AI based Early Prediction of Alzheimer's Disease using Neuropsychological Tests

Author

MUHAMMAD ALI WARIS KHAN

MS-21 (EE)

Registration Number

360949

Thesis Supervisor

DR.SHAHZAD AMIN SHEIKH

DEPARTMENT OF ELECTRICAL ENGINEERING
COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY
ISLAMABAD, PAKISTAN

SEPTEMBER 2024

THESIS ACCEPTANCE CERTIFICATE

It is certified that final copy of MS/MPhil thesis written by Mr. Muhammad Ali Waris Khan (Registration No. 00000360949) Entry-2021, of (College of E&ME) has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors, and mistake and is accepted as partial fulfillment for award of MS degree. It is further certified that necessary amendments as pointed out by GEC member of the scholar have also been incorporated in the said thesis.

Signature: Shahzad.

Name of Supervisor Dr. Shahzad Amin Sheikh

Date: 12-Sep-2024

Signature (HoD):

Dr. Qasim Umar Khan

Date: 12-Sep-2024

Signature (Dean):

Brig Dr. Nasir Rashid

Date: 12 SEP 2024

DEDICATION

This thesis is dedicated to:

First and foremost ALLAH ALMIGHTY, my beloved parents and family, whose steadfast affection, encouragement, and compromises have been my biggest source of confidence and motivation, my respected supervisor, Dr. Shahzad Amin Sheikh, for his priceless advice and mentorship, and all of those who had faith in me and supported me throughout this journey.

ACKNOWLEDGEMENTS

In the name of Allah, the Most Gracious, the Most Merciful,

Above all, I want to express my sincere gratitude to Allah Ta'ala for giving me the courage, persistence, and insight I required to finish my thesis. All of this would not have been possible without HIS heavenly direction and innumerable favors. My sincere thanks goes out to my parents and family for their constant love, support, and encouragement along this journey. Their unwavering faith in my potential has served as my inspiration, and their selflessness has been the basis of my achievement.

My profound gratitude goes out to Dr. Shahzad Amin Sheikh, my supervisor, whose knowledgeable counsel, perceptive criticism, and unwavering support were crucial to the successful completion of this thesis. His perseverance, comprehension, and commitment have been crucial in molding and completing my research. I also like to express my gratitude to Sir Kamran Aziz Bhatti and Dr. Qasim Umar Khan, two GEC members, for their support, helpful criticism, and invaluable advice during my research. Their knowledge and recommendations have greatly improved the caliber of my work.

A particular thank you to my friend and colleague Saddam Gul, whose encouragement and support have been invaluable in helping me finish this thesis. His guidance and unwavering readiness to assist me at every turn have been really beneficial. He introduced me to my supervisor, Dr. Shahzad Amin Sheikh, without whom my thesis would not have been feasible, for which I am especially grateful.

Furthermore, I would want to thank everyone who has helped me in any manner with this study endeavor. Their support and encouragement are much appreciated. Finally, I would like to express my gratitude to all of my friends and coworkers who have helped and supported me morally over this difficult but worthwhile trip. Their inspiration and unity have always been a source of support.

I am grateful to everyone who helped make this accomplishment possible.

ABSTRACT

Within the field of biomedical engineering, a novel imputation model employing machine learning techniques is proposed as a unique approach for the early diagnosis of Alzheimer's disease (AD). Early identification is essential to postpone the progression of Alzheimer's disease (AD) and lessen its burden on patients. AD is characterized by progressive cognitive loss. The planned study is carried out in two stages. Phase one addresses a major obstacle in early AD identification by introducing a state-of-the-art technique for imputing missing values in clinical datasets. Data integrity is maintained using the imputation process, which guarantees the preservation of statistical properties within each feature. Phase two involves restructuring the clinical data and applying the proposed imputation model to impute the missing values. The training of a classifier model that is intended to function with unique labels based on patient prognosis comes next. With an accuracy rate of 92%, the imputation model and classifier integration show a notable improvement in early AD identification. The results highlight the usefulness of neuropsychological evaluations as reliable markers for the early detection of Alzheimer's disease (AD), made possible by cutting-edge machine learning techniques. This research contributes to the field of artificial intelligence by presenting a robust imputation framework and its practical application in enhancing early diagnostic capabilities in biomedical engineering.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ML	Machine Learning
AD	Alzheimer's Disease
CN	Cognitive Normal
MCI	Mild Cognitive Impairment
MICE	Multivariate Imputation using Chained Equations
MLP	Multi-layer perceptron
DL	Deep Learning
MVA	Majority Voting Algorithm
NM	Neuropsychological Measures
ADNI	Alzheimer's Disease Neuroimaging Initiative
MI	Mutual Information
IG	Information Gain
BL	Baseline
M06	6^{th} month
SVM	Support Vector Machine

CHAPTER 1

INTRODUCTION

Alzheimer's disease is known as an extremely impairing neurodegenerative condition due to its relentless and persistent progression and intense intervention with daily life functioning. It affects millions of patients around the world specifically in the old age. AD causes memory loss, changes in everyday behavior and impaired thinking because of its little by little but progressive deterioration of brain structure and cognition function. There is no evident cure to this disease in medical field as of now and in order for effective management and timely intervention early detection is crucial [1].

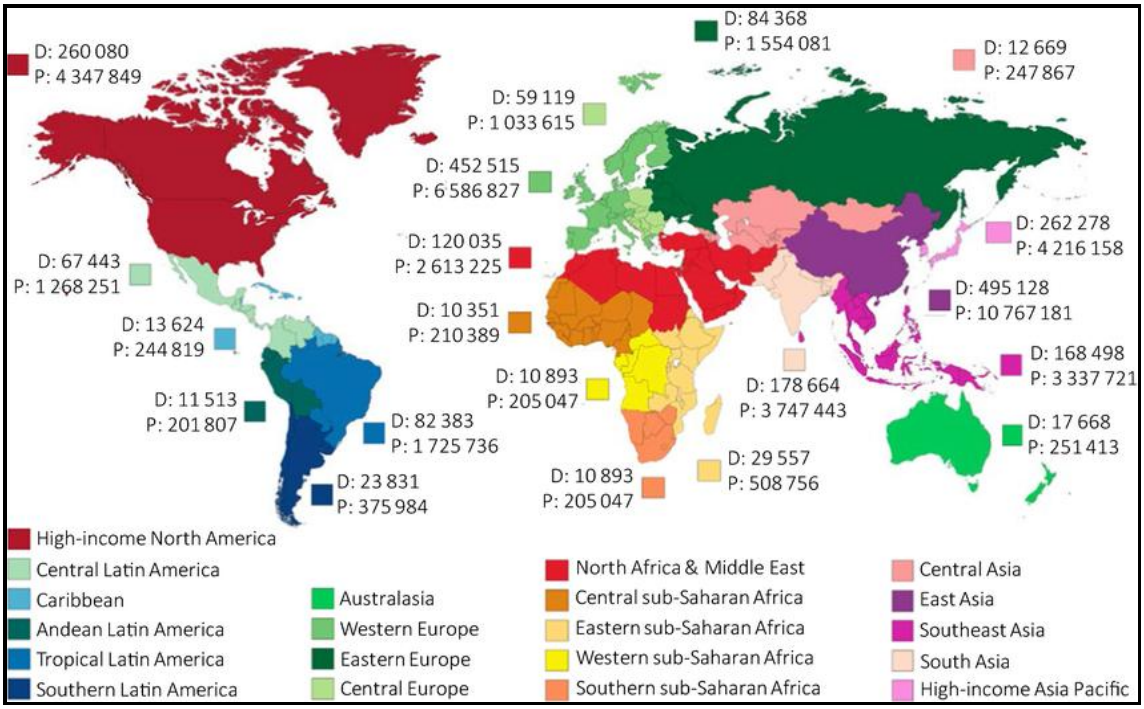

Figure 1. 1: Illustration of Mortality (D) and Incidence (P) globally for Alzheimer's Disease [1]

Advancement in medical science is very slow without the incorporation of artificial intelligence. Detection of certain disease is nearly impossible without the aid of AI. In recent years, using computed assistance and artificial intelligence (AI) for the enhanced identification of Alzheimer's disease in its initial stages has gained significant interest. A subset of AI called machine learning has shown promising results in medical field as it has the capability to extract features from data and predict accurately using the learnt features. Disease like cancer and dementia are very subtle and have no visible symptom before the disease reaches an advance stage. Using computer aided software for the purpose of early

detection of such diseases has proved its importance and has done wonders in medical science. This research focuses on the detection of Alzheimer's disease in its initial stages.

1.1 Understanding Dementia and its Underlying Causes:

Out of many diseases the world is currently trying to cure, Dementia is a very prominent disease and is a major cause of death in countries like United Kingdom, New Zealand and Bangladesh [2][16]. Neurodegeneration is a broader term which refers to the temporal loss of human brain's structure, functions and abilities. It is basically the progressive deterioration of brain cells with time. Dementia is an umbrella term that refers to the decay of mental functionality and this decay is serious enough to interfere into everyday life. In simpler terms dementia is a disease that affects thinking, memory and normal life performance also affecting the problem solving abilities of the patient. It is a severe cognition (mental actions) degrading disease and ultimately causes death of the patient. Dementia can be caused by a number of reasons including vascular dementia, age, frontotemporal disorders, lewy body dementia, Alzheimer Disease being most common accounting for 60-80% of the casses [3][4].

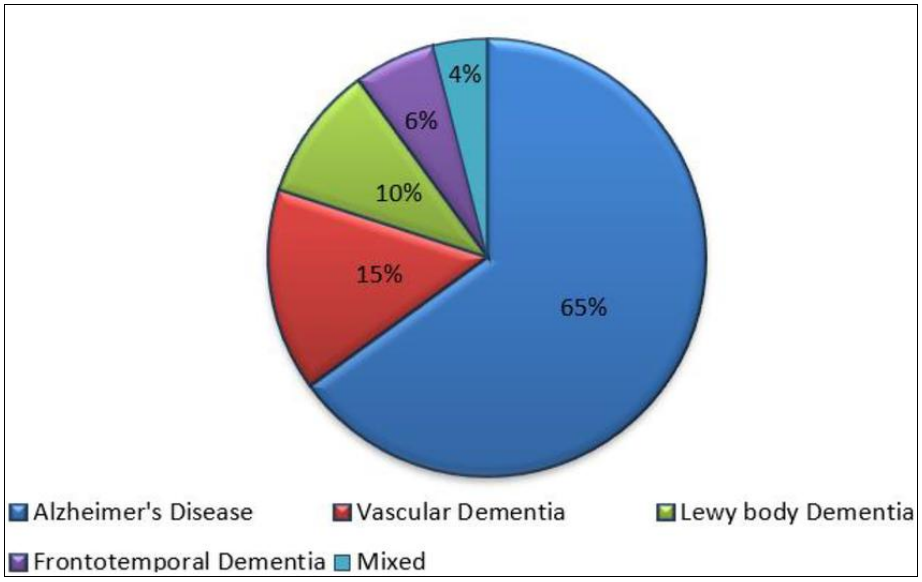

Figure 1. 2: Types of Dementia [4]

1.2 Alzheimer's disease:

Alzheimer's disease is a profoundly debilitating neurodegenerative condition and is a primary reason contributing to dementia. It mostly occurs in old age and it is an irreversible disease. In fact it is a progressive disease and worsens with time. It is a neurological disorder that

destroys brain neurons and tends to make the patient lose thinking abilities. Patients with AD start losing cognition and slowly become unable to perform simple everyday tasks which they were easily doing in routine. It is due to the neurofibrillary tangles that are caused due to excess accumulation of proteins inside brain thus making the brain incapable of performing functions that were simple to do earlier. It is usually categorized in two classes based on age, early on-set for patients diagnosed with AD before 65 and late on-set foe patients after 65 years of age. Late on-set is more common among AD patients [4][5].

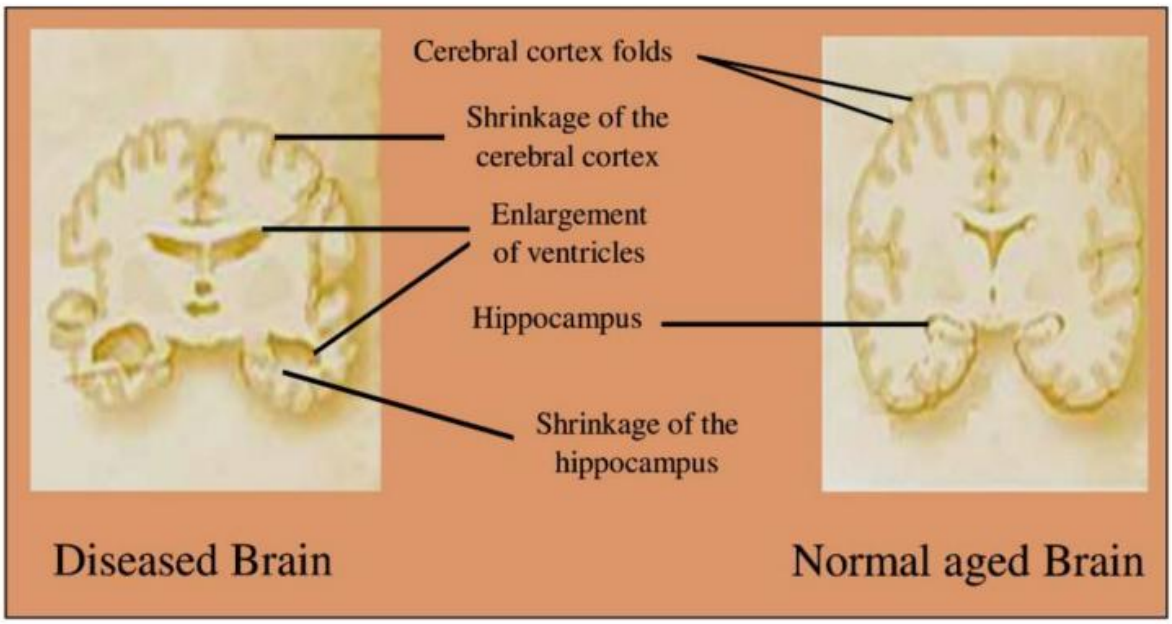

Figure 1. 3: Comparison of AD patient's brain with normal brain [4]

1.3 Why Alzheimer's Disease Occurs?

The exact reason that results in AD is not clear to medical science yet but some contributors that are significant in causing this disease are identified. The first and the foremost contributor towards AD is genetics. The presence of APOE e4 allele significantly increase risk of AD. In addition to APOE other mutations like APP, PSEN1, PSEN2 are also a risk, presence of which also indicate potential AD patient. Other than genetics the reasons that cause AD include accumulation of amyloid-Beta and Tau proteins which causes senile plague outside neurons. Tau proteins also tend to disintegrate the neuronal transport system causing neurofibrillary tangles. Other head injuries or cardiovascular diseases have also proven to be a contributor to this disease progression. Furthermore, lifestyle factors and environmental factors including diet, exercise and exposure to toxins may also add to risk of AD. Age is a major and the most significant risk factor towards AD [6].

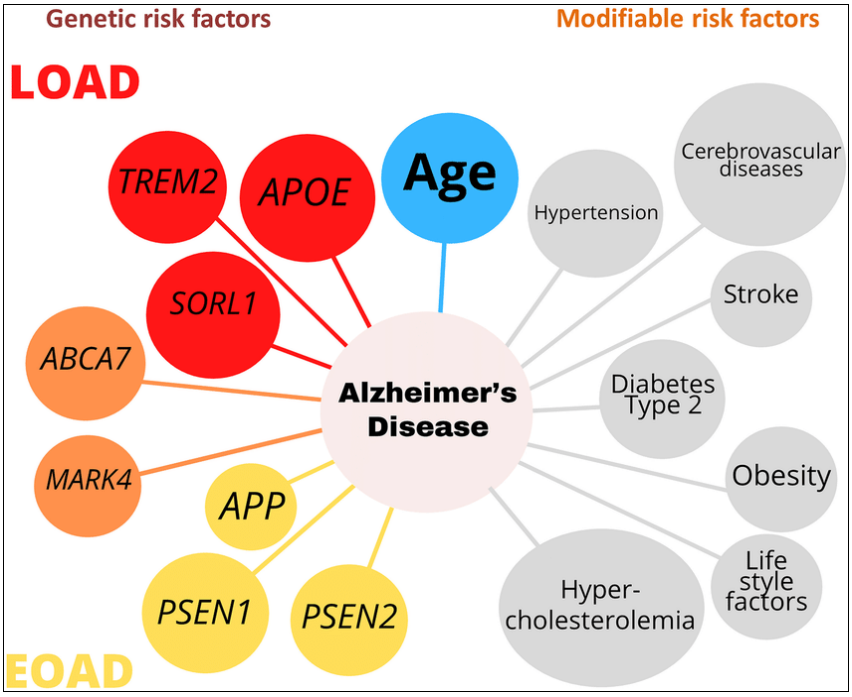

Figure 1. 4: AD early on-set (EOAD) and late on-set (LOAD) genetic and general risk factors [6]

1.4 Approached to Alzheimer's Disease Treatment:

Permanent cure regarding AD has not been developed yet, but there are several treatments that tend to temporarily slow down and delay the progression of AD to further advanced stages. Treatment of AD does not include medical operation because AD is an irreversible disease and it worsens with time and quality of life decreases with time. As time passes, the symptoms of AD are more visible and medical science has not yet developed a permanent solution to this issue. Two different types of treatments are being currently used; pharmacological treatments and non-pharmacological treatments.

