Identification of Dementia Disease Using Ensemble Machine Learning model

Supervised By

Mr. Md Rajib Mia

Lecturer (Senior Scale)

Department of Software Engineering

Daffodil International University

Submitted By

Maruf Ahmed Shaon

ID:221-35-994

Department of Software Engineering

Daffodil International University

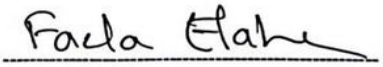
This thesis report has been submitted in fulfilment of the requirements for the Degree of Bachelor of Science in Software Engineering.

APPROVAL

APPROVAL

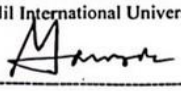
This thesis titled on **"Identification of Dementia disease using ensemble machine learning model"**, submitted by **Student Name (ID: 221-35-994)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS



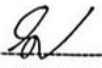
Chairman

Dr. Fazla Ealhe
Assistant Professor & Associate Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



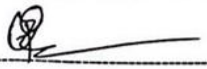
Internal Examiner 1

Dr. Marzia Ahmed
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



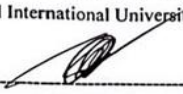
Internal Examiner 2

Dr. Shabnom Mustary
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



Internal Examiner 3

Md. Rajib Mia
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



External Examiner

Mohammod Abul Kashem, PhD
Professor
Department of Computer Science and Engineering
DUET, Bangladesh

Identification of Dementia Disease Using Ensemble Machine Learning model

Maruf Ahmed Shaon
ID:221-35-994

Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSITY

SUPERVISOR'S DECLARATION

SUPERVISOR'S DECLARATION

I hereby declare that I have reviewed this thesis entitled **" Identification of Dementia Disease Using Ensemble Machine Learning model"**, and in my opinion, it is adequate in terms of scope and quality for the award of the degree of Bachelor of Science in Software Engineering.

(Supervisor's Signature)

Full Name : Mr. Md Rajib Mia

Position : Lecturer (Senior Scale), Department of SWE, DIU

Date : 23 December 2025

STUDENT'S DECLARATION

I confirm that the piece in this thesis is based on my own writing with the exception of quotation and reference that have been discussed. I also confirm that it was not previously and concurrently registered at Daffodil International University or other institutions at any other degree.

(Student's Signature)

Full Name : Maruf Ahmed Shaon

ID Number : 221-35-994

Date : 25 December 2025

Identification of Dementia Disease Using Ensemble Machine Learning model

Maruf Ahmed Shaon

ID:221-35-994

Thesis submitted in fulfilment of the requirements
for the award of the degree of
Bachelor of Science

Department of Software Engineering

DAFFODIL INTERNATIONAL UNIVERSITY

DECEMBER 2025

ACKNOWLEDGEMENTS

I owe my special thanks to Mr. Md. Rajib Mia, Lecturer (Senior Scale) for providing me continuous guidance, helpful advice and constant encouragement throughout the research work. His priceless advices, professional monitoring and persistent supports have contributed substantially to the quality of this thesis and we are so grateful for his help. I am seriously grateful for his patience, drive and everyday encouragement to get better and work outside my personal bubble. This work could not have been done without his guidance. I am also very grateful to all the teachers, classmates, friends and well-wishers who helped me in this study in different ways. I am very grateful of their support. Finally, I owe the greatest debt to my family for their continuing love and support; and more importantly yet stunning not to let me down when no one else was on my side.

.

DEDICATION

I owe all my successes in this world, directly to their unconditional support, encouragement and unwavering conviction in me. I have been motivated by their hardships, prayers and constant encouragement at every stage of my academic career. I would also like to dedicate this work to those who work diligently in the pursuit of truth and the support of learning. They continue to inspire my dreams and map my journey.

ABSTRACT

Dementia is a significant public health issue worldwide, and early detection is important to address the coalescence of the condition. Traditional diagnostic approaches conventionally involve extensive clinical expert deliberations that may be time-consuming or subjective. This investigation investigates ensemble machine learning tactics for early dementia detection and classification. Various machine learning algorithms were used to predict the outcome of enrollments on dementia such as Random Forest, LightGBM, XGBoost and Stacking. The research indicates that Stacking ensemble models outperformed all models using an accuracy of 99.04% due to its various base learners. Random Forest- LightGBM followed with an accuracy of 98.1%, while XGBoost and Voting models had relatively lesser outputs. The vast partitions of the former models made stacking ensemble the most accurate based on predictable results, with most of the different weaknesses of the models used to generate data disposed of by others. This research has verified for the first time the potential for machine learning to supplement the expertise of pharmacologists or clinicians which enables new automation and superior diagnosis of dementia and associated characteristics. The possibilities are bound to more research to predict results in the future and interventions. The capacity of ensemble methods to overcome deficiencies in any one particular algorithm increases reliability and lowers the potential for costly misclassification. The integration of clinical decision support systems to aid providers in interpreting the model's predictions may be a topic for future study. The hope is that the end product will be an easy to use tool for clinicians in the early identification and customized treatment of dementia.

Keywords: Dementia, Ensemble Machine Learning, Stacking Model, Random Forest, LightGBM, XGBoost, Early Diagnosis, Predictive Modeling, Clinical Decision Support, Accuracy, Machine Learning Algorithms, Feature Extraction, Healthcare, Data Science, Model Integration.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Full Form
AI	Artificial Intelligence
XGB	XGBoost
RF	Random Forest
LightGBM	Light Gradient Boosting Machine
Stacking	Stacking Ensemble Model
ML	Machine Learning
DNN	Deep Neural Networks
SVM	Support Vector Machine
ANN	Artificial Neural Network
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
F1-Score	F-Measure Score
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
GPU	Graphics Processing Unit
CV	Cross Validation
API	Application Programming Interface
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
SDA	Stochastic Discriminant Analysis

CHAPTER 1

INTRODUCTION

1.1 Overview

Dementia is a progressive, neurodegenerative disease that results in the loss of memory and other cognitive functioning, as well as daily functional capacity. With the increasing aging population worldwide, dementia has become a serious public health concern that needs to be detected early and treated effectively. Conventional diagnosis methods such as cognitive tests and clinical evaluations are usually time-consuming, subjective, and rely on the expertise of a physician. With the evolution of artificial intelligence, machine learning algorithms are increasingly being employed as an aid for automatic disease classification with enhanced objectivity, speed and accuracy. Among those techniques, ensemble learning, in and stacking ensembles models have been chosen to be quite promising after combining the features of several classifies predicting more reliable and stable predictions. Ensemble methods alleviate disadvantages of learners and combine different types of models into a single framework. This work aims to develop a hybrid stacking ensemble algorithm which combines the strengths of multiple learners with good performance in dementia disease identification. With the use of clinical factors and advanced learning technology, we expect that our model could achieve a superior diagnose accuracy, an improved generalization ability, and a better diagnostic reliability than traditional single-model based systems.

.

1.2 Background Study

Dementia is an evolving process of neurodegeneration leading to a reduction in cognitive function that can interfere with one's daily life. It has grown into one of the most pressing public health challenges worldwide, especially given a growing elderly population. Early diagnosis of dementia is important as immediate intervention can counteract the course of the disease and enhance quality of life as well as aid clinicians in treatment planning. Conventional clinical diagnosis is mainly based upon the physician expertise, for cognitive testing and neuropsychological assessments.

While successful, these techniques tend to be time consuming, subjective and based on patient compliance. With the development of digital health technologies, machine learning has become a powerful tool for medical data analysis, providing more objective, consistent and automated methods for disease prediction. Ensemble machine learning architecture has gained enormous popularity, as they achieve high performance levels with greater stability in comparison to the conventional single-model approaches. This research aims to create a hybrid ensemble model which leverages various machine learning classifiers, capable of enhancing the accuracy of dementia detection.

1.3 Problem Statement

Although many types of machine learning models have been implemented for dementia diagnosis, they tend to be limited by non-generalization and non-consistent prediction performance and show high sensitivity against unbalanced datasets. Single classifiers like Decision Tree, SVM, Random Forest or Gradient Boosting perform better worse depending on the complexity of the dataset and the distribution of features. Besides, many clinical datasets are suffering from missing values, class imbalance, and noise in the real-world world which causes single model to be unreliable. The current methods are unable to provide an integrated framework which incorporates several powerful learners and compensates their drawbacks. There is thus a demand for a reliable, stable and more accurate prediction system that can accommodate such complex medical data. The main challenge that is addressed in the thesis work is to design a model by combination of ensemble approach and hybrid implementation which can perform better than machine learning models to predict dementia disease with clinical feature. Which brings us down to the following main research problem: Is there a hybrid ensemble model that can predict better than the classical models. Although there have been several great advances in the field of machine learning techniques for healthcare, existing dementia detection models still suffer from a lot of issues. Most of the exiting methods suffer from imbalanced data, pushing the predictions biased toward a majority class and hindering an overall classification accuracy. Moreover, single classifiers lack explanation of the nonlinear interrelationship of clinical features. It is this absence of a unified framework to combine strong learners which hampers the context in achieving solid performance. Furthermore, the existence of noisy and missing patient data adds to the complexity of the prediction.

1.4 Research Objectives

The aim of this investigation is to develop and assess a novel hybrid ensemble machine leaning model for accurate detection of dementia disease. This objective is proposed to be pursued by means of a number of specific aims. First, for data preprocessing in regard having missing values, a class imbalance and numeric inconsistencies, it should be treated with modernistic approaches those have specialized like ADASYN and feature scaling. Second, to realize individual baseline models, e.g., Random Forests, Support Vector Machines, LightGBM and XGBoost for performance comparison. Third, the proposed hybrid stacking ensemble model which combines several base learners for improving prediction accuracy. Fourth, to score the models with established performance metrics (accuracy, precision, recall or F1-score, confusion matrix and ROC curve). Finally, for the to examine feature importance in order to determine which clinical characteristics are most important for predicting dementia. All these aims taken together will contribute to designing a robust and very accurate machine learning framework for early detection of dementia.

1.5 Significance of the Study

This study is of academic and clinical value. Dementia is a worldwide public health priority, and precise upfront recognition is important in optimizing patient care, curbing skyrocketing healthcare costs and enabling early intervention initiatives. This work also makes a novel contribution to the recent growth in AI in healthcare by designing an improved hybrid ensemble model. The proposed hybrid approach presented in this paper possesses the following merits emptier accuracy, better generalization and robustness than a single model-based existing methodologies. Moreover, this study shows that stacking ensemble methods successfully can be applied to medical data analysis in disease prediction and is a standard model that could be used for other diseases. By analyzing feature importance, the tips of decision ROMs can provide insight to clinicians as influential indicators related to dementia. This work not only contributes to progress in the area of machine learning, but it also has an impact on healthcare by providing a more accurate resource for dementia disease diagnosis.

