11-10-2021

Machine Learning for Multiclass Classification and Prediction of Alzheimer's Disease

Solale Tabarestani
staba006@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

MACHINE LEARNING FOR MULTICLASS CLASSIFICATION AND PREDICTION OF ALZHEIMER'S DISEASE

A dissertation submitted in partial fulfillment of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

ELECTRICAL AND COMPUTER ENGINEERING

by

Solale Tabarestani

2021

To: Dean John L. Volakis
College of Engineering and Computing

This dissertation, written by Solale Tabarestani, and entitled Machine Learning for Multiclass Classification and Prediction of Alzheimer's Disease, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

_____
Mercedes Cabrerizo

_____
Armando Barreto

_____
Jean Andrian

_____
Naphtali Rishe

_____
David A. Loewenstein

_____
Malek Adjouadi, Major Professor

Date of Defense: November 10, 2021

The dissertation of Solale Tabarestani is approved.

_____
Dean John L. Volakis
College of Engineering and Computing

_____
Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2021

DEDICATION

This dissertation is entirely dedicated to my parents. The achievement of this lifelong goal would not have been possible without the unending support, love, and motivation from these loved ones.

ACKNOWLEDGMENTS

It would not have been possible to complete this dissertation and research without the support and dedication of Prof. Malek Adjouadi. I am deeply grateful to him for his guidance, enthusiastic encouragement, advice, and financial support throughout my Ph.D. program. I enjoyed working with Prof. Adjouadi for the duration of my doctoral work and had growing opportunities within and outside of the university as my research progressed. These opportunities not only include being part of the ADRC team, but internship opportunities in big companies, and serving as a lecturer for Introduction to Engineering Software Technologies for high school students. Secondly, I would like to express my appreciation for Dr. David Loewenstein and Dr. Ranjan Duara for their valuable input, insightful comments, and endless support while conducting my research.

The pursuit of a Ph.D. was not without difficulties, but ultimately it was perseverance that contributed to the achievement of an extremely distinguished lifelong accomplishment. This would not have been possible without the help and support of others that have believed in me from the very beginning. To begin, I would like to acknowledge the technical support I received from my colleagues at the Center for Advanced Technology and Education, which led to a unique and innovative research collaboration during my graduate studies. The next thing I would like to acknowledge is my sincere gratitude to my friends' caring support, whose presence, regardless of their distance brought a sense of warmth and light to my heart during those difficult times. Last but not least, I would like to express my deepest appreciation for the kind unwavering support I received from my parents throughout this journey.

ABSTRACT OF THE DISSERTATION

MACHINE LEARNING FOR MULTICLASS CLASSIFICATION AND PREDICTION OF ALZHEIMER'S DISEASE

by

Solale Tabarestani

Florida International University, 2021

Miami, Florida

Professor Malek Adjouadi, Major Professor

Alzheimer's disease (AD) is an irreversible neurodegenerative disorder and a common form of dementia. This research aims to develop machine learning algorithms that diagnose and predict the progression of AD from multimodal heterogonous biomarkers with a focus placed on the early diagnosis. To meet this goal, several machine learning-based methods with their unique characteristics for feature extraction and automated classification, prediction, and visualization have been developed to discern subtle progression trends and predict the trajectory of disease progression.

The methodology envisioned aims to enhance both the multiclass classification accuracy and prediction outcomes by effectively modeling the interplay between the multimodal biomarkers, handle the missing data challenge, and adequately extract all the relevant features that will be fed into the machine learning framework, all in order to understand the subtle changes that happen in the different stages of the disease. This research will also investigate the notion of multitasking to discover how the two processes of multiclass classification and prediction relate to one another in terms of the features they share and

whether they could learn from one another for optimizing multiclass classification and prediction accuracy.

This research work also delves into predicting cognitive scores of specific tests over time, using multimodal longitudinal data. The intent is to augment our prospects for analyzing the interplay between the different multimodal features used in the input space to the predicted cognitive scores. Moreover, the power of modality fusion, kernelization, and tensorization have also been investigated to efficiently extract important features hidden in the lower-dimensional feature space without being distracted by those deemed as irrelevant.

With the adage that a picture is worth a thousand words, this dissertation introduces a unique color-coded visualization system with a fully integrated machine learning model for the enhanced diagnosis and prognosis of Alzheimer's disease. The incentive here is to show that through visualization, the challenges imposed by both the variability and interrelatedness of the multimodal features could be overcome. Ultimately, this form of visualization via machine learning informs on the challenges faced with multiclass classification and adds insight into the decision-making process for a diagnosis and prognosis.

TABLE OF CONTENTS

CHAPTER PAGE

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

SYMBOLS AND ABBREVIATIONS

AGM	Accelerated Gradient Method
AD	Alzheimer’s Disease
ADAS	The Alzheimer's Disease Assessment Scale
ADNI	Alzheimer’s Disease Neuroimaging Initiative
APOE	Apolipoprotein E
AUC	Area Under the Curve
CDR-SB	Clinical Dementia Rating-Sum of Boxes
CN	Cognitively Normal
cFSGL	Convex Fused Sparse Group Lasso
CSF	Cerebrospinal Fluid
EEG	Electroencephalography
EMCI	Early Mild Cognitive Impairment
GRU	Gated recurrent units
ICV	Intracranial Volume
LMCI	Late Mild Cognitive Impairment
LSTM	Long-Short Term Memory

MCI	Mild Cognitive Impairment
ML	Machine Learning
MMSE	Mini-Mental State Exam
MRI	Magnetic Resonance Imaging
NIA	National Institute on Aging
nFSGL	non-Convex Fused Sparse Group Lasso
RAVLT	Rey Auditory Verbal Learning Test
PET	Positron Emission Tomography
PIB	Pittsburgh compound B
RNN	Recurrent Neural Network
ROC	Receiver operating characteristic
ROI	Region Of Interest
SVM	Support Vector Machine
TGL	Temporal Group Lasso

Chapter 1 Introduction and Research Background

1.1 Introduction

According to a March 2020 report from the Alzheimer's Association (AA), nearly 5.8 million US citizens, mostly elderly people over the age of 65 are affected by AD, a statistic predicted to reach 13.8 million by 2050. This AA report also indicates that an approximated amount of 277 billion dollars was invested in 2018 in caretaking services for patients with AD and dementia [1].

Alzheimer's Disease is a progressive and irreversible brain disorder where subtle brain changes may have started decades prior to any detectable symptoms. In its early stages, AD symptoms begin with mild cognitive decline, which can then progressively lead to more severe physical and functional impairments. Key indicators are associated with severe brain atrophy, beta-amyloid deposition, and evidence of widespread limbic and cortical neurofibrillary degeneration. In the study by [2], an interesting computational neurodegenerative disease progression score is proposed on the basis of the dynamics of the different biomarkers in AD.

Alzheimer's Disease progression is generally assessed using clinical measures, but it can also be accomplished using biomarkers involving structural magnetic resonance imaging (MRI), 18-Fluoro-DeoxyGlucose PET imaging (FDG-PET), cognitive examination, cerebrospinal fluid (CSF), and electroencephalography (EEG) [3]; [4]. Commonly used MRI biomarkers for detecting the progression of AD include cortical thickness and regional brain volume [5][6][7][8], whereas the most significant biomarkers of FDG-PET include glucose hypometabolism in neocortical brain regions [9][10][11][12]. It has also

been revealed that an increase in CSF t-tau or Phospho-Tau is a potential biomarker of disease progression [13][14][15].

Along with neuroimaging modalities, there are other unconventional measurements, known as risk factors, which are associated with Alzheimer's, such as age, genetic information, years of education, and ethnicity [16][17]. As expected, this complementary information shows that age plays a significant role in the onset of AD [18][19]. It is also well acknowledged that the most prominent genetic risk factor is the Apolipoprotein E (APOE) gene. This gene and its major alleles (E2, E3, and E4) are known to increase the risk of developing AD in individuals as young as 40 years of age [20][21].

While many studies in the literature mainly focus on disease prediction, typically relying on a single modality [22][23][24][25][26][27], recent studies have shown that incorporating biomarkers from different modalities may lead to a more accurate diagnosis [28][29][30][31][32]. New research directions have come to rely on multimodal neuroimaging data with the inclusion of other biomarkers such as cerebral spinal fluid (CSF), genetics, and neuropsychological testing. The main objectives of these research endeavors are either to discriminate patients' status via classification methods or to predict different variables using regression models. Cross-sectional and longitudinal data have been used to explore correlations between clinical neuroimaging tests, neurological exams, and biochemical measurements to monitor changes in these important biomarkers. Yet, despite much ongoing research, predicting the progression of AD, especially for early diagnosis and hence enabling the planning of treatment/ curative intervention, has remained challenging [33][34][35][36][37][38][39][40][41][42][43][44][45][46].

1.2 Research Purpose

This research endeavor seeks flexible, reliable, and precise machine learning frameworks that can precisely model the progression of AD with the ability to perform multiclass classification or prediction using regression methods. The methodology envisioned aims to enhance the prediction accuracy by effectively modeling the interplay between the multimodal biomarkers, handle the missing data challenge, and adequately extract all the relevant features that will be fed into the machine learning framework in order to understand all the subtle changes that happen in the different stages of the disease and be able to determine which role such changes play in the transition phases of the disease.

1.3 Research Problem

In order to develop such multiclass classification and prediction methods, the following tasks should be undertaken:

- **Data exploratory.** What are the important data acquisition modalities that can be used for analysis? From each modality, which biomarkers can be extracted? What are the data acquisition standards required for extracting the measurements in longitudinal studies? Which biomarkers are more important and revealing about AD diagnosis and prognosis? What is the best approach for fusing the information acquired from the different modalities? What is the relationship and interplay of what the data reveals between the different modalities?
- **Missing data challenge.** When measurements from different modalities are fused together and stored longitudinally, what is the pattern of the missing data? How can it

be addressed and overcome? Can it be restored using imputation techniques? What is the best approach for using the remaining information?

- **Developing mathematical and statistical models.** The aim is to develop machine learning techniques to model the progression of the disease while maximizing the performance metrics in classification and/or prediction tasks. The model should be trained and tuned precisely in an unbiased way to be able to utilize effectively the available data, identify and assess the important features (associating a weight to them), and perform a multiclass decision-making process and predict future trajectories of the different labels associated with the prodromal stages of the disease.

1.4 Significance of Study

Developing precise machine learning frameworks to understand disease progression is exceptionally important. Current approaches are limited to cross-sectional studies, which neglect the importance of the progressive characteristics of the disease. The nonreversible nature of Alzheimer's disease requires predicting the progression trend of the disease from the early beginning. Although no medication has been found until now that can cure or control this disease, efforts can be made towards finding the factors that can trigger the onset or provoke the patients toward a steep slope. Therefore, the main objective of this research is to capture temporal dynamics in the data and subtle changes in the multimodal biomarkers to model the progression patterns of the patients through the passage of time. Accordingly, we are focusing on longitudinal prediction to gauge the future conversion of the disease. Such intelligent methods, could fuse the biomarkers from different modalities,

model the interplay between various modalities, and predict the future condition of the patients.

1.5 Theoretical perspective and literature review

In line with the research endeavor of our study, Zhu. et al. proposed a multitask convex and non-convex fused group lasso regression for modeling the temporal relationship between multiple future time points to accurately predict the cognitive scores [47]. However, the temporal dependency assumption cannot be guaranteed in reality [48]. In 2016, Moradi et al. have studied the relationships between AD-related structural atrophy within the brain MRI with RAVLT cognitive measures over a period of 3 years. They utilized an elastic net algorithm for modeling the atrophy in MRI [49]. Wang et al. have presented a multi-layer, multi-target regression model for clinical multivariate prediction in AD [7]. This model can simultaneously handle the nonlinear relationship between MRI neuroimaging biomarkers and cognitive assessment scores. They employed matrix elastic nets to investigate the Inter-correlations between multiple test scores. Using non-smooth $\ell 2,1$-norm loss function is shown to add robustness to their proposed multi-target prediction model.

To address the sparsity in the data and model the cognitive scores in five future time points in longitudinal data, Huang et al. proposed the soft-split sparse regression-based Random Forest (RF) model. Focusing on the MRI regions of interests (ROI) volumetric features, they have defined the most discriminate regions along with the future score of the patients only based on the baseline data. Although they provided predictions for multiple

future time points, they have relied on the features of the prior time points for every prediction rather than only based on the baseline data. This means that the model cannot predict the trend of the patient's progression on the diagnosis time. Moreover, they have used a single modality and modeled the relationship between patients' MRI and disease progression, however, multimodal data for this type of research could considerably improve the results reported through the single modality modeling process [48]. A longitudinal observational study for the progression of AD is carried out to investigate the effect of the baseline characteristic on the AD progression and to compare the three criteria of Clinical Dementia Rating-Sum of Boxes (CDR-SB), Mini-Mental Status Examination (MMSE) scores, and the Lawton Instrumental Activities of Daily Living (IADL) questionnaire IADL [50].

There are some other studies that have exploited regression modeling on the longitudinal data for diagnosis and prognosis purposes [51]-[52]-[53]. However, none of them have considered the complex relationship of the samples in the baseline in multiple modalities and while also considering the various change patterns based on the diverse profile of the patients at the diagnosis time. This study is one of the first attempts that encodes the progression of the disease and also predicts the disease trend over a period of 2 years only based on the baseline data. In the proposed model, cognitive score prediction is carried out for 5 future time points.

Despite various studies that have modeled a general regressor capable of predicting one future time point, we developed a technique to model the input features extracted from various modalities of MRI, PET, CSF, and genetic tests. These features are used to predict the cognitive scores of the patients at multiple subsequent time points up to 24 months of

progression. In order to exploit the inherent relationships between the baseline samples adequately, we propose a combined classification and regression approach. In summary, the contribution of this research objective can be described in two folds. The first aim is to estimate the clinical test scores at multiple time points in the future using only multi-modal data available at baseline. Second, we present patients' profile-specific regressors, which rely on the fact that patients who are identified to be in different stages of the disease at the baseline, will follow different disease progression curves over time.

Chapter 2 Profile-Specific Regression Model for Progression Prediction of Alzheimer's Disease Using Longitudinal Data

2.1 Introduction

Many of the AD studies focus on finding neuroimaging-based markers to predict the progression of the disease, which could prove beneficial in creating care plans for the individual patients, and in developing intervention techniques, therapeutic or curative, to possibly delay the progression of the disease [54]. To plan effectively for such types of interventions it is essential to understand and delineate the different stages of AD and be able to develop concise methods to predict and detect the disease in its earliest manifestation. A convenient strategy for diagnosis is to conduct neuropsychological tests that can be used to identify abnormalities associated with the disease. One such assessment developed to determine mental abnormalities is the Mini-Mental State Examination (MMSE). Many studies have shown a reliable correlation between these clinical scores and the prognosis of AD [24].

Early reliable diagnosis of AD through imaging and volumetric calculations, cognitive tests, genetic data, and all other biomarkers is crucial to finding prospective treatments. However, this line of research still remained challenging especially in longitudinal studies due to missing data [55].

Despite the importance of the longitudinal studies for progressive diseases like Alzheimer's disease and Parkinson's, the number of such studies remains limited due mainly to data access and the difficulty and cost associated with data acquisition and related issues to patient follow-up [56]. In line with the research endeavor of our study, Zhu. et al.

proposed a multitask convex and non-convex fused group lasso regression for modeling the temporal relationship between multiple future time points to accurately predict the cognitive scores [47]. However, the temporal dependency assumption cannot be guaranteed in reality [48]. In 2016, Moradi et al. have studied the relationships between AD-related structural atrophy within the brain MRI with RAVLT cognitive measures over a period of 3 years. They utilized an elastic net algorithm for modeling the atrophy in MRI [49]. Wang et al. have presented a multi-layer multi-target regression model for clinical multivariate prediction in AD [7]. This model is able to simultaneously handle the nonlinear relationship between MRI neuroimaging biomarkers and cognitive assessment scores. They employed matrix elastic nets to investigate the inter-correlations between multiple test scores. Using non-smooth $\ell_{2,1}$-norm loss function is shown to add robustness to their proposed multi-target prediction model.

To address the sparsity in the data and model the cognitive scores in five future time points in longitudinal data, Huang et al. proposed the soft-split sparse regression-based Random Forest (RF) model. Focusing on the MRI regions of interests (ROI) volumetric features, they have defined the most discriminate regions along with the future score of the patients only based on the baseline data. Although they provided predictions for multiple future time points, they have relied on the features of the prior time points for every prediction rather than only based on the baseline data. This means that the model cannot predict the trend of the patient’s progression on the diagnosis time. Moreover, they have used a single modality and modeled the relationship between patients' MRI and disease progression, however, multimodal data for this type of research could considerably improve the results reported through the single modality modeling process [48]. A

longitudinal observational study for the progression of AD is carried out to investigate the effect of the baseline characteristic on the AD progression and to compare the three criteria of Clinical Dementia Rating-Sum of Boxes (CDR-SB), Mini-Mental Status Examination (MMSE) scores, and the Lawton Instrumental Activities of Daily Living (IADL) questionnaire IADL [50].

Some other studies have exploited regression modeling the longitudinal data for diagnosis and prognosis purposes [51]-[52]. However, none of them have considered the complex relationship of the samples in the baseline in multiple modalities while also considering the various change patterns based on the diverse profile of the patients at the diagnosis time. This study is one of the first attempts that encodes the progression of the disease and also predicts the disease trend over a period of 2 years only based on the baseline data. In the proposed model, cognitive score prediction is carried out for 5 future time points.

2.2 Methods & Materials

In this section, the data used in this study is introduced and the problem is described. Then, the model and potential challenges in predicting temporal cognitive scores are discussed.

2.2.1 Data

Recently the ADNI-QT has been released as the largest longitudinal dataset. This dataset encourages researchers to implement new techniques for accurate prediction of the future status of the subjects. The dataset includes 1,458 distinct individuals (341 NC, 255 EMCI, 529 LMCI, and 333 AD) examined every 6 months during 11 years period. For every visit

multiple measurements have been collected including neuroimaging tests (MRI, PET), demographic data (Age, Sex, Education and Ethnicity), genetic information (APOE4), CSF (ABETA, TAU, PTAU), and cognitive impairment assessment tests (FAQ), Alzheimer's Disease Assessment Scale cognitive total score (ADAS), Mini-Mental State Exam score (MMSE) and Rey Auditory Verbal Learning Test (RAVLT). Although Alzheimer's diagnosis is not possible without a brain biopsy, ADAS, MMSE, RAVLT scores are widely used in clinical and research studies as a disease progression indicator. Sample data point curation pipelines in our work are presented in Figure. 2.1.

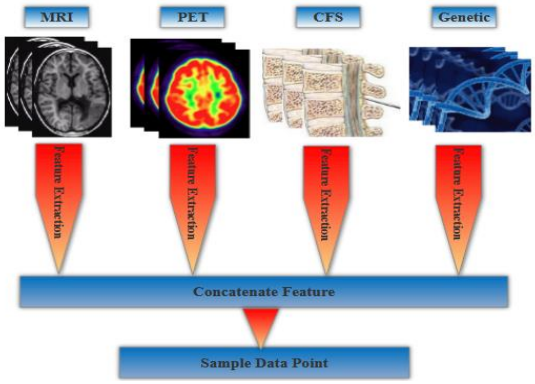


Figure 2.1. Multimodal sample data point

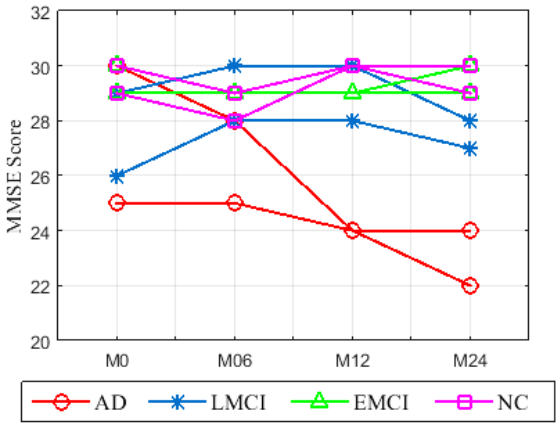


Figure 2.2. Patterns of variation of MMSE score for different classes of AD

2.2.2 Problem Description

The main theme of this objective is to precisely predict the progression of AD at the different prodromal stages of AD using uniquely only the information acquired at baseline. Understanding the disease progression is the cornerstone of developing effective and timely treatment plans that may be subject-specific. However, as mentioned earlier, most of the studies on AD have focused on the diagnosis of the disease in cross-sectional data or, but less so, on predicting disease progression on, at most, a singular subsequent point in time using temporal data. These studies neglect the gradual progression of the disease and may miss out on dependency between consequent time points and consequently on the rate of change at these different time points. Considering the fact that the patterns of progression highly depend on the time of the disease diagnosis, age of the patient, and some other unknown factors, a single regressor can hardly describe the behavior of the disease accurately for all patients with wide ranges of baseline data. Even for patients with identical diagnosis at baseline, tracking the MMSE score during that specific time is a complex task. Figure 2.2 illustrates the variation of MMSE scores for several patients in four different

classes of the disease (AD, LMCI, EMCI, and NC) during the two years. It is clear that the progression of AD, even for patients with the same initial diagnoses, does not follow a steady trend. There are sharp declines, occasional reverse improvements, and also steady periods in different patients, which add to the complexity of building an effective regression model.

Therefore, exploiting the intrinsic relationship in the baseline data will minimize regression modeling errors. On the other hand, relying on the physician diagnosis, which is prone to human error, may skew the statistical analysis. Additionally, having a high number of features from various modalities and tests, known as the "Curse of Dimensionality", may also cause collinearity in the dataset. Collinearity hinders correct modeling of disease behavior and complicates the model, making it vulnerable to overfitting. Furthermore, temporal data usually suffers from a high number of missing tests and measurements as many patients may stop their participation in the study or they may not take all the clinical tests consistently at some of the time points. These phenomena add sparsity to the dataset.

To overcome this missing data challenge, this research focuses on both cross-sectional analysis and longitudinal data modeling by proposing a combined technique. The main part of the model consists of training multiple repressors with subjects at differing stages of impairment severity using baseline data after selecting the least sparse informative features. These regressors are trained separately to predict cognitive scores at future stages. While every future time point is modeled solely based on baseline data, subjects who are missing data at some of the time points will still be preserved in the study.