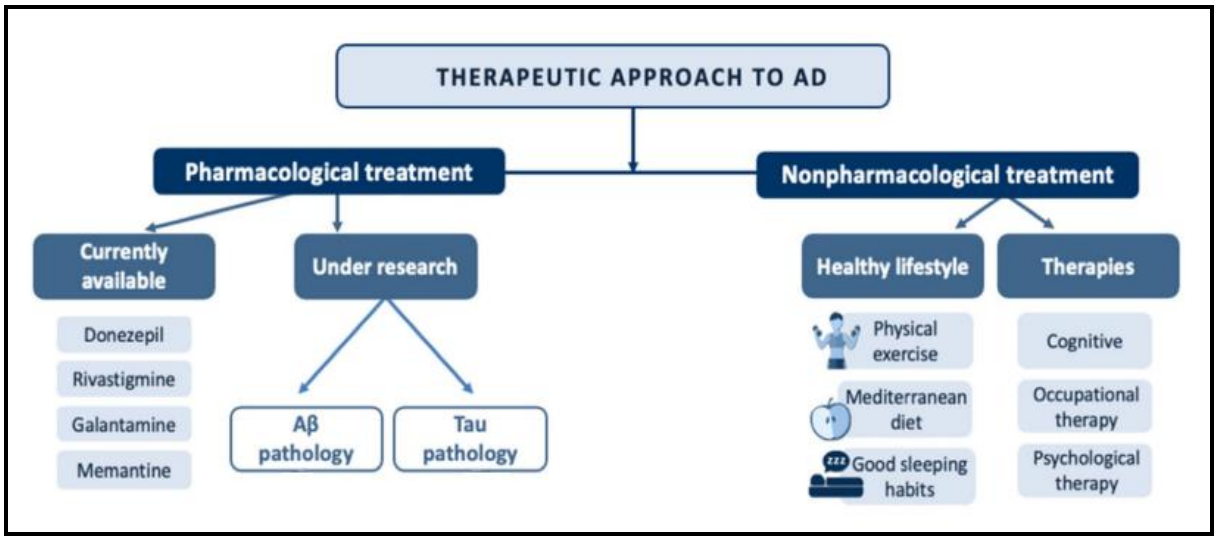

Figure 1. 5: Treatments of AD [7]

Pharmacological treatments are used in which drugs like donepezil and memantine are prescribed to the patients which regulate neurotransmitters to maintain cognition. Second type of treatments is non-pharmacological treatments which include cognitive therapy, dietary changes and physical exercises. If a patient is diagnosed at an early stage, care and proper attention can delay the progression to the next stage. Thus all these treatments are only helpful if a patient is identified successfully at an early stage [7].

1.5 Motivation:

This research's motivation stems from the critical need to enhance early diagnostic methods for Alzheimer's disease. Early identification is vital for effective management, as interventions are most beneficial during the early stages of AD. Traditional diagnosis approaches, including clinical assessments and neuroimaging, have limitations including high costs, limited accessibility, and the need for specialized expertise. By incorporating machine learning and artificial intelligence, this research is focused on developing an accurate method for early detection of Alzheimer's disease that is more efficient economically as well as practically.

Large data sets can be handled using machine learning algorithms, which also enhance classification accuracy by identifying minute trends and patterns. This attribute is extremely helpful in the context of the condition that is being discussed, namely Alzheimer's disease, because the disease is heterogeneous and can vary greatly in symptoms and course from person to person. To provide a thorough evaluation of illness risk, a strong machine learning model can integrate a variety of clinical data, such as demographics, medical histories, and cognitive test results.

1.6 Conventional Diagnosis of AD:

Conventional diagnosis of Alzheimer patients is made by regularly inspecting and monitoring decline of memory and cognition function at regular follow up intervals. A clinician manually keeps a track of a patient's history and tests performed on the patient and in this way the treatment of a patient is decided. Conventional diagnosis of AD may include cognition assessment tests, neuropsychological evaluations, clinical testing and medical images technology. Clinical testing involves all tests conducted on the patient by the clinician verbally or by giving some simple everyday tasks to the patient and evaluating by comparing

with performance of a normal person on those tests. These may include physical exams in order to test a patient's cognition level. Neuropsychological evaluation is testing the patients language, problem solving, question understanding, logic development skills. It further helps understand patient's cognition level. The last technique used in conventional diagnosis is use of modern image technology. Medical images like MRI scans and PET scans help visually see the brain and its structure and inspect the structural decline of brain cells. PET scans help visually see the accumulation of amyloid Beta or Tau proteins which help doctors better understand a patients current condition. CSF is also an expensive but useful technology used to detect the increasing ratios of this protein in the patient [8] [9].

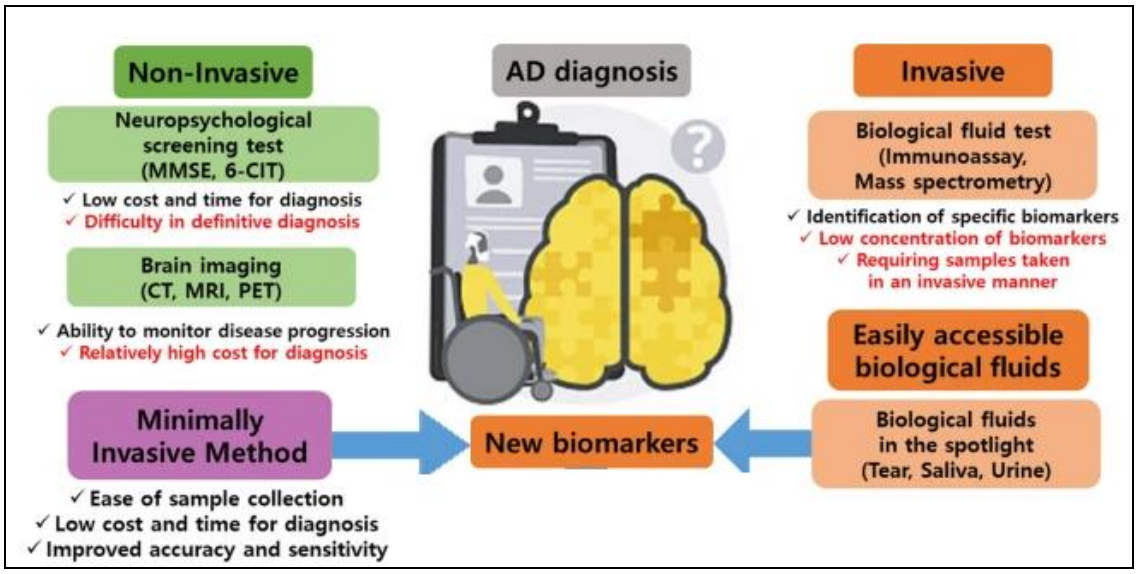

Figure 1. 6: Current and future AD diagnostic methods [9]

1.7 Progression of Alzheimer's Disease:

Alzheimer's disease is categorically divided in 3 stages. The initial stage is called Cognitive Normal (CN). At this stage the patient has normal cognition and has no complain of any impairment. No visible symptoms appear and the patient acts completely normal in his daily life as well as any clinical tests performed, but early accumulation of amyloid beta and tau protein may have started in this stage. Such patients are diagnoses using the image technology using MRI and PET scans. It is also known as pre-clinical stage.

Second stage is the Mild Cognitive Impairment. MCI is an intermittent stage and minor symptoms of memory loss and change in brain structure can be observed. MCI is further divided into 2 stages EMCI and LMCI which are early and late MCI stages. Some medical experts also consider a stage in between CN and MCI and that is called Subjective Cognitive Decline (SCD). If this stage is considered CN refers to complete normal patients, SCD refers

to initial stage with some accumulation of proteins and MCI is the intermittent stage. If patient is diagnosed at an early stage the prognosis is good.

Last stage is called AD. This is the final stage and notable symptoms can be observed, the patient himself as well as the people around him starts to feel decline in his mental abilities due to the obstructed language, decline in memory and loss of abilities to perform routine tasks. This stage is sometimes further divided into 3 sub stages called Mild Dementia due to AD, Moderate dementia due to AD and Severe dementia due to AD.

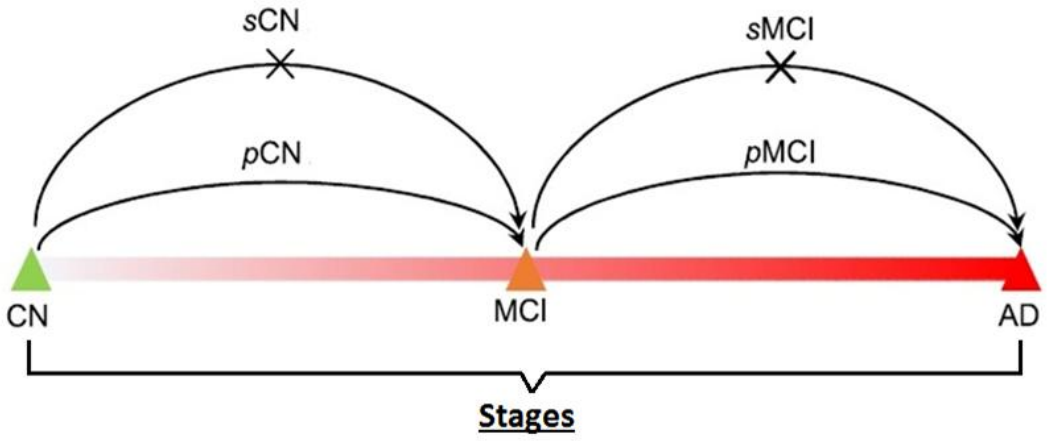

Figure 1. 7: Stages and Classification of AD [10]

1.8 Classification of Alzheimer's Disease:

Each stage of Alzheimer is further divided in 2 classifications i.e. Stable and Progressive. Any patient that was initially diagnosed at an early stage but during the follow-up visits progressed to an advanced stage is called progressive patient and similarly if a patient retains his diagnosis throughout the follow-ups he is classified as a stable patient. This type of classification is very essential if accurate early prediction of this disease is required. A patient progresses from one stage to the next he definitely follows some specific variation in his observed medical history. The main goal of dividing patients into further classification is to observe this variation. Upon successful understanding of this variation, a successful prediction can be made that a patient is a progressive patient if he follows the observed trend. [5].

1.9 Early Detection and Need of AI:

The idea of identifying Alzheimer's disease at initial stages is crucial for several reasons. Firstly it allows timely intervention resulting in slowing down disease progression. The only cure available for AD patients is to delay and slow down the advancement of the disease and hinder the progression to the next stage. Secondly it provides patients with opportunities to participate in clinical trials helping in research advancements. It also helps patient to plan his future and make informed decisions for his family at a stage he is able to make sane decisions. Conventional methods of detecting Alzheimer's disease have already been discussed but the limitations to such diagnosis include observer variance and environmental and genetic factors that add bias and pollute the diagnosis. In order to make patient's data reliable and independent of the variances and to assist the clinician making a better diagnosis out of the loads of data available computer assistance is required. In the conventional diagnosis the clinical himself manually keeps a track to patient's history and manages its future treatments but doing this for a number of patients with a huge history is humanly impossible. Incorporation of Artificial Intelligence in the process of diagnosing an Alzheimer patient has proved to be an efficient way of managing data. By using AI past patient's data can be used to predict a current patient's future by recognizing the trends, which is impossible for a human to do out of the gigantic amount of data available [11] [12]. This study is aimed at using AI models for identification of Alzheimer patients at an early stage.

1.10 Objectives of this Study:

This research is aimed to incorporate artificial intelligence into early detection of Alzheimer disease. This research is focused at developing a model which is capable of predicting a patient's future diagnosis using clinical data only. The aim is to develop a Clinical Decision Support System (CDSS) that assists the clinicians in early prediction of AD patient's diagnosis keeping it economical. Incorporation of other biomarkers like MRI scans and PET scans may include the efficiency of the system but also adds to the cost. The proposed model uses only clinical data for this prediction making it feasible for low-income countries like Pakistan where MRI scans or CSF extraction can be very expensive. Other than the economical drawback many newer patient hesitate going directly to invasive tests like genetic or CSF but are ready to take clinical assessments making the proposed model more useful.

Contributions of this paper include:

1. Using only clinical data to make efficient models for predicting AD without need of other biomarkers.
2. Development of robust ML models for accurate predictions based on classifications (stable/progressive) instead of stages (CN/MCI/AD).
3. Targeting CN to MCI progressive patients specifically to predict AD at an early stage.
4. Devising novel imputation techniques to resolve issue of missing data.

1.11 Thesis Organization:

Chapter 1: Introduction

This chapter provides the background, motivation, problem statement, objectives, scope, significance, thesis organization, and contributions to the field.

Chapter 2: Literature Review

Reviews existing literature on Alzheimer's disease, machine learning in medical diagnostics, and early detection methods. This chapter provides a comprehensive overview of previous studies, highlighting gaps in the literature and justifying the need for this research.

Chapter 3: Materials

Describes the research design, data collection methods used in the study. It also details the souring of data and the description of data used in this research.

Chapter 4: Methodology

This chapter details the implementation process, data imputation techniques, model training, and testing. This chapter provides a step-by-step guide to the development and validation of the machine learning and imputation models.

Chapter 5: Results and Discussion

This chapter discusses the results achieved by the research, including descriptive statistics, model performance, and comparative analysis. This chapter includes detailed tables and figures to illustrate the findings. This chapter also interprets the results and compares them with existing literature.

Chapter 6: Conclusion

Summarizes the findings, contributions, practical applications, and provides final remarks, discusses their implications, highlights limitations, and offers recommendations for future research.

Chapter 7: Future Work

This chapter lists all future work recommendations.

Chapter 8: References

This chapter lists all academic papers, books, and other resources cited in the thesis. This section follows the appropriate citation style as per the guidelines.

Figure 1. 8: Flow of the Thesis

CHAPTER 2

LITERATURE REVIEW

Researchers are trying to devise new techniques and methods that prove to be effective and are accurate for early prediction of AD. Early prediction means that researchers are striving to develop CDS systems that are capable of predicting a patients future diagnosis based on the history follow ups of that patient. If a patient is successfully diagnosed as CN at this point in time but may progress to MCI in future, the progression can be delayed and informed decisions can be made.

2.1 Early Detection in Literature:

Primarily early identification of Alzheimer's disease rely on clinical evaluation tests, neuroimages including Magnetic Resonance Imaging (MRI) and PET scans and other biomarker assessment like genetic data and CSF extraction. In literature many different combinations of such analysis is proposed. Clinical evaluations include cognitive tests like Alzheimer's Disease Assessment Score (ADAS), Mini-Mental State Examination (MMSE), FAQ, MoCA etc. MRI and PET scans are useful for preclinical stage diagnosis as it is helpful to detect accumulation of amyloid Beta and tau protein even before any symptoms start to appears. In literature, many researchers have also incorporated artificial intelligence to achieve accuracy. Many researchers have used machine learning models like SVM and Random forest for classification purposes. Some also have used multiple ML algorithms and performed an ensemble upon them to achieve results free from bias. Deep learning has also been used in literature but is application is very limited. Currently researchers are using multiple predictors/biomarkers to train ML models. M Ibrahim et al. [13] used Genetic data in combination with clinical data to perform binary classification between CN and MCI/AD patients. A Gamal et al. [17] used MRI images and devised a 3 class classification model for CN, MCI and AD. In 2021, Sidra M et al. [19] used a combination of clinical and MRI data to target MCIs and MCIp patients. The aim of all these researches was early detection of AD. It is observed in all these researches that all authors used multi-biomarker models to develop models for early prediction.