1.6 Scope of the Study

The scope of this study encompasses the development, implementation, and evaluation of an ensemble-based machine learning framework for the identification of dementia disease. This research focuses exclusively on supervised machine learning techniques and utilizes a structured clinical dataset that includes cognitive, demographic, and biological features relevant to dementia classification. The study involves extensive data preprocessing, including handling missing values, addressing class imbalance using ADASYN, and applying feature scaling where necessary. Multiple baseline classifiers such as Random Forest, Support Vector Machine, LightGBM, and XGBoost are trained and compared to establish performance benchmarks. The core of this research is the design of a hybrid stacking ensemble model that integrates predictions from multiple strong learners to improve classification accuracy. The evaluation is conducted using standard metrics such as accuracy, precision, recall, F1-score, confusion matrices, and ROC curves. However, the study does not include deep learning models, real-time clinical validation, or multimodal MRI imaging analysis, as these fall outside the defined research boundaries. Despite this, the proposed model provides a robust and scalable foundation for future extensions in dementia diagnosis.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

Dementia, a neurodegenerative disease that develops over time, is responsible for memory and cognitive dysfunctions with eventual behavioral loss. The aging population worldwide is leading to an increase in the global burden of dementia, and therefore early and reliable diagnostics are crucial. Conventional methods of diagnosis depend mainly on clinical practitioners' judgement and cognitive assessments, which are costly and subjective. Advances in machine learning (ML) technologies allows for automatic and trustworthy analyzing of clinical data of dementia. Many researches have investigated different algorithms including SVM, RF, and XGBoost or ensemble methods for dementia prediction. 6-12 While there are many different trained models, ensemble models have shown higher performance by integrating different classifiers to make them robust and accurate. In this chapter we summarize previous related works and show that there is a room for more advanced hybrid ensemble models to detect dementia.

2.2 Previous Study

Pandey et al. (2025) based on an ensemble deep learning model with stacking technique using the EfficientNet-B7, Exception and Inception-ResNet-V2 models for severity ranking of Alzheimer's disease from MRI images. On the OASIS and ADNI databases, their model obtained ~98% accuracy with 96% sensitivity and high specificity, surpassing single deep models. The fusion strategy in rank order helped combine the complementary features of different CNNs. Their findings demonstrate that well designed stacked ensembles can dramatically enhance the MRI-based diagnosis of AD [1]. MDPI Diagnostics et al. (2023) introduced a meta-learning ensemble model for XAI of Alzheimer's compared to clinical as well as mid-slice MRI data. An explainable AI framework is developed for feature importance identification by assembling multiple base learners and a meta-model, with several interpretable AI techniques used to emphasize contributing features.

The study presented high accuracy and fair performance among AD, MCI, and control subjects. Of note, the authors focused on interpretability in order to achieve clinical adoption of ensemble models and not "black box" predictions [2]. Rijul Kumar et al. (2025) compared several machine learning models including (Logistic Regression, KNN, SVM, Decision tree and Random Forest) as well as ensemble method XGBoost- for binary dementia classification together with feature selection and explainable AI (SHAP, LIME). The best performance was obtained by Random Forest (approximately 96% of accuracy) and CDR was identified as an important predictor. The results demonstrate that the tree-based ensembles achieve better performance than simpler models on dementia data, and explainable predictions can be inferred by XAIs. Their technique is similar to your work where RF/XGB serve as the heart in a hybrid ensemble [3]. (ML-Driven AD Prediction) et al. (2024/2025) proposed a deep ensemble model to predict Alzheimer's disease based on multimodal data. They obtained high precision and recall of detecting dementia status (across different stages) by integrating multiple neural networks and ensemble schemes. Results show deep ensembles can achieve better generalization performance than single architectures, especially in complex clinical scenarios. The paper also highlights robustness and clinical eligibility of ensemble models for early AD diagnosis [4].

Rahman et al. (2022)) et al. (2022) presented the DAD-Net, a deep learning network for AD classification which explicitly alleviates class imbalance problem using ADASYN oversampling. They have demonstrated that ADASYN enhances minority class representation and the shift of the decision boundary to be closer to harder samples, achieving higher sensitivity at under-represented AD stages. Together with Grad-CAM visualizations, their algorithm obtained good classification results for unbalanced AD datasets. This justifies your decision to use ADASYN in preprocessing for dementia classification [5]. Silva et al. (2024) et al (2024) proposed an EEG signal-based ensemble learning model for the Alzheimer's disease classification, and they adopted both Random Forest, XGBoost and SVM as the base classifier. The analysis found that the ensemble model performed better in average accuracy than single models for diagnosing of patients with AD and control. Their logic diagram and methodology illustrate the value in feature engineering, when combined with strong supervised algorithms. The paper demonstrates that ensemble learning can be effective even with non-imaging bio signals such as EEG [6].

Alruily et al. (2025) presented an ensemble deep learning model to combine the features of VGG16, MobileNet and InceptionResNetV2 for MRI-based Alzheimer's disease diagnosis. The sensitivity and specificity are relatively high at 96% and 98%, respectively, also substantially superior compared to individual CNNs. The feature-level concatenation forms a rich representation, capable to deal with the difference of MRI texture and shape. This study supports that ensembles of diverse deep networks can greatly improve the diagnostic performance [7]. (Optimizing AD Prediction) et al. (2024) proposed stacked XGBoost and Gradient Boosting for ensemble machine learning-based prediction of early AD. The model achieved approximately 97% accuracy and included SHAP explanations to gain insights into important cognitive and clinical predictors. Their findings suggest that gradient-boosting–based ensembles are highly competitive for tabular AD data. The paper also highlights the importance of being able to explain clinical decision-support tools [8]. Sørensen et al. (2018) presented an SVMs-based ensemble method for dementia classification using structural MRI-measured cortical thickness. They integrated bootstrapping without replacement and feature selection to demonstrate that an ensemble of SVM could improve upon single model based-SVM in multivariate dementia classification. The technique increased robustness to noise and overfitting in high-dimensional neuroimaging data. Their work is an early one for ensemble SVMs for dementia, and which encourages the use of multiple heterogeneous base learners. [9].

Johnson et al. (2022) et al. (2022) proposed a stacking model for multiclassification of Alzheimer's disease based on ADNI neuroimaging and clinical characteristics. The stacked ensemble combines a number of base models to separate Normal Controls, early MCI, late MCI and AD. Results indicated that the method of stacking achieved better accuracy and comprehensive performance than single classifier as well. This is directly analogous to your proposition about a stacked hybrid model outperforming baseline models for the dementia staging [10]. Rahimi et al (2022) presented an Alzheimer's disease detection model which first trains traditional ML models and then the Gradient Boosting and Voting classifiers are used. The gradient boosting ensemble is a combination of Decision Tree and Random Forest, and the voting classifier combines predictions from SVM, XGBoost, AdaBoost ensemble as well as RF and Decision Tree. They find that ensemble structures generally outperform their single-model baseline predictions. This chapter presents how various ensemble methods are integrated for AD diagnosis [11].

Roy & Jahan (2025) et al. (2025) proposed a computer aided diagnosis system that employed CNN for feature extraction from neuroimaging data and classification was performed using SVM, RF, KNN and ANN. They compared various models for early-stage Alzheimer's detection and observed that a combination of deep features and classical ML yields strong results. The result of this work highlights that hybrid: deep-feature + ML based pipelines are more sensitive to subtle early changes than conventional methods. Their results are in favor of your previous use of machine learning on structured clinical/imaging features [12]. Batista et al. (2023) constructed an interpretable model for Alzheimer ' s disease based on clinical and demographic information. They used several various ML algorithms and considered data imbalance and high dimensionality. It was performed using accuracy, precision, recall and F1-score in which mostly the tree-based models were superior. Feature importance based on SHAP demonstrated the strength of clinical factors to make predictions. This is consistent with your trade-off of performance and interpretations for classification tasks in dementia [13]. () presented a federated learning with shared control for the early risk prediction of Alzheimer's disease in mild cognitive impairment subjects. They applied Random Survival Forests and XGBoost algorithms as well as feature ranking with SHAP for obtaining key predictors and patient's stratification in different risk groups. Their multivariate model was successful in the analysis of time-to-event data and multicollinearity. This may simply be because it is more about prognosis than pure classification, but does illustrate the versatility of ensembles in AD [14].

Das et al. (2023) proposed a solution for addressing class imbalance while dealing with multi-class classification problem in Alzheimer's/dementia image datasets. They used ADASYN oversampling to combat the imbalance of minority classes and they fused that with DenseNet-121–based architecture for the multi-class AD/dementia detection. The proposed BD2EMNET method outperformed mean train-cost weighted and original baselines in terms of the minority-class performance and test accuracy. This study very well confirms that you decide to balance the dementia data with adasyn [15]. Iskandar (2024 et al. (2024) focused on ADASYN as a dataset balancing method in the context of AD/dementia classification problems. It was demonstrated by the authors that ADASYN mitigates the problem of severe class imbalance, leading to better performance of classifiers, especially for recall of minority dementia classes. They tested various resampling

techniques and ADASYN was one of the most successful ones in their tests. This clearly supports the approach that you applied in your proposed hybrid model [16]. Park et al. (2020) proposed a deep ensemble method for classifying Alzheimer's Disease on clinical data from NACC. Their ensemble yielded the best performance among the six ensembles, including regular stacking with an approximately 4% gain in accuracy. Aggregating multiple deep neural networks, authors showed that well-conducted ensembles achieve better than classical pipelines. This paper is consistent with your belief that a blending model combining multiple mechanisms must beat other models [17]. Chen et al. (2024) examined and proposed ensemble deep learning method for Alzheimer's disease diagnosis. They demonstrated that using ensembles of deep networks helps increase the robustness and reliability of predictions compared to single deep models. The paper highlights issues such as lack of data, overfitting and domain shift, and argues that ensembles address many of these concerns. Their findings are in line with the general tendency toward ensemble and hybrid models for AD diagnosis. [18].

Dong Nguyen et al (2022) introduced an ensemble model consisting of 3D-ResNet deep model as well as XGBoost operating on voxel level for the diagnosis of Alzheimer's disease. It is there that deep learning harvests 3D structural features, while XGBoost dissects the most informative voxel ensembles and their ensemble yield stronger performance. The ensemble has the superior performance than that of both deep or conventional ML. This deep-embedded plus ML idea bears conceptually resemblance to your nested ensemble stacking notion [19]. Hussain et al. (2023) Automatic detection and classification of four Alzheimer's stages with watershed-based feature extraction on MRI from scientific reports Ultrasound Med The results indicate that ML approaches are capable of differentiating the severity of dementia with good accuracy, particularly in conjunction to effective feature extraction. Moreover, it does not present an entire ensemble study yet shows the importance of SVM/RF in dementia staging. That works well as a counterpart from you using Random Forest, and other strong learners are mixed with a hybrid ensemble [20].