2.2.3 Model Description

The high dimensionality of feature space in the original data can add complexity to clustering, classification, and regression modeling. In order to simplify this problem, after data preprocessing, we apply a feature selection technique that selects highly informative features. Consequently, a Multi-class Multi-Layer Perceptron (MLP) model is employed to detect strongly related patterns in baseline subjects. Afterward, class-specific regression models are trained to precisely track the direction of the data points in each class of subjects. This approach attempts to classify subjects based on hidden patterns of the features at the baseline and then models them. Due to the fact that different classes of subjects undergo different developmental progression trends, the prediction performance benefits from these classification techniques. The correlation between similar observations will be integrated into the regression model to enhance the prediction accuracy rather than training a single model across all individuals. The flowchart of the proposed model is presented in Figure. 2.3.

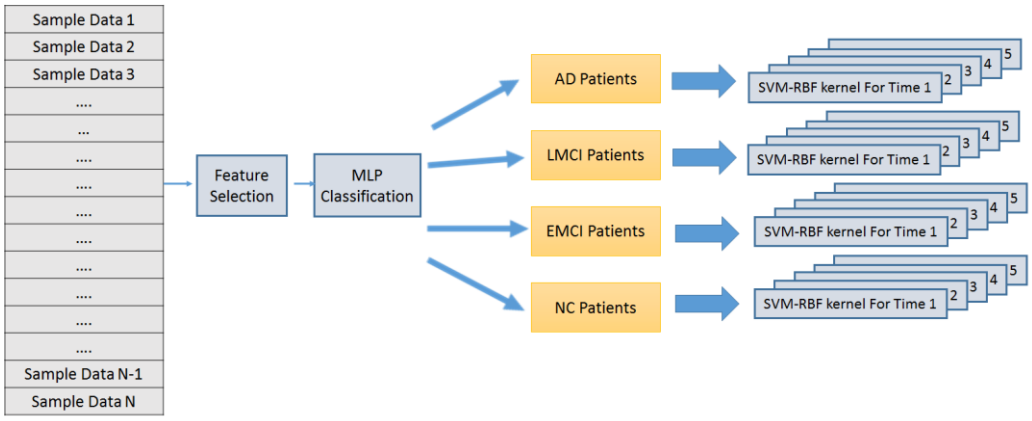


Figure 2.3. Illustration of the proposed framework

2.2.4 Feature Selection

In order to protect the prediction model against overfitting and reduce computational complexity, feature selection has been incorporated into our model. In general, model performance may be affected by the number of features in the feature set and the strategy employed is in selecting only the pertinent ones. L1-based feature selection is used to reduce data sparsity. Exhaustive experiments with the aid of error analysis using the L1 norm helped to identify the 31 most dominant features of the dataset.

2.2.5 Multiclass Classification

Multi-Layer Perceptron (MLP) is a subset of feedforward artificial neural networks. These supervised machine learning techniques consist of three or more layers of multiple processing nodes: an input layer that feeds data from external inputs, middle hidden layers that perform calculations and adjust the weights for the model, and an output layer that produces the result. This algorithm has garnered attention in recent years and has been extensively applied in various types of classification and regression modeling. The idea proposed here utilizes this algorithm but for multi-class discrimination of baseline data, overcoming the limitations imposed by using binary classifications.

2.2.6 Support Vector Machine with Radial Basis Function

Linear regression algorithms cannot describe the complex relationship between input feature spaces and future clinical scores. In order to overcome this limitation, we adopted a radial basis function (RBF) kernel in SVM to fulfill the restriction of nonlinear arbitrary

kernels. In the SVM kernel, the input layer is mapped to a high dimensional (almost infinitely dimensional) nonlinear hidden space. The RBF kernel relies on the radial or Euclidean distance from the origin yielding a linear combination of radial basis functions.

Considering the feature space from N observations as the input vector $X = \{x_1, x_2, \dots, x_N\}$, the regression problem can be defined as finding a function $f(x) = \langle w, \varphi(x) \rangle$ which tries to fit the target vector $Y = \{y_1, y_2, \dots, y_N\}$. The input space x is transformed to a high-dimensional nonlinear feature space $\varphi(x)$. The kernel of Radial Basis Function (RBF) SVM can be expressed as

$$K_{RBF}(x, x_i) = \exp\left[-\gamma \|x - x_i\|^2 / 2\sigma^2\right] \qquad (1)$$

where the γ parameter measures the similarity between each pair of points in the training set and is defined as the inverse of the standard deviation of the RBF kernel, thus restricting the size of the kernel. In Gaussian distribution, the parameter σ is defined as the standard deviation. Considering the complex nature of AD progression, we adopted this nonlinear kernel as the regression algorithm.

2.3 Results & Discussion

In this study, we selected subjects who underwent 3D volumetric imaging, genetic tests, and neuropsychological test have been selected from the ADNI-QT dataset. Subjects whose ABETHA, PTAU, or TAU have been reported out of range or those whose MMSE scores were missing have been eliminated from further analysis, leaving us 333 AD patients, 529 LMCI, 255 EMCI, and 341 NC with 38 features. From this feature set, we also excluded

the predictive biomarkers such as ADAS11, ADAS13, CDRS, and diagnosis labels (DX). The MMSE score is the target vector to be predicted with our method. The demographic characteristic of the participants is summarized in Table 1.

All samples have been normalized via removal of the mean and scaling to unit variance. After applying the L1-norm feature selection technique, which leaves only the most informative features, the Multi-layer Perceptron technique is applied to classify the baseline data into four subject classes (NC, EMCI, LMCI, and AD).

Table 2.1. Subjects' demographics considered for this study

SUBJECTS	AD	LMCI	EMCI	NC
MALE	182	326	147	173
FEMALE	151	203	108	168
AGE (MEAN±SD)	75.2+7.7	74.1+7.4	71.7+7.4	74.9+5.7
MMSE(MEAN±SD)	23.2+2.0	27.1+1.8	28.2+1.6	29.1+1.1

Since the initial stage of a subject highly impacts the pace and slope of the progression of Alzheimer's disease, modeling all of the patients with different initial patterns and tests does not converge to a coherent model. Referring to Figure 2.2, which represents the change of the MMSE scores of several patients during the time, it can be clearly seen that the AD progression, even for patients with the same initial diagnoses, does not follow a steady trend. There are sharp declines, occasional reverse improvements, and also steady periods in different patients, which add complex regression modeling. In order to address this issue, we adopted an architecture that models each class of the disease separately. Patients who have a similar initial profile in the baseline are separated into different sub-groups. Initially,

we attempted to categorize similar patient profiles with unsupervised techniques, such as a non-linear clustering algorithm of Gaussian Mixture, to avoid relying on the expert's diagnosis, which can be prone to human error. Our model incorporates information relating to the covariance of the data and the centers of the latent Gaussians to address the noisy nature of the data, and in order to exploit the underlying affinity between individuals. Moreover, to handle missing labels and prevent discarding a considerable number of observations with no diagnosis, one can take advantage of this unsupervised technique over the supervised technique. However, even with four relatively balanced classes, the outcome cluster of the GM model does not comply with the predefined classes provided by the expert. Therefore, we adopted the supervised classification technique and repeated our experiments to discriminate baseline patterns.

A hidden layer of size 100, a logistic sigmoid activation function, and an adaptive learning rate with a stochastic gradient descent solver have been selected as the best hyperparameters after comprehensive trial and error. For the prediction of cognitive scores per six selected periods, a single SVM RBF regression kernel has been trained over the training samples at various times. Several linear and non-linear regression kernels, including Lasso, Ridge, Gradient Boosting, and Elastic net have been investigated. We selected the SVM-RBF regression kernel due to the high accuracy across each time interval and overall classes. For each kernel, the hyperparameters have been tuned separately for optimum prediction using grid search. Comparative results are presented in Table 2.

Table 2.2. Comparative RMSE score assessments of the proposed method vs. other linear and nonlinear methods over five different future time points

	T1	T2	T3	T4	T5
LINEAR	5.9358	3.3933	3.3886	3.1353	4.1521
RIDGE	4.1499	5.8083	4.9830	9.3006	4.2753
XGB	6.2441	4.3137	8.3372	5.1598	5.6455
SVM-POLYNOMIAL	4.2936	3.9122	10.5797	4.1523	4.1006
ELASTIC NET	2.8610	3.0897	4.5425	6.2158	4.0159
SVM RBF	2.8514	2.8792	3.1049	3.2513	3.2263

Regression performance is evaluated using a 10-fold cross-validation procedure on the longitudinal cohort. Simulation results are depicted in Table 3. Our model showed a decrease of approximately 0.12,0.08,0.09,0.08 in terms of RMSE for 6, 12, 18, and 24 months respectively over the baseline model which does not incorporate feature selection and classification schema (basic model). Using feature selection and data normalization prior to training the model and testing individual regression kernel for each class of subjects in baseline achieved higher performance.

Table 2.3. Comparison of the prediction accuracy of the proposed method assuming four classes of Alzheimer's disease and for five future time points in terms of RMSE

		T1	**T2**	**T3**	**T4**	**T5**
AD	**Proposed**	**1.8410±0.27**	**2.0529±0.35**	**2.7727±0.23**	**2.3878±0.26**	**3.8658±0.56**
	Basic	1.8836±0.36	2.1270±0.44	2.7897±0.52	2.6244±0.35	3.8946±0.72
CN	**Proposed**	**1.6315±0.30**	**2.4977±0.43**	**2.1134±0.29**	**2.3679±0.42**	**3.1017±0.60**
	Basic	1.6397±0.35	2.5833±0.54	2.1537±0.36	2.5688±0.51	3.1574±0.66
EMCI	**Proposed**	**1.4460±0.14**	**2.0305±0.37**	**3.2103±0.74**	**2.9154±0.39**	**4.2818±0.28**
	Basic	1.5012±0.27	2.2035±0.62	3.3713±0.44	3.6721±0.47	4.5496±0.56
LMCI	**Proposed**	**1.3998±0.18**	**2.3637±0.39**	**2.4096±0.17**	**2.2104±0.33**	**3.5681±0.59**
	Basic	1.4800±0.24	2.3859±0.52	2.4935±0.45	2.2934±0.61	3.6426±0.74

Chapter 3 Longitudinal Prediction Modeling of Alzheimer Disease Using Recurrent Neural Networks

3.1 Introduction

The complex nature of AD biomarkers and the heterogeneity of measurements obtained from various imaging modalities are some of the obstacles faced in seeking effective early detection and planning therapeutic protocols [57]-[58].

In addressing the barriers impeding AD research, scientists have proposed statistical and machine learning techniques for robust diagnosis. Until recently, most efforts were dedicated to modeling the disease at a single time point using cross-sectional datasets [59]], [[60]. However, these approaches could not provide enough information about the future status of patients. At later stages of AD, where the brain has already suffered from atrophy, treatment would be too late to be effective. Early diagnosis of the disease allows for early intervention and facilitates the development of effective healthcare services. This initiates a new line of research aiming at enhancing the effectiveness of treatment by predicting the onset of the disease before the occurrence of acute neurodegeneration. The objective of these studies is to leverage temporal information from longitudinal data to model the progression of AD. Multiple classification and regression models have been proposed to predict disease progression and level of disease severity. The feature space is either based on the information available at baseline or a concatenation of features from multiple previous time points [61], [62], [63], [64]. The integration of features into a single observation window creates a high dimensional input space which is not only difficult to deal with but also disregards temporal connections between consecutive time points [65],

[66]. With the gradual nature of AD progression, these methods could not efficiently exploit the longitudinal information.

Recurrent Neural Networks (RNNs), introduced in 1986, recently gained popularity due to the intrinsic power in learning long-short term dependencies of sequenced data. These networks share information between series of data points through an additional hidden set of parameters. RNNs are now being implemented in modeling the progression patterns of chronic diseases [67], [68]. In [66], Nguyen et al. trained an RNN-LSTM network over a span of seven years to predict multiple AD biomarkers for one subsequent time point. In another study, Wang et al. applied an RNN architecture with LSTM cells to predict the global staging of the Clinical Dementia Rating (CDR) score of the next visit using previous records [69]. Aghili et al. utilized LSTM and GRU models to classify AD subjects using longitudinal records of data over an 11-year period [70].

Using the inherent correlations of sequential data, RNNs proved their potential in predicting AD-related biomarkers for a future time point. Although effective, these studies limit themselves to predicting at only a single future interval. The model introduced in this chapter, broadened the scope and application of the RNNs by predicting the progression of AD over multiple future time points simultaneously. Employing three records of data for each subject, the RNN surpassed other machine learning methods not only in estimating the categorical variable for a multiclass classification task but also in assessing the numerical value of the AD biomarker.

Furthermore, two variations of RNN, GRU, and LSTM, are investigated for the challenging task of drawing the delineation boundary of subjects in a multiclass classification scenario and also for predicting the trajectories of cognitive scores for the next two years.

3.2 Methodology

3.2.1 Recurrent Neural Network (RNN)

Processing sequences of data, RNNs have the capability to effectively incorporate temporal dependencies in longitudinal data. Figure 3.1 illustrates an RNN with data sequences of k time steps. At each time point (t_i), besides the input features (X_{t_i}), the internal state (memory) of the cell from the previous time step ($h_{t_{(i-1)}}$) are fed to the cell. Thus, unlike feedforward neural networks, RNNs can identify patterns hidden in sequences of data. However, due to a lack of long-term memory in basic RNNs, each time point is mainly affected by previous intervals in close vicinity. Therefore, they are not capable of leveraging long-term relationships in historical data and older information tends to fade away. This setback is known as “vanishing gradient” in which the network gradually forgets older traces.

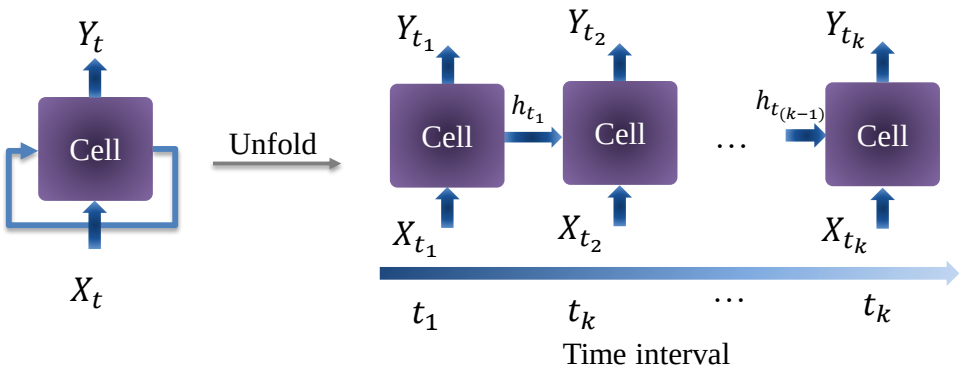


Figure 3.1. Recurrent Neural Network architecture

To address this issue, GRU and LSTM-based RNN architectures with the capability of capturing long-term memories have been proposed [71], [72]. The structure of LSTM and GRU cells as the building blocks of an improved version of RNN is shown in Figure 3.2. In an LSTM cell, three gates are denoted by sigmoid functions (σ), decide whether the previous cell state (C), the input (X), and the output (h) need to be passed to the next time step. This will make the memorizing capability of the cell to be more intelligent and durable. The following equations describe the operation principle of an LSTM cell.

$$
\begin{aligned}
f_{t_k} &= \sigma\left(W_f\left(X_{t_k}, h_{t_{k-1}}\right) + b_f\right) \\
i_{t_k} &= \sigma\left(W_i\left(X_{t_k}, h_{t_{k-1}}\right) + b_i\right) \\
\hat{\imath}_{t_k} &= tanh\left(W_{\hat{\imath}}\left(X_{t_k}, h_{t_{k-1}}\right) + b_{\hat{\imath}}\right) \\
C_{t_k} &= C_{t_{k-1}} * f_{t_k} + \hat{\imath}_{t_k} * i_{t_k} \\
o_{t_k} &= \sigma\left(W_o\left(X_{t_k}, h_{t_{k-1}}\right) + b_o\right) \\
h_{t_k} &= o_{t_k} * \tanh\left(C_{t_k}\right)
\end{aligned}
\tag{1}
$$

where t_k refers to the k^{th} time step; X_{t_k}, C_{t_k}, and h_{t_k} represent the input, state, and output

of the cell at the k[th] time step; and f_{t_k}, i_{t_k}, and o_{t_k} are the outputs of the forget, input, and output gates. Also, *W* and *b* are the weights of the neural networks.

In the gating mechanism of GRU, two gates known as reset and update gates determine the amount of the current input and output of the previous time step that needs to be preserved. With the same notations of X_{t_k} and h_{t_k} as the input and output of the cell for the k[th] time step, the mathematical equations of a GRU cell are summarized as follows.

$$
\begin{aligned}
z_{t_k} &= \sigma\left(W_z\left(X_{t_k}, h_{t_{k-1}}\right) + b_z\right) \\
r_{t_k} &= \sigma\left(W_r\left(X_{t_k}, h_{t_{k-1}}\right) + b_r\right) \\
\hat{h}_{t_k} &= tanh\left(W_{\hat{h}}\left(X_{t_k}, r_{t_k} * h_{t_{k-1}}\right) + b_{\hat{h}}\right) \\
h_{t_k} &= \left(1 - z_{t_k}\right) * h_{t_{k-1}} + z_{t_k} * \hat{h}_{t_k} \\
h_{t_k} &= o_{t_k} * \tanh\left(C_{t_k}\right)
\end{aligned}
\tag{2}
$$

where z_{t_k} and r_{t_k} are the outputs of the update and reset gates.

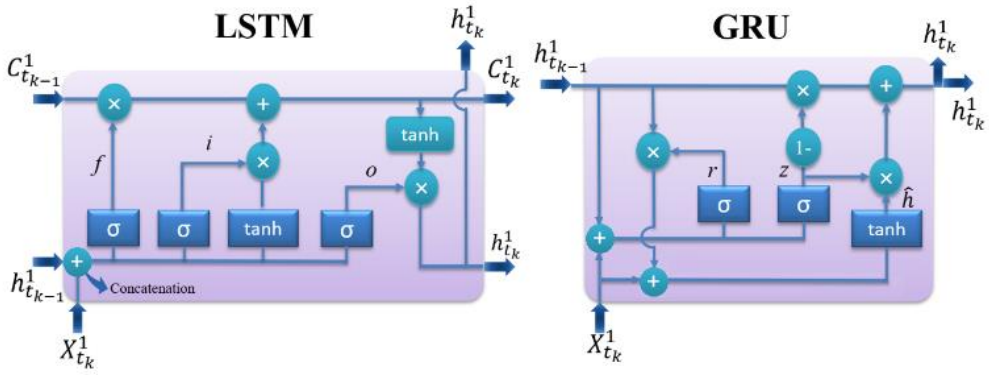


Figure 3.2. The structure of LSTM and GRU cells

:

3.2.2 Feature Selection

Referring to previous studies [70], which shed light on the possible overfitting of RNNs on the original feature space, feature analysis, and ranking has been performed on the data. Consequently, to address the highly correlated features, *L1* feature selection was employed to extract the most important features. Using the *L1* method, 25 features with the highest variance in the feature space have been selected. The correlation matrix (heat map) of the features is illustrated in Figure 3.3.

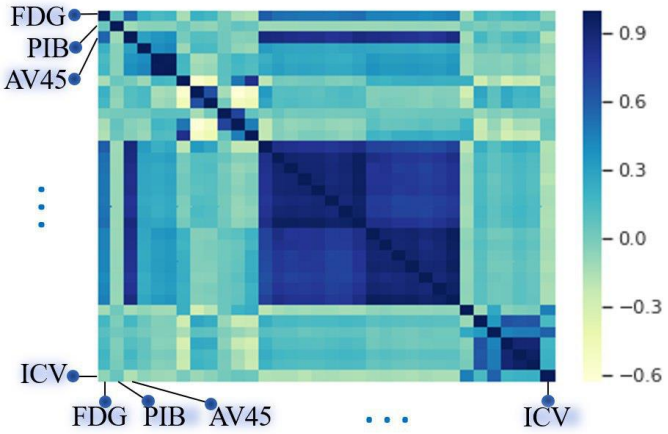


Figure 3.3. Heat-map of features used in this study

3.2.3 Longitudinal AD Prediction using RNN

The proposed framework uses the memorization capability of the LSTM/GRU cell to capture historical dependencies from three records of subjects to predict the progression of AD at three next future time points. Therefore, a many-to-many RNN architecture with LSTM/GRU cells has been developed to carry out two tasks of longitudinal multiclass classification and regression.

The structure of the network for the LSTM case is demonstrated in Figure 3.4. In the developed network, the three inputs (X_{t_1}, X_{t_2} and X_{t_3}) represent the feature space associated with three-time points of M_0 (Baseline), M_6 (after 6 months), and M_{12} (after 12 months). The information is transferred from one time point to the next one using the cell state (C) and output (Y). The outputs Y_{t_i} are the Mini-Mental State Examination (MMSE) score for regression model or status of patients (CN, MCI, and AD) for the classification model. The time steps t_4 and t_5 are associated with the future time points M24 (24 months after the baseline) and M36 (36 months after baseline). The next section discusses the material and experimental results.

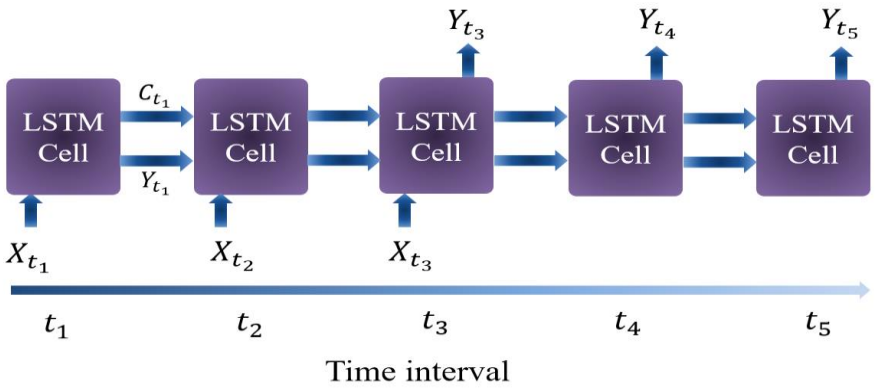


Figure 3.4. The RNN architecture used to predict the progression of AD using historical data

3.3 Results and discussions

3.3.1 Data

Longitudinal medical records from 1458 subjects (341 CN, 255 EMCI, 529 LMCI, and 333 AD) have been incorporated into this dataset. During an 11-year study, each patient has been recalled for a follow-up visit every six months. These subjects have undergone

several medical screening tests including MRI, PET, genetic tests, CSF tests, and cognitive impairment assessments. At each visit, an expert monitors the test results and updates the diagnosis for the participants. This categorical diagnosis (AD, MCI, NC) is used as the label for the multiclass classification experiment proposed in this study and the numerical Mini-Mental State Examination (MMSE) scores, an indicator of the AD cognitive impairment, with a range of 0-30 is adopted for the regression experiment. Characteristics of the dataset used in this study are summarized in Table 1.

