A Mini-Project 1 Report on

Alzheimer's Disease Detection using Ensemble Machine Learning Algorithms

Submitted to the Department of Computer Science & Engineering, GNITS in the partial fulfillment of the academic requirement for the award of B.Tech (CSE) under JNTUH, Hyderabad

By

K.Yasaswitha (22251A0516)

under the guidance of

Mrs. R. Mamatha
Assistant Professor

Department of Computer Science & Engineering
G. NARAYANAMMA INSTITUTE OF TECHNOLOGY & SCIENCE
(Autonomous) (For Women)
Approved by AICTE, New Delhi & Affiliated to JNTUH, Hyderabad
Accredited by NBA & NAAC, an ISO 9001:2015 certified Institution
Shaikpet, Hyderabd-500104
July 2024

G. Narayanamma Institute of Technology & Science

(Autonomous) **(For Women)**

Approved by AICTE, New Delhi & Affiliated to JNTUH, Hyderabad
Accredited by NBA & NAAC, an ISO 9001:2015 certified Institution
Shaikpet, Hyderabd-500104

Department of Computer Science & Engineering

Certificate

This is to certify that the Mini-Project 1 report on **"ALZHEIMER'S DISEASE DETECTION USING ENSEMBLE MACHINE LEARNING ALGORITHMS"** is a bonafide work carried out by **K.Yasaswitha (22251A0516)** in the partial fulfillment for the award of B.Tech degree in Computer Science & Engineering, G. Narayanamma Institute of Technology & Science, Shaikpet, Hyderabad, affiliated to Jawaharlal Nehru Technological University, Hyderabad under our guidance and supervision for the academic year 2023-2024.

The results embodied in the Mini project work have not been submitted to any other University or Institute for the award of any degree or diploma.

Internal Guide	**Head of the Department**
Mrs. R. Mamatha	**Dr. A. Sharada**
Assistant Professor, CSE	**Professor and Head, CSE**

Acknowledgements

We would like to express our sincere thanks to **Dr. K. Ramesh Reddy**, Principal, GNITS, for providing the working facilities in the college.

Our sincere thanks and gratitude to **Dr. K. Rama Linga Reddy**, Professor & Dean Academics, Department of ETE, **Dr. M. Seetha,** Professor & Dean R&D, Department of CSE, **Dr. N. Kalyani,** Professor & Dean of Innovation and Incubation Department of CSE, GNITS, for all the timely support and valuable suggestions during the period of our project.

We extend our heartfelt gratitude to **Dr. A. Sharada,** Professor & Head, Department of Computer Science and Engineering, GNITS, for their unwavering support and invaluable guidance throughout our project, providing timely assistance and insightful suggestions.

We are extremely thankful to **Mrs. K. Sindhura,** Assistant Professor overall project coordinator, Department of CSE, GNITS for all the valuable suggestions and guidance during the period of our project.
We are extremely thankful to **Mrs. K. Sindhura,** Assistant Professor**, Mrs. G. Mythili Sharvani,** Assistant Professor, Department of CSE, GNITS, project coordinators for their encouragement and support throughout the project.

We are extremely thankful and indebted to our internal guide, **Mrs. R. Mamatha,** Assistant Professor, Department of CSE**,** GNITS for her constant guidance, encouragement and moral support throughout the project.

Finally, we are extremely thankful to all the faculty members and staff of CSE Department who helped us directly or indirectly, parents and friends for their cooperation in completing the project work.

K.Yasaswitha (22251A0516)

ABSTRACT

It is predicted that there are around 55 million people across the globe who suffer from dementia, over 8.8 million are from India. Alzheimer's disease (AD) is a chronic and irreversible brain illness for which no effective treatment exists. There is a new case of Alzheimer's disease being discovered globally in every four seconds. It is predicted that it is going to be affected by to around 78 million people by 2030 in the world. It is the leading cause of dementia in older adults.

Alzheimer's is caused by the combination of age-related changes in the brain, along with genetic, environmental, and lifestyle factors. Alzheimer's disease primarily affects the hippocampus and cerebral cortex in the brain, leading to memory loss and cognitive decline. However, available medicines can only slow the progress of the disease. Thus, stopping and regulating the progression of AD depends greatly on its early identification. Early detection of Alzheimer's is thus very important. Cognitive Impairment, Genetic Markers, Intracranial Volume (ICV) are among the important features that can be used in the detection of Alzheimer's. The main reason Alzheimer's is challenging is because it has many causes and involves complex changes in the brain, making it hard to diagnose and treat; it remains a challenge to medical fraternity.

The main aim of this project is to leverage Ensemble Machine Learning Models to enhance performance using data cues for the early detection of Alzheimer's disease. Machine Learning approaches, such as Decision Tree, Random Forest, Support Vector Machine, Gradient Boosting, and Voting classifiers, have proven increasingly useful for Alzheimer's detection. We employ these algorithms to identify the optimal parameters for prediction and demonstrate that our ensemble method, which combines multiple features from the dataset, is robust and computationally efficient. Our results show that the accuracy of the ensemble model surpasses that of each individual algorithm, highlighting its effectiveness for Alzheimer's disease prediction.

Table of Contents

List of Figures

List of Tables

Acronyms

- **MRI:** Magnetic Resonance Imaging
- **AD:** Alzheimer's Disease
- **OASIS:** Open Access Series of Imaging Studies
- **MMSE:** Mini-Mental State Examination
- **CDR:** Clinical Dementia Rating
- **eTIV:** Estimated Total Intracranial Volume
- **nWBV:** Normalized Whole Brain Volume
- **ASF:** Atlas Scaling Factor
- **SES:** Socioeconomic Status

1. INTRODUCTION

1.1 Background Of The Study

Every three seconds, someone develops dementia. Lee and her colleagues recruited over 4,000 seniors (ages 60 and older) from the Longitudinal Aging Study in India (LASI) for interviews and neuropsychological tests. A subsample of 2,528 participants received expert-reviewed dementia diagnoses. Using a logistic regression model, they predicted dementia status for 28,949 seniors, estimating a 7.4% prevalence among adults over 60 in India, equating to about 8.8 million people. This is significantly higher than the 2020 Dementia in India report's estimate of 5.3 million.

Worldwide, at least 55 million people are believed to be living with Alzheimer's disease or other dementias. If breakthroughs are not discovered, rates could exceed 152 million by 2050.

Dementia:

Dementia is not a specific disease. It’s an overall term that describes a group of symptoms associated with a decline in memory or other thinking skills severe enough to reduce a person’s ability to perform everyday activities. Alzheimer’s disease accounts for 60 to 80 percent of cases. Vascular dementia, which occurs after a stroke, is the second most common dementia type. But there are many other conditions that can cause symptoms of dementia, including some that are reversible, such as thyroid problems and vitamin deficiencies.

Dementia is a general term for loss of memory and other mental abilities severe enough to interfere with daily life. It is caused by physical changes in the brain. Alzheimer's is the most common type of dementia, but there are many kinds.

Alzheimer’s Disease:

Alzheimer’s disease is the most common type of dementia. It is a progressive disease beginning with mild memory loss and possibly leading to loss of the ability to carry on a conversation and respond to the environment. Alzheimer’s disease involves parts of the brain that control thought, memory, and language. It can seriously affect a person’s ability to carry out daily activities. The exact causes of Alzheimer's disease aren't fully understood. But at a basic level, brain proteins fail to function as usual. This disrupts the work of brain cells, also called neurons, and triggers a series of events. The neurons become damaged and lose connections to each other. They eventually die. Scientists believe that for most people, Alzheimer's disease is caused by a combination of genetic, lifestyle and environmental factors that affect the brain over time. In less than 1% of cases, Alzheimer's is caused by specific genetic changes that almost guarantee a person will develop the disease. In these cases, the disease usually begins in middle age. The development of the disease begins years before the first symptoms. The damage most often starts in the region of the brain that controls memory. The loss of neurons spreads in a somewhat predictable pattern to other regions of the brain. By the late stage of the disease, the brain has shrunk significantly. Alzheimer's disease starts out by affecting the part of the brain that's responsible for communication and memory. Because of this disability, AD patients are at a greatly increased risk of developing these cognitive deficits such as verbally, visually, and in writing as shown in Fig 1.1.

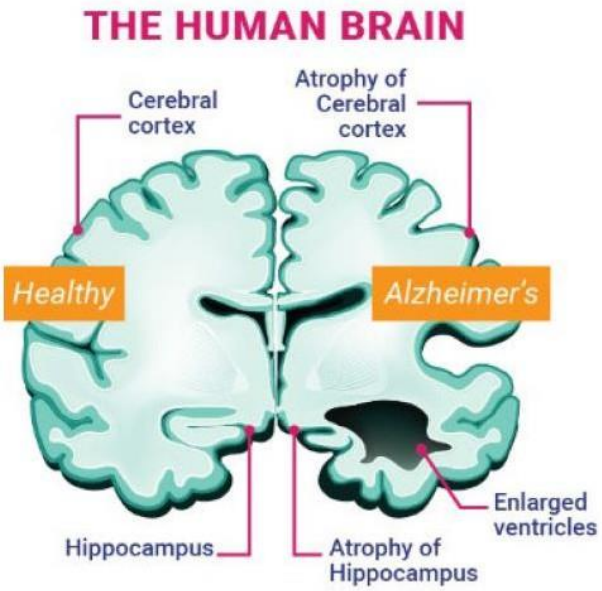

Fig 1.1: Normal brain vs Alzheimer's brain

The facts and figures related to AD as given by the National Library of Medicine show that about 6.5 million Americans of age 65 or above are suffering from this chronic disease. It is assumed that this figure can rise to 13.8 million by the year 2060. Alzheimer's disease was listed as the sixth most frequent cause of death in the United States in 2019 and was ranked 7th in 2020 and 2021 when the whole world was experiencing COVID-19. In America, AD is ranked as the fifth most common cause of death for people of the age group 65 and over. The percentage of the population(millions) affected by AD is shown below in Fig 1.2.

