3: Moderate demented

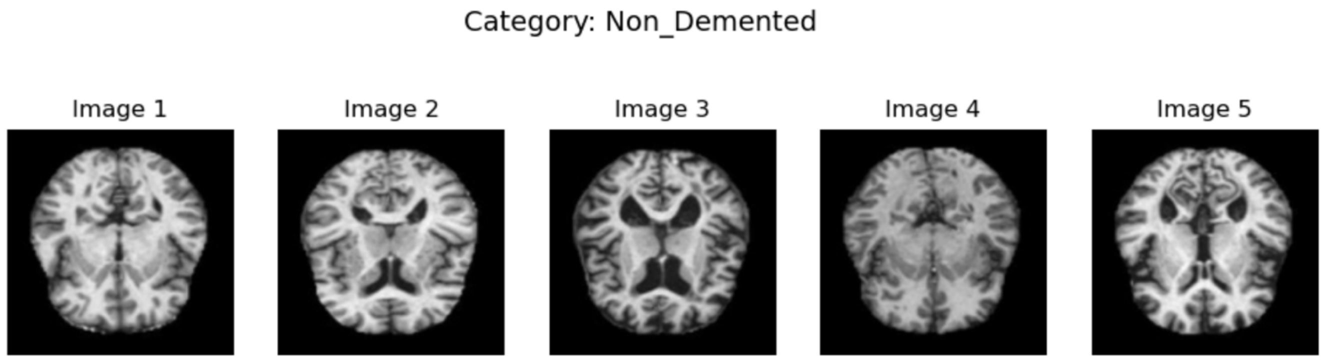

Fig 11.4: Non Demented

2.3.2)Data pre-processing

Our primary goal during the early phases of data pre-processing was to get the raw image data ready for the best possible integration with deep learning algorithms. There were multiple crucial steps in the process, all aimed at improving the dataset's uniformity and quality.

Sizing and Loading:
First, we loaded pictures from various subdirectories, each of which stood for a particular category associated with stages of Alzheimer's disease. "Moderate_Demented," "Mild_Demented," "Very_Mild_Demented," and "Non_Demented" are some of these categories. In order to guarantee uniform processing, every image was resized to a common 224 x 224 pixel size. For effective model training and interpretation, this uniformity is essential.

Normalisation:
The images' pixel values were adjusted to a uniform scale with a range of 0 to 1. This normalisation procedure, which is accomplished by dividing each pixel value by 255, encourages uniform and effective learning across all images. Standardising pixel values guarantees that, despite differences in initial pixel intensity, the deep learning model understands features consistently.
Pseudo for normalization:

Data Augmentation:
We used data augmentation techniques to diversify the dataset because we understood the value of robust model training. These methods, which include horizontal flipping, shearing, rotation, width and height shifts, and zooming, add variability and stop the model from becoming overfit to certain features in the training set.

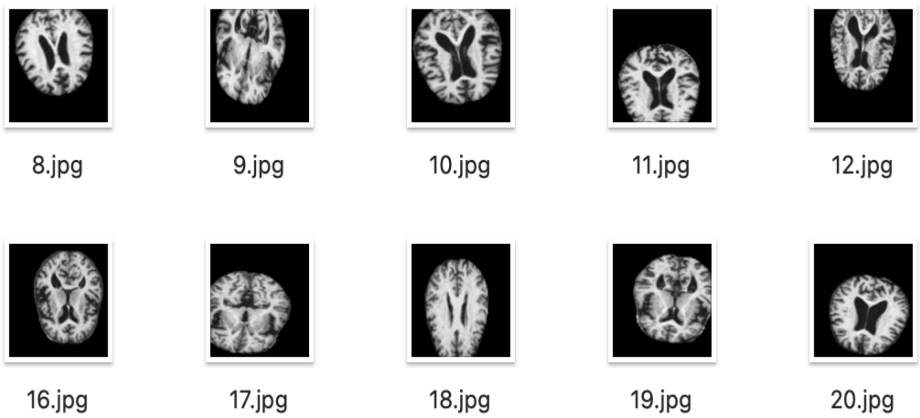

Fig 12: Augmented images

Storage of Augmented Images:

As a by product of data augmentation, augmented images were methodically kept in a special directory. For upcoming phases of model development and assessment, this well-organized storage guarantees accessibility and usability.