Table 3.1. Statistics of the Dataset Used in This Study

Category	Subjects (f/m)	Age	Education(y)	MMSE
AD	336 (150/186)	74.93±7.81	15.17 ±2.99	23.18 ± .06
MCI	864 (354/510)	73.03±7.60	15.91±2.85	27.59±1.81
CN	521 (268/253)	74.25±5.79	16.37±2.70	29.06±1.14

3.3.2 Longitudinal Data Preprocessing

Initially, the data is preprocessed to alleviate any inconsistencies caused by utilizing different data modalities and various protocols. Subjects who have participated at all five consecutive intervals including baseline, six months after the first visit (M_6), twelve months after the first visit (M_{12}), twenty-four months after the first visit (M_{24}) and thirty-six months after the first visit (M_{36}) have been considered. In the initial step of the experiments, data cleaning [17], [18], mean centering, data normalization, missing feature handling, and univariate feature analysis have been performed to discard uninformative features. Furthermore, subjects whose medical diagnosis are not reported are removed from further analysis.

3.3.3 Simulation and Results

This study evaluates the performance of two RNN variations, LSTM and GRU, on the ADNI cohort for the two tasks of classification and regression. The experiment proceeds with the selection of historical records from subjects at three intervals (baseline, M_6, and M_{12}) to predict the status of the subjects in three future time points of M_{12}, M_{24}, and M_{36}. Estimating the MMSE scores of subjects is pursued as a regression problem and predicting the diagnosis labels is defined as a multiclass classification problem. The data has been split randomly into a 75% training set, a 10% validation set, and a 15% testing set. Grid search has been utilized to select the best hyperparameters for regression and classification networks separately. In order to feed the longitudinal feature space into the RNNs, the data has been framed in the tensor form of [samples, time steps, features] which in this case is 3-time steps of the 532 samples with 34 features involving MRI, PET, Cerebrospinal fluid (CSF) and cognitive test scores as provided in Table 2.

Table 3.2. Summary of Multimodal Features Utilized in This Study

Source	Features
Cognitive tests	Everyday Cognition (ECog) questionnaire measurements, FAQ, MOCA, RAVLT, CDRSB
MRI	Ventricular volume, Hippocampus volume, Whole Brain volume, Entorhinal Cortical thickness, Fusiform, Middle temporal gyrus, ICV
PET	FDG, PIB amyloid, AV45 amyloid
Genetic	APOE4
Demographic	Age, Gender, Education
CSF	Amyloid Beta, Phosphorylated Tau, Total Tau

The performance of LSTM and GRU, implemented using the Keras deep learning library, are compared with state-of-the-art methods. It is worth noting that conventional methods

cannot incorporate historical records of subjects for enhancing prediction accuracy. This limitation has been compensated by concatenating all three historical feature sets. Competing methods are then trained on this new feature space to find an individual direct map between the feature space from past intervals with the corresponding future time points.

As for the regression experiment, RMSE and R-Correlation factors are used as evaluation metrics to compare Ridge and SVR from the *Scikit-learn* library with LSTM and GRU and the results are reported in Table 3.

Table 3.3. Regression results

Algorithm	M12		M24		M36		Total
	RMSE	Corr	RMSE	Corr	RMSE	Corr	MSE
Ridge	2.07	0.58	2.66	0.62	2.99	0.63	6.82
SVR	2.14	0.59	2.86	0.61	3.17	0.58	7.68
LSTM	1.97	**0.63**	2.33	0.69	2.54	**0.72**	5.26
GRU	1.97	**0.63**	2.33	0.69	2.54	**0.72**	5.24
Ridge + FS*	2.02	0.62	2.67	0.65	2.93	0.65	6.65
SVR + FS*	2.16	0.60	2.76	0.65	3.26	0.62	7.70
LSTM + FS*	1.85	**0.63**	2.25	0.70	2.48	0.70	4.98
GRU + FS*	**1.82**	**0.63**	**2.21**	**0.71**	**2.44**	0.70	**4.77**

*Feature selection

Similarly, the classification problem is defined as predicting the diagnosis of subjects at three future time points based on three previous intervals. For the classification task, SVM from the *Scikit-learn* library is selected as the competitive alternative to evaluate the performance of the LSTM and GRU. F-score, precision, recall, accuracy has been utilized as the classification metrics and the results are summarized in Table 4.

From Tables 3 and 4, it can be observed that the LSTM and GRU on the original feature space demonstrate lower performance in comparison to the competitive methods in some cases. Incorporating *L1* has led to a noticeable improvement in prediction accuracy, which could be associated with the overfitting of networks. Since RNNs have a high number of variables and weights, they require a larger number of samples for training. The approach investigated here employs the *L1* feature selection to overcome the limited number of samples for training an effective network, which can predict the future status of AD subjects using their historical measurements.

Table 3.4. classification results

Method	M12				M24				M36			
	ACC	PRE	REC	F1	ACC	PRE	REC	F1	ACC	PRE	REC	F1
SVM	0.66±0.04	0.44±0.05	0.66±0.05	0.52±0.04	0.61±0.04	0.38±0.04	0.61±0.04	0.46±0.04	0.61±0.03	0.38±0.04	0.61±0.03	0.48±0.04
LSTM	0.84±0.10	0.86±0.06	0.84±0.10	0.81±0.16	0.82±0.12	0.77±0.22	0.82±0.12	0.79±0.18	0.80±0.09	0.84±0.06	0.80±0.09	0.78±0.15
GRU	0.61±0.09	**0.95±0.00**	0.60±0.09	0.74±0.07	0.37±0.06	**0.99±0.00**	0.37±0.06	0.53±0.06	0.61±0.04	**0.98±0.00**	0.61±0.04	0.75±0.03
LSTM + FS	**0.88±0.03**	0.89±0.02	**0.90±0.02**	**0.89±0.02**	**0.87±0.01**	0.86±0.04	**0.87±0.02**	**0.86±0.02**	**0.88±0.02**	0.87±0.03	**0.88±0.02**	**0.87±0.03**
GRU + FS	0.68±0.09	**0.95±0.00**	0.68±0.09	0.79±0.07	0.28±0.11	**0.99±0.00**	0.29±0.11	0.43±0.13	0.51±0.08	**0.98±0.00**	0.51±0.08	0.67±0.04

Chapter 4 A Distributed Multitask Multimodal Approach for the Prediction of Alzheimer's Disease in A Longitudinal Study

4.1 Introduction

In order to study the relative temporal changes in AD, there is a need to track pathophysiological changes in a large number of observations using Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), Cognitive assessment tests (COG), and Cerebrospinal Fluid (CSF) tests. However, acquiring all these tests within a large population is costly, time-consuming, and often difficult to maintain high protocol adherence given the dropout rate and missed follow-up visits given the patients' advanced age and severity, and extent of disease progression. Consequently, there are two kinds of challenges in studying longitudinal dynamics and related patterns in medical data. The first one is due to size irregularity because of missing measurements from a specific modality. The second is due to patients missing on follow-up visits or dropping out from the study. Among the many verified assessments that can diagnose the presence of AD and scale the severity of the progression, the Mini-Mental State Examination (MMSE) and the Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog) are the most common tests used in regression-based models [73][74]. One of the earliest works in this domain was done by Tierney et al. in 1996, who used logistic regression to predict the possibility of AD progression over a period of two years [75]. The study in [76] proposed a sparse linear regression model in conjunction with a group regularization technique. The model was applied across different brain regions to select the most informative longitudinal features. Their model predicts future cognitive clinical scores among MCI subjects over a period of 24-months. Similarly, Izquierdo et al [77] predicted cognitive scores using

stochastic gradient boosting of decision trees among 1,141 individuals for whom longitudinal clinical and imaging studies were available in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. In another study (Tabarestani et al., 2019), two different variations of recurrent neural networks (RNN), namely Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) have been applied using 1458 multimodal records of subjects from the ADNI database to predict AD progression. By leveraging the patients' historical records from the previous three time points, their model could track the disease progression trends of patients at three other subsequent time points with an accuracy that outperformed methods that relied solely on the baseline records.

Multitask learning, first proposed in 1997, is shown to improve performance by extracting the relationships between multiple similar tasks through the development of a statistical model [78]. It has since attracted a lot of attention in a variety of machine learning algorithms with application domains ranging from finance to bioinformatics [79][80]. This new research trend has delivered promising performance improvement in different categories, including, but not limited to multitask learning using kernel-methods [81], interpreting task relationship [82][83], developing probabilistic and statistical models [84][85], selecting features [86][87], learning features [88][89], feature hashing [90], and task grouping [91][92].

In recent years, multitask learning has been successfully applied to longitudinal clinical data to predict the progression of neurodegenerative diseases [93][94][95][96][97]. Compared to single-task learning, multitask learning uses a regression model for predicting the future status of patients at multiple time points. The basic assumption in these models is that an inherent correlation exists among multiple records of information, which are derived from the same subjects. These studies

demonstrated that capturing this inherent relatedness could improve the generalization of the final prediction model. For example, Zhou et al. in [96] developed convex and nonconvex fused group Lasso formulation as the regularization term of the multitask learning kernel. Their model could choose the most important sets of biomarkers from different time points to model the progression of AD. Similarly, Emrani et al. employed multitask learning to predict the progression of Parkinson's disease over a period of 4.5 years [98], and Jie et al. in [99] reported that using manifold regularized multitask feature learning could yield better classification performance and could identify disease-related regions in the brain deemed important for disease diagnosis. A Sparse Group Lasso with shared Subspace Multitask learning (SGLS-MTL) has been proposed by Cao et al. [100]. Their framework uses $\ell_{2,1}$ penalty, group $\ell_{2,1}$ penalty and subspace structure to capture the correlation between the tasks, the sparse feature representation, and the shared subspaces. They have applied their SGLS multitask learning method to predict cognitive scores and to detect potential predictive MRI biomarkers. Wang et al. in [101], proposed a high-order multitask feature learning algorithm to model the longitudinal trajectories of the cognitive measures of AD subjects based on neuroimaging biomarkers. They employed a non-smooth structured sparsity-inducing norm to utilize the correlation between the adjacent tasks (prediction of cognitive measures at two subsequent time points) and the interrelations that exist between the cognitive measurements. To capture the nonlinearity in the relationship between MRI neuroimaging features and cognitive scores, Cao et al. in [102] used the $\ell_{2,1} - l_1$ norm. By combining a joint sparsity regularization term with multitask learning, their proposed model produced more accurate results. Jie et al. in [103], introduced a group regularization term to the sparse linear regression model. They have also added two smoothness regularization terms to the objective function to ensure that the model keeps

the differences between the weight vectors belonging to adjacent time-points to be small. Their proposed model leveraged the prediction performance of the MMSE and ADAS-Cog scores from other existing sparse learning based models.

The neuropathological symptoms of AD in its different stages are complex and effectively combining different modalities does augment the prospects for a more accurate diagnosis. Although many studies are dealing with multimodal datasets, only a few discussed the discrepancy in the different representations of feature domains [86][104]. On the other hand, missing a screening test on a given visit or dropping out of an entire follow-up visit results in data scarcity in the multimodal database, a drawback experienced in most longitudinal studies. Therefore, to make a reliable prediction of MMSE changes over time, a distributed multimodal multitask framework is proposed in this study to overcome these types of data scarcity problems. In multitask learning, the regularizing term presumes that an equivalent degree of importance exists in the feature space. Therefore, if a positive correlation between the features from different modalities is not found, or if the features are not linearly correlated, the process may fail to identify relevant patterns. In this case, constructing a unified multitask learning model over the concatenated information may not be the optimal approach. To address this problem, a multitask modality-specific regression framework is proposed to predict future MMSE scores for up to 48 months while relying on measurements provided at baseline. Separate multitask regression matrices are trained for each modality to ensure that the coefficient matrices select the leading features extracted from the same modality between consecutive tasks.

The objective function of each regression model uses the correlation and sparsity pattern that exists between all tasks within each modality to improve the longitudinal prediction accuracy. In the second stage of the algorithm, a gradient boosting method

is implemented to take a concatenated series of temporal predictions from different modalities and improve the overall performance of the model by predicting a final score. This segregation of modalities in multitask modality-specific regression offers the following advantages:

- Resolves issues related to nonlinear or negative correlations between different feature spaces, which could hinder the performance of multitask learning.
- Provides an error propagation-free framework through a combination of modality-specific multitask learning and gradient boosting. This approach assumes that potential errors might exist in the measurements of a specific modality that originated from capturing, processing, or extracting data. Concatenating data from different modalities will thus increase the risk of spreading this error to the fused feature space. Hence, by training separate models and performing a majority vote for the distributed models, the source of error can be detected and consequently prevented from propagating into the fused feature space.
- Overcomes the missing data challenge by projecting a high-dimensional and highly sparse input feature space into multiple low-dimensional and less-sparse spaces. This ensures that the independent coefficient matrices can collectively determine and order the most important biomarkers in the whole dataset.

It is worth noting that the motivation of the model as envisioned is to predict the trajectories of cognitive decline for subjects without any preliminary diagnosis and regard to the historical records. Thus, the applicability of the proposed framework in terms of providing prediction from baseline information makes it different from methods that need at least a few historical records to be available. For example, Zhu et al in [105] proposed a method for early diagnosis of AD by analyzing longitudinal MRI

records and constructing a new feature space from the mean and the difference between the first and last visits measurements. While involving historical records from patients in the training phase may improve the prediction accuracy, it limits the applicability of the model to only those patients with available medical records.

4.2 Background

4.2.1 Problem Description

The development of Alzheimer's Disease takes place along a trajectory spanning several years with transitions phases that vary from one patient to another. Therefore, in longitudinal AD studies, individuals repeat medical screening tests at multiple follow-up visits and their MMSE scores are recorded and analyzed at each visit. MMSE, with a range of 0 to 30, is the screening test most commonly used for memory and cognitive evaluation. While it is not intended to replace neurological diagnostic labels, it is used to validate the reliability of medical examinations or to evaluate temporal cognitive decline in people suffering from AD. Early intervention plans are effective only if the earliest manifestations of AD are identified at the onset of the disease. Therefore, predicting future trajectories of MMSE scores enables doctors to identify future pathological levels of memory and cognitive impairment. Consequently, the initial objective is to predict the MMSE scores ($\boldsymbol{b}$) of subjects, by finding the best model g, such that $g: \boldsymbol{b} = \boldsymbol{A}\boldsymbol{w}$, where $\mathbf{w}$ is the regression coefficient and $\boldsymbol{A}$ is the baseline information of the subjects. In support of the proposed approach introduced in Section 4.3, the required mathematical background is introduced in sub-sections 4.2.2 through 4.2.4.

4.2.2 Single Task Regression

Let $\boldsymbol{A} \in \mathbb{R}^{N\times P}$ be a matrix consisting of N subjects with P features describing each subject, with $\boldsymbol{b}^t \in \mathbb{R}^{N\times 1},\ t = 1,2,..,T$ defining the clinical scores of those N subjects at the t^{th} time point. The problem of predicting the clinical scores at multiple future time points could be formulated as solving T different regression models as $g^t: \boldsymbol{A} \in \mathbb{R}^{N\times P} \rightarrow \boldsymbol{b}^t \in \mathbb{R}^{N\times 1}, t = 1,2,..T$.

In the simplest form, these T regression problems can be solved using the following *Ridge* regression formula:

$$\widehat{\boldsymbol{w}}^t = \arg\min_{\ddot{\boldsymbol{w}}} \|\boldsymbol{s}\odot(\boldsymbol{b}^t - \mathbf{A}\ddot{\boldsymbol{w}})\|_2^2 + \theta\|\ddot{\boldsymbol{w}}\|_2^2 \quad (1)$$

where $\widehat{\boldsymbol{w}}^t \in \mathbb{R}^{P\times 1};\ t = 1,2,\dots,T$ are T independent coefficient vectors calculated by solving the minimization problem in Eq. (1). The $\ddot{w}$ is used as a variable under the arg min function to avoid any confusion with $\boldsymbol{w}$ (the perfect target) and $\widehat{\boldsymbol{w}}$ (the estimated target). In other words, at the last iteration, $\ddot{w}$ that minimizes the arg min function is set as the best estimate $\widehat{\boldsymbol{w}}$ (i.e., $\widehat{\boldsymbol{w}} \leftarrow \ddot{\boldsymbol{w}}$). Symbol $\odot$ defines the component-wise multiplier and vector $\boldsymbol{s} \in \mathbb{R}^{N\times 1}$ defines the missing target values; meaning that $s_n = 0$ if the target value of the n^{th} patient is missing at the t^{th} time point, and $s_n = 1$ if the target value of the n^{th} patient is available at that same time point. In Eq. (1), the $\|\ddot{\boldsymbol{w}}\|_2^2$ is the squared ℓ_2 norm of the coefficient vector $\ddot{\boldsymbol{w}}$, which is controlled by tuning parameter θ. Recall that the p norm of a vector $\boldsymbol{x} \in \mathbb{R}^{K\times 1}$ with $\boldsymbol{x} = [x_1, x_2, \dots, x_K]'$ is defined as:

$$\ell_p = \|\boldsymbol{x}\|_p = \left(\sum_k^K |x_k|^p\right)^{1/p} \quad (2)$$

The penalty term $\theta\|\ddot{\boldsymbol{w}}\|_2^2$, controls the amount of coefficient shrinkage and forces the variance to be close to zero in order to reduce the mean-squared error. Another solution

in finding g is to employ the *Lasso* regression formulated as a constrained minimization problem as follows:

$$\widehat{\boldsymbol{w}}^t = \arg\min_{\ddot{\boldsymbol{w}}} \|\boldsymbol{s}\odot(\boldsymbol{b}^t - \mathbf{A}\ddot{\boldsymbol{w}})\|_2^2 + \theta\|\ddot{\boldsymbol{w}}\|_1 \quad (3)$$

In this formula, increasing θ forces the majority of coefficients in $\ddot{\boldsymbol{w}}$, which are associated with features deemed not to be important, to be close to zero and shrink the non-zero coefficients simultaneously. The only difference between these two regression models is in squaring the ℓ_2 norm in *Ridge* regression and using ℓ_1 as the penalty terms in *Lasso* regression, which increases the sparsity of the coefficients.

4.2.3 Multitask Regression

Another way to tackle the problem of predicting cognitive scores at multiple time points is to employ multitask learning. In the single-task approach, each task is defined as predicting MMSE scores at a single time point and several independent regression models are trained separately to perform prediction for each time point. On the other hand, the multitask approach utilizes the similarities between different tasks to find a more accurate regression model that can carry out multiple prediction tasks. This means that in multi-task learning all the MMSE scores belonging to the T time points will be calculated simultaneously.

Multitask learning can be mathematically formulated as a predictor $G: \boldsymbol{A} \in \mathbb{R}^{N\times P} \rightarrow \boldsymbol{B} \in \mathbb{R}^{N\times T}$ where $\boldsymbol{B} = [\boldsymbol{b}^1, \boldsymbol{b}^2, \dots, \boldsymbol{b}^T]$ is the target values of N subjects at T time points. This multitask predictor G can be modeled using a weight matrix $\boldsymbol{W} = [\boldsymbol{w}^1, \boldsymbol{w}^2, \dots, \boldsymbol{w}^T]$ where $\boldsymbol{W} \in \mathbb{R}^{P\times T}$. In computing the $\boldsymbol{W}$ matrix, one approach is to solve the convex optimization problem as expressed in Eq. (4), also known as the convex fused sparse group Lasso (cFSGL) [96].

$$\widehat{\mathbf{W}} = \arg\min_{\ddot{\mathbf{W}}} \left\| \boldsymbol{S} \odot (\mathbf{B} - \mathbf{A}\ddot{\mathbf{W}}) \right\|_F^2 + \theta \left\| \ddot{\mathbf{W}} \right\|_{\mathbf{1}} + \lambda \left\| \ddot{\mathbf{W}} \right\|_{2,1} + \eta \left\| \boldsymbol{R}\ddot{\mathbf{W}}' \right\|_{\mathbf{1}} \quad (4)$$

where $\odot$, as defined earlier, is the component-wise multiplier and matrix $\boldsymbol{S} \in \mathbb{R}^{N \times T}$ specifies the missing target values, in which $S_{n,t} = 0$ if the target value of the n^{th} patient is missing at the t^{th} time point, and $S_{n,t} = 1$ if the target value is available. $\widehat{\mathbf{W}}$ is the estimation of the $\mathbf{W}$ achieved by solving the minimization problem. Terms θ, λ, and η are the hyperparameters that control the effect of each regularization term in the cost function and are optimized during the training phase to improve the performance of the algorithm. $\|\boldsymbol{W}\|_1$ is the Lasso penalty term and $\|\boldsymbol{W}\|_F^2$ is the squared Frobenius norm and the $\left\|\ddot{\mathbf{W}}\right\|_{2,1}$ is known as the Group Lasso penalty. Moreover, $\|\boldsymbol{R}\boldsymbol{W}'\|_1$ is the Fused Group Lasso penalty, and $\boldsymbol{R}$ is $(T-1) \times T$ sparse matrix is interpreted as a descriptor of the relatedness between different tasks. Assuming each task as a node in a graph, a relationship between every two tasks is represented by a connection between their corresponding nodes. This penalty term controls the transition between neighboring tasks and forces the transition within successive tasks to remain small (a process also known as temporal smoothness). In other words, $R_{i,j} = 0$ indicates that the task assigned to node i is not related to the task assigned to node j , while $R_{i,j} = \alpha$ indicate that task i and task j are associated with each other with a degree of α. In the proposed model, this parameter restrains the variation of predicted cognitive scores in neighboring time steps, meaning that trajectories of MMSE scores at two consecutive time points cannot have spikes. In order to solve Eq. (4), the accelerated gradient method (AGM) was used, which is available in the MALSAR package [106].

Another approach for finding the weight matrix $\boldsymbol{W}$ is to use the non-Convex Fused Sparse Group Lasso (nFSGL1) as formulated in [96]:

$$\widehat{\mathbf{W}} = \arg\min_{\ddot{\mathbf{W}}} \left\| \boldsymbol{S} \odot (\mathbf{B} - \mathbf{A}\ddot{\mathbf{W}}) \right\|_F^2 + \eta \left\| \boldsymbol{R}\ddot{\mathbf{W}}' \right\|_1 + \theta \sum_{i=1}^{P} \sqrt{\|\ddot{\mathbf{w}}_i\|_1} \tag{5}$$

where $\ddot{\boldsymbol{w}}_i$ is the i^{th} row of $\ddot{\mathbf{W}}$. The convex and non-convex Fused Group Lasso formulas allow for joint feature selection across all tasks while selecting distinct feature sets for each task.