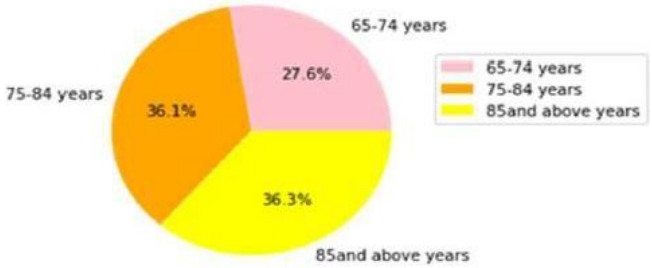

Fig 1.2: Percentage of different age groups of people affected by AD in the US

Some statistics about Alzheimer's:

- Over 55 million people worldwide are living with dementia, with Alzheimer's disease being the most common cause, accounting for 60-70% of cases.
- In the United States, approximately 6.7 million people aged 65 and older are living with Alzheimer's disease as of 2023.
- Nearly 10 million new cases of dementia, including Alzheimer's, are diagnosed each year worldwide.
- The risk of developing Alzheimer's doubles approximately every five years after age 65. Almost one-third of people aged 85 and older have Alzheimer's disease.
- Women are more likely to develop Alzheimer's than men. Nearly two-thirds of Americans with Alzheimer's are women.

- African Americans are about twice as likely, and Hispanics about one and a half times as likely, to develop Alzheimer's as older White adults.
- In 2023, the estimated cost of Alzheimer's and other dementias to the U.S. is $345 billion, including $239 billion in Medicare and Medicaid payments.
- Alzheimer's disease is the sixth leading cause of death in the United States.
- Between 2000 and 2019, deaths from Alzheimer's disease increased by 145%, while deaths from heart disease (the number one cause of death) decreased by 7%.
- As of 2023, there are over 140 active clinical trials for Alzheimer's disease treatments worldwide, exploring various therapeutic approaches including drugs, lifestyle interventions, and medical devices.

Alzheimer's Public Survey Results Commissioned by the Washington State Alzheimer's Disease Working Group

- 94.82% were very or somewhat familiar with early signs or symptoms.
- While 71% of respondents indicated they were very likely to discuss signs or symptoms with their doctor, other people expressed concerns about:
    - Loss of independence
    - Nothing to be done anyway
    - Fear of diagnosis
- The top three recommendations to raise public awareness about Alzheimer's are:
    - Physicians (1,140)
    - Public Service Announcements (952)
    - Aging/Senior Services (914)

Alzheimer's is a type of dementia that causes problems with memory, thinking and behaviour. Symptoms usually develop slowly and get worse over time, becoming severe enough to interfere with daily tasks. Alzheimer's is not a normal part of aging. The greatest known risk factor is increasing age, and the majority of people with Alzheimer's are 65 and older. But Alzheimer's is not just a disease of old age. Approximately 200,000 Americans under the age of 65 have younger-onset Alzheimer's disease (also known as early-onset Alzheimer's). Alzheimer's is the sixth leading cause of death in the United States. Those with Alzheimer's live an average of eight years after their symptoms become noticeable to others, but survival can range from four to 20 years, depending on age and other health conditions. Alzheimer's has no current cure, but treatments for symptoms are available and research continues. Although current Alzheimer's treatments cannot stop Alzheimer's from progressing, they can temporarily slow the worsening of dementia symptoms and improve quality of life for those with Alzheimer's and their caregivers. Today, there is a worldwide effort under way to find better ways to treat the disease, delay its onset, and prevent it from developing.

Types of Alzheimer's Disease

Alzheimer's disease is generally classified into two main types based on age of onset and genetic factors:

1. Early-Onset Alzheimer's Disease (EOAD):

- Age of Onset: Typically affects individuals younger than 65 years old, though it can occur as early as in one's 30s or 40s.

- Genetic Factors: Often associated with genetic mutations such as mutations in the APP (amyloid precursor protein), PSEN1 (presenilin 1), and PSEN2 (presenilin 2) genes.

2. Late-Onset Alzheimer's Disease (LOAD):

- Age of Onset: Most common form, typically occurring after age 65.

- Genetic Factors: Apolipoprotein E (APOE) gene variants, especially APOE-e4 allele, are a significant risk factor.

Classification Based on the Severity

Based on the intensity of the typical Alzheimer's symptoms, it can be classified into the following subtypes:

1. **Mild Alzheimer's**
2. **Moderate Alzheimer's**
3. **Severe Alzheimer's**

Mild Alzheimer's:

Mild Alzheimer's disease represents the early stage of this progressive neurological condition, characterized by subtle but noticeable cognitive decline. Individuals in this stage may experience mild memory lapses, such as forgetting recent conversations or events, and have difficulty with problem-solving and planning tasks. Language difficulties, spatial orientation challenges, and changes in mood or personality can also manifest. Despite these symptoms, individuals with mild Alzheimer's often retain some independence in daily activities, though they may require support and adjustments. Early diagnosis and intervention are critical to managing symptoms and planning for future care needs.

Moderate Alzheimer's

Moderate Alzheimer's disease signifies a progression beyond the early stage, where cognitive decline becomes more pronounced and daily functioning becomes increasingly challenging. Individuals typically experience worsening memory loss, including forgetting personal details and struggling with basic tasks like dressing or grooming. Cognitive impairments extend to difficulties with language, such as finding words and understanding conversations, as well as problems with spatial orientation and judgment. behavioural changes may include increased agitation, aggression, or wandering. In this stage, individuals often require more assistance with daily activities

and may need supervision to ensure their safety and well-being. Caregiver support and structured routines become crucial elements in managing the disease and maintaining quality of life.

Severe Alzheimer's

Severe Alzheimer's disease represents the advanced stage of this debilitating neurological condition, characterized by profound cognitive and functional decline. Individuals in this stage typically experience severe memory loss, often forgetting names of close family members, their own personal history, and even where they are. Communication abilities are severely impaired, with limited speech and difficulty understanding language. Basic activities of daily living, such as eating, dressing, and using the bathroom, require extensive assistance or supervision. Behavioral symptoms can include agitation, aggression, and wandering, requiring constant monitoring and specialized care. Physical capabilities decline significantly, leading to challenges with mobility and increased vulnerability to infections and other health complications. Providing compassionate care and ensuring comfort become primary goals as the disease progresses into its final stages.

1.2 Problem Statement

The project aims to address the challenge of early detection of Alzheimer's disease, leveraging machine learning algorithms to analyze dataset extracted from the Open Access Series of Imaging Studies (OASIS) through the development and evaluation of ensemble machine learning models, demonstrating that the combined approach significantly outperforms individual algorithms in prediction accuracy. With traditional diagnostic methods proving insufficient, there's a critical need for a more accurate and efficient approach to identify the disease in its early stages.

1.3 Existing systems

Existing systems for Alzheimer's detection using machine learning with tabular data focus on leveraging diverse datasets including genetic profiles, cognitive test results, biochemical markers, and demographic information. Machine learning algorithms such as support vector machines (SVM), decision trees, and logistic regression are applied to analyse these data sources. Their objective is to identify patterns and correlations indicative of Alzheimer's disease onset or progression. By integrating and analysing multifaceted data, these systems aim to improve early detection, risk prediction, and personalized treatment strategies, thereby enhancing clinical outcomes and patient care in Alzheimer's disease management [3].

1.4 Advantages & Drawbacks

Advantages of Existing System:

- ML models can detect subtle patterns in data that may indicate early stages of Alzheimer's disease before clinical symptoms appear.
- ML algorithms can integrate multiple types of data (genetic, cognitive, demographic) to enhance diagnostic accuracy compared to traditional methods.

Shortcomings of Existing System:

- Existing system for Alzheimer's detection rely on complex technology that requires extensive datasets, posing challenges for smaller healthcare facilities.
- These systems can be difficult to interpret and may occasionally produce errors. Moreover, their setup and maintenance costs can be prohibitive for smaller healthcare providers.

1.5 Proposed System

The proposed system for Alzheimer's disease detection leverages longitudinal MRI data obtained from the OASIS study, focusing on a comprehensive approach to improve accuracy and reliability in disease prediction. It incorporates various algorithms such as Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and Support Vector Machines to classify patients' data effectively. The system begins with rigorous data preprocessing steps, including handling missing values, standardizing numerical features, and encoding categorical variables to ensure data consistency and compatibility for machine learning algorithms. Key components of the system include the utilization of various machine learning models such as Decision Trees, Random Forests, Support Vector Machines (SVMs), Gradient Boosting Machines, and a Voting Classifier. These models are selected for their ability to capture different aspects of data patterns and complement each other in the ensemble framework, thereby enhancing the overall predictive performance. The system emphasizes thorough evaluation using multiple performance metrics including accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). This evaluation phase is critical for assessing the effectiveness of individual models as well as the ensemble approach in accurately identifying Alzheimer's disease across different stages and manifestations. Furthermore, the system incorporates advanced techniques like hyperparameter tuning using methods such as Grid Search or Random Search to optimize model parameters. This process aims to maximize model performance and generalizability, ensuring robustness in real- world applications. In summary, the proposed system integrates cutting-edge

machine learning techniques with comprehensive data preprocessing, thorough evaluation, and deployment strategies to develop a robust tool for early detection and classification of Alzheimer's disease using longitudinal MRI labelled data.