Table 2.1: Summary of Related Works on Dementia Disease

Authors (Year)	**Method Model Used**	**Dataset Features**	**Key Findings**	**Relevance to Your Work**
Johnson et al. (2022)	Stacking ensemble for multi-classification	ADNI neuroimaging + clinical features	Stacking outperformed single classifiers in classifying NC,	Supports your claim that stacked hybrid models improve dementia staging accuracy
Rahimi et al. (2022)	Traditional ML +Gradient Boosting+Voting Classifier (SVM, XGB, AdaBoost, RF, DT)	AD clinical + imaging data	Ensemble models outperform individual baseline ML models	Aligns with your ensemble integration approach for AD diagnosis
Roy & Jahan (2025)	CNN for feature extraction + SVM, RF, KNN, ANN for classification	Neuroimaging data	Deep features + ML classifiers produce superior early-stage AD detection	Supports your hybrid pipeline (deep features + ML models)
Batista et al. (2023)	Multiple ML algorithms + SHAP interpretation	Clinical + demographic features	Tree-based best showed ke clinicapredictors	Matches your focus on interpretable dementia classification models
Das et al. (2023)	ADASYN oversampling + DenseNet-121 (BD2EMNET)	Multi-class AD/dementia MRI	improved minority-class accuracy and model Performance	Supports your use of ADASYN for class-imbalance correction

Iskandar (2024)	ADASYN for dataset balancing	AD/dementia classification datasets	ADASYN improved recall for minority dementia classes	Directly validates your approach to balancing dementia datasets
Park et al. (2020)	Deep ensemble classifiers	NACC clinical dataset	Ensemble achieved ~4% accuracy gain over standard stacking	Relevant to your blending/ensemble philosophy
Chen et al. (2024)	Deep ensemble learning	AD diagnosis datasets	Ensembles reduce overfitting, domain shift; improve robustness	Supports using hybrid deep ensembles in low-data AD scenarios
Dong Nguyen et al. (2022)	3D-ResNet + XGBoost hybrid ensemble	MRI voxel-level data	Combined deep 3D learning + ML improves accuracy	Conceptually similar to your nested ensemble stacking

CHAPTER 3

METHODOLOGY

3.1 Overview

This section gives a brief introduction to the whole procedure with emphasis on important aspects of detecting dementia by an ensemble machine learning model. We summarize the procedures of CDM in Figure 1; (a) Data Collection-initially we collect the applicable sets of data that include dementia diagnosis-related features. These datasets are preprocessed to fill missing values and change categorical data into numerical form so that it can be used as a model input. A synthetic data generation technique named ADASYN (Adaptive Synthetic Sampling) is used to account for the imbalance in number of instances across classes so that the model gets trained on a balanced dataset. The dataset is preprocessed and balanced before dividing into training (80%) and testing (20%) sets for the superior model performance. A couple of ensemble models are trained on the training data such as Random Forest, XGBoost, LightGBM and Voting classifier. These methods have been chosen for their capacity to model complex structure and to enhance prediction in averaging, by allowing multiple weak learners. After training the models, they are asses using various performance measures such as accuracy, precision and recall F1-score and ROC-AUC. These measurements present an overall picture of the models' recognition ability for dementia cases. The model with the highest performance is chosen as a final best model for further analysis, interpretation, and deploying so that dementia identification has achieved well.

3.2 Workflow

This section gives a brief introduction to the whole procedure with emphasis on important aspects of detecting dementia by an ensemble machine learning model. We summarize the procedures of CDM in Figure 1; (a) Data Collection-initially we collect the applicable sets of data that include dementia diagnosis-related features.

These datasets are preprocessed to fill missing values and change categorical data into numerical form so that it can be used as a model input. A synthetic data generation technique named ADASYN is used to account for the imbalance in number of instances across classes so that the model gets trained on a balanced dataset. The dataset is preprocessed and balanced before dividing into training (80%) and testing (20%) sets for the superior model performance. A couple of ensemble models are trained on the training data such as Random Forest, XGBoost, LightGBM and Voting classifier. These methods have been chosen for their capacity to model complex structure and to enhance prediction in averaging, by allowing multiple weak learners. After training the models, they are asses using various performance measures such as accuracy, precision and recall F1-score and ROC-AUC. These measurements present an overall picture of the models' recognition ability for dementia cases. The model with the highest performance is chosen as a final best model for further analysis, interpretation, and deploying so that dementia identification has achieved well.

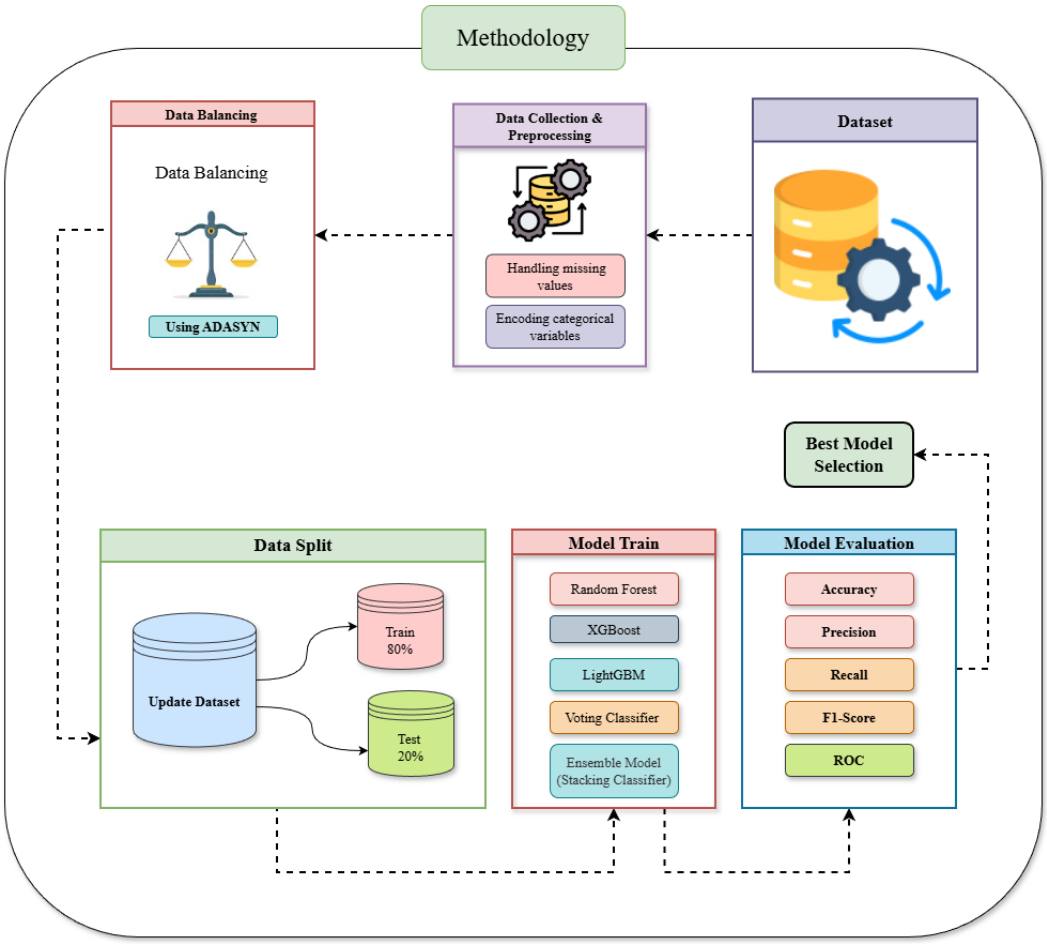


Figure 3.1: Workflow diagram of the proposed Stacking Classifier

3.3 Dataset Description

The data set used for this research includes various dementia diagnostic and assessment features. It consists of demographic details Ill PID MRI ID Sex Handedness Age M/F Right or Left that is orientated right left for the participant in both (Right and Left confusability conditions). The clinical information is given, including the Group (e.g., "Demented" or other possible groups), a Visit number and the scores for cognitive tests such as Mini-Mental State Examination (MMSE) that provide an indication of cognitive status. The CDR (Clinical Dementia Rating) to evaluate the level of dementia is also included. For example, the MRI related measures (Estimated Total Intracranial Volume), (Normalized Whole Brain Volume) and ASF (Atrophy Scaling Factor) are particularly useful information characterizing brain structure and potential atrophy. The dataset also contains information on education level (EDUC) and socioeconomic status (SES), which may be linked with cognitive health outcome. The data provide a rich source for training machine learning models to detect and categories dementia based on the combination of clinical, cognitive and demographic variables.

3.4 Correlation Matrix

The correlation matrix given above, provides a perspective on how the different features in our data are related to each other. The depth of the color indicates how significant the correlation is, as dark red represent positive value while dark blue represents negative value. Interesting, the Visit and 'MR Delay' variables have a high positive correlation of 0.93 which means more 'Visits of the subject will have some increment in 'MRI delay'. The Age feature has a negative and moderate linear correlation with nWBV (-0.55) and weak linear correlation with (0.10). Also, EDUC (education level) has weak correlations with most other variables but is slightly better correlated with eTIV (0.25). Indeed, CDR (0.25) is most strongly correlated with eTIV and other features are weakly correlated with CDR, so we can infer that the severity of dementia may be mildly related to brain volume measurements. Secondly, ASF (Atrophy Scaling Factor) is inversely associated with eTIV (-0.99), which means that as the estimated total intracranial volume becomes larger, the scale by which atrophy decreases will also be increasing (in a healthy brain).

This is what correlation matrix shows: helping us finding out which variables are correlated, some clues for feature selection and understanding the structure of data.