The joint selection of the coefficients in **W** could also be penalized in the form of $\ell_{2,1}$-norm with least square loss. Thus, finding the optimal **W** can be formulated as:

$$\widehat{\mathbf{W}} = \arg\min_{\ddot{\mathbf{W}}} \tfrac{1}{2} \left\| \boldsymbol{S} \odot (\mathbf{B} - \mathbf{A}\ddot{\mathbf{W}}) \right\|_F^2 + \lambda_1 \left\| \ddot{\mathbf{W}} \right\|_{2,1} + \lambda_2 \left\| \ddot{\mathbf{W}} \right\|_F^2 \tag{6}$$

To incorporate global and local information in the feature set with a sparse regression method, Zhu et al in [107] reformulated the objective function in equation (6) as follows:

$$\widehat{\mathbf{W}} = \arg\min_{\ddot{\mathbf{W}}} \tfrac{1}{2} \left\| \boldsymbol{S} \odot (\mathbf{B} - \mathbf{A}\ddot{\mathbf{W}}) \right\|_F^2 + \lambda_1 tr(\ddot{\mathbf{W}}' \boldsymbol{A}' \boldsymbol{L} \mathbf{A}\ddot{\mathbf{W}}) + \lambda_2 \left\| \ddot{\mathbf{W}} \right\|_{2,1} \tag{7}$$

where λ_1 and λ_2 are the regularization parameters and *tr* (.) denotes the trace operator. Here, with **R** being the adjacency matrix, the Laplacian matrix L can be defined as:

$$\mathbf{L} = \mathbf{D} - \mathbf{R} \tag{8}$$

where **D** is the symmetric diagonal matrix in which the diagonal elements D_{ii} = 1 and all the other non-diagonal entries are 0. Zhu et al. in [108] proposed an iterative method for finding the solution of multitask problem, i.e. $\boldsymbol{W}$, to reduce the number of hyperparameters that must be learned in the multitask learning problem. The objective

function in this proposed approach is to find the $\boldsymbol{w}^t$ values through the following formulation:

$$\widehat{\mathbf{W}} = \arg\min_{\ddot{\boldsymbol{w}}^t,\ \bar{\boldsymbol{w}}} \sum_t^T \alpha^t \left(\|\boldsymbol{s} \odot (\boldsymbol{b}^t - \mathbf{A}\ddot{\boldsymbol{w}}^t)\|_2^2 + \|\ddot{\boldsymbol{w}}^t - \bar{\boldsymbol{w}}\|_{2,1} \right) + \lambda_2 \left\|\ddot{\mathbf{W}}\right\|_1 \quad (9)$$

where $\bar{\boldsymbol{w}}$ is the mean vector of $\ddot{\boldsymbol{w}}^t (t = 1, 2, \ldots T) \in \boldsymbol{W}$. For each task *t*, the weights of each task are denoted as α^t are calculated automatically with the following equation:

$$\alpha^t = \frac{1}{2\sqrt{\|\boldsymbol{s} \odot (\boldsymbol{b}^t - \boldsymbol{A}\ddot{\boldsymbol{w}}^t)\|_2^2 + \|\ddot{\boldsymbol{w}}^t - \bar{\boldsymbol{w}}\|_{2,1}}} \quad (10)$$

Employing the centralized regularization in the objective function of (9) balances the variances of the coefficients in $\boldsymbol{w}^t$ by penalizing them separately using α^t.

4.2.4 Gradient Boosting

Ensemble models are effective in various prediction tasks by grouping a set of weak learners to construct a more powerful learner. Bagging and boosting are the two mainstream techniques in ensemble learning methods. The former creates independent and uncorrelated learners on subsets of data and generates the final result by voting or averaging the outcomes of independent learners. On the contrary, the latter generates a collection of weak learners, in which the predictors are trained sequentially rather than separately. In boosting methods, the goal is to utilize the error of the previous learners to develop a more efficient model for the next learner. With training the learners sequentially, subsets of data do not have the chance to concurrently affect all the learners. The algorithm invests a larger weight on the samples that were classified inaccurately, forcing the hypothesis of the next weak learners to precisely analyze those tough samples and eventually improve the performance of the model.

An extension of the boosting methods is gradient boosting, which is a supervised machine learning technique based on regression, classification, and ranking. It uses the gradient descent optimization technique to find the global or local minima of the cost function. Using a sequence of weak learners, Gradient Boosting (GB) trains a machine to fit a model on the input feature space such that each learner improves the prediction accuracy of the previous ones. Through multiple iterations, gradient boosting develops a single strong learner by combining multiple weak learners [109][110]. In the proposed method, GB constructs the final stage of the framework to improve the prediction accuracy by successively fitting a more accurate model on the residuals of the previous step. This procedure will continue until it achieves a highly accurate model. Sub-sections 4.3.3 and 4.3.4 provide more details on the role of GB in the context of the proposed framework.

4.3 Method

4.3.1 Notations and parameters

Through the rest of this chapter, matrices are denoted as bold uppercase letters, and vectors are denoted as italic bold letters. Matrices $\boldsymbol{X}_m^t \subseteq \mathbf{X}$ and $\Omega_m^t \subseteq \Omega$ are the feature space and patients' roster ID associated with the subjects who have been examined at time point t with modality test m. For these subjects, $\boldsymbol{y}^t$with $t = 1, 2, \ldots, \mathrm{T}$ are their respective cognitive scores (independent from the source of the modality). Similarly, $\boldsymbol{\mathcal{F}}$ is the risk factor matrix consisting of age, gender, years of education, and APOE4 factors for all patients. It is noted that the (') notation denotes transposition and should not be confused with t = 1, 2, …, T which defines the different time points in the longitudinal study, where T denotes the 48th month.

4.3.2 Method Overview

Tracking future MMSE scores reveals a subtle but progressive decline in cognitive levels of individuals through the different stages of AD and informs on the nature of the transition phases of the disease. However, prognostication of AD progression, regardless of the label associated with the subject at baseline, remains challenging, especially in a multimodal platform. Certain modalities have shown a relatively higher impact on the asymptomatic or symptomatic phases of AD. This promoted the use of multimodal biomarkers to improve the accuracy of identifying neurobiological and clinical symptoms of the disease. However, the interactions and correlations between the biomarkers from complementary modalities remain intricate. Furthermore, longitudinal datasets continue to suffer from the missing data challenge.

Considering the data scarcity and the discrepancy in the correlation matrix associated with the heterogeneous multimodal longitudinal dataset, we propose to utilize the modality-specific multitask coefficient matrix. These unique multitask coefficient matrices are trained over different sets of biomarkers extracted from each modality to model the temporal interaction between the baseline features and the transitions of the cognitive scores at successive time points.

The strength and capability of different modalities in tracking the progression of AD are still inconclusive. Therefore, granting equal contribution (or equal weight) to the predictive biomarkers from different modalities increases the chance of achieving better prediction accuracy. This is accomplished by capturing the complex yet the effective correlation between important modality-exclusive features and eliminating the effect of all other extraneous ones. Next, the initial outcomes of these cooperative multitask learners are fused with risk factors, which are assumed as time-invariant information. Finally, a gradient boosting kernel is trained over this new collective data representation

to leverage the prediction accuracy through ensemble learning and looking into sparse and interpretable solutions. In the next section, we will go through the setup of our multimodal-multitask model.

4.3.3 Method formulation

Suppose that $\boldsymbol{X} \in \mathbb{R}^{N \times P}$ is the multimodal feature space and $\boldsymbol{Y} = [\boldsymbol{y}^1, \boldsymbol{y}^2, \dots, \boldsymbol{y}^T]$ is representing the cognitive trajectories of these N subjects through T time steps. For each interval *t*, $\boldsymbol{X}^t \subseteq \boldsymbol{X}$ is the set of subjects who are chosen based on Ω^t, the roster ID of the population $\boldsymbol{y}^{\mathrm{t}}$. It is worth noting that some subjects may have not returned for the follow-up visit at t^{th} time point and therefore $\Omega^t < \Omega$ is possible. Considering M as the total number of modality sources, $\boldsymbol{X}^t$ and $\boldsymbol{y}^{\mathrm{t}}$ are decomposed into M subgroups, thus constructing $T \times M$ pairs of $\{(\boldsymbol{X}_m^t, \boldsymbol{y}_m^t), m = 1, 2, \dots, M,\ t = 1, 2, \dots,\ T\}$, where each pair of $(\boldsymbol{X}_m^t, \boldsymbol{y}_m^t)$ are the m^{th} single-modality measurements associated with the t^{th} time point.

The single task regression method will be extended to the $T \times M$ optimization problems to calculate $\boldsymbol{w}_m^t$ by solving equations (11) and (12).

$$\widehat{\boldsymbol{w}}_m^t = \arg\min_{\ddot{\boldsymbol{w}}} \|(\boldsymbol{y}_m^t - \mathbf{X}_m^t \ddot{\boldsymbol{w}})\|_2^2 + \theta \|\ddot{\boldsymbol{w}}\|_2^2 \tag{11}$$

$$\widehat{\boldsymbol{w}}_m^t = \arg\min_{\ddot{\boldsymbol{w}}} \|(\boldsymbol{y}_m^t - \mathbf{X}_m^t \ddot{\boldsymbol{w}})\|_2^2 + \theta \|\ddot{\boldsymbol{w}}\|_1 \tag{12}$$

where $\widehat{\boldsymbol{w}}_m^t \in R^{P_m \times 1}$ is the $\widehat{\boldsymbol{w}}_m$ estimate at the t^{th} time point.

In the multitask learning approach, the objective function will be extended to $\mathrm{G_m}: \boldsymbol{X}_m^t \rightarrow \overline{\overline{\boldsymbol{Y}}}_m$where $\overline{\overline{\boldsymbol{Y}}}_m \in \mathbb{R}^{N \times T}$ is the concatenated matrix $\overline{\overline{\boldsymbol{Y}}}_m = [\overline{\boldsymbol{y}}_m^1, \overline{\boldsymbol{y}}_m^2, \dots, \overline{\boldsymbol{y}}_m^T]$ with $\overline{\boldsymbol{y}}_m^t$ being the extended versions of their corresponding $\boldsymbol{y}_m^t$, in which the unavailable test scores of the patients are represented by zero

values. The size discrepancy in $\bar{y}_m^t$, which is a consequence of missing modalities and dropout is illustrated in Figure 4.1.

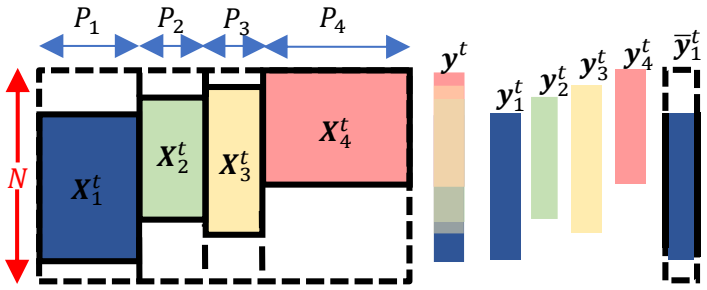


Figure 4.1. An illustrative example of size discrepancy in a longitudinal multimodal dataset. Available measurements extracted from each modality are shown with colored boxes and the missing information is displayed in the blank sections.

In this figure, patterns of missing values and arrangements of available information from four modalities are represented over a fixed time period. Using a modality-specific approach, the objective function of multitask learners will be reformulated to calculate M number of $\boldsymbol{W}_m \in R^{P_m \times T}$ where $\boldsymbol{W}_m = [\boldsymbol{w}_m^1, \boldsymbol{w}_m^2, \dots, \boldsymbol{w}_m^T]$. Thus, the cFSGL (convex Fused Sparse Group Lasso) problem can be formulated as follows:

$$\widehat{\boldsymbol{W}}_m = \arg\min_{\ddot{W}} \left\| \boldsymbol{S} \odot (\bar{\bar{\boldsymbol{Y}}}_m - \boldsymbol{X}_m^1 \ddot{\boldsymbol{W}}) \right\|_F^2 + \theta \left\| \ddot{\boldsymbol{W}} \right\|_1 + \lambda \left\| \ddot{\boldsymbol{W}} \right\|_{2,1} + \eta \left\| \boldsymbol{R} \ddot{\boldsymbol{W}}' \right\|_1 \quad (13)$$

And based on nFSGL1 (non-Convex Fused Sparse Group Lasso), the objective function will be formulated as follows:

$$\widehat{\boldsymbol{W}}_m = \arg\min_{\ddot{W}} \left\| \boldsymbol{S} \odot (\bar{\bar{\boldsymbol{Y}}}_m - \boldsymbol{X}_m^1 \ddot{\boldsymbol{W}}) \right\|_F^2 + \eta \left\| \boldsymbol{R} \ddot{\boldsymbol{W}}' \right\|_1 + \theta \sum_{i=1}^{P_m} \sqrt{\| \ddot{\boldsymbol{w}}_i \|_1} \quad (14)$$

Using a similar approach, equations (6), (7), (9), and (10) will be reformulated respectively as follows:

$$\widehat{\boldsymbol{W}}_m = \arg\min_{\ddot{\boldsymbol{W}}} \frac{1}{2} \left\| \boldsymbol{S} \odot (\bar{\bar{\boldsymbol{Y}}}_m - \boldsymbol{X}_m^1 \ddot{\boldsymbol{W}}) \right\|_F^2 + \lambda_1 \left\| \ddot{\boldsymbol{W}} \right\|_{2,1} + \lambda_2 \left\| \ddot{\boldsymbol{W}} \right\|_F^2 \quad (15)$$

$$\widehat{\mathbf{W}}_m = \arg\min_{\ddot{\mathbf{W}}} \frac{1}{2} \left\| \boldsymbol{S} \odot (\bar{\bar{\boldsymbol{Y}}}_m - \boldsymbol{X}_m^1 \ddot{\boldsymbol{W}}) \right\|_F^2 + \lambda_1 tr(\ddot{\boldsymbol{W}} \acute{\boldsymbol{X}}_m^1 \boldsymbol{L} \boldsymbol{X}_m^1 \ddot{\boldsymbol{W}}) + \lambda_2 \left\| \ddot{\boldsymbol{W}} \right\|_{2,1} \quad (16)$$

$$\widehat{\mathbf{W}}_m = \arg\min_{\ddot{\boldsymbol{w}}^t,\ \bar{\boldsymbol{w}}} \sum_t^T \alpha^t \left(\| \boldsymbol{s} \odot (\bar{\boldsymbol{y}}_m^t - \boldsymbol{X}_m^1 \ddot{\boldsymbol{w}}^t) \|_2^2 + \| \ddot{\boldsymbol{w}}^t - \bar{\boldsymbol{w}} \|_{2,1} \right) + \lambda_2 \left\| \ddot{\boldsymbol{W}} \right\|_1 \quad (17)$$

$$\alpha^t = \frac{1}{2\sqrt{\| \boldsymbol{s} \odot (\bar{\boldsymbol{y}}_m^t - \boldsymbol{X}_m^1 \ddot{\boldsymbol{w}}^t) \|_2^2 + \| \ddot{\boldsymbol{w}}^t - \bar{\boldsymbol{w}} \|_{2,1}}} \quad (18)$$

The flowchart of the proposed method in the training stage is illustrated in Figure 4.2. In this figure, step 1 represents the training process for the modality-specific regression coefficient matrices $\widehat{\boldsymbol{W}}_m$. The input space is constructed by T stack of modality-specific feature spaces, $\mathbf{X}_m^t$, t = 1, 2, …, T and the targets are their respective cognitive scores characterized as $\widehat{\boldsymbol{y}}_m^t$. At the end of the training stage, step 1 generates M modality-specific multitask learning regression coefficient matrices, $\widehat{\boldsymbol{W}}_m \in R^{P_m \times T}$ for $m = 1, 2, \dots, M$, which are comprised of $\widehat{\boldsymbol{w}}_m^t$ for $t = 1, \dots, T$ in the form of $\widehat{\boldsymbol{W}}_m = [\widehat{\boldsymbol{w}}_m^1, \widehat{\boldsymbol{w}}_m^2, \dots, \widehat{\boldsymbol{w}}_m^T]$. Consequently, using $\mathbf{X}_m^t$ as input measurements, the initial prognostications at time point t are established as:

$$\widehat{\boldsymbol{y}}_m^t = \mathbf{X}_m^t \times \widehat{\boldsymbol{w}}_m^t \quad (19)$$

for $m = 1, 2, \dots, M$ and $t = 1, 2, \dots, T$.

Modality-wise multitask coefficient matrices capture the mutual relationships between the feature spaces and cognitive score trajectories. This provides a powerful tool in obtaining the inter-modality correlations and examining the predictive power of each modality exclusively. To take advantage of the information provided from each source of modality, the outcomes of the multitask models along with risk factor parameters are combined to form the input space for gradient boosting. It is worth noting that the risk factor parameters, do not carry the unpredictable temporal pattern

as in the other biomarkers. In order to reduce unnecessary computational costs, risk factor parameters have not been processed with multitask learning models and have been added to the second stage of the model. Step 2 in Figure 4.2 shows the preparation of the data for the second stage of the method.

For the dataset used here, it is observed that if the PET measurements are available for a group of subjects, the MRI measurements are also available for that group, but the opposite is not necessarily true. Therefore, five configurations of possible modality combinations are considered in this study: (1) MRI-PET, (2) MRI-PET-CSF, (3) MRI-PET-COG, (4) PET-COG-CSF, and (5) MRI-PET-COG-CSF.

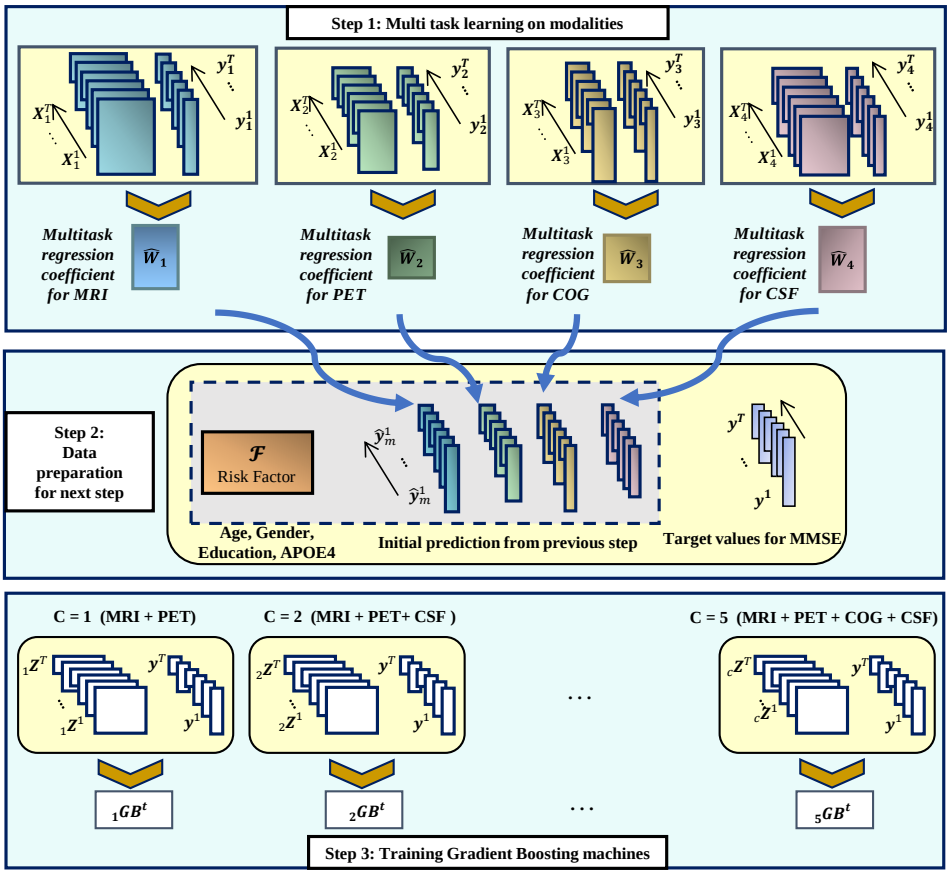
The Ω_m^t are the sets of roster IDs from subjects that have participated in test m at the t^{th} time point and ${}_c\Omega^t$ is the intersection between all Ω_m^t with respect to their availability in the c^{th} modality combination. Considering c as an indicator of the modality combination, the GB machines are developed as ${}_cGB^t : {}_c\boldsymbol{Z}^t \longrightarrow \boldsymbol{y}^t$ for $c = 1, \dots, 5$ and $t = 1, \dots, T$ over the set of ${}_c\Omega^t$. In which ${}_c\boldsymbol{Z}^t$ is the new feature space for the c^{th} GB machine and is constructed by concatenating ${}_c\widehat{\boldsymbol{y}}_m^t$ and ${}_c\boldsymbol{\mathcal{F}}^t$, which are the initial predictions and risk factors for the population of ${}_c\Omega^t$. This process has been demonstrated in step 3 of Figure 4.2.

For example, if the available modalities are MRI and PET, then c = 1. Meaning that in stage 1, only the modality-specific regression coefficient matrices of $\widehat{\boldsymbol{W}}_1$ and $\widehat{\boldsymbol{W}}_2$ can provide the initial predictions as $\widehat{\boldsymbol{y}}_1^t$ and $\widehat{\boldsymbol{y}}_2^t$. Based on their respective roster IDs, ${}_c\Omega^t$, the input space ${}_1\boldsymbol{Z}^t = [\,{}_1\boldsymbol{\mathcal{F}}^t,\ {}_1\widehat{\boldsymbol{y}}_1^t,\ {}_1\widehat{\boldsymbol{y}}_2^t]$ is constructed in step 2. Then the ${}_1\boldsymbol{Z}^t$ and their corresponding sets of cognitive scores, $\boldsymbol{y}^t$, will be used to train the corresponding ${}_1GB^t$ at step 3.

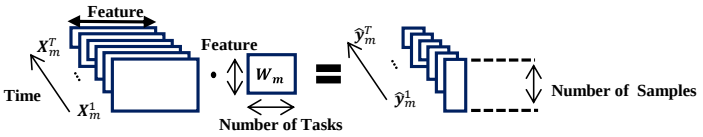
4.3.4 Test Scenario

Suppose that we want to predict the MMSE score at time point t and the patient has completed three modality tests. The available measurements from this patient are thus ($\boldsymbol{x}_1 \in R^{1\times P_1}$) extracted from MRI, ($\boldsymbol{x}_2 \in R^{1\times P_2}$) extracted from PET, ($\boldsymbol{x}_4 \in R^{1\times P_4}$) extracted from the CSF test and a vector $\boldsymbol{r}$ containing the risk factor parameters for this patient. In this scenario, the COG modality which is $\boldsymbol{x}_3$ is not available.