Feature extraction with VGG16:

We used the trained VGG16 model to extract relevant features from the pictures. We concentrated on extracting features from the final pooling layer by setting the model to exclude its top classification layers. The model is able to capture high-level representations of image content thanks to this calculated decision.

Storage of Features:

The features that were extracted were carefully saved together with the appropriate category labels. Now enhanced with extracted features, this extensive dataset serves as the basis for further tasks like classification and clustering. This procedure guarantees that the deep learning algorithms we use are capable of identifying complex relationships and patterns in the image data.

To sum up, our methodical approach to pre-processing and data selection establishes a strong basis for detailed MRI image analysis of Alzheimer's disease. Through careful selection of a wide range of datasets and application of fundamental methods like resizing, normalisation, and data augmentation, we guarantee consistency, quality, and resilience in our visual data. Our dataset is further enhanced by the addition of features that were extracted using VGG16, which enables deep learning algorithms to identify complex patterns and relationships in brain scans. The following sections will reveal the complex algorithms that were used on this prepared dataset as we move forward, providing insight into our pursuit of a complete understanding of Alzheimer's disease.

2.4) Implementation of classification and deep learning algorithms:

We now have our pre-processed MRI image dataset, so the next step in our research is to use advanced classification algorithms. Every algorithm will be put through a rigorous evaluation process using our carefully chosen dataset in order to improve accuracy and understanding. The following sections will explore the complexities of the algorithms used, revealing their special techniques, parameters, and important characteristics. By using cutting-edge imaging technology and machine learning, we hope to gain deep insights into Alzheimer's disease and advance our understanding of the condition.

Used Algorithms:
Similar to our numerical dataset, we employed a few classification algorithms here as well, including

I. Support Vector Machine (SVM):

In this case, a linear kernel was employed for its simplicity and efficiency. The data, after flattening to a one-dimensional array, was fed into the SVM model. The model achieved an accuracy of 96.25%, showcasing its ability to discern between different stages of Alzheimer's disease.

II. Logistic Regression:

Despite displaying a slightly lower accuracy of 95.16%, the logistic regression model still performed well in distinguishing between Alzheimer's disease stages.

III. Multinomial Naive Bayes:

The Multinomial Naive Bayes algorithm, which assumes multinomially distributed data, achieved an accuracy of 66.56%. While it displayed reasonable performance, Naive Bayes models might not capture complex relationships as effectively as some other algorithms

IV. k-Nearest Neighbors :

The KNN algorithm, a non-parametric method, demonstrated impressive accuracy at 98.91%. By considering the similarity of a data point to its k-nearest neighbors, the model excelled in capturing patterns in the image data, especially in distinguishing between different stages of Alzheimer's disease.

V. Decision Tree:

The Decision Tree algorithm, a versatile method for classification tasks, achieved an accuracy of 71.56%.
In addition to traditional machine learning algorithms, we incorporated advanced deep learning techniques, specifically Convolutional Neural Networks (CNN) and Artificial Neural Networks (ANN), to further enhance the model's performance and capture intricate patterns within the Alzheimer's disease image dataset.

VI. Convolutional Neural Network (CNN)

In our Alzheimer's disease classification model, we employed CNN, a specialized deep learning architecture tailored for image-related tasks. A CNN is uniquely designed to automatically learn hierarchical representations of visual features from images, making it particularly effective in discerning complex patterns and structures. The architecture comprises convolutional layers that apply filters to input images, extracting relevant features, and max-pooling layers that reduce spatial dimensions, preserving essential information.