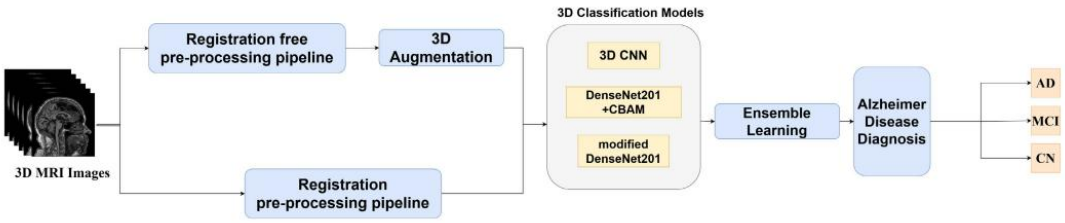

Figure 2. 1: Uni-model Classifier proposed by [17]

2.2 Past Researches on Uni-Biomarker Model:

Most of the researches conducted in the past are focused on multi model classification. Most authors use clinical data in combination of some other biomarker like MRI scans or genetic data to develop a model and achieve good accuracy. In doing so good performance models may be achieved but the dependency on other biomarkers which are very expensive increases. Maintaining follow ups of such biomarkers after every 6 months is very expensive and even impossible in some countries due to low economic structure and unavailability of proper resources. To solve this issue models that are based on only one biomarker are required. Uni-biomarker models are the need of the hour in order to avoid the barrier of economy and to implement proper medical facilitation. In literature many uni-biomarker models have been proposed. A Gamal et al. [17] used only MRI as a predictor to build a deep learning model. M Kim et al. [21] also used a uni-biomarker model using MRI scans only and developed a classifier for MCIs and MCIp patients. Sidra M et al. [15] used only clinical data as a predictor to develop a system for binary classification of MCIs, MCIp and C. Kavitha et al. [18] also used a uni-biomarker model utilizing merely clinical data to perform early prediction of MCI and AD patients in 2022.

2.3 AI for Early Detection:

Using artificial intelligence to assist clinicians by developing models that would help in early diagnosis of Alzheimer patients has shown promising results. Deep learning and machine learning has been applied to the complex data of Alzheimer patients whether it is MRI or PET scans or the clinical data. In applying ML models to Alzheimer patient data, the researchers have tried to analyze the complex trends in the data and devise models that would predict a patients future prognosis based on those trends. Application of deep learning models on

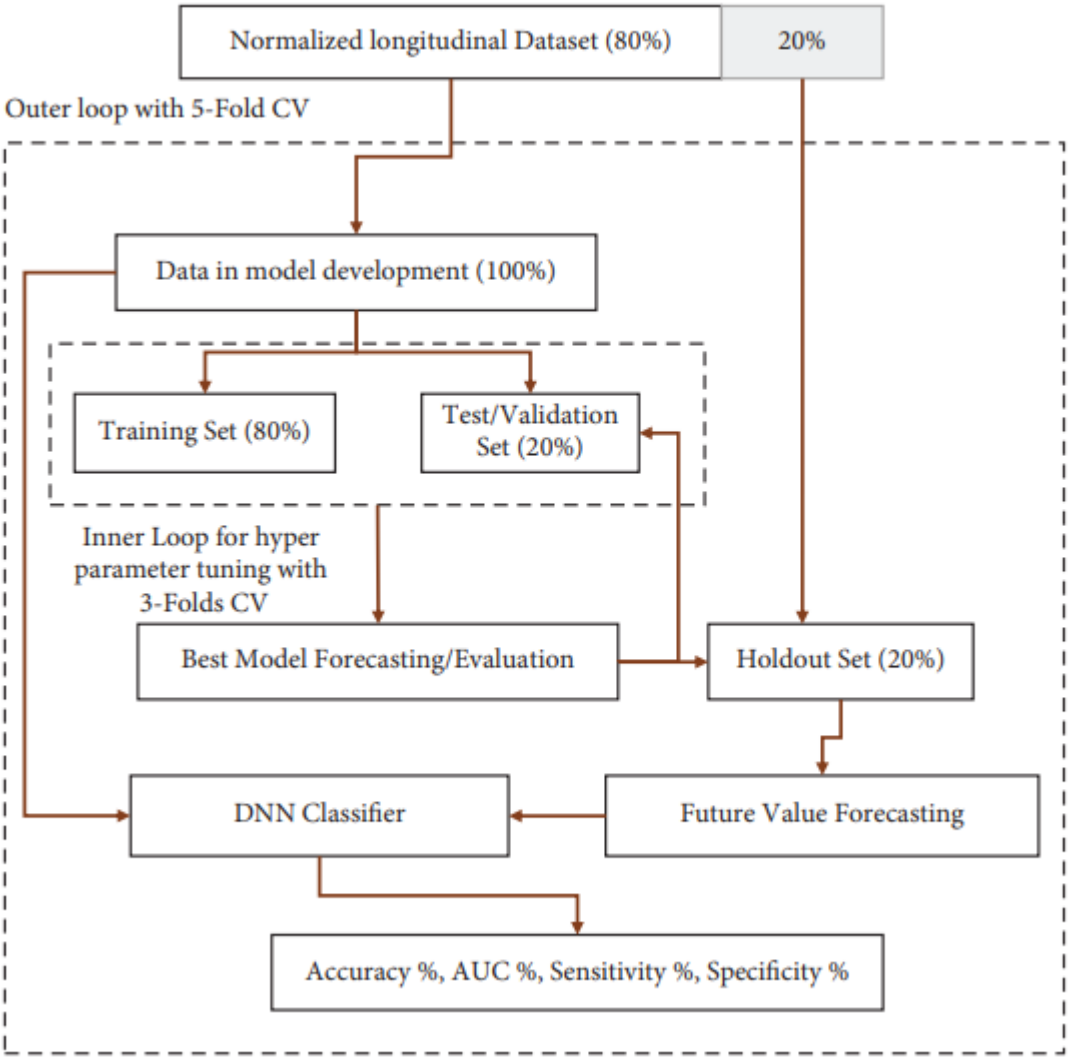

Figure 2. 2: Uni model DL classifier proposed by [15]

Alzheimer data has proved to be promising. Deep learning models are more feasible in models where neuroimages are used as predictors. Utilization of artificial intelligence for early prediction of AD is a crucial step as processing such plenty of data is impossible without computer assistance. [13] trained multiple ML models and later compared the results of all models, [17] also used modern techniques; they used DL models and performed an ensemble on them. [18] And [19] also used ML models to preform classification and [15] used DL models to achieve a good accuracy. A comprehensive summary of the researches conducted are in Table 2.1.

Table 2. 1: Researches conducted on different biomarkers using different AI techniques

Ref.	Method	Feature Ranking	Predictor	Missing Value	Validation	Accuracy %
[13]	SVM, RF, AB, MLP (results compared)	Chi-Square, Mutual Information, LASSO	NM, GENE	Multivariable Imputation by Chained Eq. (MICE)	5-Fold CV + 3-Fold CV	94
[14]	Ensemble on ML Algos (SVM, KNN, Random Forests, AdaBoost)	Logistic Regression and AUROC	NM+MRI	-	10-Fold CV	92
[15]	DNN Classifier	Students T-Test	NM	DNN Regressor	5-Fold CV	87
[17]	Ensemble in 3 DL algorithms	(ROI) not clearly defined	MRI	-	5-Fold CV	70.33
[18]	Random Forests	Information Gain	NM	Median	10-Fold CV	86.92
[19]	SVM	Students T-Test	NM + MRI	Euclidean Geometry	5-Fold CV	81
[20]	TS-SVM	L2,1 Norm	MRI	-	10-Fold CV	76.53
[21]	SVM	Students T-Test	NM + MRI	AR Parameter using LOOCV	5-Fold CV	84.29
	SVM	Students T-Test	NM	AR Parameter using LOOCV	5-Fold CV	83.26

2.4 Research Gaps:

The researches discussed have achieved good results but have some limitations. Firstly combining multiple biomarkers for devising a system definitely gives good results but adds to the overall cost of maintaining such a system. Multi-biomarker models are very precise but in order to incorporate those models in clinical settings, the data for each of those biomarkers is required. Provision of such biomarker data for every patient and for every follow up is not possible. Unavailability of data due to financial issue or missing equipment will result making the model useless. Thus a model is required that is based on one biomarker and could compete with these multi-biomarker models in terms of performance. That uni-biomarker should be clinical/neuropsychological measures preferably. The second limitation is that the classifications made on bases of stage labels like CN, MCI and AD. This kind of classification can help better identify the current status of a patient but cannot be used for

early predictions. For early prediction classification based on labels of progressor and stable patients is required. Currently the main area focused by researchers is the development of a model that predicts a patient's future diagnosis and keeping this model equally feasible for low income countries and that remains a gap.

2.5 Problem Statement:

Irrespective of the in depth understanding of the disease and its reasons, the problem of early diagnosis is still a challenge due to the multiple dimensions and complex nature of the disease. Conventional methods of diagnosis including clinical assessments, neuropsychological tests and neuroimaging evaluation lack the accuracy and precision required for early and accurate diagnosis. This delays the intervention and risk the patient to advance stages of the disease. Moreover the economical drawback of using multiple biomarkers is also not ignorable. This research focuses on devising a system based on AI that is capable of predicting a patient's diagnosis before time and label the patient as a progressor or stable patient using only clinical data and that will help clinicians make a more accurate and precise prognosis for the specified patient.

CHAPTER 3

MATERIALS

3.1 Data Availability:

In the process of performing a classification action or developing a Clinical Decision Support System (CDSS), the first and foremost requirement is the availability of data. Moreover, reliable and extensive data is a pre-requisite for any kind of Artificial Intelligence based system. With reference to the problem under discussion, availability of data related to Alzheimer's disease patients is a challenge. Large scale AD datasets are required, however the organizations dedicated at collecting and curating AD data are very limited. Out of some Alzheimer's disease data initiatives, this research incorporates datasets provided by ADNI. Few organizations that are conducting studies regarding Alzheimer data are:

1. National Alzheimer's Coordinating Center (NACC)
2. Alzheimer's Disease Data Initiative (ADDI)
3. Open Access Series of Imaging Studies (OASIS)
4. Alzheimer's Disease Neuroimaging Initiative (ADNI)

All of the mentioned organizations are collecting Alzheimer patient's data and are refining and making it available to the researchers for further advancement in the field. This research is based on ADNI data because of two major reasons:

1. ADNI is the only ongoing study that has been in operation since 2004 and has most extensive data.
2. ADNI is the only open source data available, other platforms provide data to specific authorized entities.

3.2 Alzheimer's Disease Neuroimaging Initiative (ADNI):

ADNI is an ongoing multisite (at more than one location) longitudinal (continuous or repeated measures of a particular patient over a prolong period) study. It has 55 sites in UK

and Canada. ADNI is the only open source database available for researchers. It is working since 2004 and has a large number of patients and extensive data. An additional advantage of using ADNI data is that most of the patients' enrolled with ADNI are returning patients and longitudinal data of such patients is recorded. Using longitudinal data is very essential for recognizing the trends a patient follows in progressing to advanced stages of AD. ADNI performs multiple assessment tests on patients at every 6 month interval. ADNI has divided its study in different phases. Figure 2.3 shows the different phases in which data collection is conducted.

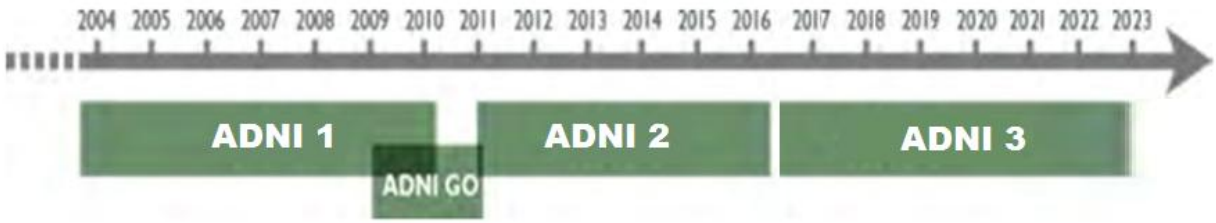

Figure 3. 1: ADNI phases

ADNI 4 is ongoing study and is currently in process. ADNI provides 5 different types of biomarkers that are:

1. Neuropsychological Measures (NM) / Clinical Data
2. MRI Data
3. PET Scans
4. Cerebrospinal Fluid (CSF) data
5. Genetic Data

These 5 different biomarkers provided by ADNI are basically the different tests and examinations conducted by ADNI on the patients. Each patient once enrolled in ADNI undergoes different assessments to record these 5 biomarker values. ADNI also labels the patient as at what stage the patient is diagnosed at the respective follow up. The initial assessment is labeled as baseline and the same evaluation is made after every 6 months. Patients follow ups may continue when ADNI announces a new phase of study. In a new phase, new patient are enrolled but old patient also fade in to the new phase.

3.3 Using Clinical Data:

The biomarkers provided by ADNI are of different natures. Some are non-invasive and completely verbal while some require extreme invasion like lumber puncture. MRI or PET scans require exposing the body to radio or gamma waves. This research uses only clinical data because of the following reasons:

1. Other biomarkers are expensive; maintaining follow-up for other biomarkers after every 6 months is not economical.
2. New patients (which require early detection) prefer clinical tests over MRI or PET scans.
3. It is the cheapest source of data for AD.
4. Clinical Data has shown promising results towards early detection.
5. Efforts are made to develop a system that predicts future diagnosis of a new patient with an economical advantage and practical usage.

By clinical data ADNI means the non-invasive tests and assessments a clinician makes by asking questions from the patient or by giving some everyday tasks to the patient. The clinical tests ADNI has included in its database include Alzheimer‟s Disease Assessment Score (ADAS), Mini-mental State Examination (MMSE), RAVLT, CDRSB, FAQ, ECog etc. In total ADNI have 30 clinical tests. These different tests are regarded as features when dealing with them in AI‟s perspective. ADNI performs these tests on each patient at each follow-up.

3.4 Understanding ADNI Data Structure:

In the ADNI database each patient is given a unique ID regarded as ‘PTID’. Follow-ups conducted every 6 month is referred to as ‘VISCODE’ which are ‘bl’ for baseline, ‘m06’ for 6th month, ‘m12’ for 12th month and so on. The database is arranged in a way that 1 row represents scores of a patient’s specific follow-up, and the last column represents the diagnosis (CN, MCI, AD) of the patient at that follow-up. ADNI’s clinical data is arranged in a .CSV file named ‘ADNIMERGE’. Within ADNIMERGE all the data available for a patient is merged. This file has scores for all clinical assessments along with the information extracted from PET scans, the numerical representation of important MRI features, genetic

features and CSF information. This file serves as a complete merged document for all information recorded for a patient.

3.5 Dataset Description:

The dataset used in this research was ADNIMERGE package. This was downloaded on 18th - March-2024 from official ADNI database website www.ida.loni.usc.edu. In order to download data from ADNI, special request is to be made to ADNI after registering to the database. ADNI analyses the request and the purpose of data requirement and approves or disapproves on that basis.