In the first step of the proposed model, modality-wise coefficient matrices will provide the most accurate predictions possible from the measurements of one modality through multitask learning. By feeding $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_4$ to their respective modality-wise coefficient matrices, the initial predictions can be calculated as $\hat{y}_1^t = \boldsymbol{x}_1 \times \widehat{\boldsymbol{w}}_1^t$, $\hat{y}_2^t = \boldsymbol{x}_2 \times \widehat{\boldsymbol{w}}_2^t$ and $\hat{y}_4^t = \boldsymbol{x}_4 \times \widehat{\boldsymbol{w}}_4^t$. Next, the initial predictions of $\hat{y}_1^t, \hat{y}_2^t, \hat{y}_4^t$ and risk factors ($\mathcal{F}$) will be concatenated to form the new feature vector ${}_c\boldsymbol{Z}^t = [\boldsymbol{r}, \hat{y}_1^t, \hat{y}_2^t, \hat{y}_4^t]$ where $c = 2$ indicates the mode for modality combination (i.e., MRI-PET-CSF). Then in the second step, gradient boosting employs a boosting approach to ensemble the outcomes from different modalities, determine the correlation among them and reduce their prediction error. The final estimation will be achieved by using the ${}_cGB^t$ machine as $\hat{y}^t = {}_2GB^t({}_2\boldsymbol{Z}^t)$. While incomplete samples with missing intervals are taken care of, through the first step of the algorithm, the second step of the proposed method deals with the missing modalities and the complex relationship between them. The gradient boosting incorporates the predictive power of salient biomarkers from each modality, models the intra-correlation between them, and adjusts the prediction error to improve the final accuracy.



(a)



(b)

Figure 4.2. (a) Flowchart of the proposed approach in the training phase, (b) Defining the dimensions in multitask formulation for step 1.

Table 4.1. Demographic characteristic of the studied subjects. valued are specified as mean±standard deviation

Category	Subjects (f/m)	Age	Education(year)	APOE (0/1/2)	MMSE
CN	206/209	74.77±5.74	16.27±2.73	300/103/11	29.07±1.12
MCI	354/510	73.03±7.60	15.91±2.85	427/340/94	27.59±1.81
AD	150/186	74.92±7.81	15.17±2.99	113/156/65	23.18±2.05
MCI to AD	2/3	78.50±2.59	16.40±2.61	1/4/0	26.00±1.58

4.4 Results and Discussion

4.4.1 Data

ADNI established the following Mini-Mental Exam (MMSE) and Clinical Dementia Rating (CDR) cut off scores to interpret the AD spectrum:

- MMSE of 30 and CDR of 0 is described as cognitively no dementia,
- MMSE of 29-26 and CDR of 0.5 is associated with questionable dementia,
- MMSE of 25-21 and CDR of 1.0 is associated with mild dementia,
- MMSE of 20-11 and CDR of 2.0 is associated with moderate dementia,
- MMSE of 10-0 and CDR of 3.0 is determined as severe dementia.

The experiments in this study used multimodal longitudinal data from 1620 subjects who were enrolled for up to 6 visits in a 4-year time span. This population consists of a total of 1620 subjects with 864 participants with mild cognitive impairment (MCI), 415 cognitively normal subjects (CN), 336 individuals with dementia (AD), and 5 participants whose status changed from mild cognitive impairment to dementia at baseline (MCI to AD conversion). All samples used in this analysis are in the range of 54.4 to 90.3 years old, with 44% female and 56 % male. The majority of the 93.24 % of the population were identified as white, 3.95% as black, and the rest were recognized

either as Asian, Indian/Alaskan, or belonging to more than one ethnicity. 76% reported their marital status as married, 12.61% as widowed, and the rest of the participants were represented as either never married or their status of marriage was recorded as unknown. Table 1 summarizes the demographic characteristics of the ADNI cohort used in this study based on the category of the disease. For the APOE column, the (0, 1, 2) values refer to the number of ε4 alleles in the APOE genotype.

4.4.2 Importance of Data Modality and Structure of the Experimental Set-Up

In preparing the data, subjects were partitioned into four categories: individuals who had completed the MRI scanning, individuals with PET scans, individuals with CSF analysis, and individuals with cognitive screening tests. The features extracted from each screening test, and the number of subjects in different time periods, are summarized in Table 2. In relation to time *t*, *t*=1 means time point at baseline or T1, *t*=2 refers to the time point at the 6th month or T6, *t*=3 refers to the time point at the 12^{th} month or T12, *t*=4 refers to the time point at 24^{th} month or T24, *t*=5 for the time point at 36^{th} month or T36 and finally for *t*=T, for the last time point at the 48^{th} month or T48. The importance of each data modality in the proposed multitask multimodal approach is reflected in the features that were selected for each modality as shown in Table 2.

Table 4.2. Summary of ADNI dataset, the number of observations in each follow-up visit and the features extracted from each modality

Source*	Number of observations						Features
	T1	T6	T12	T24	T36	T48	
MRI	1465	1333	1191	987	617	451	Ventricular volume, Hippocampus volume, Whole Brain volume, Entorhinal Cortical thickness, Fusiform, Middle temporal gyrus, and intracranial volume (ICV)
PET	1127	1009	892	714	429	335	FDG, Pittsburgh Compound-B (PIB), AV45
Cognitive Test**	1525	1357	1207	997	627	456	Rey Auditory Verbal Learning Test (RAVLT Immediate, RAVLT Learning, RAVLT Forgetting, RAVLT Perc Forgetting), Functional Activities Questionnaires (FAQ), Everyday Cognition (Ecog) scales: (EcogPtMem, EcogPtLang, EcogPtVisspat, EcogPtPlan, EcogPtOrgan, EcogPtDivatt, EcogPtTotal,EcogSPMem, EcogSPLang, EcogSPVisspat, EcogSPPlan, EcogSPOrgan, EcogSPDivatt, and EcogSPTotal)
CSF	1014	914	806	662	404	305	Amyloid Beta (ABETA), Phosphorylated Tau Protein (PTAU), and Total Tau Protein (TAU)
Risk factors	1737						Age, gender, years of education, and APOE4

* In this table MRI refers to Magnetic Resonance Imaging, PET refers to Positron Emission Tomography, COG refers to Cognitive assessment tests and CSF refers to Cerebrospinal Fluid test.

**The Mini-Mental State Examination (MMSE) and Clinical Dementia Rating Sum of Boxes (CDRSB) scores (*since initially used for labelling subjects*) and Alzheimer's Disease Assessment Score (ADAS11, ADAS13) and the Montreal Cognitive Assessment (MoCA) (*since highly correlated with MMSE*) were excluded from the feature set in the training and testing phases of the proposed prediction model.

Observe the decreasing number of observations made at subsequent time points in this ADNI longitudinal study, which highlights the missing data challenge. For this study, through the MRI imaging modality, the main features considered as the most important MRI biomarkers are extracted from seven brain regions to include Ventricular volume, Hippocampus volume, Whole Brain volume, Entorhinal Cortical thickness, Fusiform, Middle temporal gyrus and intracranial volume (ICV). Figure 4.3 illustrates these brain regions in the brain template. The PET features are single measurements of the Pittsburgh compound B (PIB), the Florbetapir (AV-45), and the fluorodeoxyglucose

(FDG) for cerebral glucose metabolism, all used as agents to image and gauge the extent of amyloid plaques at the different stages of the disease. As we are constrained to the multimodal features presented in Table 2 for this longitudinal study, future studies could involve the use of PET regional standardized uptake value ratio (SUVRs) as quantitative measures of the radiotracer uptake in regions of interest with respect to a reference region to assess how such measures, especially in disease-prone areas, relate to the MMSE score as used for prediction purposes in this study. In the features listed in Table 2, in accordance with the ADNI multisite study, FDG is the average FDG-PET of angular, temporal, and posterior cingulate, PIB is the average PIB SUVR of the frontal cortex, anterior cingulate, precuneus cortex, and parietal cortex and AV45 is the average AV45 SUVR of frontal, anterior cingulate, precuneus, and parietal cortex relative to the cerebellum.

In terms of the cerebrospinal fluid (CSF) biomarkers [111][112][113], this study considers Amyloid Beta (ABETA), phosphorylated tau protein (PTAU), and Total tau protein (TAU) as means to assess the extent of amyloid plaques in between neurons and the neurofibrillary tangles made up of tau protein within the neurons themselves, both considered to contribute to the degradation of neurons in Alzheimer's disease and other tauopathies. The other risk factors considered in this study include age, gender, level of education and Apolipoprotein E (APOE) gene. As indicated earlier, APOE with the E4 allele apolipoprotein is considered a major genetic risk factor for AD [114]. As for age and gender, it is common knowledge that age is a major risk factor in AD (since only about 5% develop symptoms of AD before the age of 65) and it is estimated that two-thirds of the 5.5 million Americans living with AD are women. Although women tend to live longer than men, we still could not conclude with certainty that this discrepancy in the larger number of women with AD is only due to longevity and

experts remain uncertain on other factors that could explain this difference. As for the level of education, there is an understanding and some studies confirm that the higher is the level of education the lower is the risk for dementia, and that cognitive reserve serves as a strength to overcome some the symptoms of AD [115] [116].

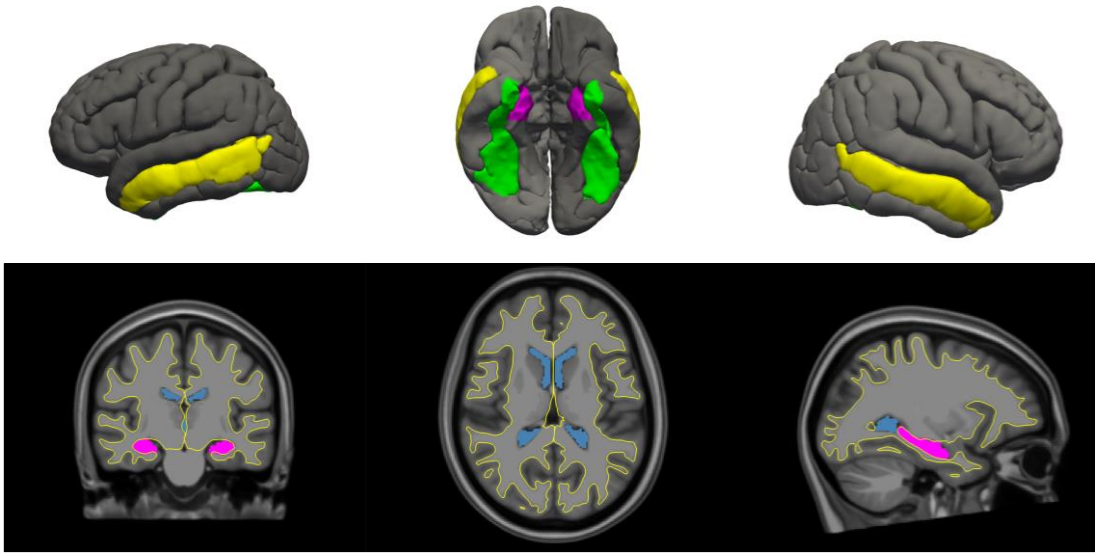


Figure 4.3. Selected MRI brain regions for tracking the progression of Alzheimer's disease. 3D mesh surface map, with purple, green, and yellow areas representing Entorhinal, fusiform, and middle temporal regions, respectively (Top). The volumetric segmentation, in which the yellow line depicts the interface between grey and white matter, and the purple and blue regions representing the hippocampus and ventricles, respectively (Bottom).

In the preprocessing step, ADAS11, ADAS13, MoCA, the Diagnosis labels (DX), and CDR were removed from the feature set since it is known that they have a high correlation with the MMSE score. We further excluded non-stable CN participants (CN to MCI or CN to AD) and subjects who are facing a reverse-phase in the progression stage (MCI to CN, AD to MCI).

Given the number of subjects considered for this study (1620), to compensate for the small sample size, nested cross-validation has been applied to our data set. From the whole dataset, 70% were randomly selected as the training set and 30% were set aside

as the testing set. This process of randomly splitting the data has been repeated 10 times to avoid any bias in the evaluation of data. For hyperparameter selection, in each of those data splits, 5-fold inner cross-validation along with exhaustive search is used to select the optimal hyperparameters for each method. For regression methods, the regularization parameters were selected in a range of $\{10^{-3}$ to $10^{3}\}$. As for the XGBoost method, the number of estimators is searched between {1 and 500}, learning rate has been searched between $\{10^{-3}$ and 1}, the number of columns used by each tree (colsample_bytree) has been searched between {0.1 to 1} and max depth has been searched between {1 and 15}.

Through the rest of this Chapter, reported values are the mean and standard deviation of the experiments in these10 different random train and test splits. It is important to mention that, feature space from every observation in both the training set and the testing set were normalized separately using the Z- score (i.e., dividing the difference between each value and the mean by the standard deviation).

4.4.3 Selecting modality-specific multitask models

The first stage of the model is focused on developing modality-specific multitask coefficient matrices. The motivation is to not confuse the multitask regression coefficients with modeling the relationship between different modalities and to preserve the maximum learning capacity to be devoted to learning the trajectories of cognitive decline. The following state-of-the-art algorithms are selected as the competing methods in the investigation of predicting clinical decline at multiple time points.

- Ridge regression
- Elastic Lasso
- Temporal Group Lasso (TGL)

- Convex Fused Sparse Group Lasso (cFSGL)
- Non-convex Fused Sparse Group Lasso (nFSGL)
- Subspace Regularized Sparse multitask learning [107]
- Parameter-free least Lasso (Zhu et al., 2018)

For single task learners, six separate regression models have been trained to predict cognitive scores for each time point. However, in multitask learning, the regression coefficients for all time points are trained together. This approach improves the efficiency of the final model by identifying and capturing the correlation between the transitions of cognitive scores at successive time points. To benchmark the performance of different methods, Root Mean Square Error (RMSE) and R correlation coefficient (denoted as *Corr* in Tables and figures that follow) are selected as the main evaluation metrics through this study. Figure 4.4 demonstrates the comparison of prediction accuracy of regression models using different sets of biomarkers. Several important empirical observations can be made from analyzing the results given in Figure 4.4.

First, single-task models yield a competitive performance at earlier time points but multitask learners significantly surpassed them at subsequent time points. This analysis found clear evidence for the superiority of multitask learners over single task learners.

Second, the sparsity and temporal sample size of each modality-specific feature space differ from each other. For each modality, the regression model which yields the highest winning rate is selected as the best predictor. The winning rate is defined here as the number of times a specific method achieves the best performance in terms of the lowest error across all intervals and highest correlation in comparison to all the other methods. It is important to emphasize that the winning models are selected during the training phase without seeing the test data.

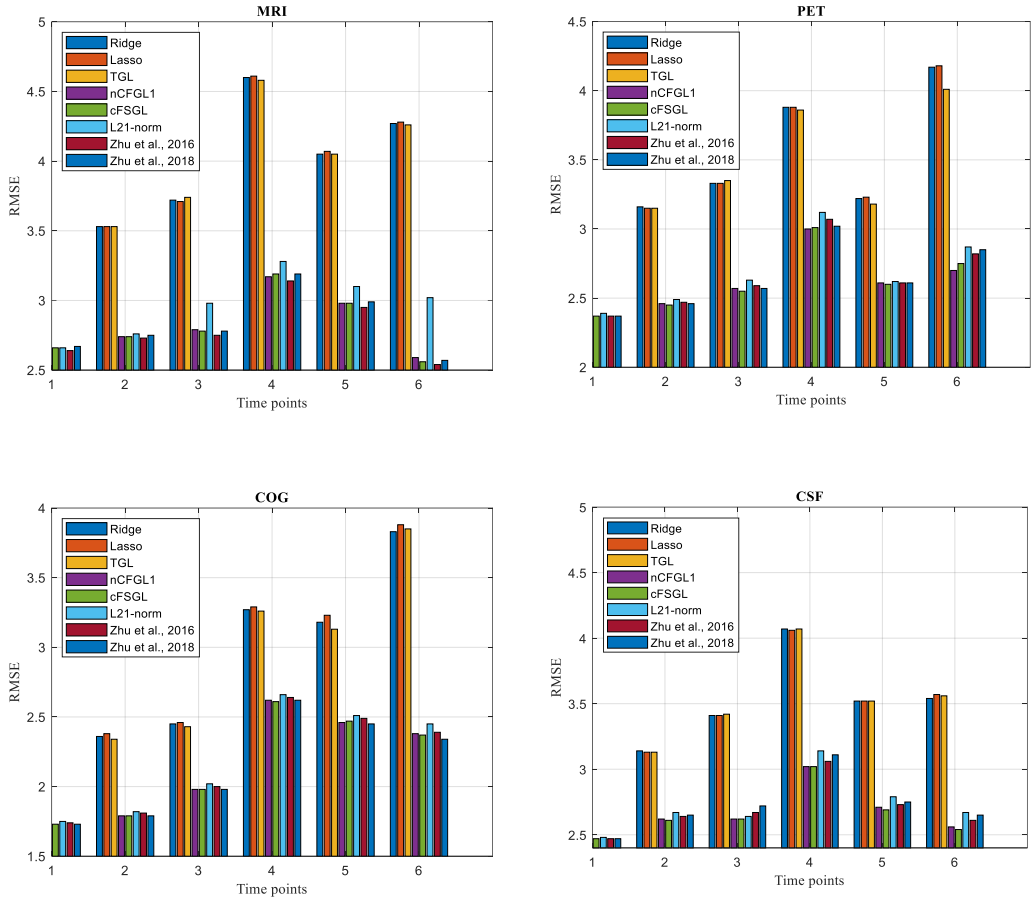


Figure 4.4. Performance comparison of different regression methods on longitudinal prediction of MMSE using different modalities.

It can be observed that cFSGL proved to be the best method for PET and CSF, just as the method in [108] yielded the best overall performance results for COG measurements, and the coefficient matrix in [107] achieved the best prediction accuracy for MRI measurements. The ℓ_2 norm regularization penalty term in cFSGL results in non-zero values in $\boldsymbol{W}$. Since the feature spaces for PET and CSF are low dimensional and less sparse, using ℓ_2 norm will help determine and keep the best predictive biomarkers. The COG modality was found to have a higher dimensionality and the pattern of features is highly sparse, which enabled the coefficient matrix in [108] to achieve better generalization than other methods.

Third, the cognitive modality achieved the smallest error in comparison to all other modalities in predicting cognitive decline. However, it must be pointed out that ADAS11, ADAS13, MoCA, CDR, and diagnosis labels were removed from the cognitive feature space to ensure that variables with a strong correlation with the MMSE label are not biasing the prediction. The scatter plot for cognitive assessment modality is shown in Figure 4.5.

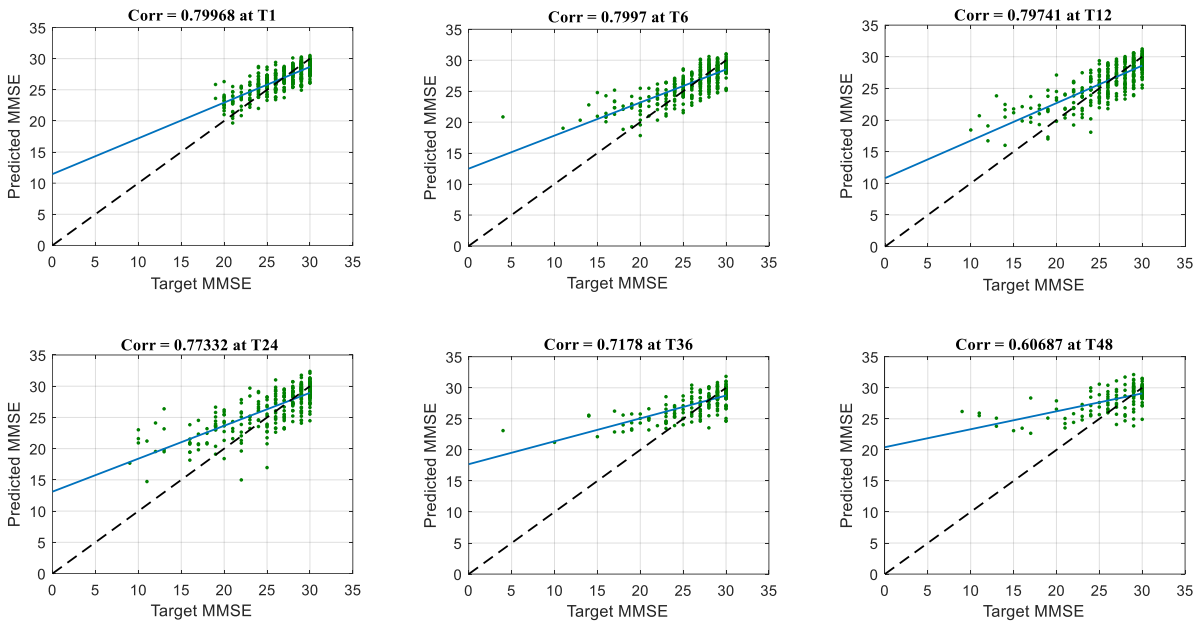


Figure 4.5. Scatter plot of predicted MMSE scores versus actual values in six time points using the cognitive assessment modality. The green line is the regression line achieved by the winning coefficient matrix and the black dashed line is the reference for perfect correlation.

Table 4.3. Hyper parameters used for tuning of Gradient Boosting

Modality Combination (C)	max_depth	Learning rate	Colsample_bytree	n_estimators
MRI_PET	2	0.07	0.98	90
MRI_PET_CSF	3	0.05	1.00	120
MRI_PET_COG	3	0.07	1.00	90
PET_COG_CSF	3	0.07	0.98	80
MRI_PET_COG_CSF	3	0.10	0.50	50

4.4.4 Final results and discussion

In order to model the complex relationship between different modalities, the outcomes of the winning predictors from Fig. 4 are combined with the risk factor measurements, as non-temporal biomarkers. These new sets of features have been utilized as the input for the gradient boosting (GB) machines. The GB machines have been trained over five combinations of modalities. Grid search has been adopted to estimate the hyperparameters of gradient boosting for different combinations of modalities. The optimal hyperparameter values for each modality have been reported in Table 3. The experimental results, in terms of RMSE, are shown in Table 4.

For all methods reported in Table 4, the training and testing sets are identical, except for the fact that the competing methods are using the conventional approach in which all features from different modalities are concatenated together. For the statistical test, the correlation coefficient between the observed and predicted values is calculated on 100 bootstrapped samples, generated from the original sample size. By testing the null hypothesis of no correlation, the significance of the correlation, p-value, is calculated for each time point.