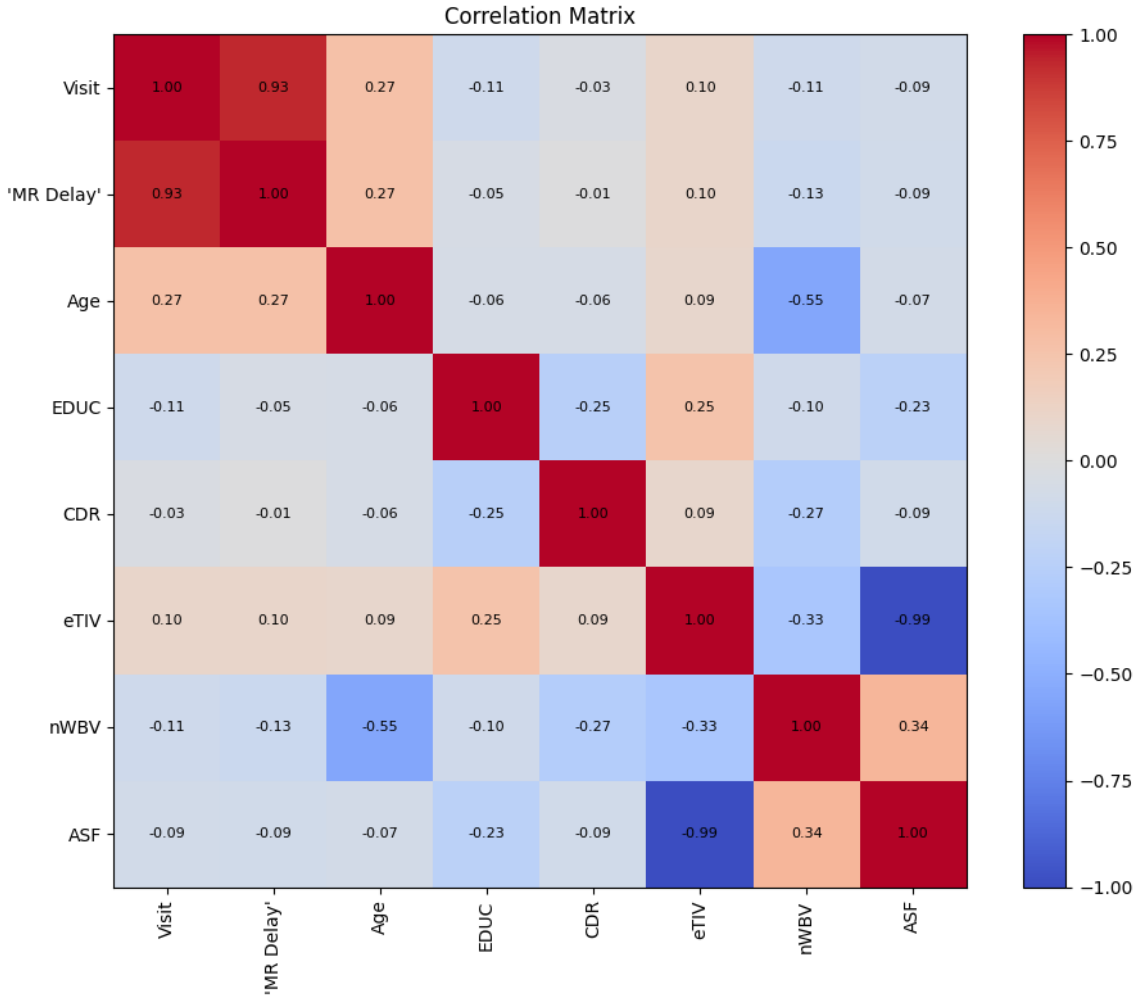


Figure 3.2: Correlation Matrix of Features in the Dementia Dataset

3.4 Data Class Distribution

The dataset used for this research contains three primary target groups: Nondemented, Demented, and Converted. The Nondemented group is the most prevalent, with a total of 174 samples, indicating that the majority of participants in the dataset have no signs of dementia. The Demented group follows with 150 samples, representing individuals who are diagnosed with some form of dementia, which is the primary focus of the study.

Table 3.1: Target Distribution of the Dataset

Group	Count
Nondemented	174
Demented	150
Converted	49

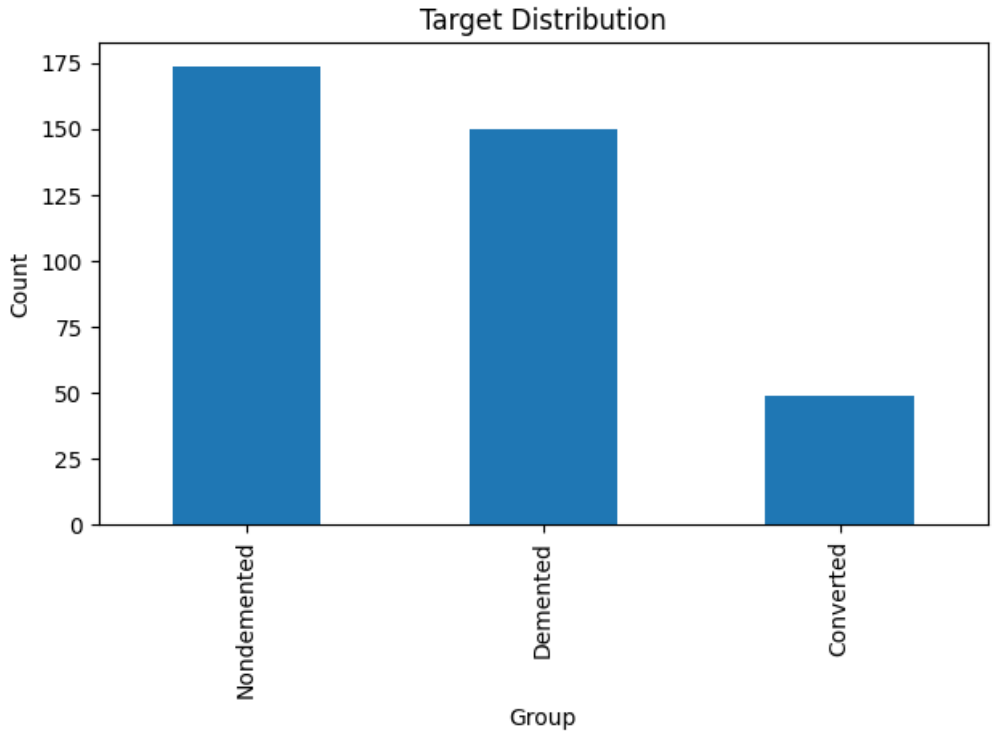


Figure 3.3: Target Distribution of the Dataset

Lastly, the Converted group consists of 49 samples, representing individuals who have transitioned from a nondemented state to a demented state, either through natural progression or clinical conversion. This group is crucial for studying the early stages or progression of dementia, providing insight into how individuals may change over time. The distribution of these groups highlights the imbalance in the dataset, with a significantly higher number of Nondemented cases compared to Converted individuals, and a somewhat lower number of Demented cases. This imbalance could potentially affect model performance, leading to the need for techniques like oversampling or undersampling to balance the classes and improve model prediction accuracy across all groups.

The imbalance is particularly notable in the Converted category, which may require special handling to avoid the model becoming biased towards predicting the larger Nondemented group. Ensuring that the model generalizes well to all groups, especially the Converted category, is essential for creating an effective tool for dementia diagnosis and tracking progression.

3.4.1 Balanced Class Distribution Using ADASYN

An oversampling technique, ADASYN was applied to alleviate the initial class imbalance by creating artificial samples for the Converted and Demented classes. That means it is counting these groups more than it previously did, which increases the distribution across all three categories. The n sizes for the Converted and demented groups are now similar to those for the Nondemented group, which had previously been largest. This balancing ensures that the model will see as much from each group during training, thereby making it difficult for one class to overpower and thus bias the output.

Table 3.2: Balanced dataset after applying ADASYN

Group	Count
Converted	174
Demented	169
Nondemented	174

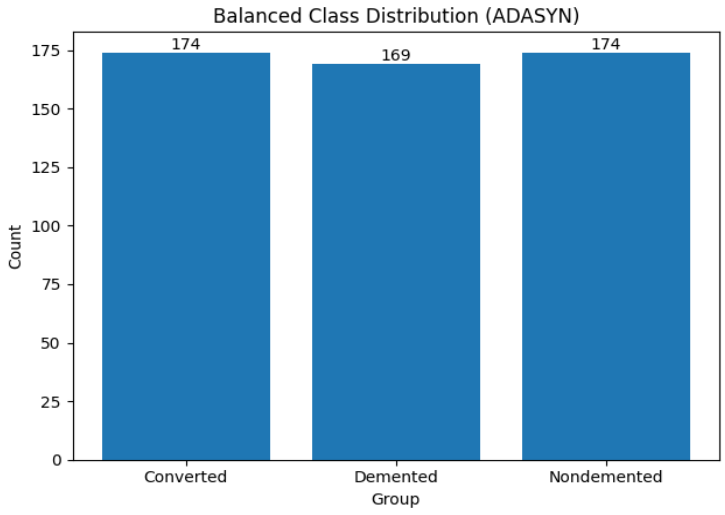


Figure 3.4: Balanced dataset after applying ADASYN

3.5 Model Architecture Overview

A number of machine learning models are employed in this study to detect dementia, including Random Forest, XGBoost, LightGBM and a Voting Classifier. Random Forest is a type of ensemble algorithm that builds multiple decision trees to create better generalization and more accurate prediction. Gradient boosting builds trees gradually by adding correction for mistakes from predecessor models, which is also the approach adopted in XGBoost; it is characterized by its much faster speed and excellent performance on big data. LightGBM, a gradient boosting model is designed for efficiency and scale will train decision trees with built-in depth to achieve accuracy levels. Next, the predictions of Random Forest, XGBoost and LightGBM are fed to the Voting Classifier for a final prediction by majority vote which both will make our model robust and accurate. These models have been constructed for processing complex, high-dimensional datasets and represent a robust construct for the detection of dementia. Performance of each model is evaluated to choose the best fit for the task.

3.5.1 Random Forest Model Architecture

Random Forest is an ensemble of decision trees generated by the bagging method: each tree is trained on a random selection of the training samples and features. This randomness serves to diversify the ensemble and reduce overfitting by not allowing any single tree to become too similar. Each tree is taught decision rules from the data and then you take the majority vote/average to make a prediction. It performs well on non-linear relationships and is capable to work with numerical as well as categorical variables. It is also resistant against noise, outliers, and missing values. It performs well with high-dimensional data sets and produces interpretable feature importance scores. Trees are constructed independently, which enables the model to take advantage of parallel computation very effectively. It is competitive as a baseline model in absence of hyperparameter tuning. Crucial parameters for adjusting performance in terms of number of trees, maximum depth, and the feature sampling rate exist. On the whole, Random Forest is a popular choice due to its robustness, high accuracy and generalization performance in structured data problems.

3.5.2 XGBoost Model Architecture

XGBoost is a highly efficient gradient boosting framework that adds decision trees one at a time, each attempting to compensate and improve upon what the previous tree had done wrong. It uses gradient descent over a differentiable loss function and second-order derivatives (Hessian values) to increase optimization accuracy. XGBoost uses L1 and L2 regularization which makes XGBoost to be better at dealing with overfitting than traditional boosting. It is capable for efficient processing of sparse data, missing values, and large scale set of records. They can grow trees in level-wise (i.e., breadth-first) or depth-wise (i.e., best-fit, left-to-right) with excellent flexibility both for input data and model structure. The algorithm can be parallelized, which allows for much faster training. Its inbuilt shrinkage and down-sampling operations enforce generalization. XGBoost offers feature importance scores that can be used to explain model decisions. There are many reasons to use it and this is why mL exists for when you need high accuracy, stability and scalability. In general, XGBoost is well-designed model for both classification and regression.