The downloaded data had the following specifications:

1. 30 clinical features
2. 10 MRI features
3. 6 PET features
4. 1 CSF feature
5. 1 Gene feature

In total, ADNIMERGE had a total of 48 features. ADNI Data for a total of 2430 patients was available. Figure 2.4 shows the visual representation of data by ADNI where black box represents the unique patient IDs as PTID, red box represents the follow up visits as VISCODE and blue box represents the features or the biomarker/tests.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	PTID	VISCODE	APOE4	FDG	PIB	AV45	FBB	ABETA	TAU	PTAU	CDRSB	ADAS11	ADAS13	ADASQ4	MMSE	RAVLT_im	RAVLT_le	RAVLT_fo
2	002_S_0295	bl	1					888.1	355.2	34.73	0	3	4	1	28	56	5	
3	002_S_0295	m06	1								0	5.33	6.33	1	28	50	7	
4	002_S_0295	m12	1					858.3	399.5	39.29	0	4.67	5.67	1	30	53	10	
5	002_S_0295	m18	1															
6	002_S_0295	m24	1								0	3.67	5.67	2	29	45	7	
7	002_S_0295	m30	1															
8	002_S_0295	m36	1					708.6	381.3	38.72	0	3.67	6.67	3	28	43	5	
9	002_S_0295	m42	1															
10	002_S_0295	m48	1					813.3	396.3	41.35	0	5	8	3	26	51	6	
11	002_S_0295	m54	1															
12	002_S_0295	m60	1	1.28568		1.4881		621.9	390.4	42.87	0	6	9	3	28	30	4	
13	002_S_0295	m66	1															
14	002_S_0295	m72	1								0	7	9	2	22	41	5	
15	002_S_0413	bl	0					1006	107.3	10.57	0	3.33	4.33	1	29	52	6	
16	002_S_0413	m06	0								0	7.33	8.33	1	29	47	8	
17	002_S_0413	m12	0					934	116.6	10.6	0	2.33	4.33	2	29	55	5	
18	002_S_0413	m18	0															
19	002_S_0413	m24	0								0	2.33	5.33	3	30	52	7	
20	002_S_0413	m30	0															

Figure 3. 2: ADNIMERGE PTID in black, VISCODE in red, Features in Blue

CHAPTER 4

METHODOLOGY

The study uses a two-phase process with the goal of increasing the precision of early Alzheimer's disease diagnosis.

Phase 1:

- Novel Imputation Method:

In this step, a sophisticated imputation method created especially for managing missing values in the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset is introduced. Specifically designed to maintain the statistical characteristics of every feature in the dataset, this method is an improvement on the Multiple Imputation by Chained Equations (MICE) methodology..

- Three-Class Classifier:

Following imputation, a classifier is trained on the data to categories patients into three groups according to ADNI labels: Cognitive Normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD).

Phase Two:

- Data Restructuring in 3D Space:

In order to better capture the intricate relationships that exist between features over time, the ADNI dataset has been reorganized into a three-dimensional space. This entails transferring patient data into a three-dimensional coordinate system, where each dimension corresponds to a different aspect of the disease's development.

- Custom Labeling Based on Prognosis:

Based on their prognosis, patients are given personalized labels that indicate whether they are expected to stabilize or advance to the next stage of Alzheimer's disease. This method of tagging is more precise and customized for each patient's path.

- Imputation of Follow-Up Data:

This reorganized 3D dataset's missing follow-up data is imputed using the unique imputation model created in the first step. By taking this step, the dataset's comprehensiveness and statistical robustness are maintained.

- Four-Class Custom Label Classifier:

Based on their specific labels, patients are then classified into four classes using a more advanced classifier. This classifier provides a more thorough and precise prediction of disease development by accounting for each patient's unique progression trends. Each phase's specific performance is broken down as follows:

4.1 Phase 1:

At first, dataset exhibited a lot of missing values; many patients had incomplete information for several follow-up visits. Missed appointments, patient resistance to particular testing or patient death are some of the possible causes of this problem. Cleaning the data to produce a trustworthy and objective dataset was a crucial first step before beginning any computational study. In order to create an error-free system that can manage unknown data efficiently, this phase is essential.

4.1.1 Data Cleaning:

4.1.1.1 Feature Filtering:

Features with insufficient data availability were identified and excluded. For instance, features related to PET imaging such as FBB (Florbetaben) and PIB (Pittsburgh Compound B) had less than 4% of data available and were thus removed from the dataset. In total, five such features were filtered out.

4.1.1.2 Non-Longitudinal Data:

Patients who only had baseline data without any follow-up information were excluded, as non-longitudinal data is not useful for early prediction of Alzheimer's disease progression.

4.1.1.3 Unlabeled Diagnoses:

Patients who lacked a labeled diagnosis at all follow-ups were also removed since their actual diagnosed stage was unknown.

4.1.1.4 Corrupted Data:

Patients with corrupted data, indicated by inconsistent labeling such as being diagnosed with AD (Alzheimer’s disease) initially, labeled as CN (Cognitively Normal) in subsequent visits, and then again labeled as AD, were excluded to maintain data integrity.

After the data cleaning process, 43 features were finalized for further analysis, which included:

Clinical Features: 30 features encompassing various clinical assessments.

MRI Features: 7 features derived from MRI scans.

PET Features: 4 features obtained from PET imaging.

CSF Feature: 1 feature from cerebrospinal fluid analysis.

Genetic Feature: 1 feature related to genetic information.

4.1.2 Pre-processing:

After data cleaning, the primary challenge encountered was addressing the issue of missing data, as illustrated in Table 2, which details the extent of missing data per feature. Prior to training any machine learning models, it is crucial to effectively handle these missing values. Various techniques employed by researchers to address this issue include:

1. **Mean/Median Imputation:** This technique uses the column's mean or median to fill in any missing values.
2. **Last Observation Carried Forward (LOCF):** Reusing the most recent data point available to impute missing values.
3. **Next Observation Carried Backward (NOCB):** Utilizing the subsequent accessible data point to impute missing values.

4. **Row Deletion:** Delete all rows that have missing values.

Table 4. 1: Missing Data Percentages

Features	**Missing Data Percentage**
CDRSB	28.46182
ADAS11	30.38607
ADAS13	31.0011
ADASQ4	30.16685
MMSE	30.15467
RAVLT_immediate	30.90367
RAVLT_learning	30.89758
RAVLT_forgetting	31.09244
RAVLT_perc_forgetting	31.53696
LDELTOTAL	42.49787
DIGITSCOR	76.85422
TRABSCOR	32.95579
FAQ	28.48618
MOCA	54.77408
EcogPtMem	52.6428
EcogPtLang	52.74632
EcogPtVisspat	53.25174
EcogPtPlan	52.86201
EcogPtOrgan	53.72671
EcogPtDivatt	53.10559
EcogPtTotal	52.70369
EcogSPMem	52.38095
EcogSPLang	52.34442
EcogSPVisspat	53.47704
EcogSPPlan	52.89246
EcogSPOrgan	54.39654
EcogSPDivatt	53.61101
EcogSPTotal	52.39313

Although these techniques are widely used, they are not without limits. The dataset is drastically reduced when all missing data rows are removed, leaving inadequate information for model training. The performance of the model may be impacted by bias introduced by imputation using the mean or median. When there are a few missing follow-up visits, LOCF and NOCB work well; however, when data is absent for several consecutive visits, they fall short.

The Multiple Imputation by Chained Equations (MICE) technique is a more effective method for imputing missing values. MICE is useful for imputing missing values in datasets where numerous variables have associated missing values, even if it is rarely employed in Alzheimer's disease (AD) research. MICE is especially useful for complex, multi-variable datasets since it repeatedly imputes one variable using the others.

This study proposes a new imputation model based on the MICE algorithm. The purpose of this model is to impute missing values properly while maintaining the distribution and statistical integrity of each feature. This method guarantees the imputed dataset's objectivity and stability, offering a strong basis for later machine learning model training and analysis.

4.1.3 Proposed Imputation Method:

As shown in Figure 4.1, the suggested strategy for imputing missing values combines a number of procedures in a certain order. This multi-step procedure ensures a more accurate and efficient handling of missing data by addressing possible problems that can occur throughout the imputation process. Direct application of MICE to the Alzheimer's dataset is not always successful since there are times when complete feature data is missing, which causes problems with the MICE algorithm's convergence. A preliminary imputation phase is added in order to lessen this issue. Pre-filling the data with a variety of methods fills in any gaps that can impede the MICE process. The method makes sure that any problematic data is handled effectively by putting these preparatory techniques into practice before using MICE. This makes the data more stable and dependable for the MICE algorithm. By combining the benefits of several imputation techniques, this hybrid approach provides a strong and all-encompassing way to handle missing values in Alzheimer's disease data.

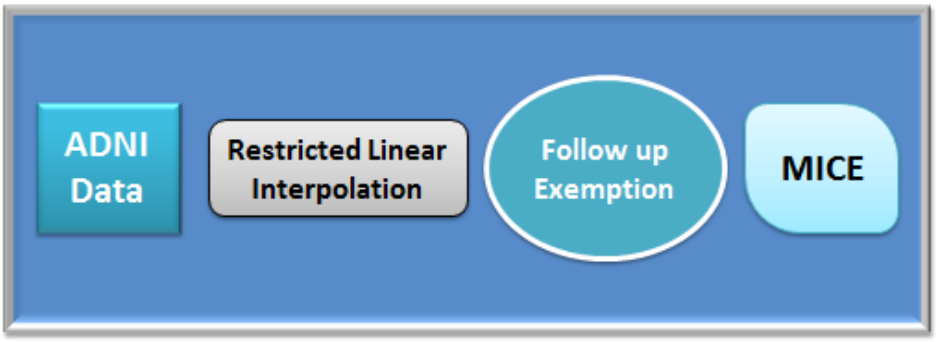

Figure 4. 1:Proposed Imputation Model

The following are the specific steps of the suggested imputation method:

4.1.3.1 Restricted Linear Interpolation

The technique of restricted linear interpolation involves averaging the values of adjacent cells (i.e., the preceding and subsequent follow-up visits) in order to fill in any missing data. To ensure optimal precision, this operation is limited to cells that have both neighboring values available. Although linear interpolation is well-known for its ability to calculate missing values precisely and accurately, it is used in this case in a limited way to increase precision. In order to prevent biases from other features interfering with the interpolation, this step is carried out column-wise and feature-wise.

4.1.3.2 Follow-Up Exemption

Rows having a significant amount of missing data are handled in the following step. A threshold of 20 missing values was established, out of the total 43 features that were accessible. Rows that exceeded this cutoff were not included because MICE finds it inefficient to handle follow-ups that have more than 20 missing features. Since MICE uses the available features to predict missing values, this step makes sure that follow-ups with significantly missing data do not compromise the imputation process.

4.1.3.3 Multivariate Imputation by Chained Equations (MICE)

The MICE technique uses other properties in a dataset to forecast missing values by utilizing machine learning models. The following are the steps in MICE:

1. **Initialization of Missing Values:** The initialization of missing values.

2. **Column Replacement:** Inserting the original missing cells into each column one at a time, starting with column A.
3. **Model Training:** Using column A as target variable using additional variables to train a machine learning model.
4. **Prediction:** Making predictions about the absent values in column A using the learned model.
5. **Iteration:** Update the dataset with the predicted values by going through steps 2 through 4 again for each column.
6. **Convergence:** Finishing the initial iteration and going on to complete other iterations until convergence.

For imputation, all 43 features were utilized, as including more features helps the machine learning model better identify data trends and make more accurate predictions. LOCF and NOCB were used as initialization tools, with LOCF filling missing cells with the last available value and NOCB filling them with the next available value if no prior value exists. A linear regression model was implemented for training and predicting missing values, with 10-fold cross-validation incorporated to avoid over fitting and achieve optimal results. MICE was set for 10 iterations.

An example is provided to illustrate the working of MICE, highlighting its iterative process and the incremental improvement in imputation accuracy with each iteration.

4.1.3.4 MICE Example:

Consider the dataset in Table 4.2 as an example for analyzing the implementation of MICE. The dataset contains three distinct variables: A, B, and C. To utilize MICE for missing value imputation, the dataset must satisfy certain criteria:

1. **Multiple Variables**: The dataset must include more than one variable or feature.
2. **Pervasive Missing Data**: Missing data should be present across all or most variables.
3. **Interrelated Variables**: The variables must exhibit interrelationships.

The dataset in question meets these requirements, containing three variables (A, B, and C) with nine independent data points each. Missing data is present within each variable: variable A has missing data at serial numbers 1, 4, and 7; variable B at 2, 5, and 8; and variable C at 3, 6, and 9.

Table 4. 2: Example Dataset for MICE

Sr.	A	B	C
1	N/A	1	0.1
2	20	N/A	0.2
3	30	3	N/A
4	N/A	4	0.4
5	55	N/A	0.5
6	80	8	N/A
7	N/A	8	0.8
8	85	N/A	0.8
9	90	9	N/A

4.1.3.4.1 Step 1: Initialization

The initial step involves filling missing values using LOCF/NOCB. Table 4.3 illustrates the initialized missing values, which are highlighted.

Table 4. 3: Missing values initialised

A	B	C
20	1	0.1
20	**1**	0.2
30	3	**0.2**
30	4	0.4
55	**4**	0.5
80	8	**0.5**
80	8	0.8
85	**8**	0.8
90	9	**0.8**

4.1.3.4.2 Step 2: Replacing Target Column

The target column, the column being imputed, is replaced with its original values containing missing data points. Table 4.4 shows column A replaced with the original data, where other columns have no missing values.

Table 4. 4: Replacing target column with original column

A	B	C
N/A	1	0.1
20	**1**	0.2
30	3	**0.2**
N/A	4	0.4
55	**4**	0.5
80	8	**0.5**
N/A	8	0.8
85	**8**	0.8
90	9	**0.8**

4.1.3.4.3 Step 3: Preparing training dataset

In this step, the data is split into training and test datasets. Rows with missing values in the target column are separated. Table 4.5 shows the training and test datasets.

Table 4. 5: Dataset split for training

A	B	C
20	**1**	0.2
30	3	**0.2**
55	**4**	0.5
80	8	**0.5**
85	**8**	0.8
90	9	**0.8**

A	B	C
N/A	1	0.1
N/A	4	0.4
N/A	8	0.8

4.1.3.4.4 Step 4: Machine Learning Model Training:

The complete dataset (left side of Table 4.5) is used to train a machine learning model, with A as the target variable and B and C as predictors. The trained model is then used to predict missing values of A in the test dataset (right side of Table 4.5). In Table 4.6, x, y, and z represent the imputed values for A.

Table 4. 6: Imputed values for feature A

A	B	C
X	1	0.1
20	**1**	0.2
30	3	**0.2**
Y	4	0.4
55	**4**	0.5
80	8	**0.5**
Z	8	0.8
85	**8**	0.8
90	9	**0.8**

4.1.3.4.5 Step 5: Repeat for next column

This process is repeated for all other features.

Table 4. 7: Dataset prepared for feature B imputation

A	B	C
X	1	0.1
20	N/A	0.2
30	3	**0.2**
Y	4	0.4
55	N/A	0.5
80	8	**0.5**
Z	8	0.8
85	N/A	0.8
90	9	**0.8**

Table 4.7 shows the dataset prepared for imputing feature B, where A now contains predicted values. By following the same steps, the values for B can be predicted. Table 4.8 represent the training dataset for feature B and in table 4.9, J, K and l represent the imputed values for B.

Table 4. 8: Training data for feature B imputation

A	B	C
X	1	0.1
30	3	**0.2**
Y	4	0.4
80	8	**0.5**
Z	8	0.8
90	9	**0.8**

Table 4. 9: Imputed Values for B

A	B	C
20	**J**	0.2
55	**K**	0.5
85	**L**	0.8

Repeating same steps leads to imputation of C.

4.1.3.4.6 Step 6: Repeat for n Iterations:

This iterative process is repeated for multiple iterations until the MICE algorithm converges. Each iteration involves reimputing all columns, refining the imputed values. Convergence is achieved when the imputed values stabilize, with minimal differences observed between iterations. A predefined threshold for acceptable differences determines the stopping criterion for the iterations.

4.1.4 Proposed Model:

The proposed model for Phase 1, depicted in Figure 4.2, begins with the utilization of the imputed dataset. The initial step involves ranking the 30 clinical features based on their relevance and contribution towards AD diagnosis. This ranking process employs four distinct techniques to ensure a comprehensive assessment of feature importance:

1. **Mutual Information**: This technique measures the mutual dependence between two variables. It captures the amount of information obtained about one variable through the other, thus identifying features that have the highest dependency on the target variable.

2. **LASSO (Least Absolute Shrinkage and Selection Operator)**: This regression analysis technique improves the statistical model's interpretability and prediction accuracy by performing regularization and variable selection. By putting a limit on the regression coefficients' absolute sizes, it effectively reduces some of them to zero and chooses a subset of the features.