1.6 Methodology

This section outlines the methodology employed for detecting Alzheimer's disease using longitudinal MRI data, focusing on data preparation, model selection, training, and evaluation as shown in Fig 1.3.

Data Collection and Preprocessing

The dataset used in this study comprises longitudinal MRI data from the OASIS study, including demographic information and neuroimaging metrics. The initial step involved loading the dataset and performing preliminary checks for missing values and duplicates. Missing values in numerical features such as age and cognitive scores were imputed using appropriate statistical methods (e.g., median for age) to ensure data completeness. Categorical variables like gender were encoded to numerical format for compatibility with machine learning algorithms.

Feature Engineering and Scaling

Feature engineering was crucial to enhance the predictive power of the models. Numerical features such as Mini Mental State Examination (MMSE), Estimated Total Intracranial Volume (eTIV), Normalized Whole Brain Volume (nWBV), and Atlas Scaling Factor (ASF) were standardized using techniques like StandardScaler from scikit-learn. This step ensures that all features contribute equally to model training without biases due to differing measurement scales.

Model Selection and Training

Several machine learning algorithms were evaluated for their effectiveness in classifying dementia stages based on longitudinal MRI data. These algorithms include Decision Trees, Random Forests, Support Vector Machines (SVM), Gradient Boosting Machines, and an ensemble approach using a Voting Classifier. Each model was trained on the pre-processed dataset to learn patterns indicative of different stages of Alzheimer's disease progression.

Model Evaluation

Evaluation of model performance was conducted using various metrics such as accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These metrics provide insights into how well models classify different stages of Alzheimer's disease. Cross-validation techniques were employed during model evaluation to ensure robustness and generalizability of the models.

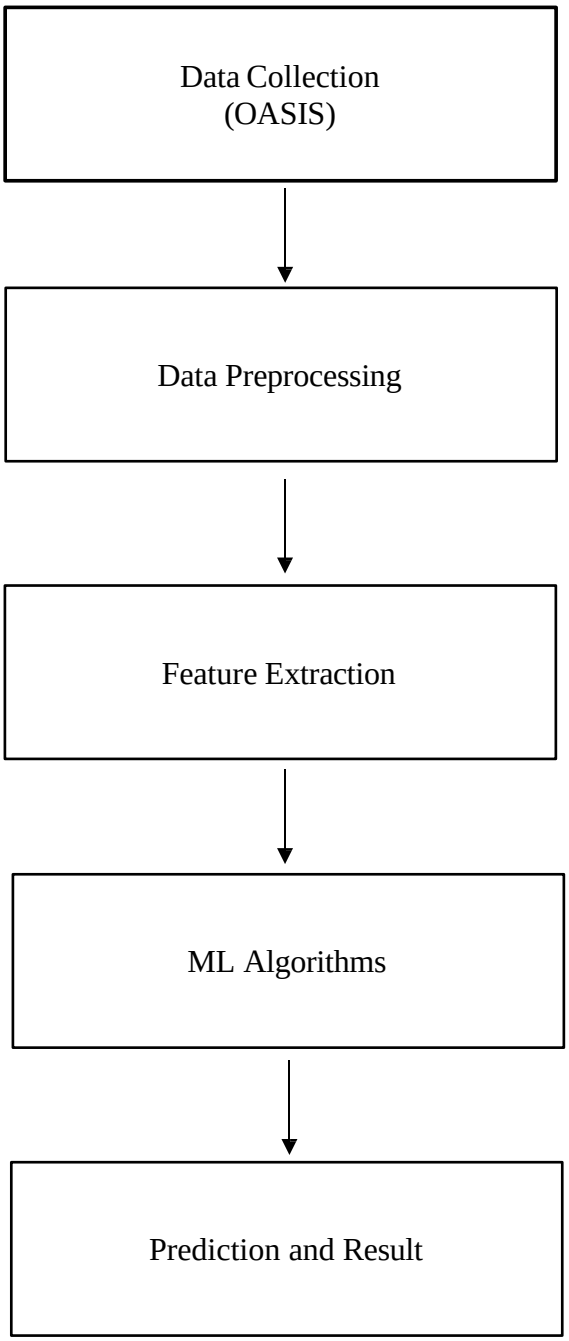

Fig 1.3: Flow diagram of Alzheimer's detection

1.7 Objectives of the project

- To develop a machine learning model capable of accurately detecting Alzheimer's disease by analyzing longitudinal MRI data.
- To preprocess the longitudinal MRI data by handling missing values, scaling features, and splitting the dataset into training and testing sets.
- To improve efficiency which is targeted through comparing machine learning model performance across metrics like accuracy, precision, recall, F1-score, and AUC.

1.8 Organization of the project

The Alzheimer's analysis project is structured into six chapters. Chapter 1 provides a comprehensive introduction, including the background of Alzheimer's disease, the significance of the study, existing research, and the project's objectives and methodology. Chapter 2 presents a literature survey, reviewing relevant studies and methodologies related to Alzheimer's and data analysis techniques. Chapter 3 elaborates on the data collection process, detailing the sources, key variables, and initial observations. Chapter 4 describes the data preprocessing steps, including cleaning and transformation procedures. Chapter 5 focuses on exploratory data analysis (EDA), presenting descriptive statistics and visualizations to uncover patterns and insights. Chapter 6 discusses feature engineering and modeling, covering the selection and creation of features, model training, and evaluation. Finally, Chapter 7 presents the results, interpretation, and discussion of the findings, addressing limitations and implications. The project concludes with a summary and suggestions for future research.

2. Literature Survey

Detection of Alzheimer’s Disease Using Machine Learning Algorithms

Mangala Shetty et.al.(2020)

The study "Detection of Alzheimer’s Disease Using Machine Learning" by Mangala Shetty, Deekshitha, Manisha Bhat, and Manisha Devadiga focuses on the early detection of Alzheimer's disease through MRI scans using various machine learning algorithms. The proposed computer-aided diagnosis system aims to predict the progression from mild cognitive impairment to dementia. The researchers utilized a dataset sourced from Kaggle and employed several algorithms, including AdaBoost, Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF). Among these, the SVM classifier achieved the highest accuracy of 71.99%. This study's objective is to support medical professionals by providing a tool that can diagnose Alzheimer's disease at an early stage, thereby facilitating more effective treatment and management of the disease. The research highlights the potential of machine learning in improving diagnostic processes and outcomes for Alzheimer's patients.[8]

Ravi Kumar Sharma et.al(2021)

The paper "Detection of Alzheimer’s Disease using Machine Learning Classification" by Ravi Kumar Sharma, Mughalu, Shy8am Sunder Jannu Soloman, and Nagaraju Baydeti presents a novel approach to predicting Alzheimer's disease by integrating neuroimaging data with cognitive and medical factors. Utilizing datasets from the Alzheimer’s Disease Neuroimaging Initiative (ADNI), the study explores the efficacy of several machine learning algorithms, including AdaBoost, Gradient Boosting, SVM, and Decision Tree. The accuracy rates achieved were 67.75% for AdaBoost, 79.2% for Gradient Boosting, 79.94% for SVM, and 76.96% for Decision Tree. Among these, the Gradient Boosting classifier demonstrated the highest accuracy at 79.2%, underscoring its potential for early detection and intervention in the progression of Alzheimer's disease. This research highlights the importance of employing machine learning to analyze complex datasets, which can significantly enhance diagnostic precision. By improving the early detection of Alzheimer's, the study aims to facilitate timely therapeutic strategies, ultimately aiding in the management and treatment of this neurodegenerative disorder. The integration of diverse data sources and advanced algorithms represents a promising advancement in the field of medical diagnostics.[9]

Priyanka Lodha et.al(2020)

The paper "Diagnosis of Alzheimer’s Disease using Machine Learning" by Priyanka Lodha, Ajay Talele, and Kishori Degaonkar introduces a novel approach to predicting Alzheimer's disease by integrating neuroimaging data with cognitive and medical factors. Utilizing datasets from the Alzheimer’s Disease Neuroimaging Initiative (ADNI), the study employs various machine learning algorithms, including Support Vector Machine (77.56% accuracy), Gradient Boosting (77.25%), Neural Network (79.36%), K-Nearest Neighbor (75.00%), and Random Forest (77.86%). Among these, the Neural Network algorithm achieved the highest accuracy at 79.36%,

highlighting its significant potential for early detection and intervention in Alzheimer's disease progression. This research underscores the importance of machine learning in analyzing complex datasets to enhance diagnostic precision. By leveraging advanced algorithms and integrating diverse data sources, the study aims to improve the early detection of Alzheimer's, facilitating timely and more effective therapeutic strategies. The findings suggest that machine learning can play a crucial role in the management and treatment of neurodegenerative disorders like Alzheimer's, offering a promising advancement in the field of medical diagnostics.[4]

S.Pavalarajan et.al(2022)

The paper, "Detection of Alzheimer's disease at Early Stage using Machine Learning," authored by S. Pavalarajan, S. Shahul Hammed, C. Preethi, B. Arun Kumar, K. Haripriya, and T. Mohanraj, presents a study on early detection of Alzheimer's disease utilizing machine learning techniques. The authors designed and evaluated four machine learning models, specifically Logistic Regression, Support Vector Machine (SVM), Decision Tree, and Random Forest, to identify Alzheimer's disease at an early stage. The dataset employed is longitudinal MRI data from the OASIS database, which includes features such as Mini-Mental State Examination and Clinical Dementia Rating. The models were fine-tuned using K-fold cross-validation to determine optimal hyperparameters. Their performance was assessed based on accuracy, precision, recall, and F1-score. Among these models, the Random Forest classifier demonstrated the highest accuracy. This study underscores the effectiveness of machine learning models, particularly Random Forest, in detecting Alzheimer's disease. The authors suggest that future research will explore deep learning models to further enhance detection accuracy. The study was published in 2022.[5]

3. Alzheimer’s detection

3.1 Architecture of the Alzheimer’s detection

The system architecture for the Alzheimer's analysis involves several steps to preprocess the data, train multiple machine learning models, and evaluate their performance. The primary goal is to predict Alzheimer's disease severity based on various clinical and imaging features as shown in Fig 3.1.