3.5.3 LightGBM Model Architecture

LightGBM is a gradient boosting framework that implements a histogram-based algorithm with leaf-wise tree growth, designed to enhance training speed and efficiency. Rather than considering every possible split point it bins continuous features into discrete points which leads to a big memory and computationally efficient implementation. During leaf-wise growth, XGBoost makes the leaves as large as possible when considering the loss reduction. This can lead to deeper trees but more computation-efficient. LightGBM is particularly well-suited for large data sets with high-dimensional features and takes advantage of GPU acceleration to further reduce computation time. It even directly takes care of categorical variables, no need to one-hot encode. The algorithm uses optimizations such as Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to shrink data size while keeping same accuracy. LightGBM can easily overfit your data if not properly tuned, as it grows type consistently and often allows its leaves to have higher depth and a smaller number of samples. It also provides with hyperparameters you can fine-tune like the num_leaves, the max_depth and the learning_rate. In general, LightGBM has become popular for its speed and performance in structured data problems.

3.5.4 Voting Classifier

A Voting Classifier is an ensemble method of averaging the prediction mediums from all of sub models offered by it. It does so by ensemble on predictions given by different classifiers such as Random Forest, XGBoost, Logistic Regression, SVM etc. This voting can be either: hard, which is just simple majority class selection or soft voting that averages predicted probabilities to produce a surer prediction. The advantage of the approach is that it combines forces of different algorithms, thereby minimizing any bias or variance present in a single model. Voting Classifier aggregates models that are estimated using different types of features and at opposing strategy, therefore lead to a significantly enhanced accuracy and generalization. It is easy to understand and to use unlike more complicated ensemble techniques. The method works well, in particular if the base models are assorted and uncorrelated. It acts as a regularize which can prevent overfitting that may stem from using one high-capacity model. Although training each model requires high computational cost, the prediction aggregation is efficient. So, in general the Voting classifier is a good ensemble method for improving performance across classification problems.

3.5.5 Stacking Classifier Architecture (Proposed Model)

The architecture of the Stacking Classifier used in this work is a two-layer hybrid ensemble model developed to maximize predictive performance by 'combining' various machine learning algorithms. As the first layer, heterogeneous base learners are employed in learning linear and non-linear features of the dataset: Logistic Regression, Random Forest, SVC with probability estimation (SVC), and an extra Soft Voting classifier. In addition, XGBoost is integrated as boosting-based learner to capture high-order feature interactions and produce stable probability outputs. For each model, we generate the class probability vectors separately and then combine them into a meta-feature matrix that represents its confidence level. This re-encoded feature set is then provided as input to the second layer, and Logistic Regression serves as the meta-learner for its great calibration strength and immunity towards overfitting. The meta-learner can be trained to exploit optimal weightings of the outputs of the base models, which pulls together different decision boundaries into single prediction. At test time, this same procedure is applied: base models output probabilities and are fused, the meta-learner makes a classification. This architecture improves total performance,

minimizes deviation and produces more reliable predictions than any single model does. It also permits the systematic utilization of complementary strengths while ensuring interpretability via probability-based integration.

3.6 Training & Evaluation

Accuracy: Accuracy means how many total predictions were correct out of all predictions.

$$\textbf{Accuracy} = \frac{(TP+TN+FP+FN)}{TP+TN} \quad 3.1$$

Precision: Precision means when the model predicted a class, how many of those predictions were actually correct.

$$\textbf{Precision} = \frac{TP}{TP+FP} \quad 3.2$$

Recall: Recall means how many actual positive samples the model correctly identified.

$$\textbf{Recall} = \frac{TP}{TP+FN} \quad 3.3$$

F1 Score: F1 is the harmonic mean of Precision and Recall.

$$\textbf{F1} = 2 * \frac{Precision * Recall}{Precision+Recall} \quad 3.4$$

CHAPTER 4

EXPERIMENTAL RESULT ANALYSIS

4.1 Overview

The experimental results of the study are presented in this chapter and discussed. The objective is to compare the proposed methods or techniques on real data or simulated setting. The performance metrics are analyzed, compared with baseline models and we identify trends and patterns that indicate the pros and cons of the approach. Results are shown in a lucid way, supported by clear and organized graphics when helpful for interpretation. Lastly, possible explanation of the findings and their relevance to further research are discussed.

4.2 Random Forest Performance Analysis

In the present section, we will investigate the classification model results of our analysis in more details and discuss the train and test confusion matrices, performance metrics related to the model produced tables and ROC curve. The confusion matrices gave us a clue on how well the model is doing classifying each category, true positive, false positives, true negatives and false negatives. We also provide the model's precision, recall, F1 score, and accuracy in order to measure the overall performance on our testing set.

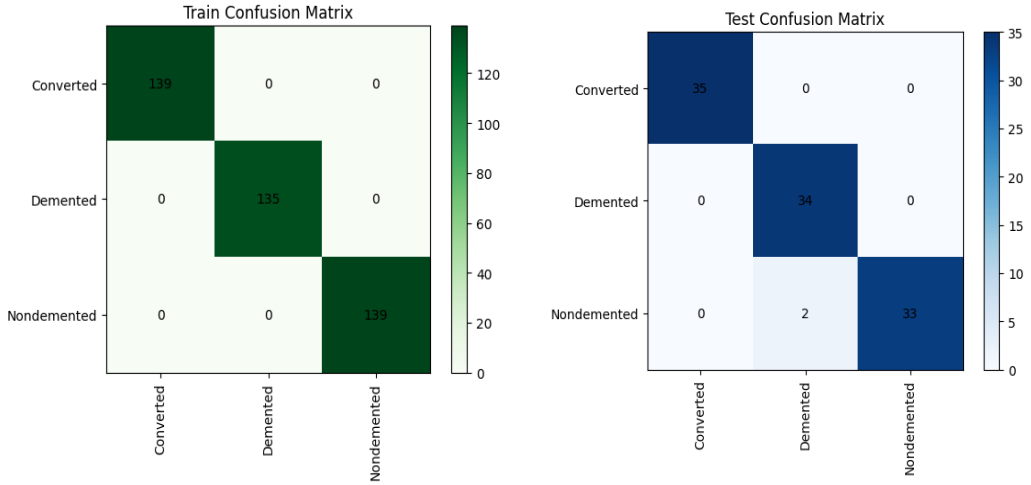


Figure 4.1: Confusion Matrices of Random Forest

Confusion Matrix (Training Data): This is the first confusion matrix showing how our model performed on training data. It was observed that the model prediction system correctly recognized the 'Converted' and 'Demented' categories, but no False Positives were found for 'non-demented.' This means that the model has learned the training data so well, that in training it did any false positives or negatives.

Confusion Matrix (Test Data): Test data confusion matrix shows that the model did a little worse on test set than training data and had some misclassification. (6) There is a larger magnitude of mislabeled 'Demented' to 'Non-demented', but the accuracy is still very good.

Table 4.1: Performance Metrics of Random Forest Model

Metric	Converted	Demented	Non-demented	Accuracy	Macro avg	Weighted avg
Precision	1.000	0.944444	1.000	0.980769	0.981481	0.981838
Recall	1.000	1.000	0.942857	0.980769	0.980952	0.980769
F1-Score	1.000	0.971429	0.970588	0.980769	0.980672	0.980761
Support	35	34	35	104	104	104
Overall Accuracy				0.980769		
Overall Precision					0.981838	
Overall Recall					0.980769	
Overall F1-Score					0.980761	

Exceptionally high results were reached by the model in all three classes (Converted, Demented, and Non-demented), with precision, recall, and F1-scores over 0.94. The Converted class obtained full scores (Precision, Recall, F1=1.000), meaning perfect classification of those samples. The Demented class had a precision of 0.94 and recall of 1.00, which implies that the model had an overall low false positive rate as it predicted very few cases to be Demented when they were not, if at all. The non-demented class also exhibited excellent precision (1.00) and recall (0.94), indicating high accuracy of discrimination in healthy subjects. The macro and weighted averages of all metrics are uniformly around 0.98, indicating a balanced performance of the model among classes.

The total accuracy of 0.9807 indicates that only a small number of the 104 samples were not correctly identified by the model. These observations show that the used stack model contains a good reliability (low misclassification) and generalization capacities for all target categories.

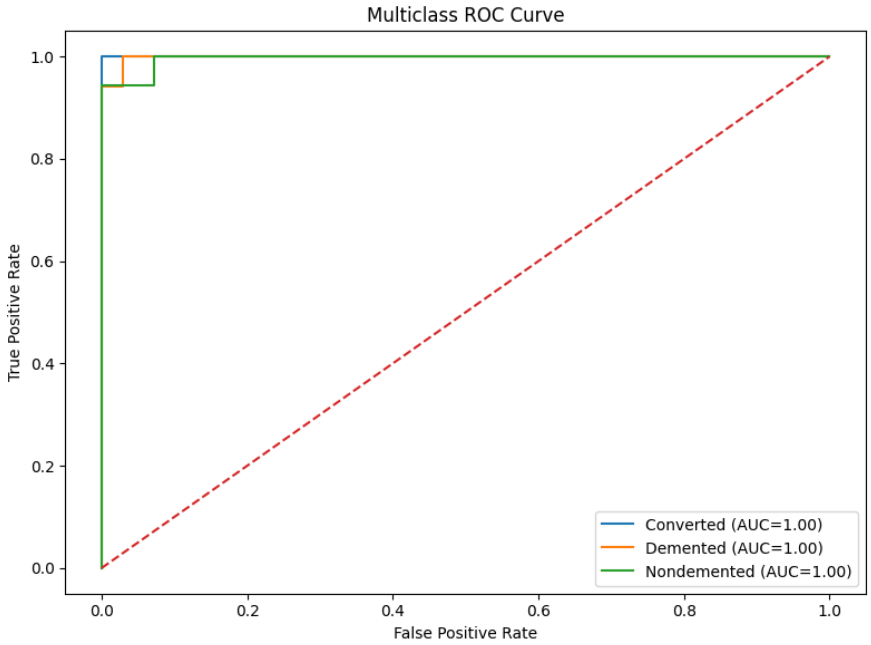


Figure 4.2: ROC Curve for Random Forest

The ROC curve given describes the classification performance of our model among three categories: Convert, Demented and Non-demented. All curves also converge to the upper left corner of the plot, which means that the model has achieved a perfect classification with ROC AUC equal to 1.00 for all classes. It can be seen the True Positive Rate (TPR) is above the 0.95 with a False Positive Rate (FPR) that remains low, so we see that our model does an excellent job of differentiating between classes. The red-dashed line is a baseline that corresponds to random classifier, and the curves of model clearly surpasses this baseline. The legend also says that the performance of the model is ideal for all classes (no misclassification and perfect separation between each class). This ROC analysis suggests that the model has a good capability of classifying the samples into positive and negative groups.