3. **Chi-Square Test**: This statistical test evaluates how independent category variables are from one another. Finding features that significantly impact the outcome and evaluating if there is a substantial correlation between the feature and the target variable are two areas in which it is especially helpful.

4. **Student's T-Test**: To ascertain whether two groups' means differ noticeably from one another, this test contrasts their means. It aids in locating characteristics that exhibit a notable variation in means among various target variable classes.

Following the application of these ranking procedures, the top 10 features were chosen in accordance with their overall cumulative importance across all techniques. This selection makes sure that the most useful and instructive features are kept for model training, which improves the prediction model's effectiveness and precision.

The dataset was divided into training and testing subsets in a 70/30 ratio prior to starting the model training process. By dividing the data in this way, the model is tested on hypothetical data, resulting in an objective assessment of its performance. K-fold cross-validation was then used to further partition the training subset, which comprised 70% of the data, with k set

to 10. Using this method, the training data is divided into ten equal sections, or folds. Nine folds are utilized to train the model in each iteration, with one fold serving as the validation set. Each fold serves as the validation set once during the ten repetitions of this process. K-fold cross-validation has a number of benefits. It maximizes the utilization of available data by guaranteeing that each and every data point in the training set is used for both training and validation. By ensuring that the model functions effectively on various subsets of the training data, it also helps to mitigate overfitting by enhancing the model's generalizability.

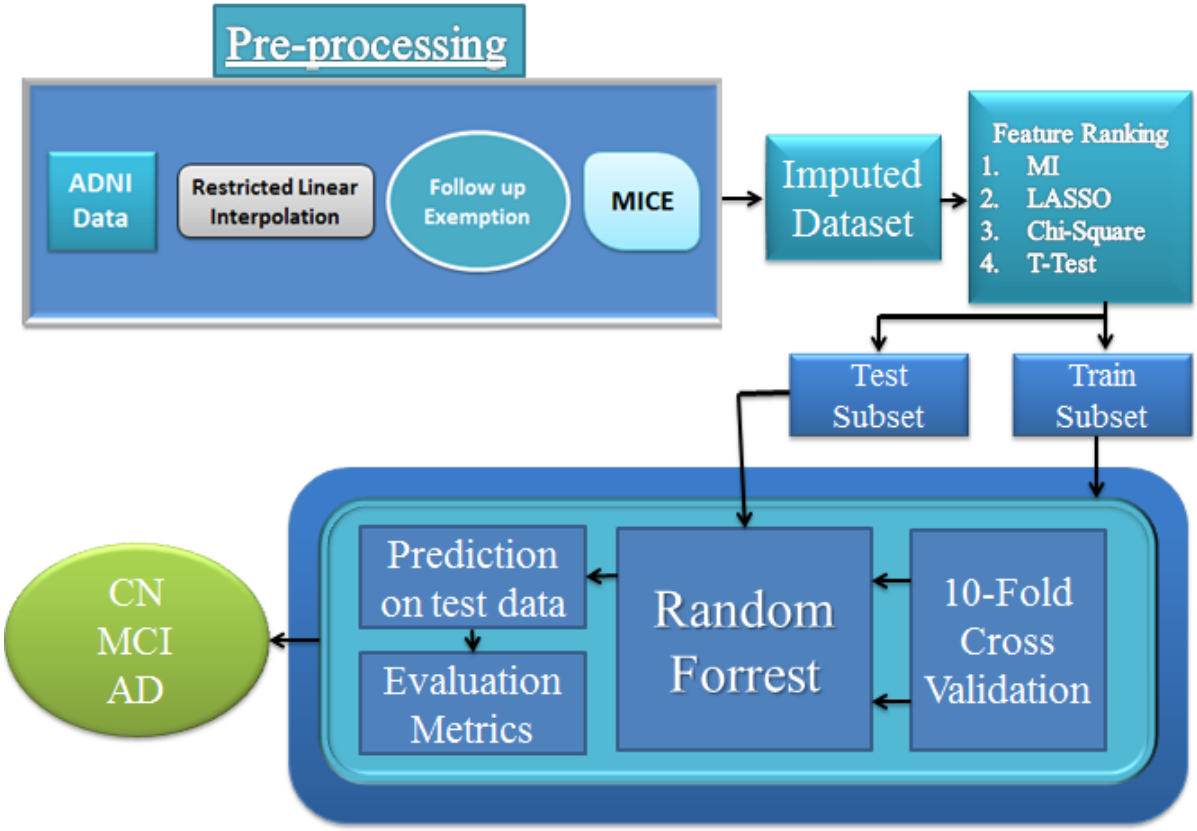

Figure 4. 2: Proposed Model of phase 1

Fitting the random forest method to the chosen features is the task of the model training phase. The k-fold cross-validation approach is used to train and validate the algorithm. Metrics including accuracy, precision, recall, F1 score, and area under the ROC curve (AUC-ROC) are used to assess the algorithm's performance.

Afterwards, the 30% testing group is subjected to final testing using the trained model. In order to confirm the model's robustness and dependability in forecasting outcomes for new patients, this last evaluation offers an estimate of the model's performance on unseen data.

To summarize, the first phase of the proposed model involves a robust imputation model proposition followed by thorough feature ranking and selection process, which is then followed by a methodical training and validation approach utilizing k-fold cross-validation. By using the most pertinent variables and thoroughly training and validating the model, this approach guarantees the production of a dependable and accurate predictive model for the early identification of Alzheimer's disease.

4.1.5 K-Fold Cross Validation:

A reliable statistical method for evaluating the effectiveness and generalizability of machine learning models is K-Fold Cross-Validation. Compared to a straightforward train-test split, it offers a more accurate estimation of a model's predicted performance.

4.1.5.1 Working Principle:

The process of K-Fold Cross-Validation entails dividing the dataset into k folds of equal size. The following is a summary of the procedure:

1. **Dataset Partitioning**: The dataset is split up into k folds, or about equal-sized chunks.

2. **Model Training and Validation**:

   Iteration: The remaining k−1 folds are utilized for training, while one fold is reserved as the validation set for each iteration (i).

   Model Evaluation: The model undergoes training on k−1 training folds and is assessed on the validation fold.

3. **Performance Metrics**: Every fold has a performance log.

4. **Aggregating Results**: To give a general evaluation of the model's efficacy, the performance indicators from each fold are averaged after k iterations.

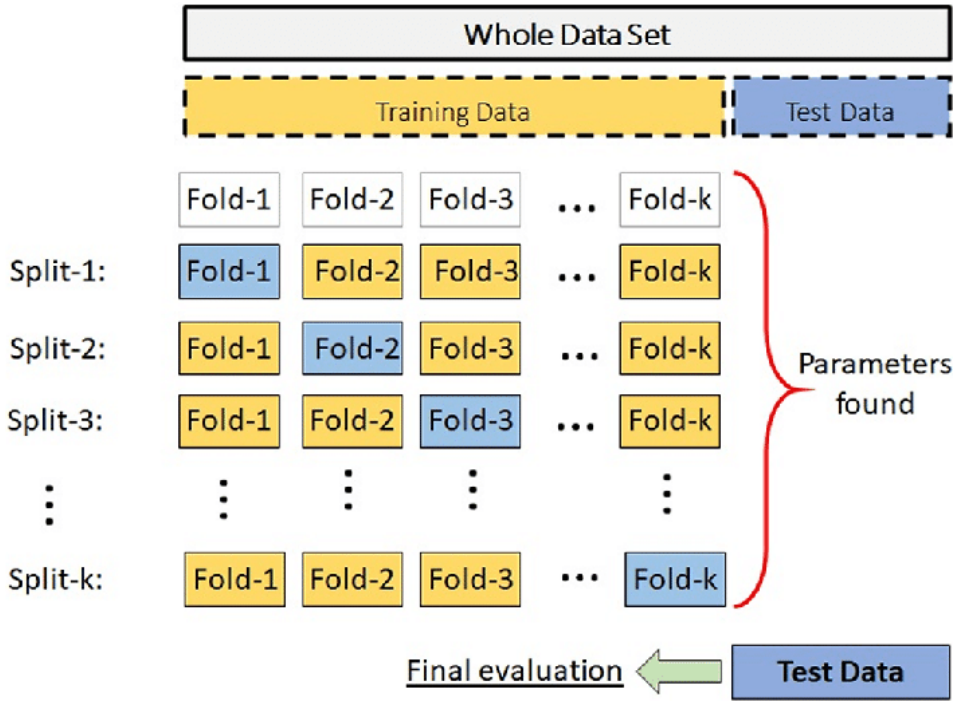

Figure 4. 3: K-Fold Cross Validation

4.1.5.2 Advantages:

Since each data point is used for both training and validation, every instance is tested, which is the main benefit of k-fold cross-validation. This approach is useful for:

1. **Reducing Over fitting**: K-fold cross-validation lowers the likelihood of overfitting to any specific subset by training the model on several subsets of the data.

2. **Utilizing Data Efficiently**: By frequently training and verifying the model on various data subsets, it makes the most use of the available data.

4.1.5.3 Practical Usage

In practical use, k-fold cross-validation is employed to:

1. **Tune Hyper parameters**: By offering a trustworthy approximation of model performance across many configurations, it aids in the selection of the optimal hyperparameters.

2. **Assess Model Robustness**: This provides information on the model's stability and resilience in relation to various data subsets.

3. **Estimate Generalization Error**: It offers a more precise approximation of the model's performance on hypothetical data.

4.2 Phase 2:

Validating the proposed imputation model for managing missing values was the main goal of the first phase. Phase 2 involves a structural transformation of the imputed Alzheimer's Disease Neuroimaging Initiative (ADNI) data into a three-dimensional (3D) plane. The intrinsic variation in follow-up data availability amongst patients, which adds to the missing values, makes this change necessary.

The imputation model that was built and validated in Phase 1is reapplied to accommodate these newly presented missing data. This model is essential for maintaining data integrity and ensuring accurate representation in the 3D space. Figure 4.4 illustrates the comprehensive workflow of Phase 2, depicting each step of the restructuring and imputation process.

4.2.1 Detailed Process of Phase 2

[1] Custom Labeling

[2] Data Restructuring in 3D Space:

[3] Identification of Missing Values:

[4] Application of Imputation Model:

[5] Validation and Quality Check:

[6] Class Classifier

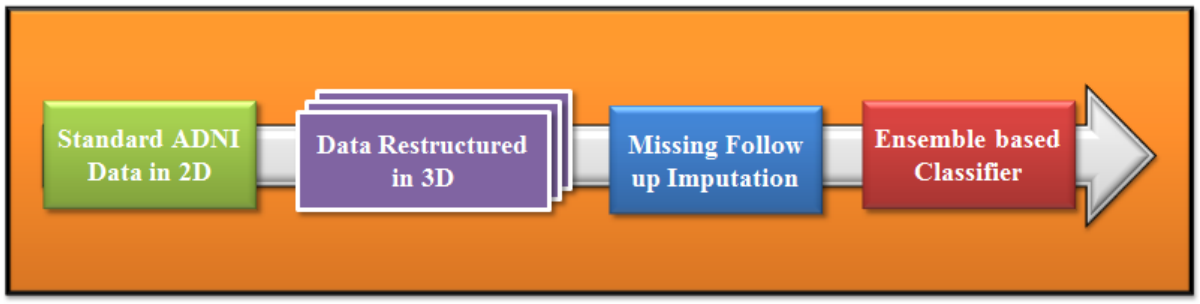

Figure 4. 4: Flow of Phase 2

4.2.2 Custom Labeling

In the phase 2, the final imputed data from the last phase was used as a starting point. In this phase a classifier based on custom labels was implemented. CNs patients are the patients that were initially diagnosed with CN and retained their diagnosis till the last follow-up. CNp patients were initially diagnosed as CN but later progressed to MCI, similarly MCIs patients retain their diagnosis till the latest follow up and are diagnosed as MCI and MCIp patients were initially diagnose as MCI patient but progressed to dementia due to AD with time. Patients labeled as progressing patient follow a specific trend in the biomarkers and ML models can be used to identify those trends. Accurate identification of such trends can lead to accurate prediction and early detection of AD.

4.2.3 Restructuring Data:

After assigning these custom labels to each patient, the first step is to restructure the data. Originally ADNI provides 3 dimensional data but to make the data easily accessible and usable, the data is compressed in a single excel the file which converts it into 2 dimensions. Figure 4.4 shows the labeled 3 dimensions of the data. Black block shows the different patients as 1st dimension. The blue box shows the different follow up visits a patient has as 2nd dimension and the red box shows the 3rd dimension which is the different features. A single cell from the sheet in figure 4.4, can be represented using 3 different coordinates. Each cell represents a specific feature value of a specific follow up of a specific patient. Thereby each cell can be represented in a 3 dimensional co-ordinate system with different patient represented as y-axis, different follow up of each patient as x-axis and different feature as z-axis.

Now to convert this data back to 3 dimensional data, the 3rd dimension has to be extracted to the z-axis. The different patients are kept along y-axis and their respective follow ups along x-axis. This creates a 2D sheet for one feature. Similarly 2D sheets can be constructed for all other features and layered on top along the z-axis. Each feature would be represented as a 2D slice of the restructured data frame. Within each 2D slice, 1st and last columns are common that represent different patients and their respective custom label. All feature sheets have the same patients arranged in ascending order and with same custom label.

	PTID	VISCODE	CDRSB	ADAS11	ADAS13	ADASQ4	MMSE	RAVLT_im	RAVLT_le	RAVLT_fo	RAVLT_pe	LDELTOTA
2	002_S_0295	bl	0	3	4	1	28	56	5	3	23.0769	12
3	002_S_0295	m06	0	5.33	6.33	1	28	50	7	2	16.6667	
4	002_S_0295	m12	0	4.67	5.67	1	30	53	10	2	14.2857	15
5	002_S_0295	m18										
6	002_S_0295	m24	0	3.67	5.67	2	29	45	7	2	16.6667	10
7	002_S_0295	m30										
8	002_S_0295	m36	0	3.67	6.67	3	28	43	5	2	22.2222	15
9	002_S_0295	m42										
10	002_S_0295	m48	0	5	8	3	26	51	6	2	15.3846	6
11	002_S_0295	m54										
12	002_S_0295	m60	0	6	9	3	28	30	4	3	37.5	13
13	002_S_0295	m66										
14	002_S_0295	m72	0	7	9	2	22	41	5	0	0	14
15	002_S_0413	bl	0	3.33	4.33	1	29	52	6	5	41.6667	12
16	002_S_0413	m06	0	7.33	8.33	1	29	47	8	4	30.7692	
17	002_S_0413	m12	0	2.33	4.33	2	29	55	5	6	42.8571	17
18	002_S_0413	m18										
19	002_S_0413	m24	0	2.33	5.33	3	30	52	7	10	76.9231	11
20	002_S_0413	m30										
21	002_S_0413	m36	0	1.33	4.33	3	29	53	7	4	30.7692	14

Figure 4. 5: Original ADNI data with labeled 3 dimensions

4.2.4 Imputing Missing Data:

It was found that inconsistent follow-up visit data from patients led to more missing values, making it difficult to keep the dataset consistent. A standard follow-up time of M96 (8 years) was set for all patients in order to solve this problem. This consistent time period made it easier to imputationally fill in the gaps in the follow-up data.

These missing data were handled by using the custom imputation model that was created in Phase 1. In particular, the model imputed missed follow-ups on an individual basis based on patient data that was previously available. The Multivariate Imputation by Chained Equations (MICE) algorithm, which was set up to operate with a linear regression model, was used in the imputation process. Ten-fold cross-validation was used in this configuration, and ten cycles of iteration were performed to guarantee the correctness and robustness of the imputation. The data was divided into ten subsets for the cross-validation process, with nine of the subsets being used as training sets and the tenth subset as a validation set. This methodology facilitated a thorough assessment of the imputed data's dependability and the model's performance. The imputation model was able to fill in the gaps in the follow-up data with effectiveness due to the numerous iterations and validation, which also produced a comprehensive and consistent dataset for further analysis.