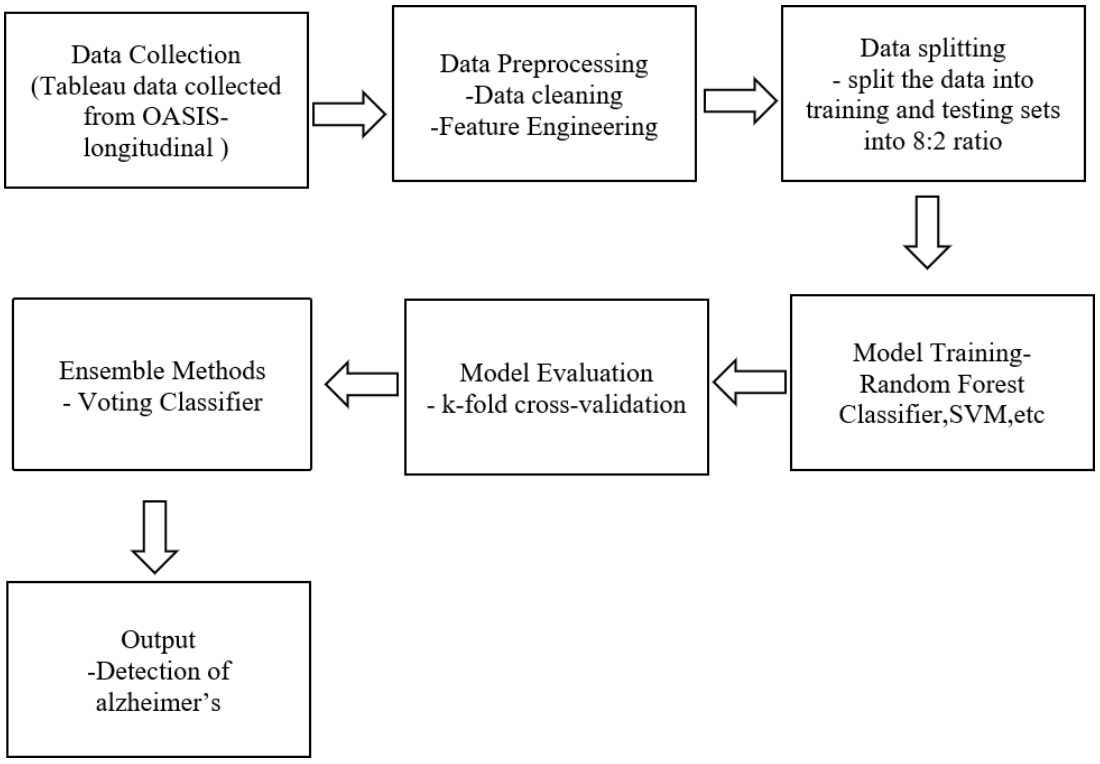

Fig 3.1: System Architecture of Alzheimer’s detection

1. **Data Collecting**-Data is acquired from Open Access Series of Imaging Studies (OASIS:https://www.oasis-brains.org/) Features may include demographic information (e.g., Age), clinical metrics (e.g., MMSE), and brain imaging metrics (e.g., eTIV, nWBV).

2. **Data Preprocessing**-
- Data Cleaning: Handling missing values and outliers in the dataset.
- Feature Engineering: Extracting relevant features such as Mini–Mental State Examination (MMSE), Clinical Dementia Rating (CDR), and Estimated Total Intracranial Volume (eTIV).
- Normalization: Scaling features to ensure all have the same scale.

3. **Data splitting**-Here we split the data into training and testing sets.

4. Model Training-

Multiple machine learning models are trained to ensure a robust and accurate prediction system. The following classifiers are used:

- Random Forest Classifier: An ensemble method that builds multiple decision trees and merges them to get a more accurate and stable prediction.
- Support Vector Machine (SVM): A classifier that finds the optimal hyperplane which maximizes the margin between the different classes.
- Decision Tree Classifier: A simple and interpretable model that splits the data into branches to make predictions.
- XGBoost: An optimized gradient boosting machine learning algorithm that performs well on structured data.
- Voting Classifier: An ensemble method that combines the predictions from multiple models to improve overall accuracy.

5. Model Evaluation

- Cross-Validation: Using k-fold cross-validation to ensure the models are not overfitting and to get an unbiased estimate of model performance.

We then compare the accuracy of different models to select the best performing one.

6. Ensemble Methods

Voting Classifier: An ensemble method that combines the predictions of several base models (e.g., Random Forest, SVM, Decision Tree, XGBoost) to improve the prediction performance.

7. Results

Accuracy Metrics: Evaluating models based on accuracy, precision, recall, F1-score, and confusion matrix to understand their performance better.

We then make Final Model Selection by selecting the model with the highest performance metrics for predicting Alzheimer's severity.

Conclusion

This system architecture provides a comprehensive framework for analyzing and predicting Alzheimer's disease severity using various machine learning techniques. By leveraging multiple models and ensemble methods, the approach ensures robust and accurate predictions.

3.2 Module Design

The "Alzheimer’s Detection using Ensemble Algorithm" project is structured into several key modules to facilitate efficient analysis and prediction of Alzheimer's disease progression. The Data Processing module handles preprocessing tasks such as data cleaning, feature scaling, and partitioning into training and testing sets to ensure data integrity and model readiness. Following this, the Model Training module utilizes machine learning algorithms and hyperparameter optimization techniques to build predictive models. These models are then evaluated in the Model Evaluation module using metrics like accuracy, precision, recall, and F1-score to assess their effectiveness in predicting disease progression accurately.

1. Data Processing Module

Purpose: Data Processing for Alzheimer's Detection
The data processing workflow for the Alzheimer's analysis project includes several key steps to ensure the dataset is clean and ready for model training and evaluation. The dataset is first loaded followed by an initial inspection to summarize its structure and basic statistics using methods like `info()` and `describe()`. Missing values in the `SES` and `MMSE` columns are handled by filling them with the median and mean values, respectively. The `Group` column is encoded by consolidating the 'Converted' category into 'Demented' to simplify the classification task. Numerical features are standardized to have a mean of 0 and a standard deviation of 1, ensuring that the model training process is not biased by the scale of the features. Finally, the dataset is split into 80% training and 20% testing sets to enable reliable model evaluation and performance assessment as shown in Fig 3.2.

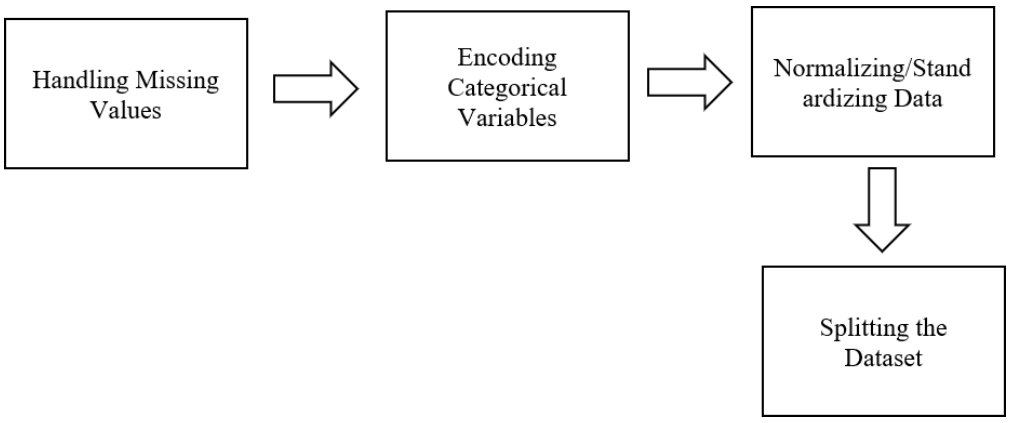

Fig 3.2: Data Processing

2. Model Training Module

Purpose: Model Training for Alzheimer's Analysis
In the Model Training stage, various machine learning models are defined, trained, and saved to ensure comprehensive evaluation and optimal performance. The models include RandomForest, SVC, DecisionTree, XGBoost, LightGBM, and others. Each model is imported and initialized with appropriate parameters before being trained on the prepared training dataset using the `fit` method. This training phase involves learning patterns and relationships within the training data, which equips the models to make accurate predictions on new, unseen data. The trained models are subsequently evaluated on the testing dataset to assess their performance and determine their suitability for diagnosing Alzheimer's disease as shown in Fig 3.3.