The proposed model achieved a correlation coefficient of 0.82 (p = 6.20e-47) at T1, 0.86 (p = 4.18e-62) at T6, 0.80 (p = 1.18e-41) at T12, 0.81 (p = 1.82e-38) at T24, 0.79 (p = 6.11e-20) at T36 and 0.76 (p = 1.44e-15) at T48 on the test data. The coefficient of determination is another statistical metric to evaluate the accuracy of regression models. This parameter presents the percentage of the variation in the dependent variable (predicted value) that can be described by the independent variable (target value). The coefficient of determination for the proposed model is 0.67 at T1, 0.73 at T6, 0.64 at T12, 0.66 at T24, 0.62 at T36, and 0.58 at T48. Fig. 6 shows the scatter plots of predicted MMSE scores versus the actual scores with correlation values reported

within each scatter plot. Colors are representing groups of subjects belonging to different stages of AD. The progressive nature of AD results in a steady, though uncertain slope in terms of cognitive decline. Patients who are diagnosed with late stages of AD at baseline have a higher chance to encounter a steep descent to severe cognitive decline within the following 48 months. Therefore, at the time points with an unbalanced population, in terms of the cognitive score distribution, individuals with a severely low MMSE score are detected as outliers. For example, according to Fig. 6, there are very few subjects with a cognitive score of less than ten, which makes it difficult for the system to keep track of all values. It should be pointed out that considering a weighting scheme of the distributions at the different stages of the disease and at different time points could help in improving the prediction accuracy of the trajectories in cognitive decline [117].

Table 4.4. Comparison of the results from our proposed method with other existing methods on longitudinal multi modal data. The error has been reported using RMSE metric in six different future time points.

Method	Modality	Time Points					
		T1	T6	T12	T24	T36	T48
Ridge	MRI, PET, COG, CSF	1.90±0.47	2.33±0.68	2.43±0.74	3.17±0.73	3.20±0.83	4.05±0.90
Lasso	MRI, PET, COG, CSF	1.83±0.37	2.34±0.64	2.45±0.53	3.11±0.70	3.15±0.74	4.00±0.76
TGL	MRI, PET, COG, CSF	1.93±0.43	2.32±0.45	2.42±0.55	3.22±0.67	3.10±0.82	3.87±0.93
nCFGL1	MRI, PET, COG, CSF	1.81±0.55	2.31±0.58	2.41±0.67	3.28±0.46	3.49±0.59	4.06±0.70
cFSGL	MRI, PET, COG, CSF	1.88±0.85	2.33±0.64	2.40±0.73	3.20±0.68	3.03±0.86	3.61±0.78
$\ell_{2,1}$-norm	MRI, PET, COG, CSF	1.89±0.75	2.34±0.52	2.38±0.76	3.24±0.59	3.08±0.67	3.64±0.69
[160]	MRI, PET, COG, CSF	1.87±0.52	2.31±0.66	2.32±0.50	3.27±0.62	2.98±0.96	3.56±0.87
[161]	MRI, PET, COG, CSF	1.86±0.53	2.27±0.61	2.38±0.64	3.23±0.57	3.02±0.84	3.42±0.64
Proposed	MRI, PET	2.02±0.26	2.30±0.32	2.88±0.36	3.06±0.35	2.51±0.31	2.60±0.32
	MRI, PET, CSF	1.95±0.39	2.22±0.31	2.81±0.30	2.92±0.33	2.51±0.37	2.72±0.30
	MRI, PET, COG	1.60±0.27	1.79±0.23	2.30±0.23	2.41±0.35	2.53±0.32	2.20±0.30
	PET, COG, CSF	1.63±0.20	1.80±0.28	2.25±0.20	2.38±0.25	2.41±0.26	2.38±0.29
	MRI, PET, COG, CSF	1.62±0.24	1.78±0.22	2.24±0.24	2.38±0.21	2.28±0.22	2.19±0.15

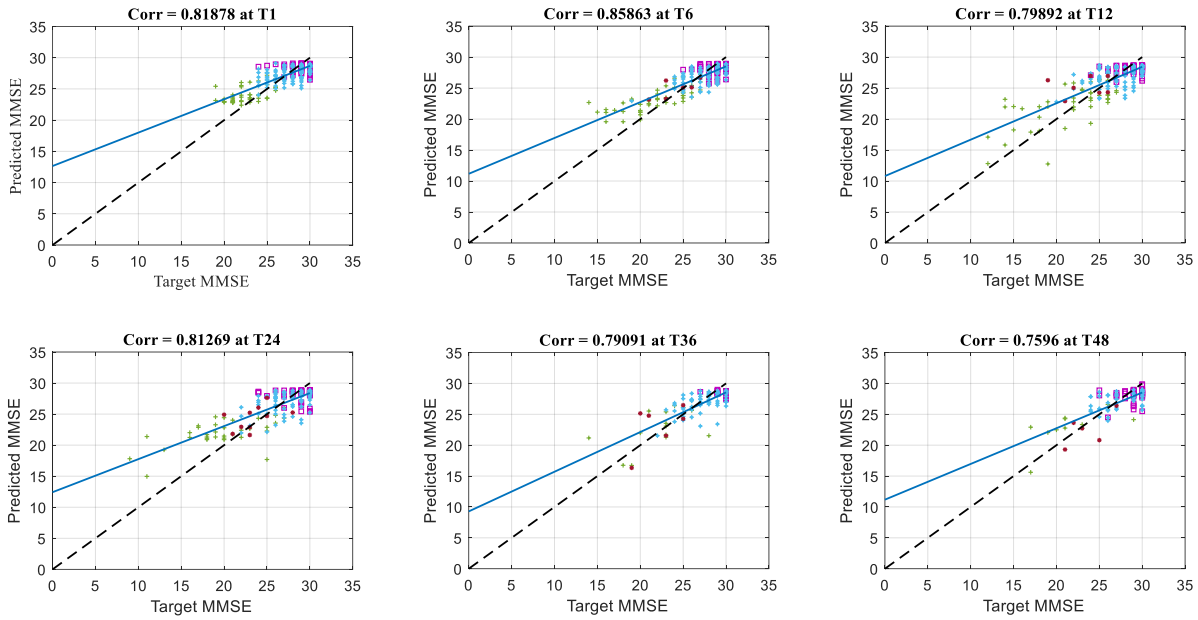
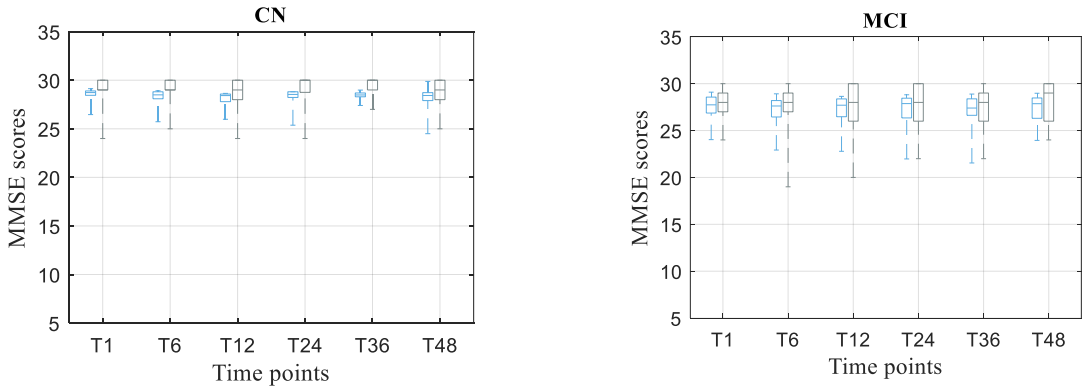


Figure 4.6. Scatter plot of predicted MMSE scores versus actual values at six different time points. The blue line is the fitted regression line achieved by the proposed model and the dashed black line is the perfect correlation. Red squares (□) are the CN group, blue plus signs (+) are the MCI group, red asterisks (*) are the MCI converter group and green plus signs (+) are the AD group.

Since the focus is on predicting the trajectories of MMSE scores, the longitudinal distributions of predicted versus actual target MMSE scores for each group are provided in Figure 4.7.



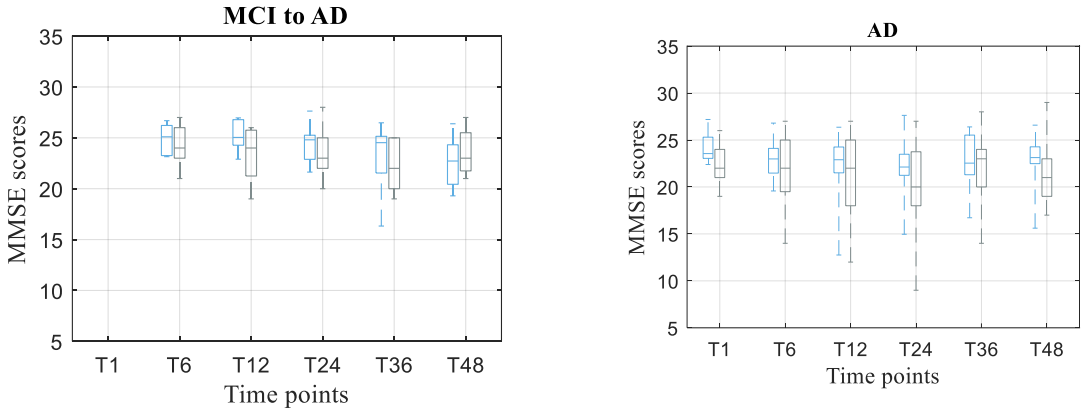


Figure 4.7. Longitudinal trajectories of MMSE scores through 6 time points for each category of disease. In each figure, boxplots in blue are used for the distribution of predicted MMSE scores, and black boxplots are used for the distribution of target MMSE scores.

To further evaluate the superiority of the proposed model, following the approach described in [103], paired *t*-test has been performed on the residuals of the proposed method and each of the competing methods. The results summarized in Table 5 show that except

Table 4.5. Comparison of p-values obtained from residuals of the proposed method and the competing methods using the combination of modalities of MRI, PET, COG, CSF

	Ridge	**Lasso**	**TGL**	**nCFGL1**	**cFSGL**	**$\ell_{2,1}$norm**	**[160]**	**[161]**
T1	0.063	0.083	0.386	0.386	0.501	0.086	0.029	0.032
T6	0.007	0.011	0.003	0.001	< 0.001	< 0.001	0.024	0.013
T12	0.004	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.002	< 0.001
T24	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
T36	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.010	0.012
T48	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.013	0.010

for the baseline, the proposed method for all other five future time points demonstrates statistical significance, with all p-values less than 0.05, proving its effectiveness.

Since independent models are separately trained over each feature space, our model brings the following advantages: (1) feature scarcity from one particular modality would not be an issue for the other regression models; (2) any error within the data of one modality could be prevented from propagating through other modalities; (3) the model could be easily extended to include other modality sources with little adjustments to consider sparsity patterns of the measurements; (4) the proposed model applies to a wide variety of subjects with any combination of modality sources, without being restricted to their baseline diagnosis or their historical records; and (5) the robustness and flexibility of the presented framework in handling missing data preserve enough information to monitor and predict MMSE trajectories with relatively high accuracy.

Chapter 5 A Tensorized Multitask Deep Learning Network for Progression Prediction of Alzheimer's Disease

5.1 Introduction

According to the report published in 2020 by the Alzheimer Association (AA), nearly 13.8 million people will be affected by this disease by 2050 [118]. Due to the irreversible nature of AD, its early diagnosis is of paramount importance. Researchers and clinicians agree that in the latter stages of the disease, brain atrophy has already ensued, and no medication or intervention could potentially be of benefit to the patient in reversing the outcome but could nonetheless slow down the progression of the disease. Understanding the early clinical symptoms of AD and carefully assessing the subtle changes characterizing the different disease states could yield more effective plans for patient-specific drug treatment and therapeutic interventions.

Extensive research has focused on using different machine learning techniques for the diagnosis and prognosis of AD. However, the rather low accuracy of classification results and regression techniques in delineating converter from non-converter groups and Mild Cognitive Impairment (MCI) from Cognitively Normal (CN) draws attention to the diversity and heterogeneity of the potential features that could be extracted from the multimodal and multiclass AD datasets **[**119**]**.

A retrospective of previous studies on multimodal datasets reveals some inconsistencies in modeling the relationship between features from the different recording modalities. Although several linear methods have been previously reported in the literature with the ability to linearly fuse the information from different modalities [120], several authors have also suggested different approaches to fuse the multimodal biomarkers. Tong et al. [121] used a nonlinear graph fusion process to combine pairwise similarity matrixes that were derived from each modality. They analyzed the efficiency

of their unified graph in terms of area under the curve (AUC) for the binary classification of AD-CN, MCI-CN, and also in a three-way classification. In terms of regression models, Wang et al. [122] addressed the issue of the nonlinear relationship that exists between the cognitive assessment scores and the clinical biomarkers using a matrix elastic-net kernel, which was embedded through a multilayer network. They evaluated their model by predicting the Fluency Test, Rey's Auditory Verbal Learning Test (RAVLT), Trail making test (TRAILS). The study proposed in [123] explores the merits of least square regression (LSR), multi-target ridge regression, and multi-target low-rank regression model with trace-norm regularization (MR-Trace) to assess the temporal correlations between imaging and cognitive data. In [48], Huang et al. have taken a supervised nonlinear approach for predicting clinical scores. Their proposed method consisted of a sparse regression-based random forest (RF) algorithm, which uses longitudinal information from multiple time points to predict future cognitive scores. They also used their method to sequentially predict missing target scores at the middle time points to improve data availability for the final model. Wei et, al. proposed a classification method to distinguish non-converter MCI (MCI-NC) versus converter MCI (MCI-C) by using an SVM classifier over features that are a combination of FreeSurfer-derived MRI features and nodal features derived from the thickness network [124]. Early detection of the disease is investigated in [125] to detect subjects as healthy, subjective cognitive decline (SCD) or amnestic mild cognitive impairment (aMCI) based on SVM and features extracted from white matter. In order to predict MCI-to-AD conversion authors of [126] developed an extreme learning machine (ELM)-based grading method to efficiently fuse multimodal data and predict the conversion within 3 years. A tool named PredictND is developed by Tolonen et, al as an attachment to clinical decision support systems to categorize the subjects as healthy

or as one of the four different kinds of AD based on multimodal data [127]. Six binary classification groups (AD vs. HC, MCIs vs. MCIc, AD vs. MCIc, AD vs. MCIs, HC vs. MCIc, and HC vs. MCIs) are considered in [128] using multimodal features from structural, diffusion, and functional neuroimaging data and the APOE Genotype. They showed that the (left/right) precentral region can be considered the most significant region. Furthermore, they found that FDG, AV45-PET, and rs-fMRI were the most important neuroimages.

Deep learning architectures have been recently used in the reported literature for the diagnosis or prognosis of Alzheimer's disease [129][130][131][132][133][134]. Jo et, al reviewed many of them regarding the diagnosis and prognosis of AD and AD stages [135]. In terms of classification, Liu et al. [136] proposed a cascaded convolutional neural network (CNN) to learn multimodal patch-based features from different regions of the brain. Using MRI and PET images, their deep 3D-CNN algorithm could achieve good binary accuracy in differentiating AD vs CN, progressive MCI vs CN, and stable MCI vs CN, but no multiclass classification was performed. The challenge of efficient feature representation and its effect on prediction accuracy has been addressed in several clinical studies. For example, Suk et al. [137] made use of a stacked auto-encoder to model the complex pattern embedded in the features to enhance the classification accuracy of AD vs MCI [137]. In another study, Liu et al. [138] proposed the use of an auto-encoder to extract high-level features from the available modalities. Then a zero-masking method was employed to train the network in a way that it could reconstruct missing features from available modalities and passing them to the Softmax layer to classify the subjects into four groups of AD, non-converter MCI (MCI-NC), converter MCI (MCI-C), and CN. In another study, Liu et al. [139] also proposed the use of stacked auto-encoders for AD/MCI classification using neuroimaging

modalities. Jha et al. [140] proposed a sparse auto-encoder for binary classification of AD from cognitively normal (CN) subjects. The use of Recurrent Neural Networks has been proposed by Wang et al. [141] to predict a future stage of the patient using historical clinical records. Liu et al. [142] proposed a combination of CNN and Recurrent Neural Network (RNN) for feature extraction and classification. Considering the large size of PET images, instead of using 3D CNN, they employed 2D CNN to extract features from 2D PET slices. The extracted features then were used by gated recurrent unit (GRU) for classification of AD and MCI subjects from CN. A classifier based on CNN networks and regularization is proposed in [143] to distinguish early MCI versus CN subject while using structural MRI and diffusion tensor imaging (DTI) images as the multi-modality input data.

The correlation between the categorical and numerical variables brings the potentially open question of whether jointly learning approaches could leverage the learning performance of classification and regression tasks. Liu et al. in [144] proposed the use of a convolutional neural network for joint regression and classification tasks. Using landmark detection, their method could identify the most informative patches around the selected MRI landmarks and then automatically extract features for the training phase. They also incorporated demographic information as additional information. Using their deep multi-task multi-channel learning (DM^2L) framework, they reached an accuracy of 51.8% in a four-class classification process. Multimodal feature fusion has also been explored by Zhu et al. [145] through canonical features used as regressors to select features. Their proposed sparse multitask learning process could predict ADAS-Cog, MMSE, and AD stages simultaneously. A deep polynomial network (DPN) was used by Shi et al. [146] for both binary and multiclass classification tasks, where they used two stacked DPN (SDPN) to extract high-level features from

MRI and PET and then used another (SDPN) to fuse the information from these two modalities. Using their two-stage stacked deep polynomial network, they obtained an accuracy of 55.34% in multiclass classification with higher accuracies obtained when using binary classification. Multilayer multi-target regression (MMR) framework has been proposed by Zhen et al. [147] to encode the inter-target correlation as well as the relationship between the input and output space via low-rank learning. They evaluated their model over 18 datasets and proved that their method could outperform almost all the reported performances in terms of average Relative Root Mean Squared Error (RRMSE) over the target variables.

In the study by Daoqiang et al. [148], a Multi-Modal Multi-Task (M3T) learning framework is used for the prediction of multiple clinical variables from a multimodal dataset. In their proposed approach, each task is defined as the prediction of a single cognitive score. First, a multitask feature selection method selects a subset of features that are most relevant to the prediction task. Then, in order to fuse the selected features derived from different modalities, they adopted a multiple-kernel scheme. Finally, for predicting each clinical score, separate support vector regression (SVR) was trained over the multimodal fused data. They validated their model over two cognitive scores of MMSE and ADAS-cog. With similar objectives, Zhu et al. [149] utilized a matrix-similarity-based loss function combined with group lasso to select the best features for both classification and regression tasks. By exploiting the high-level information in the target response matrix, their feature selection method could preserve the information in the predicted response matrix. Moreover, Wang et al. [150] proposed an 8-layer CNN with a rectified linear unit (ReLU) and max-pooling to classify AD patients from healthy control subjects in the OASIS dataset. Liu et al. adopted ensemble learning [139] to fuse multiple neuroimaging modalities, where a four-layer neural network was

utilized for binary classification of AD/MCI and AD/CN. The first layer of their proposed model fuses the prediction information from each modality. This fused information is then utilized in the next layers to improve classification accuracy.

In this study, we propose a novel neural network architecture, structured as a Kernelized and Tensorized Multitask network (KTMnet) process, to predict two joint tasks of classification and longitudinal prediction simultaneously. This network uses dense layers to first extract features from each modality separately, then uses gaussian kernel layers and tensorization over the modality fused feature space to nonlinearly map the data in low-dimensional space to a high dimensional space. Empirical results show an enhanced performance of the proposed method in comparison to all other related methods reviewed in this article, especially when delineating the challenging group of MCI (converters and non-converters) from CN in multiclass classification.

5.2 Materials and Method

5.2.1 Subjects

The clinical data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies, and non-profit organizations, as a 60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of

sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

A total number of 1117 individuals consisting of 632 males and 485 females were considered for this study. The average age is 73.84 with total average years of education of 16.04. The average MMSE score of the population is 27.44 at baseline and 27.06, 26.82, and 26.02 at the next 6 and 12, and 24 months, respectively. At each follow-up visit, participants were labeled as AD, MCI (Mild Cognitive Impairment), and CN, and those participants from the MCI stage that are converting to AD are labeled as the MCI to the AD group. The demographics of the subjects used in this study are given in Table 1.

Table 5.1. Demographic characteristics of subjects used in this study. Label f/m stands for the number of females in comparison to males. Age, years of education, MMSE, and CDR of subjects in each category are presented by mean ± standard variation of that variable.

Parameter	Value	Total	Alzheimer	MCI-C	MCI-NC	Control
Subjects	number	1117	157	191	441	328
Gender	f/m	485/632	73/84	75/116	184/257	153/175
Age	year(mean±std)	73.84±7.07	76.77±6.99	73.86±7.47	70.85±7.19	75.01±5.71
Education	year(mean±std)	16.04±2.78	14.63±3.15	16.09±2.74	16.09±2.63	16.36±2.68
MMSE	number(mean±std)	27.43±2.46	23.24±1.96	27.23±1.75	28.30±1.59	29.15±1.01
CDR	number(mean±std)	1.25±1.36	3.98±1.51	1.62±0.92	1.24±0.74	0.03±0.13

5.2.2 Problem Description

In longitudinal AD studies, disease progression can be gauged via screening the categorical or numerical labels of participants through time. The categorical labels in ADNI are AD, MCI (including the converter and non-converter groups), and CN. On the other hand, there are also numerical measurements needed to assess cognitive impairment, which augment the in-depth analysis of the data. Mini-Mental State Examination or MMSE is the best-known clinical AD predictor that is accepted and

used worldwide. While predicting the diagnosis labels is accomplished through classification methods and predicting the numerical value of neuropsychological test scores is performed through regression models, the underlying features for both tasks are constructed from similar sets of measurements. This relationship between these two types of modeling methods motivated researchers to train these highly interrelated tasks of regression and classification through multitask learning.

To model the progression of AD, a time frame of 24 months has been considered in this study to assess the conversion prospects of the MCI group into AD. Therefore, only those subjects that completed a baseline scan (M0) and showed up for a follow-up visit 6 months later (M6), 12 months later (M12), and 24 months later (M24) were considered. Studying longitudinal AD cohorts could improve our understanding of AD pathogenesis. While most patients that have been diagnosed as belonging to the intermediate stage of MCI have been known to progress towards the AD stage, there is some evidence that some of them might stabilize at the MCI stage. However, the different conversion slopes for the different individuals suggest that this stable group is converting into AD in a much longer time frame. Figure 5.1. shows the number of subjects in each category of AD over 24 months.