4.2.5 Classification Model:

A full 3D dataset with 8 years of complete data for 1800 patients and 30 features was ready after missing value imputation. Using this data, a classifier model has to be trained. Based on the longitudinal data, a machine learning model that can identify the patient as a progressor or a stable patient is needed. For this purpose four independent machine learning models were trained for each feature. The proposed model uses 2 stage ensemble for the final label prediction. Each feature is used to train a separate ML models that predict the label of the patient based of that specific feature. Four different ML models were trained for each feature:

1. SVM
2. Random Forrest
3. Multilayer Perceptron (MLP)
4. Gradient Boosting

Each of the 4 models predicts a label for the patient based on that specific feature. The final label for this feature is obtained using the stage 1 MVA. Similarly 4 models were trained for all features and stage 1 MVA devises the label based on its respective feature. For combining the labels from all features and deciding on the final label, a 2nd stage MVA is implemented.

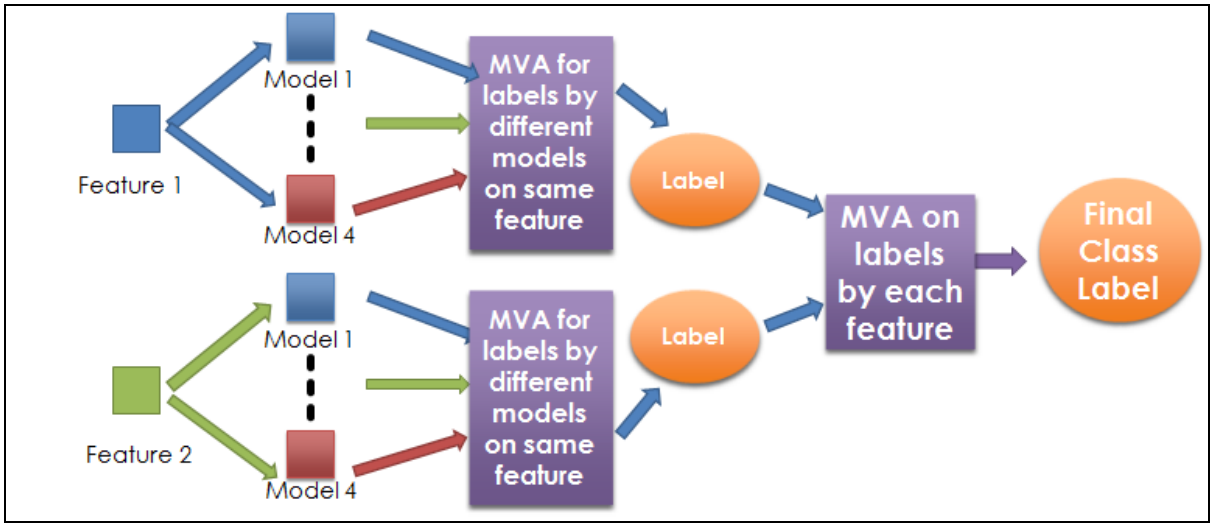

Figure 4. 6: 2 Stage MVA Classifier

4.2.5.1 Majority Voting Algorithm:

In ensemble learning, several models' predictions are combined to provide a final prediction that is more trustworthy and robust. Majority Voting involves training a collection of different machine learning models—each using a different set of algorithms or techniques—on the same dataset. When it comes time to predict, every model votes for what will happen, and the majority's choice determines the final prediction.

The majority voting algorithm is frequently applied in two ways:

1. Hard Voting
2. Soft Voting

1. Hard Voting

In the hard voting method, every trained model votes for the label it believes to be correct. These votes are totaled for each model to determine the final classification. In particular, the number of predictions from each model is tallied, and the class with the most votes is chosen as the final output. This technique efficiently increases classification accuracy by utilizing the combined decision-making ability of several models. Hard voting was used in this study in order to satisfy the conditions for ensemble learning, as shown in Figure 4.5, which shows the two-stage Majority Voting Algorithm (MVA) model. This graphic illustrates how the final label is obtained by averaging predictions from many models through hard voting.

2. Soft Voting

In contrast, soft voting does not provide discrete class predictions; instead, it makes predictions about the probabilities for each class. A probability distribution over all potential classes is provided by each model, and these probabilities are then averaged over all models. The ultimate forecast is made for the class with the highest average likelihood. Soft voting is a useful tool for making decisions when models have different levels of confidence since it allows the certainty of each model's predictions to be taken into account. This method reduces the effect of individual model biases and integrates probabilistic outputs to improve forecast reliability overall.

Ensemble learning relies on both soft and hard voting mechanisms, which offer complimentary approaches for combining model predictions to get more reliable and accurate results. The specifics of the data and the intended balance between classification confidence and precision will determine whether to use soft or hard voting.

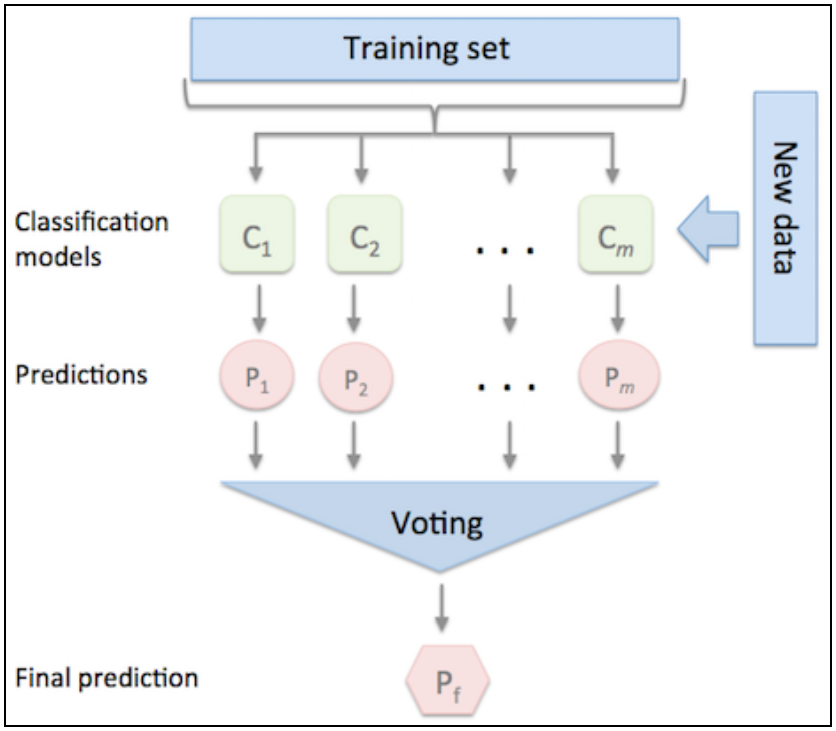

Figure 4. 7: Majority Voting Algorithm

CHAPTER 5

RESULTS

The findings from using the suggested model in both study phases are presented in this chapter. It includes a thorough examination of the performance of the classifier, feature ranking, and imputation procedure. The comparative analysis and assessment metrics demonstrate how well the suggested approaches handle missing information, choose important features, and correctly categorize Alzheimer's disease stages.

5.1 Phase 1:

5.1.1 Preprocessing/Imputation:

The correctness of values imputed by the suggested model is verified by evaluating the classifier's achieved accuracy and comparing it with models found in existing literature. Examining the distributions and statistical characteristics of the features both before and after imputation will help determine how accurate the imputed values are. The distribution and statistical characteristics are maintained if the imputations are exact and correct. This meticulous review confirms the consistency of the imputed data and shows how accurate it is with following procedures.

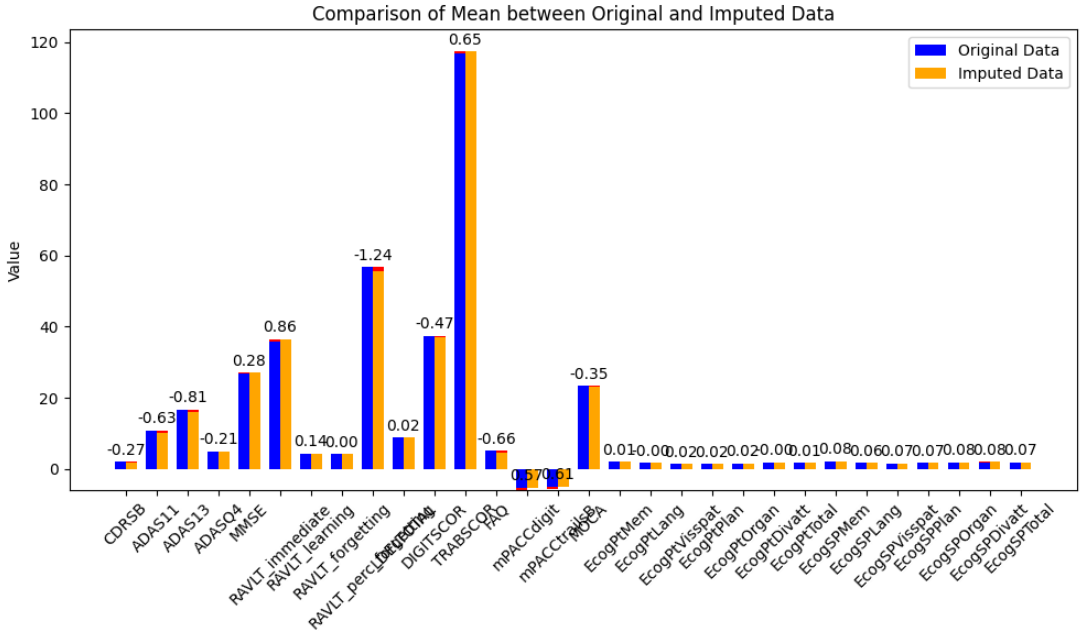

Figure 5. 1: Mean of Original v/s Imputed Features

Figures 5.1, 5.2, and 5.3 present a comparison of the mean, median, and standard deviation of the features before and after imputation. The figures show a minimal variation, suggesting that the features' statistical qualities are not significantly affected. Figure 5.4 and 5.5 show the distributions of CDRSB and ADAS11 before and after imputation. The comparison shows that, other from scaling along the y-axis, the distributions stay mostly identical.

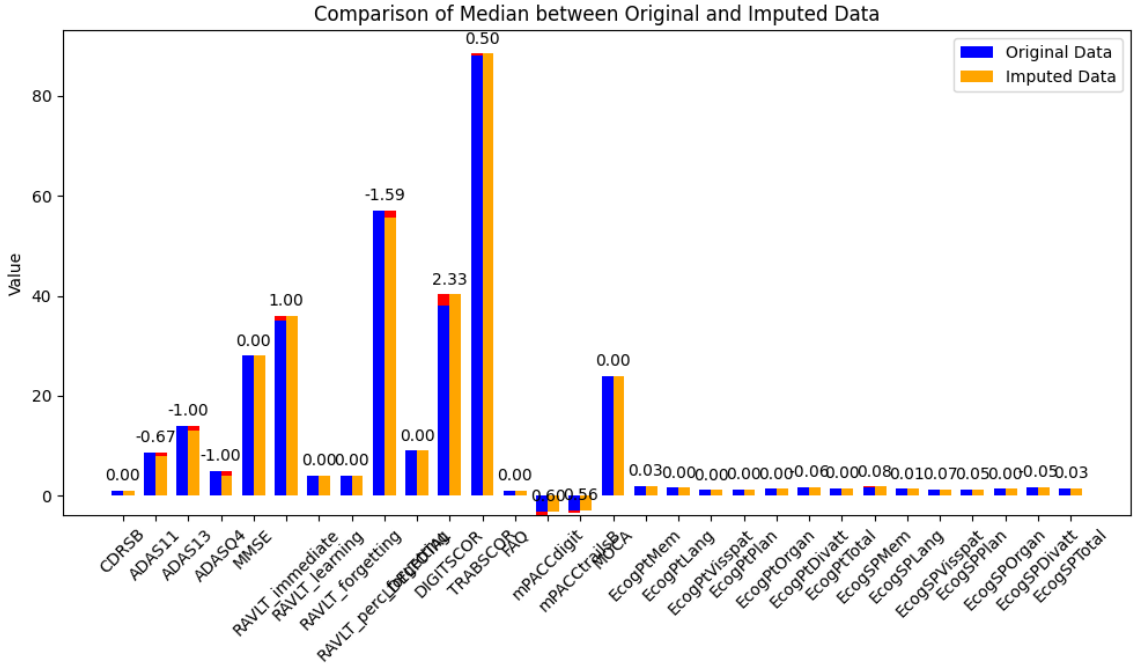

Figure 5. 2: Median of Original v/s Imputed Features

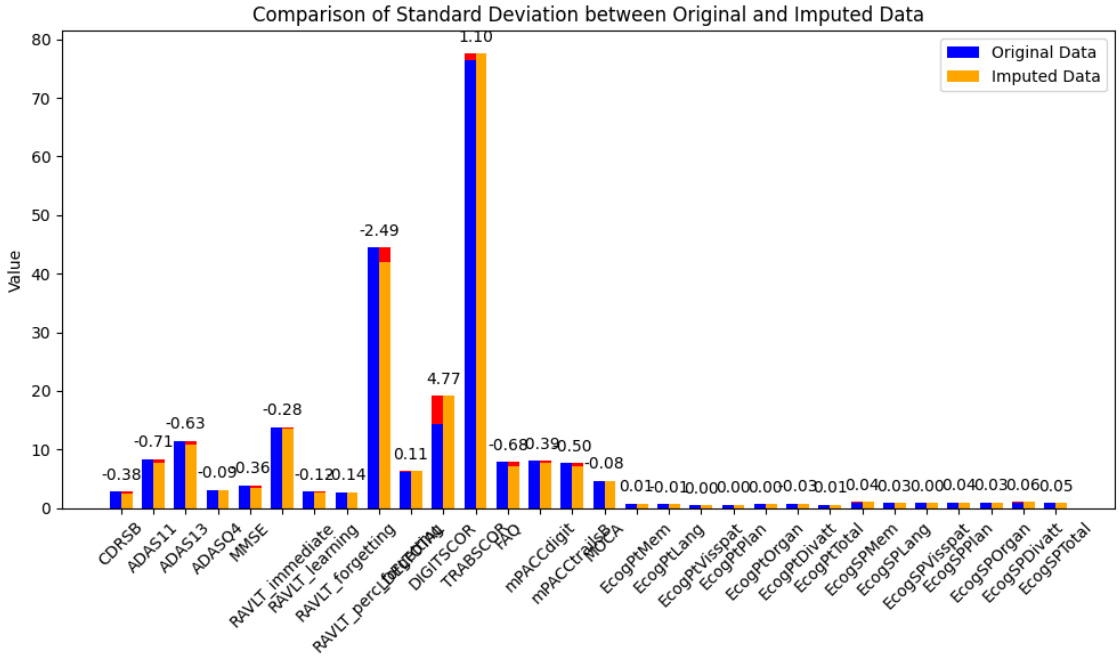

Figure 5. 3: Standard Deviation of Original v/s Imputed Features

A consistent trend across all characteristics is seen, indicating a consistent effect on feature distributions following imputation. These results validate the imputation process's robustness and its capacity to maintain the integrity of feature statistics and distributions.

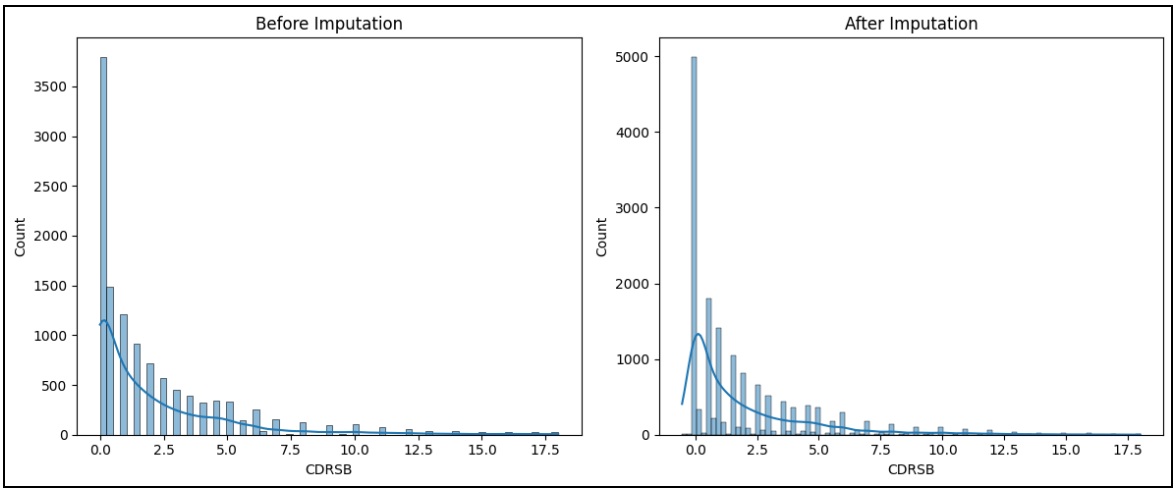

Figure 5. 4:CDRSB Histogram Before and After Imputation

Histograms are a graphical depiction of the distinct values and the frequencies that correspond to them inside a feature. The y-axis in a histogram shows the frequency of each value, while the x-axis shows the range of values that make up a feature. The values appended to the feature cause the y-axis to be scaled following imputation. An upward shift along the y-axis denotes an increase in the frequency of a certain value. Consistent observations of this phenomenon have been observed, notably with regard to ADAS11. Validating the accuracy of the imputed values is done by looking at the distribution and statistical characteristics of all the features that are kept.