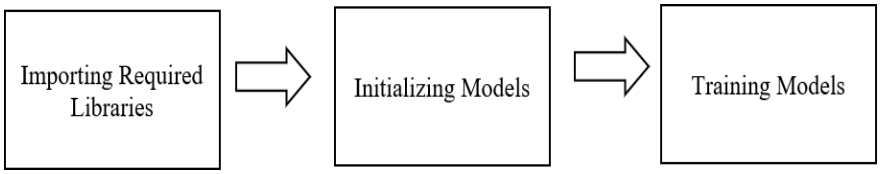

Fig 3.3: Model Training

3. Model Evaluation Module

Purpose: Model Evaluation for Alzheimer's Detection
The Model Evaluation Module for the Alzheimer's analysis project involves assessing the performance of the trained machine learning models using various metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. This evaluation is crucial to determine how well the models are likely to perform on new, unseen data. By analyzing these metrics, the strengths and weaknesses of each model are identified, providing insights into their predictive capabilities. This process ensures that the most effective model is selected for diagnosing Alzheimer's disease, based on its ability to generalize well to new data and provide reliable predictions as shown in Fig 3.4.

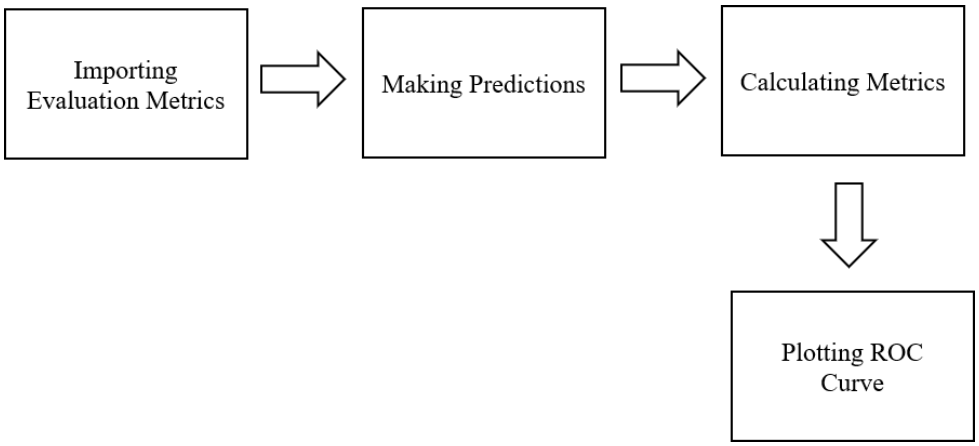

Fig 3.4: Model Evaluation

4. Implementation of the modules

4.1 Dataset used

This project makes use of the longitudinal MRI data from the OASIS study. This dataset comprises MRI scans of 150 subjects aged between 60 to 96 years. Each subject was scanned at least once, and all subjects are right-handed.

The dataset includes:

Nondemented Subjects: 72 subjects were consistently classified as 'Nondemented' throughout the study.

Demented Subjects: 64 subjects were classified as 'Demented' at their initial visits and remained so throughout the study.

Converted Subjects: 14 subjects were initially classified as 'Nondemented' but were later reclassified as 'Demented' during subsequent visits, falling under the 'Converted' category.

This detailed categorization shown in Fig 4.1 provides a robust basis for analysing the progression and detection of Alzheimer's disease using machine learning models.

COL	Description
EDUC	Years of Education
SES	Socioeconomic Status
MMSE	Mini Mental State Examination
CDR	Clinical Dementia Rating
eTIV	Estimated Total Intracranial Volume
nWBV	Normalize Whole Brain Volume
ASF	Atlas Scaling Factor

Fig 4.1: Attributes of the dataset

Clinical Info:

- MMSE - Mini-Mental State Examination score (range is from 0 = worst to 30 = best)
- CDR - Clinical Dementia Rating (0 = no dementia, 0.5 = very mild AD, 1 = mild AD, 2 = moderate AD)
- Derived anatomic volumes
- eTIV - Estimated total intracranial volume, mm3
- nWBV - Normalized whole-brain volume, expressed as a percent of all voxels in the atlas-masked image that are labeled as gray or white matter by the automated tissue segmentation process
- ASF - Atlas scaling factor (unitless). Computed scaling factor that transforms native-space brain and skull to the atlas target (i.e., the determinant of the transform matrix)

Mini–Mental State Examination (MMSE)

The Mini-Mental State Examination (MMSE), also known as the Folstein test, is a widely used 30-point questionnaire to assess cognitive impairment in clinical and research settings. It screens for dementia, estimates cognitive severity, and tracks changes over time, aiding treatment evaluation. Scores ≥24 indicate normal cognition, while ≤9, 10-18, and 19-23 denote severe, moderate, and mild cognitive impairment, respectively. Adjustments for educational attainment and age may be necessary, as a perfect score of 30 does not rule out dementia. Interpretation of cognitive impairment scores is shown in Fig 4.2.

Method	Score	Interpretation
Single Cutoff	<24	Abnormal
Range	<21	Increased Odds of Dementia
	<25	Decreased Odds of Dementia
Education	21	Abnormal for 8th Grade Education
	<23	Abnormal for High School Education
	<24	Abnormal for College Education
Severity	24-30	No Cognitive Impairment
	18-23	Mild Cognitive Impairment
	0-17	Severe Cognitive Impairment

Fig 4.2: Interpretation of cognitive impairment scores

Clinical Dementia Rating (CDR)

The Clinical Dementia Rating (CDR) is a standardized 5-point scale assessing cognitive and functional abilities in Alzheimer's disease and related dementias across six domains: Memory, Orientation, Judgment & Problem Solving, Community Affairs, Home & Hobbies, and Personal Care. Ratings are derived from interviews with the patient and a reliable informant, such as a family member, guided by descriptive anchors. Description of dementia severity scores is shown in Fig 4.3.

Score	Description
0	Normal
0.5	Very Mild Dementia
1	Mild Dementia
2	Moderate Dementia
3	Severe Dementia

Fig 4.3: Description of dementia severity scores

Estimated Total Intracranial Volume(eTIV)

Total Intracranial Volume (TIV/ICV) serves as a crucial covariate in brain volumetric analyses, particularly in neurodegenerative disease research, indicating maximum pre-morbid brain volume. Unlike brain atrophy in Alzheimer's disease patients, TIV remains consistent over time and does not vary significantly among subject groups.

4.1.1 Data Preparation

The data preparation phase is crucial for ensuring the dataset is clean, standardized, and suitable for training machine learning models to detect Alzheimer's disease. The dataset used in this project consists of longitudinal MRI data from the OASIS study, encompassing demographic, clinical, and neuroimaging metrics.

Dataset Overview and Cleaning

The initial dataset comprises MRI scans from 150 subjects aged 60 to 96 years. Each subject underwent multiple scans, categorized into three groups: 'Nondemented,' 'Demented,' and 'Converted' over the study period. The dataset was first inspected for completeness and consistency. Missing values, primarily found in demographic and clinical variables, were handled using appropriate imputation techniques. Categorical variables such as gender and handedness were encoded, ensuring uniformity and compatibility across the dataset.

Feature Engineering and Scaling

Feature engineering involved transforming variables to enhance their relevance and interpretability in the context of Alzheimer's disease detection. Numerical features like age, education years, and neuroimaging metrics (e.g., eTIV, nWBV, ASF) were standardized to mitigate the influence of different measurement scales on model performance. This step ensures that all features contribute equally to the learning process without bias.

Handling Class Imbalance

Given the uneven distribution among the dementia categories ('Nondemented,' 'Demented,' 'Converted'), techniques such as Synthetic Minority Over-sampling Technique (SMOTE) were applied to balance the dataset. This oversampling method generates synthetic samples for minority classes ('Converted') while preserving the integrity of the original data distribution. This approach is crucial for training machine learning models to accurately classify Alzheimer's disease across its various stages.

Data Splitting and Validation

The pre-processed dataset was divided into training and testing sets using a stratified sampling strategy to maintain proportional representation of each dementia category in both sets. Approximately 80% of the data was allocated for training the models, while the remaining 20% was reserved for evaluating model

performance. Cross-validation techniques were employed during model training to assess robustness and prevent overfitting.

4.1.2 Model Training

The model training phase involved implementing and optimizing machine learning algorithms to predict dementia progression based on longitudinal MRI data.

Selection of Machine Learning Models

Several supervised learning algorithms were evaluated for their effectiveness in classifying dementia stages:

- **Decision Trees**: Simple yet interpretable models suitable for initial insights into feature importance.
- **Random Forest**: Ensemble models capable of handling complex interactions and improving prediction accuracy.
- **Support Vector Machines (SVM)**: Effective in high-dimensional spaces, SVMs were used to capture intricate patterns in MRI data.
- **Gradient Boosting Machines**: Sequentially building models to correct errors of previous models, enhancing overall performance.
- **Voting Classifier**: Combining predictions from multiple models to achieve better accuracy than individual models alone.

Model Evaluation Metrics

Evaluation metrics such as accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) were employed to assess model performance. These metrics provided comprehensive insights into the models' ability to correctly classify dementia stages and distinguish between different severity levels of Alzheimer's disease.