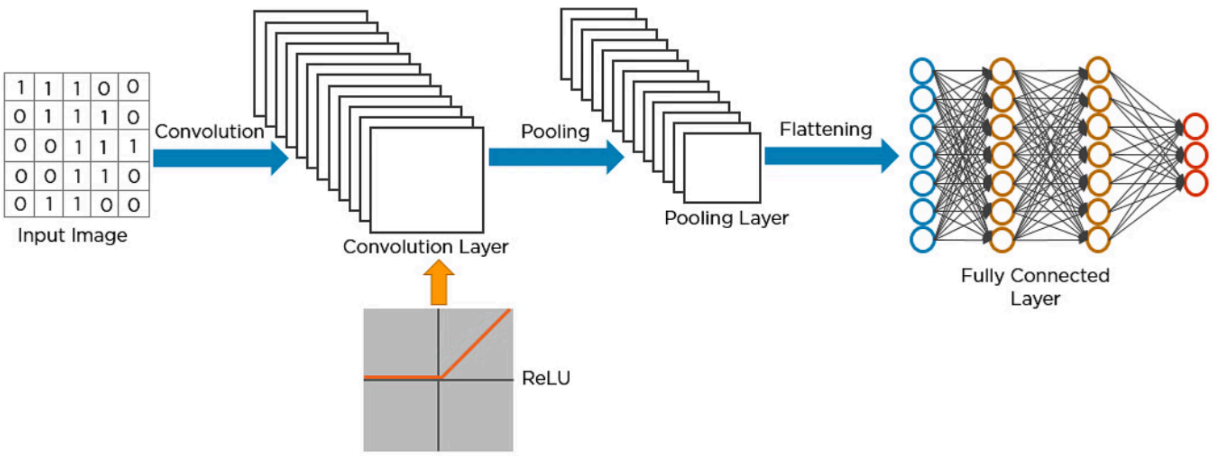

Fig 13: Implementation of CNN

In our implementation, the CNN model consists of three convolutional layers, each followed by max-pooling, creating a hierarchical feature extraction mechanism. The Rectified Linear Unit activation function is applied to introduce non-linearity and capture intricate relationships within the data. The final layers include a flattening operation to convert the extracted features into a one-dimensional array, a dense layer with ReLU activation for further feature refinement, and a dense output layer with softmax activation, producing probabilities for each class in our data.
To enhance the model's performance, we utilized the Adam optimizer and categorical cross-entropy loss function during compilation. The model was trained for 10 epochs on the training dataset, with validation on the test set after each epoch. The accuracy metric was employed to assess the model's performance.
[11] The CNN's strength lies in its ability to automatically learn intricate patterns and features from images, contributing significantly to the classification of Alzheimer's disease stages based on the provided dataset. This architecture's hierarchical feature extraction and pattern recognition capabilities played a pivotal role in achieving high accuracy and robust performance in our classification model. The Accuracy of CNN Model is 0.875

VII. Artificial Neural Network (ANN))

In our Alzheimer's disease classification model, the Artificial Neural Network (ANN) assumes a crucial role in discerning intricate patterns within the dataset. The ANN architecture consists of flattened input layers, followed by densely connected layers with Rectified Linear Unit (ReLU) activation functions. This design enables the network to automatically learn and extract complex hierarchical representations of features present in the Alzheimer's disease images. The multiple hidden layers with varying numbers of neurons facilitate the learning of diverse and abstract features, enhancing the model's ability to capture nuances in the input data.