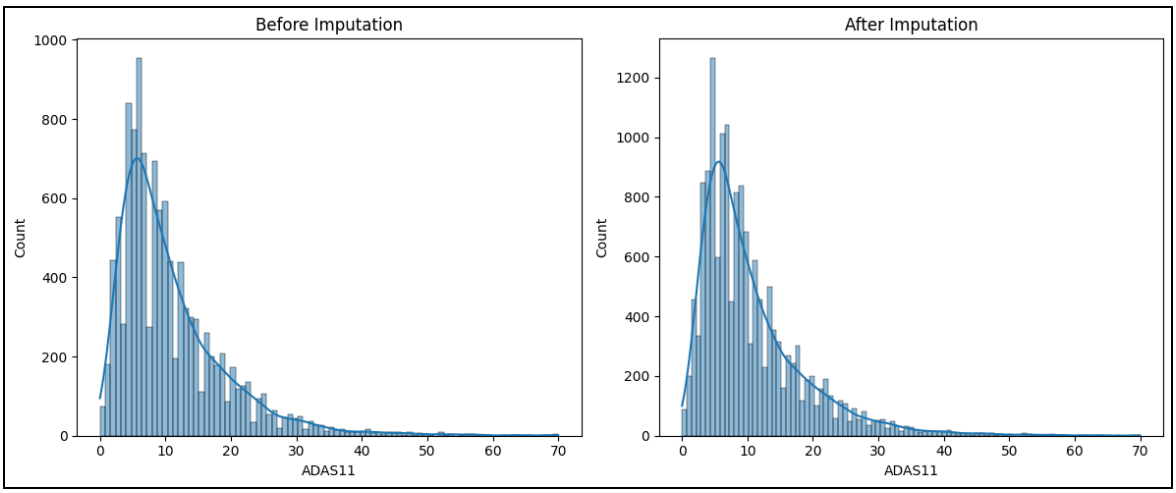

Figure 5. 5: ADAS11 Histogram Before and After Imputation

5.1.2 Feature Ranking:

Top 10 features ranked by various techniques are enumerated in table 5.1, providing insight into the selection process. The ranking techniques utilized in this research include Mutual Information, LASSO, Chi-Square, and Student's T-Test, each coming up with salient predictors that contribute most towards early Alzheimer's disease detection. The use of ranking techniques, to extract most relevant features, facilitates the training of machine learning models and highlights the robustness of the analytical approach.

Table 5. 1: Top 10 ranked features according to 4 different techniques

Sr	MI	Chi-Square	Lasso	T-Test
1.	CDRSB	TRABSCOR	CDRSB	AGE
2.	mPACCdigit	RAVLT_perc_forgetting	EcogSPPlan	EcogPtDivatt
3.	mPACCtrailsB	FAQ	EcogSPDivatt	RAVLT_forgetting
4.	FAQ	ADAS13	EcogPtOrgan	EcogPtOrgan
5.	ADAS13	mPACCdigit	EcogPtLang	EcogPtLang
6.	LDELTOTAL	mPACCtrailsB	EcogPtTotal	EcogPtVisspat
7.	ADAS11	ADAS11	EcogSPLang	EcogPtPlan
8.	MOCA	DIGITSCOR	EcogSPVisspat	EcogPtTotal
9.	MMSE	CDRSB	EcogPtMem	EcogPtMem
10.	ADASQ4	RAVLT_immediate	LDELTOTAL	EcogSPLang

5.1.3 Classifier Results:

Table 5.2 presents a comprehensive comparison of accuracy achieved by various ranking techniques implemented within the study. These techniques are meticulously assessed to gauge their effectiveness in feature selection and subsequent model training. In contrast, Table 5.3 provides a detailed examination of accuracy, contrasting the performance of the proposed model with existing models outlined in the literature. Specifically, accuracy of the proposed model is highlighted, showcasing its performance in Alzheimer's disease classification tasks.

Leveraging a Random Forest framework and a similar experimental setup applied to data imputed by the Phase 1 model, notable enhancements in accuracy are observed. This validation not only corroborates the precision of the imputed values but also underscores the robustness of the Phase 1 model.

Table 5. 2: Accuracy Comparison between ranking techniques

	Random Forest
Complete Data	0.90
Mutual Information	0.88
Chi-square	0.88
LASSO	0.89
T-Test	0.76

Consequently, the proposed model for missing value imputation is seamlessly integrated into Phase 2 for further validation and analysis. Figure 5.6 shows the confusion matrix for phase 1 where label 0 represent CN, 1 represent MCI and 2 represent AD:

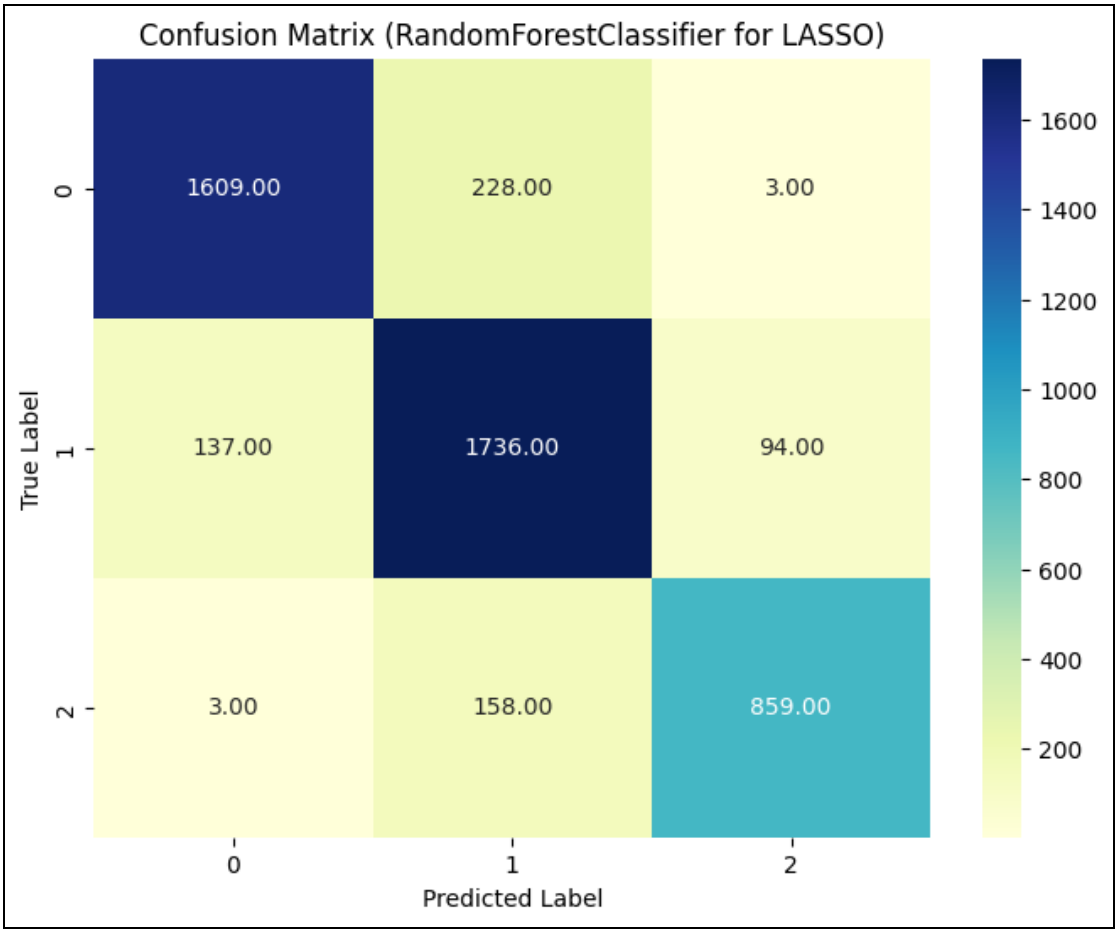

Figure 5. 6: Confusion Matrix for Phase 1 Classifier

Table 5. 3: Accuracy Comparison with literature

Sr.	Missing Values Imputation	Feature Ranking	Classifier Model	Predictor	Validation	Accuracy %
1. [15]	DNN Regressor	T-Test	DNN	NM	5-Fold	87
2. [18]	Median	MI	Random Forrest	NM	10-Fold	86
3. [19]	Euclidean Geometry	T-Test	SVM	NM+MRI	5-Fold	81
4. [21]	Auto Regressor	T-Test	SVM	NM	5-Fold	83
5.	**Proposed**	**LASSO**	**Random Forest**	**NM**	**10-Fold**	**89**

5.2 Phase 2:

5.2.1 Data Restructuring:

Figure 5.7 depicts the restructured data, serving as a visual representation of one feature within the dataset. This representation is extended to other features, each occupying a slice in the third dimension. Subsequently, individual models are trained for each feature, and an ensemble approach is employed to amalgamate the outputs for final class labeling.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	PTID	bl	m06	m12	m18	m24	m30	m36	m42	m48	m54	m60	m66	m72	m78	m84	m90	m96	Calculated Diagnosis	
2	002_S_029	0	0	0	0	0	0	0	0	0	0	0	0	0					CNs	
3	002_S_041	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	CNs	
4	002_S_055	0	0	0	0	0	0	0											CNs	
5	002_S_068	0	0	0	0	0	0	0	0	0	0	0	0.053488	-1.32314	0.029317	0	0		CNs	
6	002_S_072	0.5	0.5	3	1.5	6	4	2	2.75	3.5	4.25	5	5	5	5	5	6.173962		MCIp	
7	002_S_078	0.5	1	1	1.5	1.5	1.5	1.5	2.75										MCIs	
8	002_S_095	2	1.5	3.5	5	8													MCIp	
9	002_S_107	2.5	3.5	3.5	5	8	7	6											MCIp	
10	002_S_115	1.5	1.5	1.5	0.5	2.5	2.5	2.5	2	1.5	1.25	1	1	1	0.75	0.5		0.5	MCIs	
11	002_S_126	0	0	0	0	0	0	0	0	0	0	0	0	0		0		0	CNp	
12	002_S_126	1.5	0	1	0.5	1.5	1.25	1		0.5	0.75	1	1.5	2					MCIp	
13	002_S_128	0	0	0	0	0	0	0	0	0	0	0	0	0		0		0	CNs	
14	002_S_201	0.5	0	0	0	0	0	0	0			0							CNp	

Figure 5. 7: Restructured Data (Single Feature)

5.2.2 Custom Labeling:

Figure 5.8 is a visual bar plot analysis of the number of patients available per label training. It is observed that patients with stable diagnosis are more in both CN and MCI cases and progressor patient for CN are the least with around 200 patients.

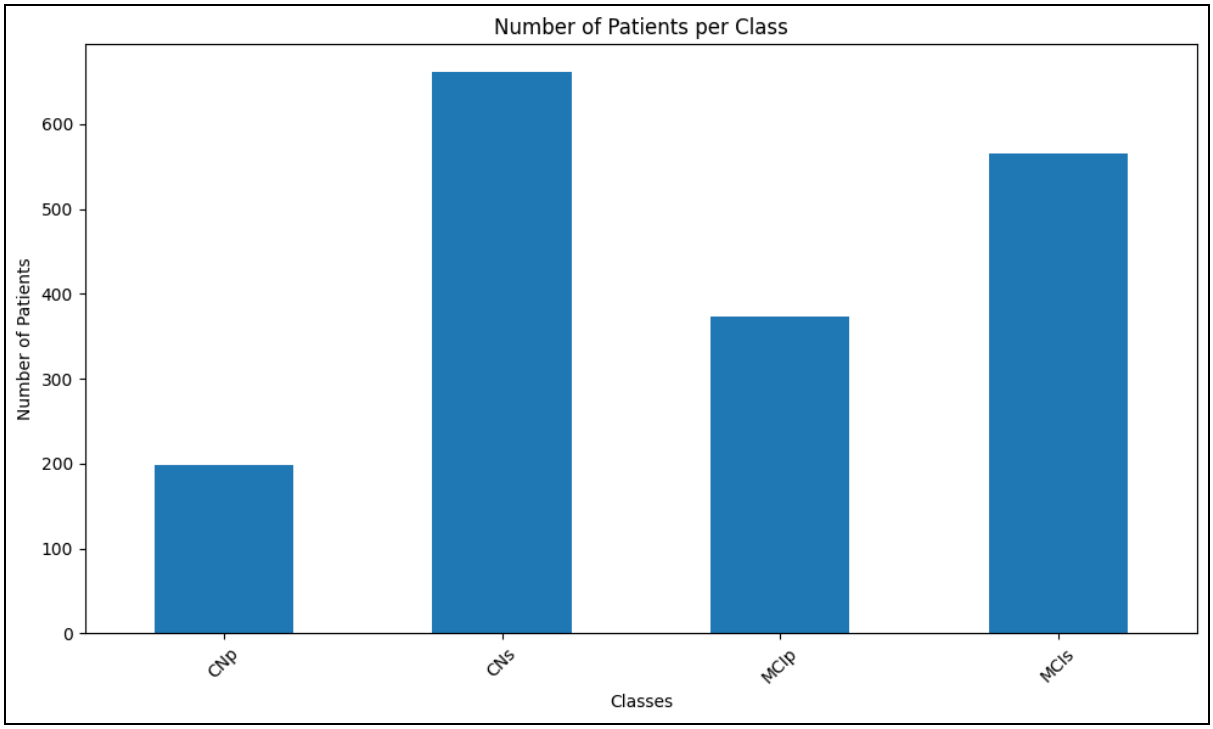

Figure 5. 8: Bar Plot for number of patients per label

5.2.3 Classifier Results Comparison:

Table 5.4 presents a comparative analysis between the proposed model for early Alzheimer's prediction and various models from existing literature. The classification task is evaluated using both the top 5 and top 10 ranked features, providing a comprehensive overview of performance differences among various modeling approaches.

The proposed model has demonstrated notable performance compared to many existing models in the literature across several metrics. Notably, the model's prediction accuracy surpasses that of many contemporary methods, while also exhibiting reduced overall complexity. This is particularly significant as most prior research focuses on binary classification tasks, such as distinguishing between MCI and MCIp patients or MCI and AD patients. In contrast, the proposed model effectively performs four-class predictions, accurately classifying CNs, CNp, MCIs, and MCIp, a capability that is relatively uncommon in existing research.

The fact that the suggested model only relies on five or ten features without sacrificing predictive performance further emphasizes how effective it is. This simplified method preserves overall accuracy and, in certain situations, improves it while reducing model complexity. The feature set reduction shows that a more parsimonious model can nevertheless

achieve good prediction performance, leading to more effective and understandable outcomes in tasks involving the classification of Alzheimer's disease. This result emphasizes how well the suggested approach works to minimize needless complexity while producing reliable forecasts.

Table 5. 4: Accuracy comparison phase 2

Sr.	Missing Values Imputation	Feature Ranking	Classifier Model	Predictor	Classes	Validation	Acc %
1. [15]	DNN Regressor	T-Test (17-features)	DNN	NM	MCIs / MCIp	5-Fold	87
2. [18]	Median	MI	Random Forrest	NM	MCI / AD	10-Fold	86
3. [19]	Euclidean Geometry	T-Test (5-features)	SVM	NM+MRI	MCIs / MCIp	5-Fold	81
4. [20]	-	L 2,1 Norm	TS-SVM	MRI	MCIs / MCIp	10-Fold	76
5.	**Proposed**	**MI (5 Features)**	**Ensemble on ML models**	**NM**	**CNs/CNp MCIs/MCIp**	**10-Fold**	**91**
		MI (10 Features)			**CNs/CNp MCIs/MCIp**	**10-Fold**	**92**
		MI (10 Features)			**MCIs / MCIp**	**10-Fold**	**96**

The confusion matrix for the ensemble classification utilizing the top 5 features for the custom labels is displayed in Figure 5.9. For CNs, CNp, MCIs, and MCIp, respectively, the encoded values are 0, 1, 2, and 3.