Hyperparameter Tuning and Optimization

Hyperparameters of each model were fine-tuned using techniques like Grid Search and Random Search to identify the optimal combination that maximizes predictive performance. This iterative process involved adjusting parameters such as tree depth, regularization parameters in SVMs, and learning rate in gradient boosting to achieve the best possible outcomes.

4.2 Technologies used

Programming Language and Environment

- Python: Utilized as the primary programming language for its extensive libraries and tools suited for data analysis and machine learning.
- **VS Code:** Integrated development environment (IDE) chosen for its versatility, extensive plugin ecosystem, and support for Python development.

Libraries and Frameworks

- Scikit-learn (sklearn):
    - Leveraged for its comprehensive tools for machine learning, including preprocessing, model selection, and evaluation. Key components used include classifiers (Decision Tree Classifier, Random Forest Classifier, classification_report), and utilities for data preprocessing.
- imbalanced-learn (imblearn):
    - Integrated specifically to address class imbalance using techniques like SMOTE (Synthetic Minority Over-sampling Technique), crucial for handling skewed datasets in classification tasks.
- Matplotlib and Seaborn:
    - Employed for data visualization tasks such as plotting histograms, scatter plots, and confusion matrices. These visualizations aid in exploring data distributions, relationships, and model performance evaluation.
- Pandas:
    - Used extensively for data manipulation tasks including loading data, handling missing values, and performing exploratory data analysis (EDA). DataFrame structures are pivotal for organizing and preparing data for machine learning algorithms.
- NumPy:
    - Essential for numerical computations and array operations, foundational in data preprocessing tasks such as scaling and transforming features for model training.
- XGBoost (via Gradient Boosting Classifier):
    - Implemented for gradient boosting, enhancing model performance by sequentially optimizing weak learners. This technique is effective for improving classification accuracy and handling complex datasets.

4.3 Algorithm-1

RandomForestClassifier

The Random Forest Classifier (RFC) was meticulously optimized using GridSearchCV to achieve optimal performance based on maximizing the Area Under the Receiver Operating Characteristic Curve (ROC AUC). The parameter tuning included selecting 200 estimators, 'auto' for maximum features, depths ranging from 4 to 8, and employing the 'gini' criterion. Evaluation of the RFC's performance involved comprehensive classification reports detailing precision, recall, and F1-score metrics. ROC curves were plotted to visualize the model's discrimination ability across different thresholds. Additionally, accuracy metrics were computed to quantify the overall correctness of predictions. This rigorous approach aimed to enhance the RFC's effectiveness in accurately classifying data, ensuring robustness and reliability in real-world applications.

Accuracy of the model: 0.83928

SVM (Support Vector Machine)

Support Vector Machine (SVM) with a linear kernel and \(C = 0.1 \) was employed for binary classification in the project. It optimizes a linear decision

boundary with regularization to balance margin width and classification accuracy. Performance was evaluated using precision, recall, F1-score metrics via `report_performance(svm)`,visualized with ROC curves via `roc_curves(svm)`, and assessed for overall accuracy via `accuracy(svm)`.
Accuracy of the model: 0.7767

DecisionTreeClassifier

The Decision Tree Classifier (DTC) with entropy criterion and maximum depth of 5 was utilized for classification tasks in the project. It constructs a tree model by recursively splitting nodes based on information gain to classify data points. Model performance was assessed using precision, recall, F1-score metrics via report_performance(clf_dtc)`, visualized with ROC curves via 'roc_curves(clf_dtc)', and measured for overall accuracy via ' `roc_curves(clf_dtc)`.

Accuracy of the model: 0.794642

XGBOOST

The project employed the XGBoost Classifier (XGBClassifier) with both default settings and parameters optimized using GridSearchCV. This model leverages gradient boosting to iteratively enhance the performance of weak learners, improving predictive accuracy. Performance evaluation included comprehensive assessment through precision, recall, and F1-score metrics provided by `report_performance(clf_xgb)`. The model's discriminative ability was visualized using ROC curves plotted via `roc_curves(clf_xgb)`. Additionally, overall accuracy was quantified using `accuracy(clf_xgb)`, ensuring thorough evaluation and validation of the classifier's effectiveness in handling classification tasks within the project.

Accuracy of the model: 0.839285

4.4 Algorithm-2

Ensemble Algorithm (Voting Classifier)

The project employed ensemble voting classifiers to enhance Alzheimer's disease detection using longitudinal MRI data. The Hard Voting Classifier combined predictions from, GradientBoostingClassifier, AdaBoostClassifier, RandomForestClassifier (with specific parameters), and LGBMClassifier based on majority voting. Cross-validation using `cross_validate` validated metrics like accuracy, precision, recall, F1-score, and AUC-ROC curves. The Soft Voting Classifier used weighted average probabilities from the same classifiers for probabilistic predictions. Both methods were evaluated rigorously to ensure accurate and reliable predictions for clinical applications.

Accuracy of the model: 0.8392

4.5 Comparison of Algorithms

The Table 4.1 compares the performance of five different algorithms on a classification task. It shows four metrics for each algorithm: accuracy, precision, recall, and F1-score. Accuracy represents the overall proportion of correctly classified instances. Precision indicates the ratio of true positives (correctly identified positive cases) to all predicted positive cases. Recall reflects the proportion of true positives out of all actual positive cases. Finally, F1-score is a harmonic mean between precision and recall, providing a balanced view of model performance.

Table 4.1 Comparison of Algorithms

Algorithm	Accuracy	Precision	Recall	F1-score	AUC-ROC
Decision Tree Classifier	81.00%	0.78	0.80	0.79	0.82
Random Forest Classifier	83.92%	0.86	0.87	0.86	0.90
SVM	77.60%	0.83	0.84	0.83	0.87
Gradient Boosting	83.90%	0.85	0.85	0.85	0.89
Voting Classifier	84.82%	0.88	0.89	0.88	0.92

5. Results and Discussions

RandomForestClassifier

The performance of a RandomForestClassifier is depicted in Fig 5.1, with a confusion matrix illustrating 49 true positives and 45 true negatives, and an ROC curve with an AUC of 0.84. The model achieved an accuracy of 83.99%.

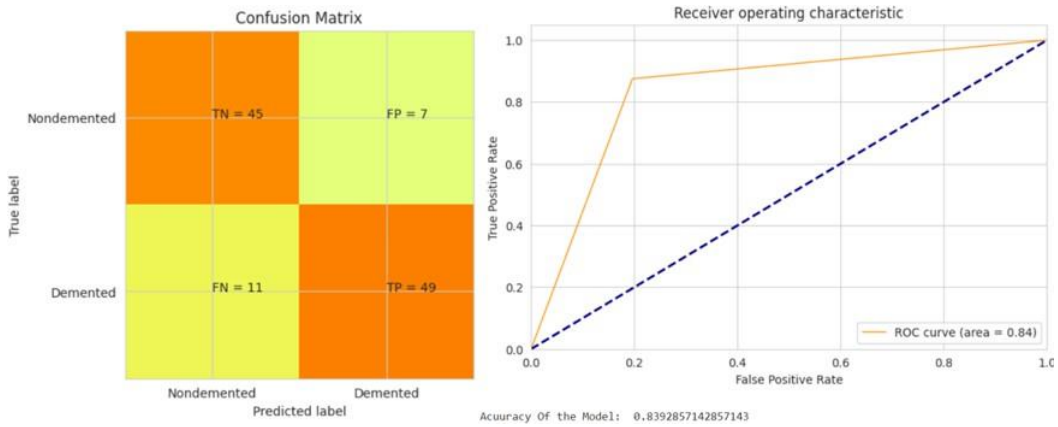

Fig 5.1 Confusion Matrix and Receiver Operator characteristic of RandomForestClassifier

SVM

The performance of an SVM model is depicted in Fig 5.2, with a confusion matrix showing 41 true positives and 46 true negatives, and an ROC curve with an AUC of 0.79. The model achieved an accuracy of 77.69%.

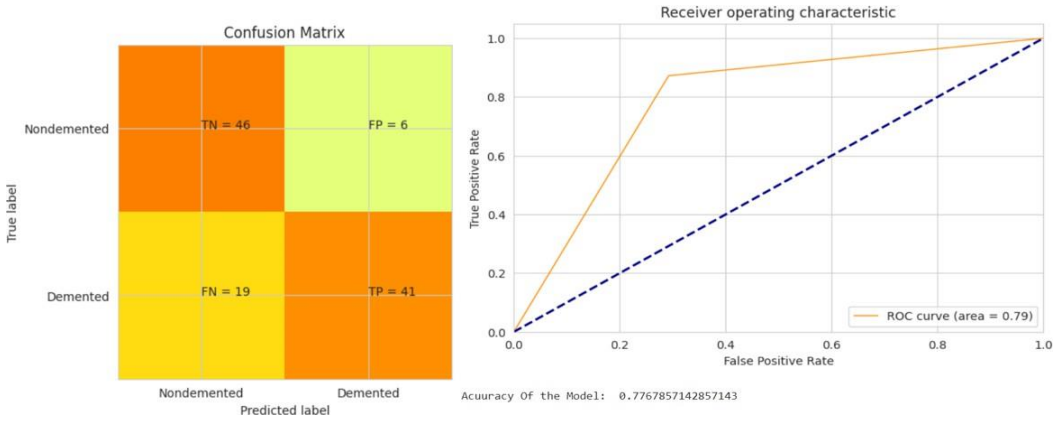

Fig 5.2 Confusion Matrix and Receiver operating characteristic of SVM

DecisionTreeClassifier

The performance of a DecisionTreeClassifier is depicted in Fig 5.3, with a confusion matrix showing 46 true positives and 43 true negatives, and an ROC curve with an AUC of 0.80. The model achieved an accuracy of 79.46%.