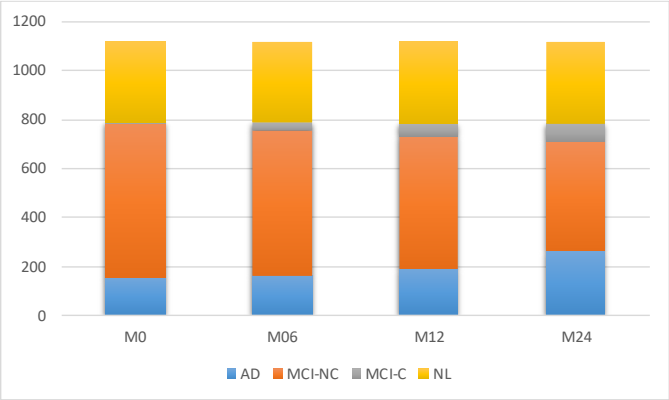
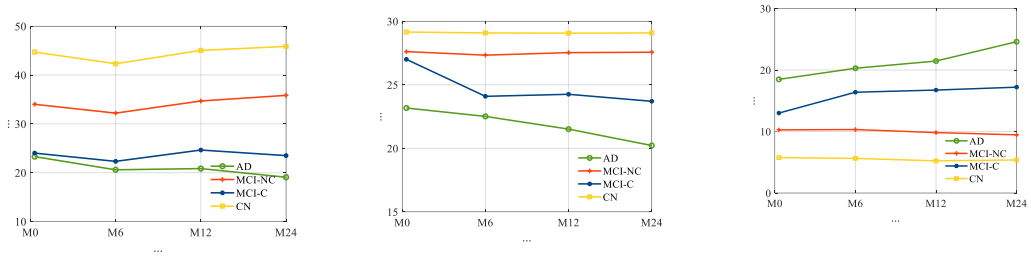


Figure 5.1. Number of subjects in each of the four subgroups of AD at different time points

The average longitudinal changes of neuropsychological test scores for the 4 subgroups are shown in Figure 5.2. It can be observed that for AD and MCI-C populations, the mean of the neurological cognitive test is decreasing by 13% and 12.7%, respectively through time, which means that the health status of the subjects is declining continuously through time, demonstrating the importance of predicting the cognitive decline of AD prone population as early as possible.



(A) (B) (C)

Figure 5.2. The average trajectories of (A) RAVLT, (B) MMSE, and (C) ADAS11 score for subjects for four different classes of AD

5.2.3 Problem formulation

The proposed Kernelized and Tensorized Multitask network (KTMnet) as shown in Figure 5.3. is structured to effectively estimate the progression of Alzheimer's disease by predicting the categorical and numerical labels simultaneously. Let $\boldsymbol{y}_r$ be the sets of longitudinal neuropsychological test scores (MMSE), for the regression task (Task 1) and $\boldsymbol{y}_c$ be the sets of categorical labels for the classification task (Task 2). The input space for both tasks is the multimodal features of $\{\boldsymbol{x}_{m_1}, \boldsymbol{x}_{m_2}, \boldsymbol{x}_{m_3}, \boldsymbol{x}_{m_4}, \boldsymbol{x}_{m_5}\}$, in which the vector $\boldsymbol{x}_{m_i}$ comprises the extracted measurements from modality *i*. Note that in this study, input features are extracted from MRI, PET, CSF, cognitive measurements, and risk factors at baseline. Hence, vectors $\boldsymbol{y}_r$ and$\boldsymbol{y}_c$ for this study can be established as $\boldsymbol{y}_r = [Score_{M0}, Score_{M6}, Score_{M12}, Score_{M24}]'$ and $\boldsymbol{y}_c = [AD, MCI-C, MCI-NC, CN]'$, where *MCI-C* and *MCI-NC* define the MCI converter and non-converter groups, with the prime symbol (') defining the transpose function. The risk factor parameters considered are age, years of education, sex, and APOE4. The overall objective function of the proposed multitask network could be modeled as an algorithm in which $\boldsymbol{y}_r = E_r(\boldsymbol{x}_{m_1}, \boldsymbol{x}_{m_2}, \boldsymbol{x}_{m_3}, \boldsymbol{x}_{m_4}, \boldsymbol{x}_{m_5})$ and $\boldsymbol{y}_c = E_c(\boldsymbol{x}_{m_1}, \boldsymbol{x}_{m_2}, \boldsymbol{x}_{m_3}, \boldsymbol{x}_{m_4}, \boldsymbol{x}_{m_5})$ with E_r and E_c being the corresponding estimators. The architecture of the proposed KTMnet method is shown in Figure 5.3. The proposed network is a series of operations as defined through equations (1) and (2). Feature representation, modality fusion, and tensorization have been incorporated in an end-to-end artificial neural network to harness the advantage of performing regression and classification tasks jointly in a unified framework. This multitask framework aims to make use of the features extracted from each modality through modality fusion and tensorization to secure a higher prediction accuracy. First, the feature vectors of each modality would be

extracted by $F_{m_i}$and then all the features from different modalities will be fused by function f. Next, a 3D tensorization (T) is applied to the fused feature vector to represent higher-order relations between features. Finally, tensor features will be extracted by F and will be fed to the regressor function f_r and classifier function f_c as in equations (1) and (2) below.

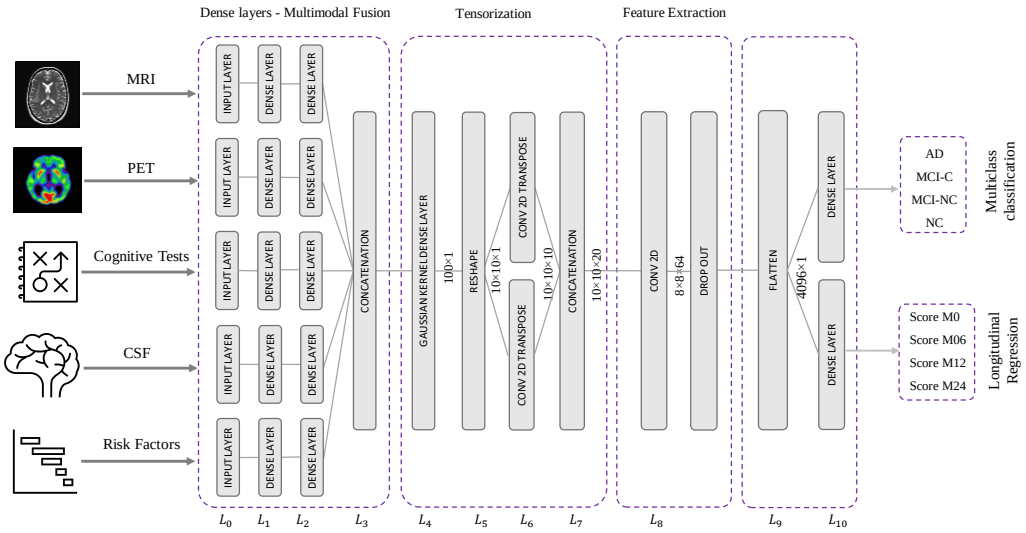


Figure 5.3. Design architecture of the proposed network

$$Task\ 1: \hat{\boldsymbol{y}}_r = f_r(F\left(T\left(f\left(F_{m_1}(\boldsymbol{x}_{m_1}), F_{m_2}(\boldsymbol{x}_{m_2}), F_{m_3}(\boldsymbol{x}_{m_3}), F_{m_4}(\boldsymbol{x}_{m_4}), F_{m_5}(\boldsymbol{x}_{m_5})\right)\right)\right) \quad (1)$$

$$Task\ 2: \hat{\boldsymbol{y}}_c = f_c(F\left(T\left(f\left(F_{m_1}(\boldsymbol{x}_{m_1}), F_{m_2}(\boldsymbol{x}_{m_2}), F_{m_3}(\boldsymbol{x}_{m_3}), F_{m_4}(\boldsymbol{x}_{m_4}), F_{m_5}(\boldsymbol{x}_{m_5})\right)\right)\right) \quad (2)$$

The loss function that has been employed to calibrate jointly the longitudinal regression and classification tasks is as follows:

$$Loss = \alpha \times MSE\ (\boldsymbol{y}_r, \hat{\boldsymbol{y}}_r) + \beta \times \mathrm{l}(\boldsymbol{y}_c, \hat{\boldsymbol{y}}_c) \quad (3)$$

in which y is the target value and $\hat{y}$ is the value predicted by the network. The MSE is the mean square error for the regression task defined as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_{r_i} - \hat{y_{r_i}})^2 \quad (4)$$

And the categorical cross-entropy of $\mathrm{l}(y_c, \hat{y}_c)$ is defined as:

$$\mathrm{l}(y_c, \hat{y}_c) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_{c_i}\log \hat{y}_{c_i} + (1 - y_c)\log(1 - \hat{y}_c)\right] \quad (5)$$

with N being the number of observations with c defining the number of categories assigned to the class label.

5.2.4 Network Architecture

The proposed framework makes use of artificial neural networks to accomplish the tasks of multiclass classification and longitudinal regression simultaneously. This network relies on convolutional neural layers to jointly perform the processes of tensorization and feature extraction. Given the schematic diagram of the network shown earlier in Figure 5.3., the main properties of the proposed network are as described in the following subsections.

5.2.5 Modality fusion

The relational correlation of features within each modality and between the different modalities remains an important subject in developing robust prediction algorithms. The importance of using and fusing relevant information from different modalities to improve classification is well documented in the literature, and some studies have shown significant improvement in comparison to relying on a single modality. For this reason, modality fusion has also been considered in the proposed network to incorporate the advantages of intra-modality and inter-modality feature representation. First, the network starts by transforming the raw features into a primary single modality representation space using fully connected layers. Two fully connected layers of L0 and L1 are then used to transform the extracted features from MRI, PET, CSF,

neurocognitive measurements, and risk factor parameters into an initial intra-modality feature-space representation. Let n_{mod} be the length of the input feature vector of named modality mod, then L0 is the input layer for each modality with n_{mod} nodes. These single modality features are then processed via two fully connected layers of L1 and L2 with $2 \times n_{mod}$ and n_{mod} nodes followed by linear activation function layers. The intermodality feature space is then initiated by integrating the previous fully connected layers in L3, which concatenates the outputs of the L2 layer to create the new feature vector.

5.2.6 Tensorization

Complementary and shared information found in features from different modalities is an essential part of reliably modeling the progression of neurodegenerative diseases. However, concatenating the features from different modalities and processing them using a simple network will not consider the inhomogeneity of the multimodal dataset. Therefore, it is reasonable to transform the feature space into a higher dimensional receptive field to enable the network to find more meaningful relationships.

Kernel methods can be used to tackle problems with linearly inseparable or nonlinear problems. A nonlinear mapping function can map linearly inseparable data in a low-dimensional space into a high-latitude space where it becomes possible to linearly separate the mapped data. The Gaussian kernel function is a representative function that is commonly used and is also adapted in neural networks [151], [152].

Tensorization is defined as transforming or mapping the lower-order data to higher-order data to improve the process of generalization afforded at this higher-order [153], [154]. This means that when the data is not providing a satisfactory feature representation in a lower-dimensional space, transferring it to a higher dimensional

space may improve the data analysis with the potential for retrieving hidden information in that same data. As an example, a vector can thus be reshaped into a 2D matrix or a 3D tensor of any arbitrary shape with width, height, and depth of W×H×D dimensions. Similarly, a matrix can also be reshaped into a higher-order tensor, by reshaping each column to a tensor of order K and stacking the results along the K+1th dimension.

In this proposed design, kernel function and tensorization are both used to extract higher-order features from fused multimodal features. In this new architecture, a dense layer with a Gaussian kernel is used for kernelization and a convolutional neural network is used for tensorization. In this way, a tensor with the size of 10×10×20 is generated through the following procedure:

- L4 uses Gaussian dense layer to assist tensorization.
- L5 reshapes the 100-node output vector of layer L4 to create a 2D 10×10 tensor.
- L6 performs 2D transpose convolutional filtering with a kernel size of 3×3, a stride of 1, padding type of "same", and linear activation function along with:
- 10 kernels with a dilation rate of 1
- 10 kernels with a dilation rate of 2
- Concatenation of the outputs from the two above dilation layers

5.2.7 Feature extraction

In this step, more predictive features are extracted from the generated tensor. Since the feature extraction part is also based on 2D convolutional filtering with the network being trained in an end-to-end fashion, there is not a strong distinction for separating the network into the tensorization part and feature extraction part. The extracted feature at the end of this stage is still a tensor. For this reason, 2D convolutional filtering is

performed in L8 by using 64 filters with a kernel size of 4×4 and applying the ReLU activation function. A dropout rate of 10% is implemented to randomly deactivate the connection between the neurons during the training phase to overcome any potential for overfitting.

5.2.8 Classification and longitudinal regression

This last component of the network is dedicated to classification and regression. For this reason, L9 flattens the output of the L8 layer to build a vector with the size 4096×1. The output of the L12 layer is connected via two fully connected networks with an L1 regularizer to the two output layers (i.e. $\boldsymbol{y}_r$ and $\boldsymbol{y}_c$) in L10. Four nodes are assigned for the regression part, which has a ReLU activation function, and four nodes are assigned for the classification part with a Softmax activation function.

5.2.9 Optimizer selection

In deep learning, choosing the right optimization method is key to tuning an accurate model. During the training, weights are iteratively updated until the network converges to a minimum cost function. Small learning rates will keep updating the weights with smaller steps, which could consequently lead to a minimal loss function. Updating the weights by taking large scales comes with the risk of skipping over the optimal weights. Still, some measure of caution should be taken when assuming smaller steps, as there is a risk of being trapped into some local minima.

With the proposed network, after testing several common optimization methods for training, the adaptive Adam algorithm has been selected as the optimization method. Adam, developed by Kingma and Ba [155], is one of the most common and adaptive optimizers used in deep learning applications, which uses adaptively approximations of

lower-order moments to yield an efficient, robust, and easy-to-tune solution. The adaptive learning rate is estimated by retaining an exponentially decaying average of previously squared gradients along with keeping the exponentially decaying averages of past gradients. Using this optimization approach with a learning rate of 0.001 and with exponential decay rates for the moment estimates β_1 and β_2 of 0.9 and 0.999, respectively, resulted in a robust trained network for longitudinal prediction and multiclass classification with higher accuracy in comparison to other optimizers.

In summary, the proposed structure of the network accomplishes both classification and longitudinal regression tasks by enabling the network to utilize the complementary/shared information in the extracted features. Integrating these two challenging tasks within a unified framework elevated the accuracy and robustness of the model by considering the inter-relatability between tasks. For training the network, an end-to-end learning process has been used to learn from both feature representation and modality fusion simultaneously to address both regression and classification tasks.

Regularization and dropouts were used to minimize the likelihood of overfitting in layers L4, L8, and L9. Feature dimensionality reduction is exploited to implicitly select and extract features between L1-L2 and between L9-L10. While all network layers from L1 to L10 are extracting features, the main part of the tensorization process is assumed to take place in layers L5 through L8 based on transposed and dilated convolutional filtering.

5.3 Preprocessing and Experimental Setup

5.3.1 Preprocessing

The procedure for predicting disease progression requires considering additional constraints. Subsequently, only the subjects that have a baseline scan and who showed up for a follow-up visit at 6, 12, and 24 months later, were considered in this longitudinal data collection.

The following preprocessing steps are performed in this analysis:

- Excluding all subjects whose cognitive score or diagnosis label has not been reported.
- Excluding the Aβ, P-tau, or Tau values, which are reported out of range (e.g., >1300 or < 80 for Tau).
- Removing the predictive biomarkers of ADAS13, MoCA, and CDR, which are found to be highly correlated with the status or label of the subjects. This was done so as not to bias favorably our longitudinal regression results which involve predicting future MMSE scores.
- Performing mean centering and normalization of training and test data using mean and variance of training data (z-score).

At the end of these preprocessing steps, a total number of 1117 subjects (328 CN, 191 MCI-C, 441 MCI-CN, and 157 AD) were considered for this study.

5.3.2 Experimental Setup

Empirical evaluations were conducted on the Intel Xeon E7 with NVIDIA QUADRO M6000 GPU. The proposed network is implemented in Python with the Keras library [156] using the TensorFlow backend [157]. For hyperparameter selection, a split of 15% of the data has been dedicated to 3-fold cross-validation trials, where the

set of hyperparameters that achieved the minimum bias and variance has been selected. The hyperparameters are the number of kernels used in L9, L10, L11 in the range of {256, 128, 64, 32, 16, and 8}and β in the range of {10, 20, ..., and 200} in which grid search has been performed. After hyperparameter selection, similar to the approach utilized in [136], [137], [158], 10-fold cross-validation trials were performed on the remaining 85% of data to avoid the occurrence of bias within a lucky partitioning. In each round of training set, 5% of data has been utilized for supervising the training process to prevent the network from overfitting. A batch size of 150 and the number of epochs of 200 were set for this process. We performed two sets of experiments to analyze the contribution of this work for each of the prediction tasks for evaluation purposes.

5.3.2.1 Task 1: Regression task for prediction of disease progression

In the following experiments, the first task of our KTMnet model is the longitudinal prediction of trajectories of the MMSE score. The neuroimaging modalities of MRI and PET, the cerebrospinal fluid (CSF) biomarkers, genetic information, and cognitive assessment tests have been used to create the multimodal data. Since the state-of-the-art algorithms used different performance metrics, to benchmark our method with other methods, the performance of the network is measured by the following common metrics:

The Root Mean Square Error, which is defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2} \quad (6)$$

The R correlation coefficient with the formula given below:

$$R(Y, \hat{Y}) = \frac{\sum_{i=1}(\hat{Y}_i - \bar{Y})\,(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}(\hat{Y}_i - \bar{Y})^2}\sqrt{\sum_{i=1}(Y_i - \bar{Y})^2}} \quad (7)$$

With $\hat{Y}$ defining the predicted values, Y being the real values, N is the number of observations and $\bar{Y}$ is the average of the real values in Y. The RMSE metric measures the standard deviation of the residuals between the predicted and actual targets, while the correlation coefficient metric measures the weight of similarity between them. Low RMSE and high correlation coefficient are desirable, conveying how well the predictive model is approximating the targets.

5.3.2.2 Task 2: classification task for prediction of disease status

For the classification task, the subjects were grouped according to the diagnosis label defined by ADNI as AD, EMCI, LMCI, and CN. The diagnosis label has also been tracked and labeled 24th months after their first visit and subjects are then labeled as MCI converter group (MCI-C) if they have been diagnosed as MCI in the baseline and their diagnosis status has progressed into AD. The MCI Non-Converter group (MCI-NC) label is assigned to subjects whose diagnosis label did not change after 24 months. The network is trained to perform a 4-way classification (along with the longitudinal regression task) to predict the subjects' class labels after 24 months. In this second test using the features at baseline, the aim was to predict the probability of converting from MCI to AD, 24 months ahead of time.

5.4 Results and Discussion

5.4.1 Prediction Results

The prediction results for the MMSE test scores at baseline and at time points of 6 months, 12 months, and 24 months are summarized in Table 2. The proposed model demonstrated a total RMSE of 2.32±0.52 and a correlation of 0.71±5.98 with a p-value of 5.09e-10 for predicting MMSE throughout the 24 months after baseline. Figure 5.4. shows the scatter plots of predicted MMSE values versus the actual target values at time points T0, T6, T12, and T24.

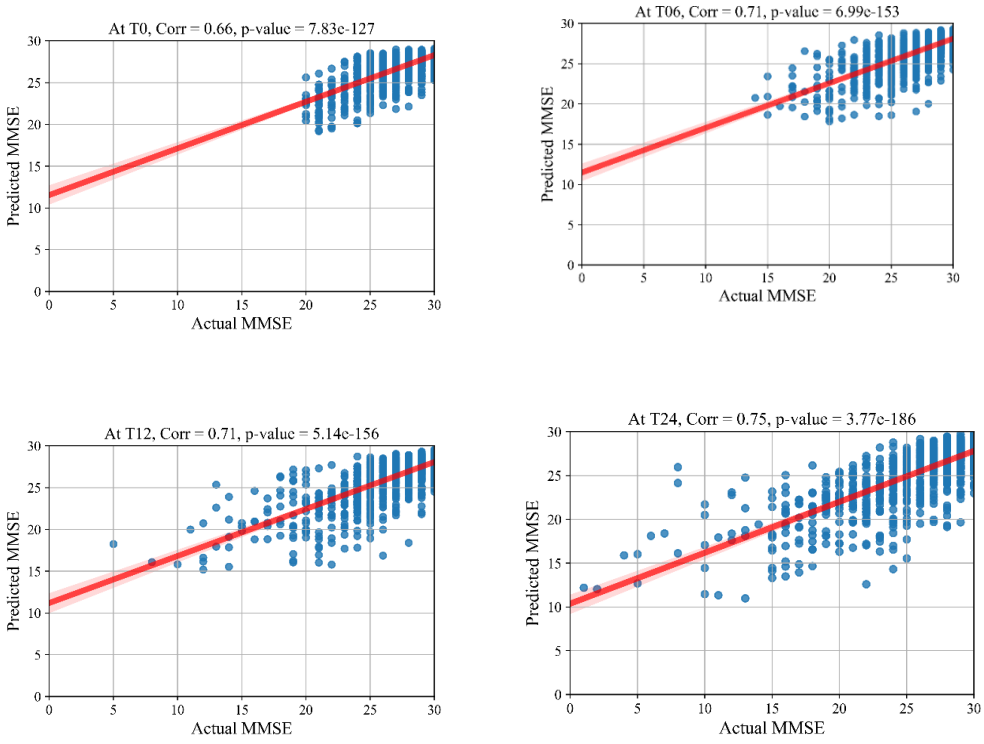


Figure 5.4. Scatter plots of predicted MMSE values

Table 5.2. Comparison of longitudinal regression performance of the proposed network in contrast to other methods reported in the literature

Study	Data	Subjects	T0		T06		T12		T24	
			RMSE	Corr	RMSE	Corr	RMSE	Corr	RMSE	Corr
[145]	MRI+PET	202	1.80±0.13	0.57±0.23	-	-	-	-	-	-
[144]	MRI+DEM	1984	2.37	0.57	-	-	-	-	-	-
[52]	MRI	755	2.37±0.19	0.57±0.05	-	-	-	-	-	-
[159]	MRI	445	1.75±0.20	0.75±0.08	2.31±0.29	0.79±0.10	2.48±0.40	0.79±0.12	3.00±0.38	0.83±0.06
[93]	MRI+PET+CSF	186	2.11±0.35	0.65±0.27	-	-	-	-	-	-
SVR	Multimodal*	1104	1.75	0.42	2.02	0.54	2.52	0.54	3.12	0.51
KTMnet	Multimodal*	1117	1.79±0.12	0.66±0.81	2.10±0.15	0.71±0.92	2.42±0.28	0.71±0.41	2.97±0.45	0.75±3.10

* Multimodal here refers to using MRI, PET, DEM, CSF, and cognitive measurements without the inclusion of ADAS11, ADAS13, and CDR-SB

5.4.2 Multiclass Classification Results

In this experiment, the results of the multiclass classification considering the four groups of AD, MCI-C, MCI-NC, and CN are shown in Table 3 with a comparison to other competing methods in the literature.