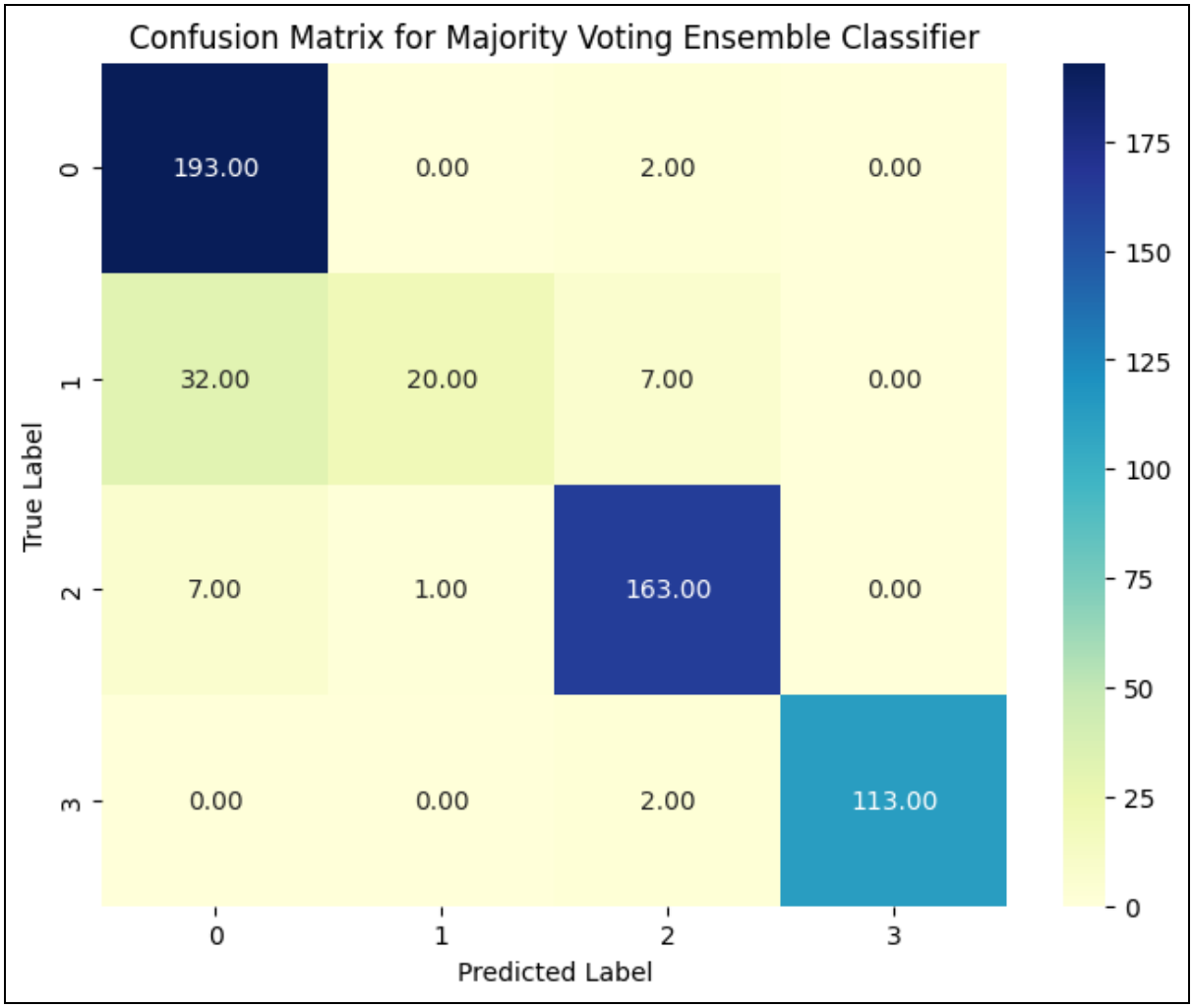

Figure 5. 9: Confusion Matrix of phase 2 classifier

CHAPTER 6

CONCLUSION

By creating and validating innovative predictive models and imputation methods, the research described contributes significantly to the field of early Alzheimer's disease diagnosis. The accuracy of categorization is significantly improved by the suggested imputation model, confirming the accuracy and dependability of imputed values. This enhancement highlights how well the suggested methodology works to handle the problem of missing data, which is essential to the reliability of predictive analytics in clinical contexts.

6.1 Phase 1:

The updated MICE algorithm was put through a thorough testing process in the first step and contrasted with conventional imputation techniques. Performance measures like computing efficiency, statistical integrity preservation, and imputation accuracy were carefully assessed. This comparison study demonstrated the improved accuracy and computing efficiency of the updated MICE algorithm, confirming its efficacy. Furthermore, cross-validation techniques were used to train and validate the classifiers, and common metrics including accuracy, precision, recall, and F1-score were used to assess the classifiers' performance. A different test set was used to evaluate the models' capacity to generalize to previously encountered data, thereby confirming their robustness and dependability.

6.2 Phase 2:

The imputation model was used prior to a classification framework specifically designed for early Alzheimer's disease prediction in the second part of the study. This phase's usage of custom labels showed the model's capacity to correctly categorize patients into a variety of groups, such as CNs, CNp, MCIs, and MCIp. The model's capacity to classify data into many classes is a significant improvement over current binary classification methods, suggesting that it could be used for more sophisticated early Alzheimer's disease detection. The research also removes dependency over multi-biomarkers data, such as MRI, PET scans, and CSF biomarkers, in addition to the primary contributions of enhanced imputation and classification models, in order to improve predictive power and enrich the feature set. By

utilizing this combination, the imputation's performance is improved while also addressing the complex nature of Alzheimer's illness.

The suggested approach has important practical ramifications for healthcare workflows. Clinical settings can easily incorporate the sophisticated imputation and classification approaches, which may result in more individualized treatment regimens and better patient results. The model's practicality is demonstrated by its capacity to generate accurate predictions based only on clinical data, a component that has been neglected in previous studies.

Rigid data quality assurance procedures are another component of the thorough methodology used in this investigation. In order to ensure the consistency and integrity of the data, the dataset underwent thorough cleaning and quality tests to remove biases and abnormalities. The innovative imputation technique was designed to work with the cleaned dataset efficiently, correcting missing values while maintaining each feature's statistical properties. Maintaining the dataset's longitudinal character, which is crucial for monitoring illness progression over time, received extra care. The temporal element of the dataset was preserved by optimizing the imputation approach to handle time-series data.

By using sophisticated cross-validation approaches, the cleaning and imputation algorithms' robustness and dependability were confirmed. The imputed dataset's performance was assessed using a variety of machine learning algorithms, demonstrating its effectiveness in forecasting the course of Alzheimer's disease.

In conclusion, this study lays a strong basis for precise and trustworthy predictive modeling in the identification and tracking of Alzheimer's disease development. The suggested approaches support the development of early detection techniques by addressing issues with missing data and improving classification accuracy. They also have the potential to have a substantial influence on clinical procedures and patient care.

CHAPTER 7

FUTURE WORK

The advancement of AI in Alzheimer's disease research can lead to several opportunities for future research that expand upon the findings of this study. In an effort to increase predictive accuracy and usefulness in clinical settings, the suggested models and approaches could be improved in the ways listed below.

7.1 Integration of Deep Learning Approaches:

The prediction power of the models created in this study could be greatly increased by implementing deep learning techniques. Because they can capture temporal dependencies and sequential patterns, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are particularly well-suited for modeling longitudinal data. These neural network topologies could provide more dynamic and nuanced insights into how Alzheimer's disease progresses, enhancing the precision of early diagnosis and prognosis.

7.2 Development of Advanced Predictive Models:

Developing sophisticated predictor models to predict patients' future scores on different clinical tests could be the focus of future research, building on the imputation model presented in this study. This methodology would facilitate the preemptive detection of illness progression and the customization of personalized treatment regimens. Using forecasting and time-series analysis methods should improve the models' ability to anticipate future illness trajectories and give clinicians important new knowledge.

7.3 Multi-Modal Data Fusion:

The prediction power of the models could be improved by incorporating more multi-modal data sources, such as enhanced neuroimaging modalities, wearable sensor data, and genetic information, into the feature set. When these many forms of data are integrated with the current clinical data, it could lead to a more thorough understanding of Alzheimer's disease and increase prediction accuracy. To maximize model performance, multi-modal data fusion

methods such as feature concatenation and ensemble learning approaches could be investigated.

7.4 Implementation of Transfer Learning:

In order to facilitate the adaption of models to new datasets or populations with limited data, transfer learning approaches could be used to utilize pre-trained models on similar tasks or datasets. This strategy may hasten the creation of reliable predictive models and improve their applicability to various patient populations and therapeutic contexts. The performance of classification and prediction tasks could be further improved by fine-tuning pre-trained deep learning models with the use of the imputed dataset.

7.5 Enhanced Imputation Techniques:

Subsequent investigations may examine the enhancement of imputation methodologies, encompassing the creation of hybrid models that merge numerous imputation strategies or integrate sophisticated statistical models. Examining the use of generative models for imputation tasks, like Generative Adversarial Networks (GANs), may provide creative ways to handle intricate patterns of missing data and enhance the quality of the data.

7.6 Exploration of Explainable AI:

The predictability and transparency of the results could be improved by integrating explainable AI approaches into the models. By offering insights into the models' decision-making process, techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) could help build patient and physician confidence and understanding.

7.7 Longitudinal Studies and Data Collection:

Longitudinal studies with longer follow-up times and more data collection could be used to confirm the predictive usefulness and accuracy of the models. Compiling information on lifestyle factors, therapy responses, and other variables may improve the models' capacity to represent the course of the disease and the outcomes of individual patients over time. By following these suggestions, future investigations can expand on the groundwork laid by this study, enhancing the field of Alzheimer's disease prediction and assisting in the provision

of more efficient and customized patient care. To fully realize the potential of predictive models and enhance early detection and management of Alzheimer's disease, it will be imperative to incorporate sophisticated techniques, multi-modal data, and real-world validation.

REFERENCES

[1] Cano, P. Turowski, et al., "Nanomedicine-based technologies and novel biomarkers for the diagnosis and treatment of Alzheimer's disease: from current to future challenges," J. Nanobiotechnol., vol. 19, p. 122, 2021. doi: 10.1186/s12951-021-00864-x.

[2] G. Cheung, E. To, et al., "Dementia prevalence estimation among the main ethnic groups in New Zealand: a population-based descriptive study of routinely collected health data," BMJ Open, vol. 12, p. e062304, 2022. doi: 10.1136/bmjopen-2022-062304.

[3] G. Livingston, et al., "Dementia prevention, intervention, and care: 2020 report of the Lancet Commission," Lancet, vol. 396, pp. 413-446, Jul. 30, 2020. doi: 10.1016/S0140-6736(20)30367-6.

[4] P. Girotra, T. Behl, A. Sehgal, S. Singh, S. Bungau, "Investigation of the Molecular Role of Brain Derived Neurotrophic Factor in Alzheimer's Disease," J. Mol. Neurosci., vol. 72, pp. 173-186, 2022. doi: 10.1007/s12031-021-01824-8.

[5] H. Braak, D. R. Thal, E. Ghebremedhin, K. Del Tredici, "Stages of the pathologic process in Alzheimer disease: age categories from 1 to 100 years," J. Neuropathol. Exp. Neurol., vol. 70, no. 11, pp. 960-969, Nov. 2011.

[6] L. K. Pradhan, P. K. Sahoo, S. Chauhan, S. K. Das, "Recent advances towards diagnosis and therapeutic fingerprinting for Alzheimer's disease," J. Mol. Neurosci., vol. 72, pp. 1143-1165, 2022. doi: 10.1007/s12031-022-02009-7.

[7] V. García-Morales, A. González-Acedo, et al., "Current Understanding of the Physiopathology, Diagnosis and Therapeutic Approach to Alzheimer's Disease," Biomedicines, vol. 9, article 1910, 2021. doi: 10.3390/biomedicines9121910.

[8] Juganavar, A. Joshi, T. Shegekar, "Navigating Early Alzheimer's Diagnosis: A Comprehensive Review of Diagnostic Innovations," Cureus, vol. 15, no. 9, e44937, Sep. 09, 2023. doi: 10.7759/cureus.44937.

[9] I-H. Oh, W-R. Shin, J. Ahn, J-P. Lee, J. Min, J-Y. Ahn, Y-H. Kim, "The present and future of minimally invasive methods for Alzheimer's disease diagnosis," Toxicol. Environ. Health Sci., vol. 14, pp. 309-318, Sep. 13, 2022. doi: 10.1007/s13530-022-00144-7.

[10] Y. Liu, L. Yue, S. Xiao, W. Yang, D. Shen, M. Liu, "Assessing clinical progression from subjective cognitive decline to mild cognitive impairment with incomplete multi-modal neuroimages," Med. Image Anal., Oct. 2021. doi: 10.1016/j.media.2021.102271.

[11] Milne, "Dementia screening and early diagnosis: The case for and against," Health Risk Soc., vol. 12, no. 1, pp. 65-76, Feb. 2010. doi: 10.1080/13698570903509497.

[12] Subasi, "Use of artificial intelligence in Alzheimer’s disease detection," in Artificial Intelligence in Precision Health, Elsevier Inc., 2020. doi: 10.1016/B978-0-12-817133-2.00011-2.

[13] M. E. Ebrahim, J. S. Jemimah, F. Abuhantash, et al., "Predicting early Alzheimer’s with blood biomarkers and clinical features," Sci. Rep., vol. 14, p. 6039, 2024. doi: 10.1038/s41598-024-56489-1.

[14] P. Ramya, C. Ramesh, O. S. Rao, "Predicting the Transition from Mild Cognitive Impairment to Alzheimer’s Disease using Cognitive Tests and MRI Measures of Demographic Data with an Ensemble Model," Int. J. Intell. Syst. Appl. Eng., vol. 12, no. 2, pp. 250-268, 2023.

[15] Akhtar, S. Minhas, N. Sabahat, et al., "A Deep Longitudinal Model for Mild Cognitive Impairment to Alzheimer’s Disease Conversion Prediction in Low-Income Countries," Appl. Comput. Intell. Soft Comput., vol. 2022, article ID 1419310, 2022. doi: 10.1155/2022/1419310.

[16] Naheed, M. Hakim, M. S. Islam, et al., "Prevalence of dementia among older age people and variation across different sociodemographic characteristics: a cross-sectional study in Bangladesh," Lancet Neurol., doi: 10.1016/S1474-4422(21)00018-6.

[17] Gamal, M. Elattar, S. Selim, "Automatic Early Diagnosis of Alzheimer's Disease Using 3D Deep Ensemble Approach," IEEE Access, vol. 10, pp. 1-1, 2022. doi: 10.1109/ACCESS.2022.3218621.

[18] C. Kavitha, V. Mani, S. R. Srividhya, et al., "Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models," Front. Public Health, vol. 10, pp. 1-1, 2022.

[19] S. Minhas, A. Khanum, A. Alvi, et al., "Early MCI-to-AD Conversion Prediction Using Future Value Forecasting of Multimodal Features," Hindawi, vol. 2021, article ID 1-1, 2021.

[20] Y. Zhu, M. Kim, X. Zhu, et al., "Long range early diagnosis of Alzheimer's disease using longitudinal MR imaging data," Med. Image Anal., doi: 10.1016/j.media.2020.101759.

[21] S. Minhas, A. Khanum, F. Riaz, et al., "Predicting Progression from Mild Cognitive Impairment to Alzheimer's Disease using Autoregressive Modelling of Longitudinal and Multimodal Biomarkers," IEEE J. Biomed. Eng., doi: 10.1109/JBHI.

[22] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, "Multiple imputation by chained equations: what is it and how does it work?," *Int. J. Methods Psychiatr. Res.*, vol. 20, no. 1, pp. 40–49, 2011, doi: 10.1002/mpr.329.