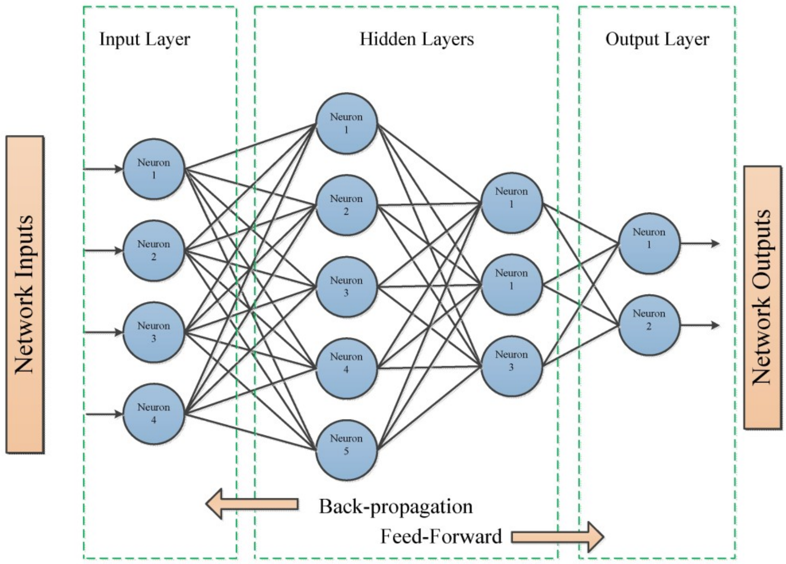

Fig 14: Implementation of ANN

The final output layer, equipped with a softmax activation function, produces probability distributions across the four disease stages: Mild-demented, Moderate-demented, Very-Mild-demented, and Non-demented. During training, the Adam optimizer and categorical cross-entropy loss function are employed to optimize the model's weights and biases, ensuring robust learning. The ANN undergoes training for 10 epochs, with performance evaluation based on accuracy metrics to gauge its effectiveness in classifying Alzheimer's disease stages. This ANN architecture serves as a potent tool in our model, providing a framework for systematic feature learning and contributing to the model's overall classification accuracy. The Accuracy of ANN Model is 0.8086.

2.5)Model evaluation

Following the training phase, an in-depth evaluation was carried out to determine how well the developed Alzheimer's disease classification model performed using data that had never been seen before. To ensure the model's generalizability, real-world scenarios were simulated using a different test set from the training data. A variety of assessment metrics were used in order to give a thorough grasp of the model's advantages and disadvantages.

2.5.1. Accuracy:
It represents the overall correctness of the model's predictions , In Alzheimer's disease detection, accuracy measures how well the model correctly identifies both positive and negative instances among all instances.
High accuracy indicates the model's ability to make correct predictions across different stages of Alzheimer's disease.

Accuracy = Number of Correct Predictions/Total Number of Predictions

2.5.2. Precision:
Precision quantifies the accuracy of positive predictions , In Alzheimer's detection, It measures how many of the predicted positive cases are actually true positive cases .High precision suggests that when the model predicts Alzheimer's, it is often correct, minimizing false positives.

Precision=(True Positives)/(True Positives + False Positives)

2.5.3. Recall (Sensitivity):
Recall measures the model's ability to capture all positive instances , In Alzheimer's detection, recall assesses how well the model identifies true positive cases among all actual positive cases. High recall indicates that the model effectively identifies individuals with Alzheimer's disease, minimizing false negatives.

Recall= (True Positives) / (False Negatives + True Positives)

2.5.4. F1 Score:
It is the harmonic mean of precision and recall , In Alzheimer's disease detection, F1 score provides a balanced measure of the model's accuracy, considering both false positives and false negatives.
A high F1 score indicates a model that performs well in terms of both precision and recall, striking a balance between avoiding false positives and false negatives.

F1-Score=2×(Precision * Recall)/(Precision+Recall)

6. Confusion Matrix:

It provides a detailed breakdown of the model's predictions, highlighting true positives, true negatives, false positives, and false negatives. This matrix aids in understanding the specific areas where the model excels or requires improvement.

2.5.7 Receiver Operating Characteristic (ROC) Curve :

In binary or multiclass classification scenarios, the ROC curve illustrates the trade-off between true positive rate and false positive rate. It helps visualize the model's discriminatory ability across different thresholds.