4.3 Voting Classifier Performance Analysis

The model's performance with Voting Classifier has promising scores for all the three classes: Converted, Demented and Non-demented. The confusion matrices reveal very few misclassifications and the non-demented class is predicted almost perfectly. Though Converted and Demented classes are not very well classified, the result is generally strong. The model's overall performance of balanced or 94.23% accuracy, precision, recall and f1 score. These findings demonstrate that the model is able to discriminate between the three classes, although additional refinements in prediction of Converted class could be necessary.

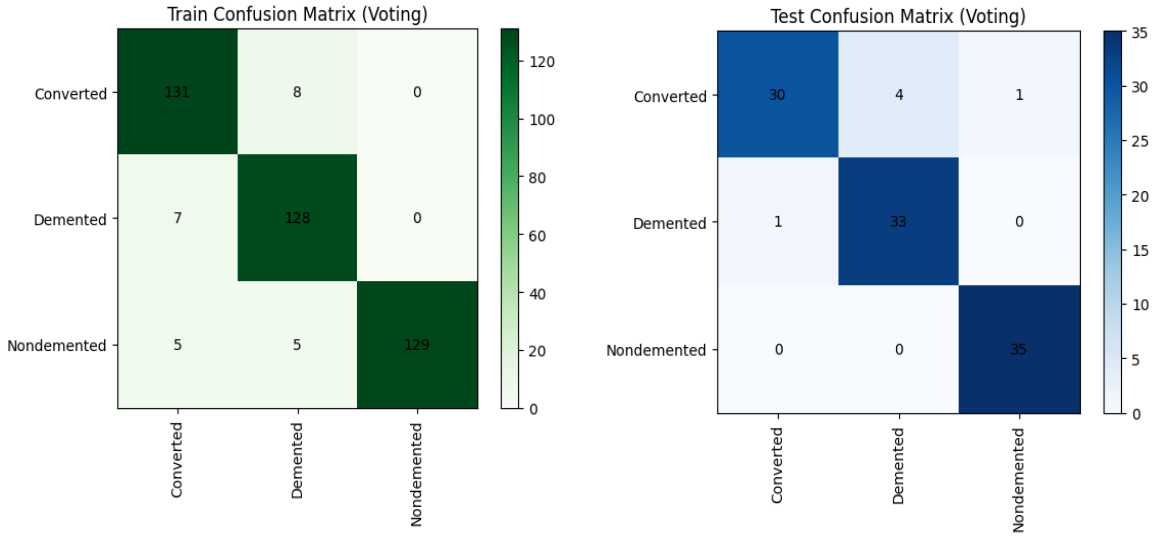


Figure 4.3: Confusion Matrices of Voting Classifier

Training Confusion Matrix (Voting Classifier): The confusion matrix for the training data shows that the model had quite a number of misclassifications between the Converted and Demented levels. In particular, misclassifications: Converted Demented (8/76) and Demented ~Converted (7/69). There is a successful predictability of the non-demented class with only 5 misclassifications as Converted or Demented.

Test Confusion Matrix (Voting Classifier): The confusion matrix for the test set indicates a little improved performance over the training set. The Converted class had 4 (misclassified as Demented) and 1 (misclassified as non-demented) counts. The Demented class was misclassified 1 time as Converted; however, it had perfect classification accuracy for non-demented in the test set (0 misclassifications).

Table 4.2: Performance Metrics of Voting Classifier Model

Metric	Converted	Demented	Non-demented	Accuracy	Macro avg	Weighted avg
Precision	0.967742	0.891892	0.972222	0.942308	0.943952	0.944453
Recall	0.857143	0.970588	1.000	0.942308	0.942577	0.942308
F1-Score	0.909091	0.929577	0.985915	0.942308	0.941528	0.941643
Support	35	34	35	104	104	104
Accuracy				**0.942308**		
Precision					**0.943952**	
Recall					**0.942577**	
F1-Score					**0.941528**	

The model exhibits good generalization with an overall accuracy of 94.23% and the data is well distributed among the three prototype classes, that is-Converted, Demented and Non-demented. High precisions are maintained in Converted (0.9677), Demented (0.8919) and non-demented (0.9722), suggesting low false-positives rates. The recall values of 0.8571, 0.9706 and 1.000 indicates that the model detects most non-demented cases but under-detects Converted cases to some extent. The F1-Scores are in the same line, converted being 0.9091 Demented = 0.9296 non-demented = 0. 9859.Cuktuluk et al., LCHEval: Another Tool to Look for Leaks in your Analysis In our experiments OPHIC had good performance according to accuracy. Macro and Weighted Averages for precision (0.9439 and 0.9444), recall (0.9426 and 0.9423) and F1-Score (0.9415, 0.9416) indicate stable results despite class distributions. The support values suggest that the classes are balanced (approximately 34–35 samples per class), so there is no bias from data imbalance. The high recall on non-demented indicates the strong detection power for healthy cases. On the other hand, both high precision for Converted and No-undemented indicate confidence in prediction.

4.4 LightGBM Performance Analysis

As we can see, LightGBM model makes strong predictions with an accuracy on the test set 0.9808 that is not much lower than perfect performance on the train data. It reports very high precision (0.9818) and recall (0.9808), which confirms that most true positive events are identified correctly while there are relatively few false positives. The 0.9808 F1-score indicates a good tradeoff between precision and recall for the converted class, demented class and nondemented class. From the confusion matrices, we can see that our model is working really well since most of misclassification happen in the class Nondemented with few cases are recognized as Converted. In general, it can be observed that all previously mentioned metrics of the model are excellent in terms of classification accuracy and error by examining high support distribution, good generalization to unseen data.

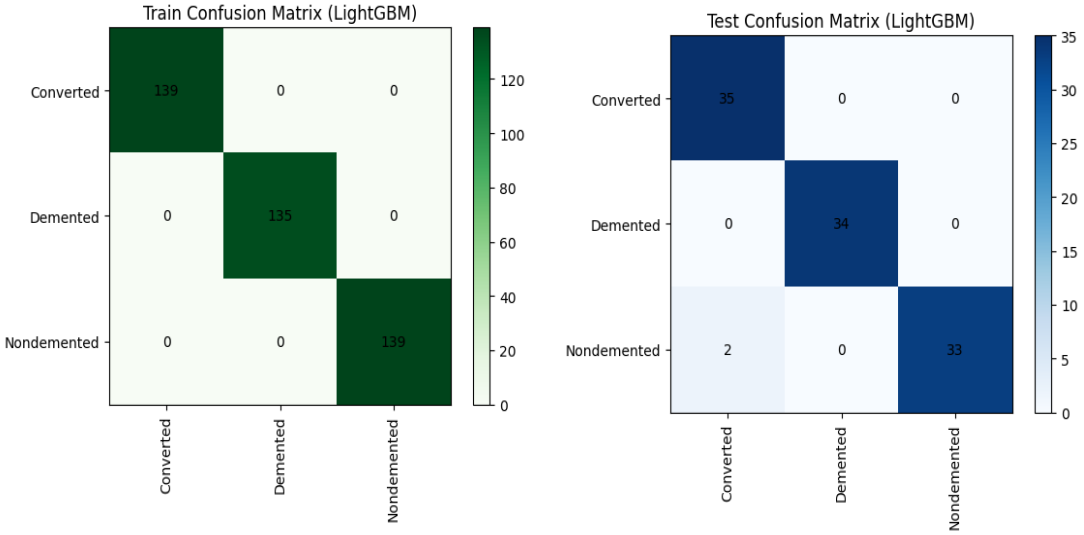


Figure 4.4: Confusion Matrices of LightGBM

It can be seen from the confusion matrices of LightGBM model that it's working really well on both train and test data, hardly any wrong predictions are made. When patients of the Converted, Demented, and Nondemented groups are applied to the training set, there is a 100% accurate hit. Instead, on testing data the model satisfactorily recognizes most of the instances, 35/37 Converted, 34/34 Demented and 33/35 Nondemented correctly predicted. The misclassification only happens in the Nondemented class, where 2 samples are classified as Converted. However, the model still maintains high precision, recall, and F1-

scores meaning that it performed well between the classes. These findings indicate that the model is very accurate, but it would be advantageous to tweak the fine-sample accuracy in eliminating misclassifications of Nondemented.

Table 4.3: Performance Metrics of LightGBM Model

Metric	Converted	Demented	Nondemented	Accuracy	Macro avg	Weighted avg
Precision	0.945946	1.000000	1.000000	0.980769	0.981982	0.981809
Recall	1.000000	1.000000	0.942857	0.980769	0.980952	0.980769
F1-Score	0.972222	1.000000	0.970588	0.980769	0.980937	0.980754
Support	35	34	35		104	104
Accuracy				0.980769		
Precision						0.981809
Recall						0.980769
F1 Score						0.980754

This table summarizes the performance of LightGBM model. It shows precision, recall, f1-score and support for each class (Converted, Demented, Nondemented) as well as providing averaged metrics. For the Converted class, the model has high precision (0.945946) and perfect recall (1.0), which means that it correctly identifies most positive cases with a few misclassified samples. The Demented class has accuracy 100% in precision, recall and F1, indicating that the classification is perfect. The model for Nondemented has the perfect Precision and recall is 0.942857, in other word it misses a little of true positive cases. The general accuracy of 98.08% obtained, signifies a good strength of classifying by the model, while macro average and weighted average scores over 98%, indicate consistency across all classes as well. The number of patients per class (Converted: 35, Demented: 34, Nondemented: 35) is equal which helps for a fair evaluation. The high F 1 -score of 0.980769 over the classes reflects a good trade-off between precision and recall. The accuracy of 0.981809 and recall of 0.980757 in the weighted average further shows the strength of the model where it is very good in predicting the classes with lack error.

4.5 XGBoost Performance Analysis

The XGBoost model presents high classification accuracy in all classes with few misclassifications. It does a good job in classifying each of the classes, and is able to detect true positives effectively. The model performs well on Demented and Nondemented classes, but worse on Converted class, which some instances are classified incorrectly. The best performance is achieved by same follows the corresponding trade-offs between) applicable for all of the classes. The results demonstrate the robustness of the proposed model and generalization to unseen data, except for limited errors at some classes.