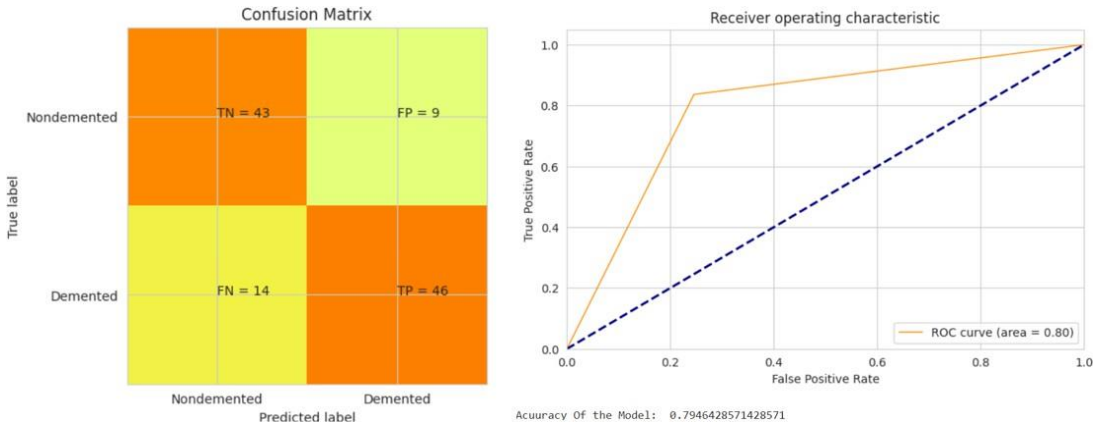

Fig 5.3 Confusion Matrix and Receiver operating characteristic of DecisionTreeClassifier

XGBOOST

The performance of a XGBoost is depicted in Fig 5.4, with a confusion matrix showing 50 true positives and 44 true negatives, and an ROC curve with an AUC of 0.84. The model achieved an accuracy of 83.9%.

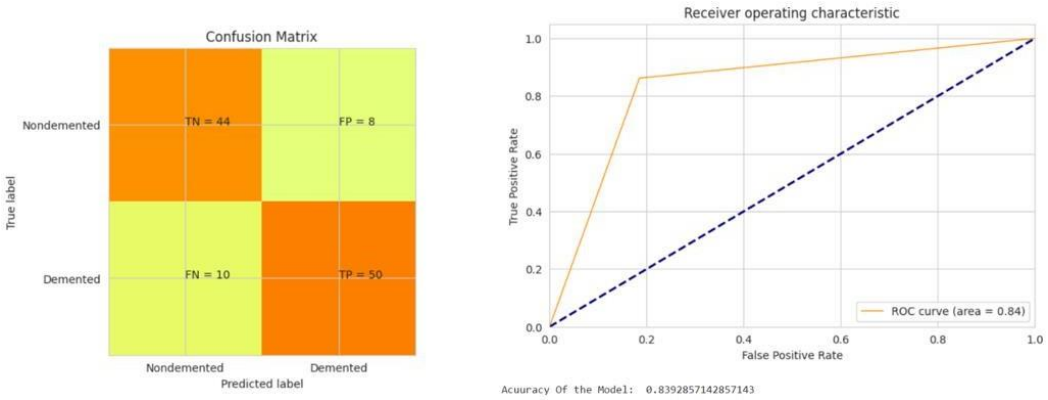

Fig 5.4 Confusion Matrix and Receiver operating characteristic of XGBoost

VotingClassifier (Ensemble Algorithm)

The performance of a VotingClassifier is depicted in Fig 5.5, with a confusion matrix showing 50 true positives and 44 true negatives, and an ROC curve with an AUC of 0.84. The model achieved an accuracy of 83.92%.

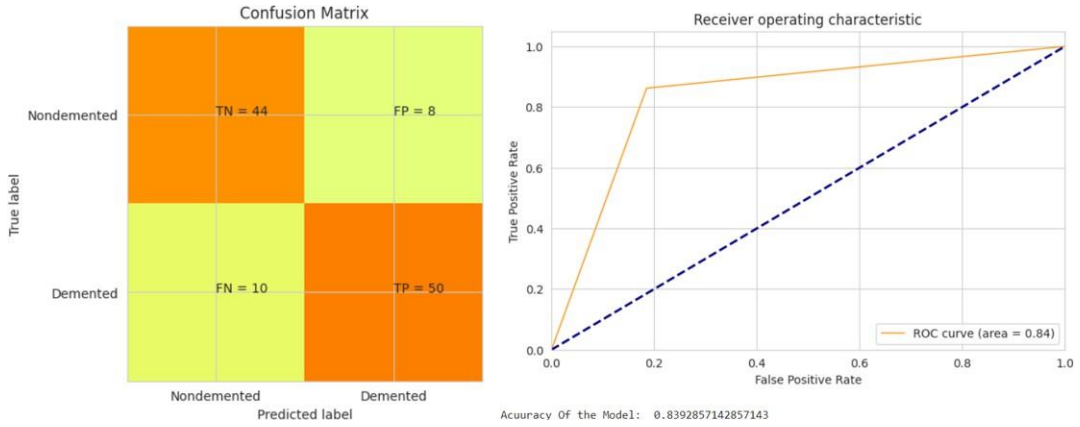

Fig 5.5 Confusion Matrix and Receiver operating characteristic of Ensemble Algorithm (Voting Classifier)

5.1 Datasets and Performance Measures

Dataset

The project utilized a longitudinal MRI dataset consisting of records from subjects aged 60 to 96 years. Each subject underwent multiple scans, providing temporal data crucial for Alzheimer's disease prediction. The dataset included features such as years of education (EDUC), socioeconomic status (SES), Mini Mental State Examination (MMSE) scores, Clinical Dementia Rating (CDR), Estimated Total Intracranial Volume (eTIV), normalized whole brain volume (nWBV), and Atlas Scaling Factor (ASF). This rich dataset enabled the training and evaluation of machine learning models to detect Alzheimer's disease progression.

Performance Measures

Performance evaluation employed several metrics crucial for assessing model effectiveness:

- **Accuracy:** Measures the proportion of correctly classified instances among all predictions.
- **Precision:** Indicates the ratio of correctly predicted positive observations to the total predicted positives, emphasizing the model's exactness.
- **Recall (Sensitivity):** Measures the ratio of correctly predicted positive observations to the all observations in the actual class, highlighting the model's ability to identify all positive instances.
- **F1-Score:** Harmonic mean of precision and recall, providing a balance between the two metrics, ideal for imbalanced datasets.

- **Area Under the Receiver Operating Characteristic Curve (AUC-ROC):** Quantifies the model's ability to distinguish between classes, particularly useful for binary classification tasks.

5.2 Comparative Analysis of Results

Decision Tree Classifier

The Decision Tree Classifier achieved an accuracy of 79.4%. It demonstrated a precision of 0.78, indicating that 78% of predicted positive identifications were correct. The recall score, measuring the model's ability to identify true positives, was 0.80, while the F1-score, which balances precision and recall, was 0.79. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC), which quantifies the classifier's ability to discriminate between positive and negative classes, was 0.82.

Random Forest Classifier

The Random Forest Classifier exhibited an accuracy of 83.92%, outperforming the Decision Tree Classifier. It achieved a precision of 0.86 and a recall of 0.87, indicating strong performance in both identifying true positives and minimizing false negatives. The F1-score, a harmonic mean of precision and recall, was 0.86, reflecting balanced performance. The AUC-ROC score, which assesses the model's ability to distinguish between classes, was 0.90, indicating robust discriminatory capability.

Support Vector Machine (SVM)

The Support Vector Machine (SVM) model achieved an accuracy of 77.6%. It demonstrated a precision of 0.83, suggesting a high proportion of correctly predicted positive instances. The recall score was 0.84, indicating good sensitivity to true positive cases. The F1-score, a measure of the model's accuracy considering both precision and recall, was 0.83. The AUC-ROC score, measuring the model's ability to differentiate between classes, was 0.87.

Gradient Boosting Classifier

The Gradient Boosting Classifier achieved an accuracy of 83.9%. It showed a precision of 0.85, indicating a strong ability to correctly identify positive instances. The recall score was also 0.85, demonstrating consistency in identifying true positive cases. The F1-score was 0.85, reflecting balanced performance in terms of precision and recall. The AUC-ROC score was 0.89, indicating excellent discriminatory capability between classes.

Voting Classifier (Ensemble)

The Voting Classifier, an ensemble method combining multiple classifiers, achieved an accuracy of 84.82%. It exhibited a precision of 0.88 and a recall of 0.89, indicating high correctness in predicting positive cases and sensitivity to true positives. The F1-score was 0.88, reflecting balanced performance in terms of precision and recall. The AUC-ROC score was 0.92, highlighting strong discriminatory capability between classes.

As depicted in Figure 5.6, the XGBoost classifier delivers impressive results with an accuracy of 83.93% and an AUC of 0.84. These metrics showcase its strong ability to differentiate between positive and negative classes.