The combination of these evaluation metrics offers a comprehensive assessment of the model's performance, ensuring a thorough understanding of its capabilities and areas for potential refinement. The results of the model evaluation contribute valuable insights to the interpretation and trustworthiness of the developed Alzheimer's disease classification model.

In summary, our methodology integrates numerical and MRI datasets to enhance Alzheimer's disease detection. Robust pre-processing and feature analysis were conducted on the numerical dataset, followed by the application of diverse machine learning algorithms. The MRI dataset underwent careful data selection, normalization, augmentation, and feature extraction using VGG16, leading to the implementation of classification and deep learning algorithms.

Moving forward, we will discuss the obtained results, conducting a comparative analysis of algorithm performance on both datasets. This following assessment aims to draw conclusions about the effectiveness of machine learning and deep learning in Alzheimer's disease detection, contributing valuable insights to the field.

3.) Discussion

We have used a combination of methods to examine the detection of Alzheimer's disease, combining MRI and numerical datasets with a wide range of machine learning and deep learning algorithms. Looking more closely at the outcomes of this thorough methodology, we need to evaluate each algorithm's performance and determine how it contributes to the accuracy of Alzheimer's diagnosis.
The following are the numerical dataset accuracy values, where the training set comprises 80% and the testing test comprises 20%.

Sno	**Model**	**Accuracy**
1	Decision Tree	70.67
2	Random Forest	77.33
3	XG Boost	84
4	Gaussian NB	73.33
5	Bernoulli NB	56
6	SVM	78.67
7	Ada Boost	80
8	KNN	73.33
9	Logistic Regression	64

Table 1: Numeric Dataset Accuracies

Further using our numerical dataset, we specifically applied cross-validation techniques in order to achieve robust Alzheimer disease detection. By reducing the likelihood of overfitting, this technique—which makes use of Python's scikit-learn library—improves the dependability and generalizability of our machine learning models.
We divided our dataset effectively for iterative training and validation using a variety of models and five-fold cross-validation. To fully evaluate the model's performance, the outcomes—which included individual scores, mean accuracy, and standard deviation.
In order to ensure that our models adjust well to the subtle differences in a variety of datasets, cross-validation is crucial in preventing overfitting. This methodical approach directs the selection of models based on consistency across various dataset subsets as well as accuracy.
In conclusion, our thoughtful application of cross-validation serves as an essential quality control tool, offering a more

The accuracy comparison table before and after cross validation is shown below.

Sno	Model	Accuracy_CrossValidation	Accuracy Validation
0	Decision Tree	0.7371	0.7066
1	Random Forest	0.7588	0.7866
2	XG Boost	0.7586	0.84
3	Gaussian NB	0.7692	0.7333
4	Bernoulli NB	0.6112	0.56
5	SVM	0.7825	0.7866
6	Ada Boost	0.7559	0.8
7	KNN	0.579	0.7333
8	Logistic Regression	0.7076	0.64

Table 2: Comparision of Accuracies with cross validation

The initial accuracy assessment revealed differences in the performance of various models on the numerical dataset prior to the implementation of cross-validation. XG Boost, in particular, showed the best accuracy at 84%, proving how useful it is for identifying Alzheimer's disease. Nevertheless, models such as Bernoulli NB showed a reduced accuracy of 56%, indicating possible limits in managing the particular features of the dataset.

Following the application of cross-validation, we noticed variations in the models' accuracy scores. XG Boost demonstrated consistent performance by maintaining its high accuracy. Additionally, models demonstrating improvements included Decision Tree and Ada Boost, demonstrating the beneficial effects of cross-validation on the generalizability and reliability of models.
This comparison demonstrates how important cross-validation is for improving machine learning models' overall performance evaluation. Through the iterative training and validation of our models on various subsets of the dataset, cross-validation reduces the possibility of overfitting and offers a more precise depiction of their efficacy in a range of data scenarios. The accuracy gains that have been observed highlight how important this method is for guaranteeing the stability and dependability of our models for detecting Alzheimer's disease.