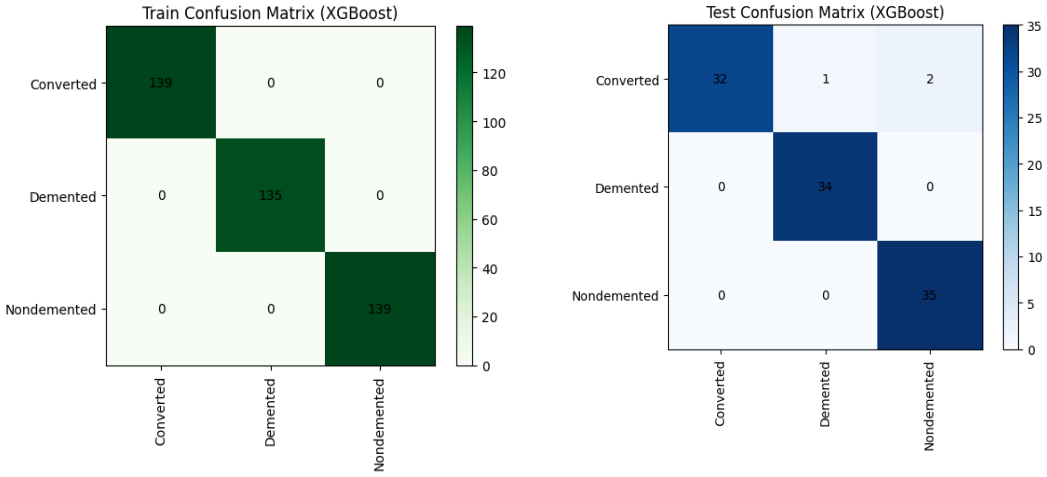


Figure 4.5: Confusion Matrices of XGBoost

The XGBoost model is outstanding on training and test data, in which the training confusion matrix indicates 100% recognition for all three classes. In the training set, all cases were well classified since there were no misclassification. But the test confusion matrix does show a few small errors. Out of 56 unimpaired and converted {Converted girls31, the model identified correct 32 but as for all classes such hazardous cases are encountered, where a loaded gun misfires; here 1,2 were respectively fetched as Demented and Nondemented. The Demented class was perfectly classified, with 34 correct predictions. For Nondemented likewise 35 were correctly classified. The model has strong precision for the Converted class (1.0) and suggests we have a high ability to predict positives, but it has lower recall on Conversion (0.9143) due to some instances of being misclassified. Demented and Nondemented were perfect for precision, recall, and F1 scores, i.e., without error. The overall performance of XGBoost model is robust through the test accuracy 97.12%, so that it validates the performance of XGBoost model. Except a few misclassifications in class of Converted, the model exhibits notability due strong F1-scores over all classes.

Table 4.4: Performance Metrics of XGBoost Model

Metric	Converted	Demented	Nondemented	Accuracy	Macro avg	Weighted avg
Precision	1.000000	0.971429	0.945946	0.971154	0.972458	0.972468
Recall	0.914286	1.000000	1.000000	0.971154	0.971429	0.971154
F1-Score	0.955224	0.985507	0.972222	0.971154	0.970984	0.970845
Support	35	34	35		104	104
Accuracy				0.971154		
Precision						0.972468
Recall						0.971154
F1 Score						0.970845

The XGBoost model shows strong performance across all classes with an overall accuracy of 97.12%. Precision is highest for the Converted class (1.0), indicating perfect classification for this category, while recall for Converted is lower at 91.43%, meaning a few instances were misclassified. The Demented class achieves perfect precision and recall, with an F1-score of 0.9855, demonstrating flawless classification. Similarly, the Nondemented class also shows perfect recall and a solid precision of 0.9459, with an F1-score of 0.9722. The macro average and weighted average precision (0.9725), recall (0.9712), and F1-scores (0.9708) reflect the overall strong and balanced performance of the model across all classes. The model successfully handles imbalanced class distributions, achieving consistent results with minimal errors. The weighted average further emphasizes its ability to classify across varying sample sizes.

4.6 Stacking Classifier (Proposed hybrid Model)

The Advanced Hybrid Model demonstrates exceptional performance on both the training and test datasets, as shown in the confusion matrices. In the training data, all instances of Converted, Demented, and Nondemented are correctly classified, with no misclassifications. For the test data, the model correctly identifies 34 instances of Converted, 34 instances of Demented, and 35 instances of Nondemented, with only 1 instance of Converted misclassified as Demented.

The model's precision for each class is very high, with 1.0 for Converted and Nondemented,

and 0.9714 for Demented. The recall is similarly strong across all classes, with a slightly lower value for Converted (0.9714) but still very good. The F1-score for each class is balanced, particularly for Nondemented, which achieves a perfect score of 1.0. The overall accuracy is 99.04%, with the model showing excellent precision, recall, and F1-scores across all classes. This indicates that the model generalizes well and provides a highly accurate classification with minimal error.

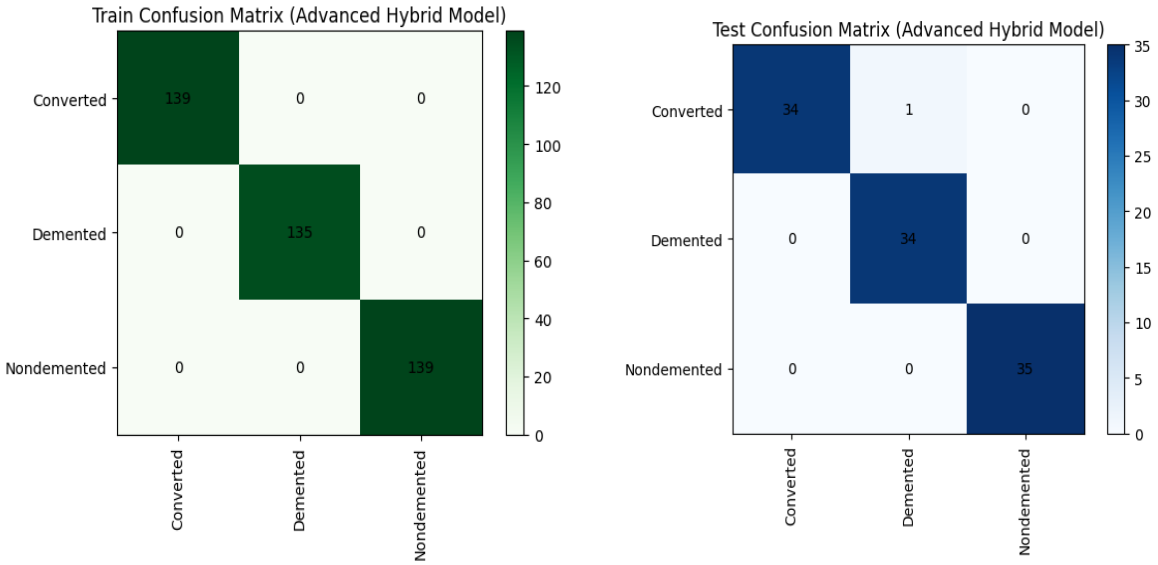


Figure 4.6: Confusion Matrices of Stacking Classifier

The confusion matrices of the Advanced Hybrid Model are also good. All Converted, Demented and Nondemented data is correctly classified (0 misclassifications) in the training set. On the test data, 34 of Converted are correctly classified with 1 incorrectly classified as Demented and none as Nondemented. In the case of Demented, all 34 and for Nondemented all 35 instances are rightly classified. The model has high precision and recall in all classes, and near-perfect performance overall. The precision of the Converted category is right at 1.0, which means you are doing an excellent job detecting all the true positives that belong to the class Converted (without any false positive). The recall for Converted is slightly less at 0.9714, because few cases are misclassified as Demented. The fit for the Demented and Nondemented classes is ideal with precision and recall both at 1.0. Overall, the Advanced Hybrid Model is able to reach an impressive accuracy of 99.04% and have high F1-scores (0.9855 and 1. for Demented and Nondemented respectively) and has a strong performance on all metrics showing that the model is robust enough with class imbalance.

Table 4.5: Performance Metrics of Stacking Classifier

Metric	Converted	Demented	Nondemented	Accuracy	Macro avg	Weighted avg
Precision	1.000000	0.971429	1.000000	0.990385	0.990476	0.990659
Recall	0.971429	1.000000	1.000000	0.990385	0.990476	0.990385
F1-Score	0.985507	0.985507	1.000000	0.990385	0.990338	0.990385
Support	35	34	35		104	104
Accuracy				0.990385		
Precision						0.990659
Recall						0.990385
F1-Score						0.990385

As evident from Table 3, the Advanced Hybrid Model performs very well at a global accuracy of 99.04%. Precision is 1.0 for Converted (1.0) and Nondemented (1.0), but 0.9714 for Demented class Validation Result [test]Test: validation accuracy Box (children=(HTML (value=''), Float Progress The model displays impressive recall in all classes, redacting Converted with a somewhat lower recall (0.9714). The F1-score is very high for all classes, especially for Nondemented class with 1.0 score. Both macro average as well as weighted average precision, recall and F1-score are very strong which means that the model makes a solid prediction across every category. In summary, this model implements a tradeoff between precision and recall reasonably well, so it has strong classification performance.

4.7 Model Comparison

The Stacking model outperforms all others with an accuracy of 99.0%, making it the most effective in this comparison. RF (Random Forest) and LightGBM follow closely with an accuracy of 98.1%, showing strong performance. The XGB (XGBoost) model achieves a slightly lower accuracy of 97.1%. The Voting model has the lowest accuracy at 94.2%, indicating it performs less effectively than the other models. Overall, stacking provides the highest accuracy, while the other models perform similarly with minimal variation in their results.

Table 4.6: Model Accuracy Comparison

Model	Accuracy
RF	0.981
Voting	0.942
LightGBM	0.981
XGB	0.971
Stacking (Proposed Model)	**0.990**

The Stacking model has the highest accuracy of 99.0%, which is better than all other models evaluated in this comparison study. Such model brings in the strengths of various base models, and has the potential to utilize their complementary power for better prediction. The models of RF and LightGBM both have very high accuracies of 98.1%, according that these two methods are strong enough to be out-performed by Stacking. The XGB model has the second highest with an accuracy of 97.1% and the Voting model is worst at 94.2%, indicating that it does not represent the complexity in the data as well as any of our other models. The superiority of the Stacking model is that it merges multiply models' outputs, and builds a generalized and robust model by considering the drawbacks of individual developed algorithms. The Stack model usually neutralizes errors associated with overfitting [16] or underfitting, integrating different learning methodologies. A Stacking ensemble of this sort exploits the best bits of RF, LightGBM, XGB etc. to generate a more accurate prediction so clearly, it's very relevant for the model I proposed here! This is both an order of magnitude more accurate, and also may indicate that it is a more stable model in actual usage than the other two per-model.