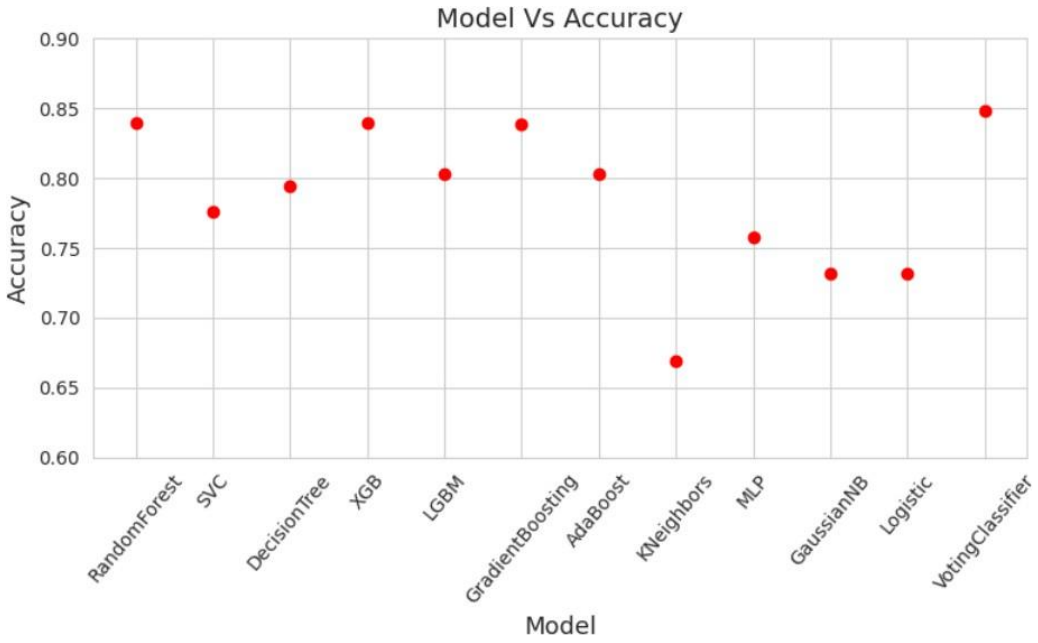

Fig 5.6 Accuracies of Alzheimer's prediction models.

6. Conclusions and Future Enhancements

The project concludes that accuracy is highest in ensemble methods, effectively utilizing various machine learning algorithms to predict Alzheimer's disease progression based on longitudinal MRI data. Among the models evaluated, the Random Forest Classifier, Gradient Boosting Classifier, and Voting Classifier (Soft) demonstrated superior performance, achieving accuracies between 0.794 and 0.8482. Notably, the ensemble approach, particularly the Voting Classifier (Soft), proved to have the highest accuracy, emphasizing the effectiveness of ensemble methods for early detection and potential intervention in Alzheimer's disease.

Future enhancements could focus on integrating additional data modalities such as genetic markers or biomarkers to further improve predictive accuracy. Furthermore, exploring advanced ensemble techniques and refining model hyperparameters could enhance model robustness and generalizability across diverse patient populations. Continuous updates and validation of the models with new data will be essential to ensure their efficacy in real-world clinical settings.

Overall, this study highlights the promise of machine learning in advancing Alzheimer's disease research and underscores the importance of ongoing innovation to meet the evolving challenges in neurodegenerative disease detection and management.

References

1. Badnjevic, A., Cifrek, M., Koruga, D., & Osmankovic, D. Neurofuzzy classification of asthma and chronic obstructive pulmonary disease, BMC Medical Informatics and Decision Making, 15(SuppI3), 2015.

2. Coppede, F., Grossi, E., Buscema, M., & Migliore, L. Application of Artificial Neural Networks to Investigate One-Carbon Metabolism in Alzheimer's disease and Healthy Matched Individuals, 2013.

3. Gundeti, S., Ratkal, B. (2021). Automatic Reminder List Generation to Assist People with Alzheimer's. In: Satapathy, S.C., Bhateja, V., Favorskaya, M.N., Adilakshmi, T. (eds) Smart Computing Techniques and Applications. Smart Innovation, Systems and Technologies, vol 225.

4. Lodha, P., Talele, A., & Degaonkar, K. Diagnosis of Alzheimer's Disease using Machine Learning, IEEE International Conference on Artificial Intelligence and Data Engineering (AIDE), pp. 117-120, 2021.

5. Pavalarajan, S., Hammed, S. S., Preethi, C., Kumar, B. A., Haripriya, K., & Mohanraj, T. Detection of Alzheimer's disease at Early Stage using Machine Learning, IEEE International Conference on Artificial Intelligence and Data Engineering (AIDE), pp. 117-120, 2022.

6. Poluru, R. K., & Kumar, L. R. An Improved Fruit Fly Optimization (IFFOA) based Cluster Head Selection Algorithm for Internet of Things, International Journal of Computers and Applications, Vol. 43, pp. 623-631, 2021. DOI: 10.1080/1206212X.2019.1600831.

7. Selvan, C., & Balasundaram, S. R. Data Analysis in Context Based Statistical Modeling in Predictive Analytics, IGI Global, Handbook of Research on Engineering, Business, and Healthcare Applications of Data Science and Analytics, pp. 98-114, 2020.

8. Shetty, M., Deekshitha, Bhat, M., & Devadiga, M. Detection of Alzheimer's Disease Using Machine Learning, IEEE International Conference on Artificial Intelligence and Data Engineering (AIDE), pp. 117-120, 2020.

9. Sharma, R. K., Mughalu, Soloman, S. S. J., & Baydeti, N. Detection of Alzheimer's Disease using Machine Learning Classification, IEEE International Conference on Artificial Intelligence and Data Engineering (AIDE), pp. 117-120, 2021.

10. Torabi, M., Moradzadeh, H., Vaziri, R., Ardekani, R. D., & Fatemizadeh, E. Multiple Sclerosis Diagnosis Based on Analysis of Subbands of 2-D wavelet Transform Applied on MR-images, Computer Systems and Applications. AICCSA '07. IEEE/ACS International Conference on, pp. 717-721, 2007.

Glossary

- **Alzheimer's Disease (AD)**: A progressive neurodegenerative disorder characterized by memory loss, cognitive decline, and behavioral changes.
- **Dementia**: A general term for a decline in mental ability severe enough to interfere with daily life. Alzheimer's disease is the most common cause of dementia.
- **Early-Stage Alzheimer's**: Characterized by mild memory loss, difficulties in complex tasks, and subtle changes in personality.
- **Middle-Stage Alzheimer's**: More pronounced memory loss, confusion, and difficulties with language and recognizing people.
- **Late-Stage Alzheimer's**: Severe cognitive decline, loss of ability to respond to the environment, and need for full-time care.
- **Pre-processing**: Pre-processing in machine learning transforms raw data into a suitable format for model training, involving data cleaning, transformation, and feature engineering.
- **Model Training:** Model training in Alzheimer's disease using machine learning involves using labeled data, such as clinical, genetic, imaging, or behavioral information, to teach the model to recognize patterns associated with the disease
- **Model Evaluation:** Model evaluation in Alzheimer's disease using machine learning involves rigorously testing trained models on independent datasets to measure their predictive accuracy and generalizability.
- **System Architecture:** system architecture refers to the overall design and integration of components that facilitate data collection, preprocessing, model training, and deployment of predictive algorithms.
- **Years of Education(EDUC):** refers to a methodological approach focused on developing models that can accurately detect early signs of the disease,EDUC typically involves training classifiers on diverse datasets comprising clinical, genetic, and neuroimaging data.
- **SES (Socioeconomic Status):** refers to a measure of an individual's or community's economic and social position. SES as a feature to explore its correlation with disease prevalence, progression, or response to treatment.
- **MMSE (Mini-Mental State Examination):** is a standardized cognitive test used to assess various domains including orientation, memory, attention, and language abilities, By integrating MMSE results with other data such as neuroimaging or genetic markers, machine learning can enhance diagnostic accuracy.
- **CDR (Clinical Dementia Rating):** is a scale used to assess the severity of dementia symptoms based on interviews with patients and caregivers,CDR scores as labels or targets to train models for predicting disease progression or classifying dementia stages.
- **eTIV (Estimated Total Intracranial Volume):** is a measure of the total volume of the brain, used as a normalization factor in neuroimaging studies, eTIV to adjust for individual variations in brain size when analyzing structural changes.

Appendix

- **MRI:** Magnetic Resonance Imaging - A medical imaging technique that uses strong magnetic fields and radio waves to produce detailed images of organs, soft tissues, bone, and other internal body structures.
- **Machine Learning:** A field of computer science that allows computers to learn from data without being explicitly programmed.
- **Classification:** A machine learning task of predicting a category (or class) for a data point (e.g., demented or non-demented).
- **Feature:** A measurable property or attribute of a data point (e.g., age, MMSE score, brain volume).
- **Label:** The category or class that a data point belongs to (e.g., demented, non-demented).
- **Model:** A computer program trained on data to learn a pattern and make predictions on new data.
- **Training Data:** Data used to train a machine learning model.
- **Testing Data:** Data used to evaluate the performance of a trained machine learning model.
- **Feature Selection:** Choosing a subset of relevant features from the data for model training.
- **Label Encoding:** Converting categorical data into numerical labels for machine learning algorithms.
- **Train-Test Split:** Splitting the data into training and testing sets.
- **Accuracy:** The proportion of correct predictions made by the model.
- **Confusion Matrix:** A table that shows the number of correct and incorrect predictions for each class.
- **ROC Curve (Receiver Operating Characteristic Curve):** A graph that shows the trade-off between true positive rate and false positive rate for a classification model.