So far, our focus has been on the accuracy values obtained from the numerical dataset. Now, let's delve into the insights derived from the accuracy assessments of the MRI image dataset. This exploration involves a comprehensive examination of the dataset after pre-processing and the application of various machine learning and deep learning models. The following accuracy values shed light on the performance of these models in the context of Alzheimer's disease detection from MRI images.

Sno	Model	Accuracy
1	SVM	0.9625
2	Logistic Regression	0.95
3	Multinomial NB	0.66
4	KNN	0.98
5	Decision Tree	0.71
6	CNN	0.87
7	ANN	0.8

Table 3: Image pre-processing accuracies

The accuracy assessments of the MRI image dataset have provided valuable insights into the performance of various machine learning and deep learning models for Alzheimer's disease detection. Notably, models such as SVM, KNN, and CNN demonstrated high accuracy, with scores of 96.25%, 98%, and 87%, respectively. These results highlight the efficacy of these models in accurately identifying patterns and features indicative of Alzheimer's disease within MRI images. On the other hand, models like Multinomial NB exhibited a lower accuracy of 66%, indicating potential limitations in handling the complexity of the image dataset. In conclusion, the robust performance of certain models underscores the promising potential of utilizing MRI images for accurate Alzheimer's disease detection, while acknowledging the variations in performance among different algorithms.

4. Concluding Remarks

In concluding our study on Alzheimer's disease identification, our innovative approach merges MRI and numerical datasets with diverse machine learning and deep learning algorithms. Our in-depth analysis of the numerical dataset highlighted the effectiveness of XG Boost, achieving an outstanding 84% accuracy, particularly notable when incorporating cross-validation.

The MRI image dataset exhibited remarkable accuracies, notably with CNN (87%), KNN (98%), and SVM (96.25%), showcasing the capability to accurately diagnose Alzheimer's using MRI images. Significantly, our model distinguishes itself by presenting a comprehensive diagnostic methodology, seamlessly integrating numerical and image data. The meticulous application of cross-validation ensures the reliability and consistency of our findings.

In essence, our study not only propels advancements in Alzheimer's detection techniques but positions our model as a promising breakthrough that surpasses certain existing research models. It provides a nuanced perspective for swift and precise diagnosis, emphasizing the potential of integrated datasets and sophisticated algorithms for enhancing healthcare outcomes.

5. Future Works

In light of the study, future research could focus on improving the numerical dataset pre-processing procedures even more. Deeper understanding might be gained by investigating sophisticated feature engineering methods unique to Alzheimer's symptoms. Further research into the effects of unbalanced datasets on model performance may open the door to more focused enhancements.

Practically speaking, working with healthcare organisations could make it easier to integrate patient data from the real world, overcoming issues with differentiating data quality and demographics. It might also be investigated how federated learning techniques might be used to enable models to learn from decentralised data sources while maintaining privacy.

It is morally required to take into account any biases in the models and dataset in order to ensure equitable representation for all demographic groups. It is increasingly important to stay up to date on ethical standards as the field develops and to actively involve stakeholders in the decision-making processes surrounding the use of models in clinical settings.

To further improve the precision, understanding, and relevance of Alzheimer's disease detection, a comprehensive strategy comprising domain-specific improvements, cutting-edge technology integration, and ethical considerations is outlined in the future trajectory.