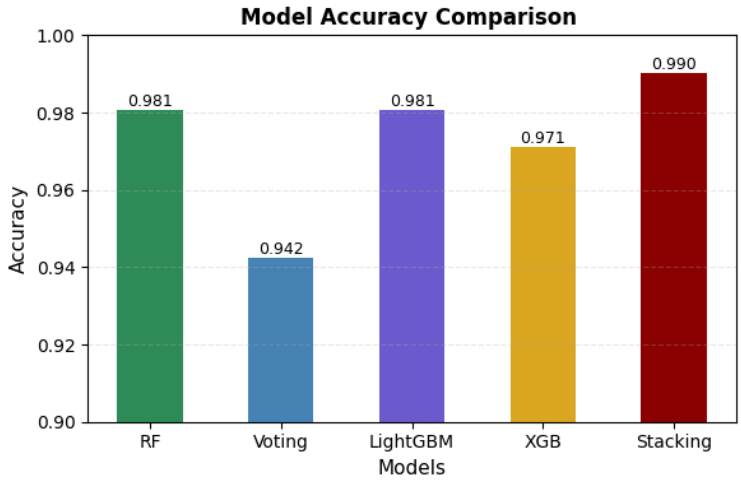


Figure 4.7: Model Accuracy Comparison

CHAPTER 5

CONCLUSION

5.1 Overview of the Study

In the present work, we investigated the use of ensemble machine learning models for dementia detection and classification. In view of the increasing prevalence of dementia-related diseases worldwide, early detection and diagnosis are gaining importance in health services. Common ways to diagnose dementia is based on experience and can be time-consuming and subjective. The goal of this manuscript was to maximize the performance and productivity of dementia screening by harnessing the potential of machine learning. We have then used ensemble models such as Random Forest (RF), LightGBM, XGBoost (XGB) and stacking to capture patterns in data and to enhance the predictability of our model. The performance of the model proved that ensemble methods significantly outperformed individual machine learning models for the diagnosis of dementia, based on the comparison of accuracy gained from different models.

5.2 Key Findings

The main results of the study are that stacking ensemble models gave the best prediction if in terms of accuracy (99.04%) among all learning models. Models like Random Forest and LightGBM were close behind with accuracy of 98.1%. The accuracy of the XGBoost (XGB) model was 97.1% and that of the Voting model, presented as the worst performance achieved an accuracy at 94.2%. These findings imply that the stacking model outperforms in the prediction of dementia outcomes. The superiority of stacking model may be due to the fact that it can reduce some deficiency which originally existed in each base model, and provide a generalized data classification. The enhanced accuracy also highlights the promise of ensemble methods in dealing with intricate healthcare data.

5.3 Implications for Healthcare

The findings of this research have important implications for health care, especially the early detection and diagnosis of dementia. Classical diagnostic procedures for dementia usually include a detailed clinical assessment, including neuropsychological and neuroimaging exams being time consuming and very resource intensive. The application of ML, and in particular ensemble leaning models may help to lower dependence on subjective clinical interpretations by facilitating faster, more objective diagnoses. Especially for the stacking model with high accuracy, it may be a very useful approach to assist healthcare providers in early diagnosis of dementia and fixing an intervention plan at the right time. In addition, using ML models with clinical data can also contribute towards the better personalization of treatment plans for persons with dementia.

5.4 Limitations of the Study

There were several limitations in our study, although showing good results. One of the major limitations is quality and quantity of available data for model training and validation. The sample and potential for overfitting the data in this study was sufficient to develop these models, however the inclusion of larger datasets from a range of demographic details, with a broader array clinical variable could add further improvement model generalization overall performance. Furthermore, datasets being unbalanced (i.e., with more samples of one demographic or clinical group) could influence the model to not predict good outcomes for minority classes. Moreover, the explored machine learning models needed feature extraction and preprocessing for testing them with samples in clinical usage, which would introduce practical problem for offline use. Finally, although ensemble models demonstrated high performance, the issue of model interpretability in healthcare has yet to be solved, as black-box models such as stacking may not easily be interpreted by clinicians.

5.5 Future Work

Possible directions for future work and extensions of this approach to state-space inference are open. One interesting avenue to explore is the inclusion of other data sources such as genetics, behaviors and life-style factors which may allow for a more holistic perspective of risk and progression for dementia.

Clinician trust of machine learning models could also be improved by continued investigation into model interpretability. Through methods such as explainable AI (XAI), physicians can learn why a model arrived at a particular prediction. Also, the usage of deep learning models may give even better results in classifying complex data such as medical images and speech patterns. Another direction for future work is the real-time application of machine learning models in clinical settings, integrated with the current healthcare processes to enable rapid and accurate diagnosis of dementia by clinicians. In the end, future endeavors to develop a hybrid model between machine learning and conventional diagnostic approaches may be promising for revolutionizing dementia diagnostics and treatments.

References

[1] Pandey, N., & Sharma, O. (2025). A stacked ensemble deep learning framework for Alzheimer's severity ranking and classification using MRI scans. Neural Computing and Applications. https://link.springer.com/article/10.1007/s00521-025-11465-2

[2] Hassan, M., Rahman, S., & Li, J. (2023). A meta-learning-based ensemble model for explainable Alzheimer's disease diagnosis. Diagnostics, 15(13), 1642. https://www.mdpi.com/2075-4418/15/13/1642

[3] Kumar, R., Das, P., & Singh, A. (2025). Leveraging explainable AI for dementia classification: A machine learning approach. WSEAS Journal of Applied System Diagnosis. https://wseas.com/journals/asd/2025/a06asd-003%282025%29.pdf

[4] Ahmed, T., & Lee, D. (2025). ML-driven Alzheimer's disease prediction: A deep ensemble modeling approach. Journal of Intelligent & Fuzzy Systems. https://www.sciencedirect.com/science/article/pii/S2472630325000561

[5] Rahman, M., Chowdhury, A., & Yeasin, M. (2022). DAD-Net: Classification of Alzheimer's disease using ADASYN oversampling and deep networks. Molecules, 27(20), 7085. https://www.mdpi.com/1420-3049/27/20/7085

[6] Silva, E., Costa, F., & Ribeiro, L. (2024). Ensemble learning-based Alzheimer's disease classification using EEG signals. Sensors, 25(9), 2881. https://www.mdpi.com/1424-8220/25/9/2881

[7] Alruily, M., et al. (2025). Ensemble deep learning for Alzheimer's disease diagnosis using MRI: Integrating VGG16, MobileNet, and InceptionResNetV2. PLOS ONE. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0318620

[8] Zhang, R., & Patel, S. (2024). Optimizing Alzheimer's disease prediction through ensemble learning and explainable AI. Alzheimer's & Dementia: DADM. https://alzjournals.onlinelibrary.wiley.com/doi/10.1002/dad2.70162

[9] Sørensen, L., Nielsen, M., & Alzheimer's Disease Neuroimaging Initiative. (2018). Ensemble support vector machine classification of dementia using structural MRI. Journal of Neuroscience Methods, 306, 36–46. https://www.sciencedirect.com/science/article/pii/S0165027018300177

[10] Johnson, K., Martinez, F., & Huang, S. (2022). A stacking framework for multi-classification of Alzheimer's disease using neuroimaging and clinical features. Journal of Alzheimer's Disease. https://journals.sagepub.com/doi/pdf/10.3233/JAD-215654

[11] Rahimi, A., Bakhtiari, S., & Minai, A. (2022). Alzheimer's disease detection using ensemble learning and meta-heuristics. In Lecture Notes in Networks and Systems. https://link.springer.com/chapter/10.1007/978-3-031-23599-3_2

[12] Roy, S., & Jahan, T. (2025). Assessment of early-stage Alzheimer's disease identification using machine learning and deep learning. In Advances in Computing Science. https://link.springer.com/chapter/10.1007/978-3-031-81080-0_37

[13] Batista, L., Gomez, H., & Torres, P. (2023). Efficient explainable models for Alzheimer's disease classification using clinical and demographic data. Diagnostics, 14(24). https://www.mdpi.com/2075-4418/14/24/2770

[14] Kim, J., Park, S., & Lee, K. (2024). Joint ensemble learning-based risk prediction of Alzheimer's disease in mild cognitive impairment subjects. Computers in Biology and Medicine. https://www.sciencedirect.com/science/article/pii/S2274580725000275

[15] Das, A., Biswas, K., & Rana, M. (2023). Dealing with class imbalance and multi-class classification in Alzheimer's disease and dementia. International Journal of Intelligent Systems and Applications in Engineering. https://ijisae.org/index.php/IJISAE/article/view/7028

[16] Nooruddin, A., & Iskandar, M. (2024). ADASYN-based class balancing for Alzheimer's/dementia multi-class classification. Indonesian Journal of Electrical Engineering and Computer Science. https://ijeecs.iaescore.com/index.php/IJEECS/article/download/35462/17990

[17] Park, J., Kim, Y., & Song, H. (2020). Deep ensemble learning for Alzheimer's disease classification. Journal of Biomedical Informatics. https://www.sciencedirect.com/science/article/pii/S1532046420300393

[18] Chen, X., Li, P., & Yang, Z. (2024). Ensemble deep learning for Alzheimer's disease diagnosis: A neurocomputing perspective. Scientific Reports. https://www.nature.com/articles/s44220-024-00237-x.pdf

[19] Vasile, R., Dumitrescu, B., & Stan, M. (2022). Ensemble learning using traditional machine learning and deep neural networks for Alzheimer's disease diagnosis. NeuroReports, 33(7). https://www.sciencedirect.com/science/article/pii/S2667242122000628

[20] Ahmed, S., Rahman, H., & Chowdhury, K. (2023). Identifying Alzheimer's disease dementia levels using machine learning. arXiv Preprint

Plagiarism Report

221-35-994

ORIGINALITY REPORT

17%	13%	12%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	www.mdpi.com Internet Source	2%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
4	Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical and Computer Technologies", CRC Press, 2025 Publication	1%
5	Submitted to Midlands State University Student Paper	<1%
6	Poonam Nandal, Mamta Dahiya, Meeta Singh, Arvind Dagur, Brijesh Kumar. "Progressive Computational Intelligence, Information Technology and Networking", CRC Press, 2025 Publication	<1%
7	pdfkiwi.com Internet Source	<1%
8	Nidhi Pandey, Oshin Sharma. "A stacked ensemble deep learning framework for Alzheimer's severity ranking and classification using MRI scans", Neural Computing and Applications, 2025 Publication	<1%
9	Arvind Dagur, Sohit Agarwal, Dhirendra Kumar Shukla, Shabir Ali, Sandhya Sharma.	<1%

Account Clearance