6.) References

1. Kavitha, C., Mani, V., Srividhya, S. R., Khalaf, O. I., & Tavera Romero, C. A. (2022). Early-stage Alzheimer's disease prediction using machine learning models. *Frontiers in public health*, *10*, 853294.
2. Liu, S., Masurkar, A. V., Rusinek, H., Chen, J., Zhang, B., Zhu, W., ... & Razavian, N. (2022). Generalizable deep learning model for early Alzheimer's disease detection from structural MRIs. *Scientific reports*, *12*(1), 17106.
3. Neelaveni, J., & Devasana, M. S. G. (2020). Alzheimer Disease Prediction using Machine Learning Algorithms. 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS).
4. Bari Antor, M., Jamil, A. H. M., Mamtaz, M., Monirujjaman Khan, M., Aljahdali, S., Kaur, M., ... & Masud, M. (2021). A comparative analysis of machine learning algorithms to predict alzheimer's disease. Journal of Healthcare Engineering, 2021.
5. Wang, J., Huang, S., Wang, Z., Huang, D., Qin, J., Wang, H., ... & Liang, Y. (2023). A calibrated SVM based on weighted smooth GL1/2 for Alzheimer's disease prediction. Computers in Biology and Medicine, 158, 106752.
6. Seixas, F. L., Zadrozny, B., Laks, J., Conci, A., & Saade, D. C. M. (2014). A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer′ s disease and mild cognitive impairment. Computers in biology and medicine, 51, 140-158.
7. Porsteinsson, A. P., Isaacson, R. S., Knox, S., Sabbagh, M. N., & Rubino, I. (2021). Diagnosis of early Alzheimer's disease: clinical practice in 2021. The journal of prevention of Alzheimer's disease, 8, 371-386.
8. Battineni, G., Chintalapudi, N., Amenta, F., & Traini, E. (2020). A comprehensive machine-learning model applied to magnetic resonance imaging (mri) to predict alzheimer's disease (ad) in older subjects. Journal of Clinical Medicine, 9(7), 2146.
9. Mittelman, M. S., Ferris, S. H., Steinberg, G., Shulman, E., Mackell, J. A., Ambinder, A., & Cohen, J. (1993). An intervention that delays institutionalization of Alzheimer's disease patients: treatment of spouse-caregivers. The Gerontologist, 33(6), 730-740.
10. Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., ... & Beckett, L. (2005). Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimer's & Dementia, 1(1), 55-66.
11. Ajagbe, S. A., Amuda, K. A., Oladipupo, M. A., Oluwaseyi, F. A., & Okesola, K. I. (2021). Multi-classification of Alzheimer disease on magnetic resonance images (MRI) using deep convolutional neural network (DCNN) approaches. International Journal of Advanced Computer Research, 11(53), 51.
12. Paulsen, J. S., Nance, M., Kim, J. I., Carlozzi, N. E., Panegyres, P. K., Erwin, C., ... & Williams, J. K. (2013). A review of quality of life after predictive testing for and earlier identification of neurodegenerative diseases. Progress in neurobiology, 110, 2-28.
13. Kapasi, A., & Schneider, J. A. (2016). Vascular contributions to cognitive impairment, clinical Alzheimer's disease, and dementia in older persons. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1862(5), 878-886.
14. Korolev, I. O., Symonds, L. L., Bozoki, A. C., & Alzheimer's Disease Neuroimaging Initiative. (2016). Predicting progression from mild cognitive impairment to Alzheimer's dementia using clinical, MRI, and plasma biomarkers via probabilistic pattern classification. PloS one, 11(2), e0138866.
15. Kato, S., Homma, A., Sakuma, T., & Nakamura, M. (2015, August). Detection of mild Alzheimer's disease and mild cognitive impairment from elderly speech: Binary discrimination using logistic regression. In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 5569-5572). IEEE.
16. Ramírez, J., Górriz, J. M., Salas-Gonzalez, D., Romero, A., López, M., Álvarez, I., & Gómez-Río, M. (2013). Computer-aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of features. Information Sciences, 237, 59-72.
17. Singh, D., & Mishra, A. (2021). Multiple-ensemble methods for prediction of Alzheimer disease. International Journal of Data Mining and Bioinformatics, 26(1-2), 99-128.