Table 5.3. Comparison of 4-way multiclass classification performance of methodologies reported in the literature using ADNI dataset

Study	Data	Subjects	Validation Method	Accuracy
[138]	MRI	758	10-fold	46.30 ± 4.24
[138]	MRI+PET	331	10-fold	53.79 ± 4.76
[145]	MRI+PET	202	10-fold	0.619 ± 1.54
[144]	MRI+PET+DEM[1]	202	Independent test	51.80
[160]	MRI+PET	202	10-fold	61.06 ± 1.40
[93]	MRI+PET+CSF	805	10-fold	53.72(max)
KTMnet	MRI+PET+CSF+COG+DEM	1117	10-fold	66.85±3.77

1 DEM stands for Demographic information (Age, Gender, and Education)

In this multiclass classification process, it is important to investigate the classification performance of the network for each category of subjects. The total classification accuracy achieved by our proposed KTMnet method is 66.85±3.77. In classifying the AD group from all other classes, the proposed network achieved a precision of 70.49%±9.33, a sensitivity of 57.21±9.41, an F1 score of 62.72±10.11, and an AUC of 94%. In classifying the MCI-C group, the network reached a precision of 45.33±7.22, a sensitivity of 50.79±9.42, an F1 score of 47.72±7.62, and an AUC of 83%. In classifying the MCI-NC group, the network reached a precision of 69.72±8.63, a sensitivity of 67.57±7.00, an F1 score of 68.16±5.06, and an AUC of 84%. In classifying the CN group, the network reached a precision of 77.89±6.62, a sensitivity

of 79.78±9.74, an F1 score of 78.10±5.89, and an AUC of 94%. Figure 5.5 illustrates the receiver operating characteristic (ROC) curves showing the capability of the network in discriminating between the four groups. This graph outlines the classification performance over all sets of possible thresholds. By varying the threshold, the observations are assigned to certain classes and the True Positive Rate on the y-axis is plotted against the False Positive Rate in the x-axis. Figure 5.6 shows the confusion matrix for contrasting the correct and incorrect predictions.

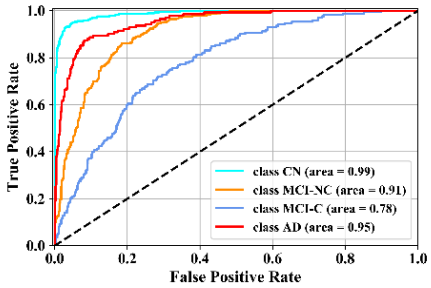


Figure 5.5. Comparison of ROC curves of the KTMnet for AD vs MCI-C vs MCI-NC vs CN

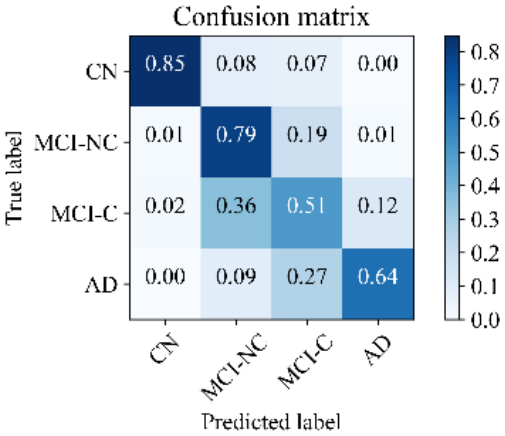


Figure 5.6. Confusion matrix of the KTMnet model

5.4.3 Discussion

The deep learning network developed in this study, together with its unique architecture, is designed to perform both tasks of multiclass classification and regression simultaneously, predicts disease progression by tracking the MMSE test scores at four consecutive future time points in a time window spanning 24 months and assessing their categorical labels as (AD, MCI-C, MCI-NC, and CN). This objective has been accomplished through extracting and fusing the complex inter- and intra-modality features, extracting hidden features by using tensorization that projects the feature space into a higher-dimensional space, and eventually modeling the feature representation through non-linear transformations.

In the reported literature, binary classification of AD patients (including the converter and non-converter groups) has been taken into consideration recently [161]-[162]-[163]-[164]. In these studies, attention was more focused on correctly classified subjects by measuring and reporting the metrics of sensitivity and specificity. However, the more challenging multiclass classification of AD cohorts using multimodal screening tests has not been fully explored for the diagnosis and prognosis of AD. This topic becomes even more challenging when progression is assessed in a population of subjects without any preliminary information about their baseline disease category. In a multiclass classification scenario, where there is no auxiliary information to reduce the number of false-positive and false-negative samples, the probability of over and under diagnosis will be increased, making it more important to use additional metrics for performance evaluation purposes. In this study, significant efforts are made to clearly and unambiguously report the performance of the proposed architecture. Table 4 summarizes specific studies that performed multiclass classification or longitudinal regression tasks for meaningful comparisons.

A noteworthy observation in this model was the see-saw effect encountered during hyperparameter searching. Although we received better results in comparison to other methods reported in the literature, the classification and regression tasks were not in sync with each other. To be more specific, the regression task was falling from its optimum point when the parameters were tuned to increase classification accuracy, and the reverse was also true when the parameters were tuned for increasing prediction accuracy. The initial expectation from this experiment was that diagnostic labels and cognitive tests should be able to substitute for one another, i.e., they should be able to transform the feature space when being used as targets for a specific model. However, this was not the case in our experiments. While setting up the experiments, we also tested for the applicability of the model to predict other cognitive tests. Among all these three cognitive scores of (MMSE, RAVLT, and ADAS11) the best results were obtained with multitasking MMSE with diagnosis labels. Thus, we focused on reporting the results of this setup only.

Moreover, different combinations of modalities have been investigated to provide for more meaningful comparisons with other reported studies. Results provided in Figure 5.7 demonstrate the influence of the different combinations of modalities in predicting the MMSE scores. Four different modality combinations have been considered, where RF signifies risk factor parameters, and C1 to C4 refers to the various combinations of the different modalities as indicated in the legend of Figure 5.7.

Also, the accuracy of the multiclass classification for predicting the progression of AD in a period of 24 months in terms of their categorical labels is shown in Figure 8. It should be noted that for the sake of uniformity, all the results reported in this study are generated using the same network shown in Figure 5.3. Therefore, the network that has been analyzed to yield the results shown in Figure 5.7 and Figure 5.8 used the

hyperparameters (optimizer, dropout rate, decay rate, hidden layer size, and so on) that have been optimized exclusively with respect to the five modalities considered (MRI, PET, CSF, COG, and DEM).

Table 5.4. Summary of prediction tasks accomplished in the literature

Method	Multitask	Classification Type	Class Name	Regression Type	Modality	Subjects
[165]	No	Multiclass	AD-MCI-CN	-	MRI	397
RELM [165]	No	Multiclass	AD-MCI-CN	-	MRI	214
[160]	No	Multiclass	AD/MCI/CN & AD/MCI-C/MCI-NC/CN)	-	MRI – PET	202
JRMI[145]	Yes	Multiclass	AD/MCI/CN & AD/ MCI-C/MCI-NC /CN	Single time point	MRI – PET	202
DM2L[144]	Yes	Binary & Multiclass	AD/MCI/CN & AD/pMCI/sMCI/CN	Single time point	MRI – Demographic	1984
DW-S2MTL[166]	No	Binary & Multiclass	AD/MCI/CN & AD/pMCI/sMCI/CN	-	MRI - PET – CSF	805
SMKMTL[52]	No	Binary	AD/ MCI-C/MCI-NC / CN	Multiple Cognitive Scores	MRI	788
SAE[138]	No	Multiclass	AD/ MCI-C/MCI-NC / CN	-	MRI & (MRI+PET)	758 – 331
SMTL[159]	No	-	AD/MCI/CN	4 time points	MRI	445
MSMT[95]	No	-	CN/ MCI/AD	4 time points	Multimodal	818
CNN[167]	No	Binary	AD/pMCI/sMCI/ CN	-	MRI + PET	397
M3T[93]	Yes	Binary	MCI-C/MCI-NC & AD/CN & MCI/ CN	2y changes of MMSE	MRI + PET + CSF	186
MSJL[149]	No	Binary	AD/CN, MCI/CN, MCI-C/MCI-NC	Single time point	MRI + PET + CSF	202

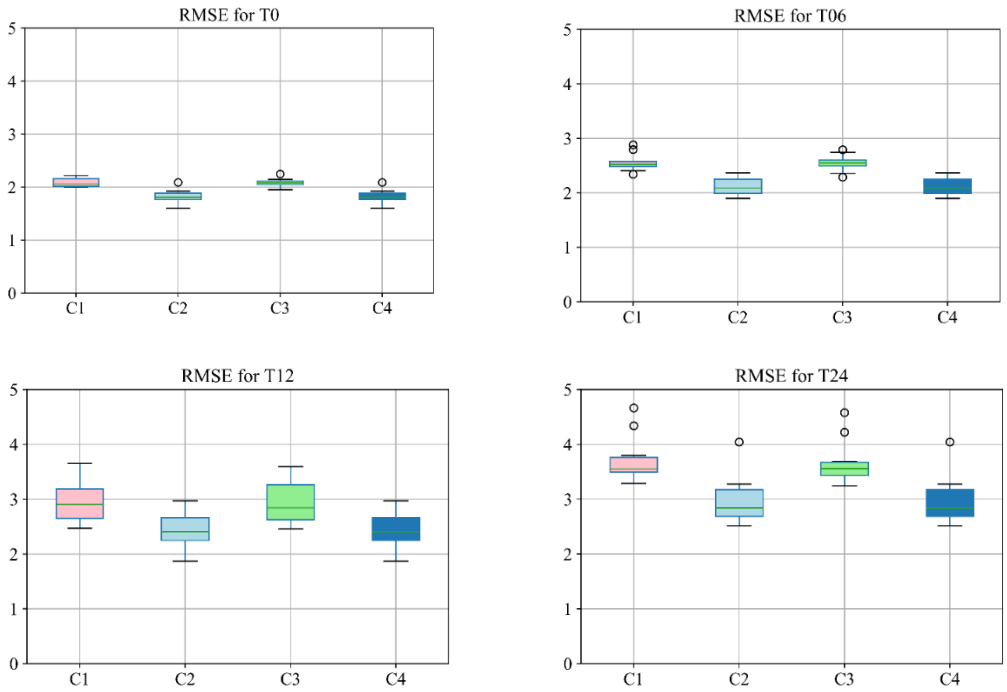


Figure 5.7. Boxplot for RMSE of mixture category of subjects using different combinations of modalities. Here C1 stands for MRI+PET+RF, C2 stands for MRI+PET+RF+COG, C3 stands for MRI+PET+RF+CSF, C4 stands for MRI+PET+RF+COG+CSF

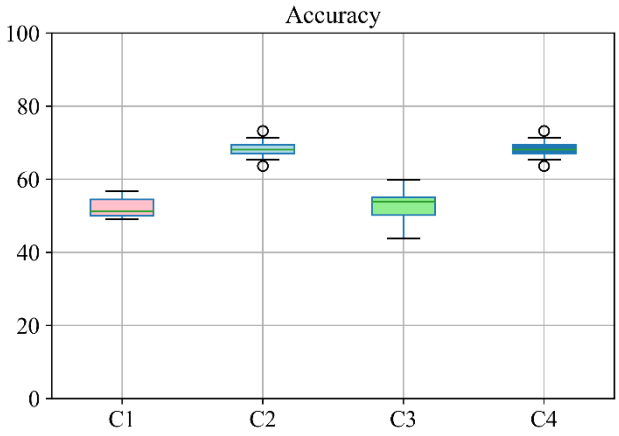


Figure 5.8. Boxplot for accuracy of multiclass classification achieved through the proposed network based on a different combination of modalities. Here C1 stands for MRI+PET+RF, C2 stands for MRI+PET+RF+COG, C3 stands for MRI+PET+RF+CSF, C4 stands for MRI+PET

Chapter 6 A Unique Color-Coded Visualization System with a Fully-Integrated Machine Learning Model for the Enhanced Diagnosis and Prognosis of Alzheimer's Disease

6.1 Introduction

Most common methods that help in the visualization and diagnosis of AD are typically accomplished through the use of web-based visual platforms, manifold decision spaces, locally linear embedding (LLE), latent profile analysis (LPA), heat maps, brain connectivity maps, specific AD neuroimaging signatures [168]–[180], and PET imaging for amyloid positivity models and the staging of amyloid burden as exemplified in [181]–[185]. There is merit to all these different venues at imaging the disease. For example, brain connectivity maps, iso maps, locally linear embedding (LLE), principal component analysis (PCA), 3D scattering transforms, latent profile analysis (LPA), and the concept of histons are all interesting methods for visualizing data, but they are more useful for dimensionality reduction and classification purposes and are not amenable to intuitive visual interpretations. For example, the histon is simply a new adaptation of the standard 2-D image histogram such that the elements in the histon are pixels classified as belonging to the same segment through image segmentation rather than having a certain grayscale of the standard histogram. On the other hand, LLE is another manifold learning dimensionality reduction method similar to iso map except that it finds a set of weights that approximate data points through local linear interpolations. Also, 3D heatmaps All these methods are indeed visual methods reflecting a decisional space that help more in the classification process than in facilitating a visual interpretation of a diagnosis and/or prognosis of the disease trajectory.

The challenges for understanding AD are in deciphering the interplay between the different biomarkers for enhanced diagnosis, multiclass classification, and regression analysis, especially as it relates to the pathogenesis of the disease [186]–[190] and its early detection [191], [192]. There is also wide-ranging deliberation on the nature of cognitive reserve [193], [194] in potentially biasing the cognitive tests and hence the diagnosis. There is also the issue of chronology in the manifestation of amyloid-beta plaques and tau tangles [195]–[197] and their synergistic effects on AD pathology. Moreover, there is the important issue of the APOE genotype [198]–[200] and its association with cognitive reserve, cortical thinning, as well as with its potential link to both amyloid-beta and tau aggregation and the cerebrospinal fluid (CSF) biomarker [201], [202]. The central aim in all these studies is in identifying the earliest manifestations of AD so that preventive measures can be undertaken and early treatment/therapeutic interventions can be considered [203]–[205]. The data used for this study involving the "QT-PAD Project Data" encompasses all these aforementioned biomarkers which are used for the machine learning design model to generate the visual outcome which is to be compared with the target image in terms of the disease trajectory.

An effective way to approach such complex and challenging issues is to process multimodal data through machine learning [206]–[208]. However, we believe that such ML models will be more informative if they included some form of visualization that could facilitate understanding of the inner workings of the ML model with regards to the resulting visual outcome. The assertation here is that the means to assess the importance of features and the interpretability of results will be enhanced [209]–[211]. It is emphasized here that although the goals of seeking (a) high accuracy in the multiclass classification and (b) prediction of disease trajectory at future time points

based on baseline features remain important, the belief here is that it would be more helpful if we understood the subtle nuances in the resulting visual outcomes as a glimpse at the inner workings of machine learning, especially when dealing with converters. Ultimately, this form of visualization via machine learning informs on the challenges faced with multiclass classification and adds insight into the decision-making process. It may also shed some light on the opacity of the black box problem associated with machine learning. Moreover, this will also augment the deliberation process by reassessing the difficult cases, like the converter cases to determine whether a misclassification is deemed an outright misclassification or if the visual outcome from the ML system should be further scrutinized in context to other biomarkers for a more complete picture as to the differences between target and ML outcome.

6.2 Methods

6.2.1 Data and Study design

The clinical data used in the preparation of this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). Only subjects that have a baseline (T0) scan (in at least one feature modality) and showed up for follow-up visits at T6, T12, and T24 have been considered in this study, leading to a total of 1123 subjects as shown in Table 1. These subjects are categorized by ADNI into the 3 classes of CN, MCI, and AD at baseline and for each of the referral sessions. The input features used for each modality, along with the number of observations made at the different time points are obtained from the "QT-PAD Project Data" AD Modelling Challenge [http://www.pi4cs.org/qt-pad-challenge] as given in Table 2. Hence, inputs to the ML model contain features from baseline including MRI and PET sequences, demographic information, and specific cognitive measurements.

The outputs of the ML network are automatically generated as an image containing colorful strips expressing disease progression at different time points. It is important to emphasize that in designing this color-coded visualization scheme, the Mini-Mental State Examination (MMSE), Clinical Dementia Rating Sum of Boxes (CDR-SB) scores, since both initially used for the labeling of subjects and Alzheimer's Disease Assessment Score (ADAS11, ADAS13), for their correlation to MMSE and CDR-SB, were excluded from the input feature space in the training and testing phases so as not to bias the ML model. Each feature set is normalized by mean normalization over its non-missing values, i.e. $\bar{\boldsymbol{f}} = (\boldsymbol{f} - mean(\boldsymbol{f})) / (max(\boldsymbol{f}) - min(\boldsymbol{f}))$ for the vector $\boldsymbol{f}$ including all non-missing values of a given feature f. Then the missing values of $\bar{\boldsymbol{f}}$ are set to 0 to make sure they do not affect the network training phase.

Table 6.1. Study population and subgroups

Categories based on diagnosis					Categories based on conversion		
	Number of samples #			Total		#	Description
	NC	AD	MCI		AD	163	Stable Dementia
Baseline	331	163	629	1123	MCI	442	Stable MCI
6th month	331	195	597	1123	CN	329	Stable Normal
12th month	332	243	548	1123	MCIc	181	MCI converter to AD
24th month	334	342	447	1123	others	8	Others (e.g. MCI to CN)
						1123	Total

Table 6.2. ADNI dataset with the features extracted from each modality/source

				Number of subjects: 1123	
Modality	**Feature**	**Minimum Value**	**Average Value**	**Maximum Value**	**Number of missed values at baseline**
MRI	Ventricular volume	5650.0	39420.220	145115.0	39
	Hippocampus volume	3091.0	6798.67	10769.0	158
	Whole Brain volume	738813.0	1022118.21	1443990.50	18
	Entorhinal Cortical thickness	1426.0	3507.23	5896.0	160
	Fusiform	8991.0	17354.76	26280.0	160
	Middle temporal gyrus	9375.0	19545.76	29435.0	160
	Intracranial volume (ICV)	1116279.11	1536383.48	2072473.30	8
PET	'FDG'	0.69	1.24	1.707168	321
	Pittsburgh Compound-B (PIB)	1.18	1.53	1.89	1116
	'AV45'	0.83	1.19	2.02	614
Cognitive Test	RAVLT immediate	7.0	35.59	71.0	3
	RAVLT learning	-2.0	4.29	11.0	3
	RAVLT forgetting	-5	4.35	13.0	3
	RAVLT percforgetting	-100.0	57.37	100.0	4
	Functional Activities Questionnaires (FAQ)	0.0	3.73	30.0	4
	Montreal Cognitive Assessment (MoCA)	10.0	23.78	30.0	616
	Everyday Cognition (Ecog): 'EcogPtMem'	1.0	2.12	4.0	613
	Ecog: 'EcogPtLang'	1.0	1.73	4.0	612
	Ecog: 'EcogPtVisspat'	1.0	1.37	4.0	614
	Ecog: 'EcogPtPlan'	1.0	1.40	4.0	612
	Ecog: 'EcogPtOrgan'	1.0	1.48	4.0	624
	Ecog: 'EcogPtDivatt'	1.0	1.79	4.0	615
	Ecog: 'EcogPtTotal'	1.0	1.67	3.82	612
	Ecog: 'EcogSPMem'	1.0	2.01	4.0	615
	Ecog: 'EcogSPLang'	1.0	1.56	4.0	614
	Ecog: 'EcogSPVisspat'	1.0	1.38	4.0	622
	Ecog: 'EcogSPPlan'	1.0	1.50	4.0	616
	Ecog: 'EcogSPOrgan'	1.0	1.57	4.0	638
	Ecog: 'EcogSPDivatt'	1.0	1.78	4.0	621
	Ecog: 'EcogSPTotal'	1.0		3.89	614
CSF	Amyloid Beta (ABETA)	203.0	852.54	1697.0	449
	phosphorylated tau protein (PTAU)	8.21	27.45	94.86	338
	Total tau protein (TAU)	81.54	284.98	816.9	337
Risk Factors	Age	55.0	73.93	91.4	0
	years of education	6.0	15.92	20	0
	APOE4	0	0.56	2	0

6.2.2 Color coding

The adage "a picture is worth a thousand words" together with the challenge imposed by both the variability and interrelatedness of the multimodal features served as an incentive to create the ML4VisAD model. The (23×23×3) target images are color-coded and include a region of uncertainty (RU) as represented by the additional black bar. These 23×23×3 target images are exemplified in Figure 6.1. The three channels (R, G, B) are used to represent the state of the disease with different colors, AD: red, Mild Cognitive impairment-MCI: blue, and Cognitively Normal-CN: green. In this color-coded scheme, subjects that are stable over time would display a single color as in cases (a) through (c), and subjects who convert at certain time points to other states would be displayed with two or more colors as in case (d) through (g).

Trajectories of cognitive status are defined through a 24-month timeline (including baseline T0 and three referral sessions T1 (6^{th} month), T2 (12^{th} month), and T3 (24^{th} month). To assess the degree of uncertainty that could be introduced by the machine learning model, a black bar, an area reserved for uncertainty (RU) is included at the end following the T3 time point. It is emphasized that the black bar could be situated anywhere in this display and is used solely to estimate the degree of uncertainty that the machine learning algorithm injects into the ML visual outcome through its many inner computations. This should mean that the black bar in the output image should remain unchanged or unaffected if the machine learning is stable and has performed its task reliably. For any other noted change, the difference in the black bar between target and output image can provide an estimate of what we refer to as the degree of uncertainty that may have permeated the decision-making process. The size 23x23 of the RGB image could have been of any *NxN* dimension. In the discussion section, we provide a Figure to explain that a larger target image (e.g., 45×45) would provide added

resolution in the output image with smoother transition phases. It is noted, however, that the higher *N*, the more convolutional layers are needed in the ML model, and hence the more processing time would be needed as detailed in section 2.

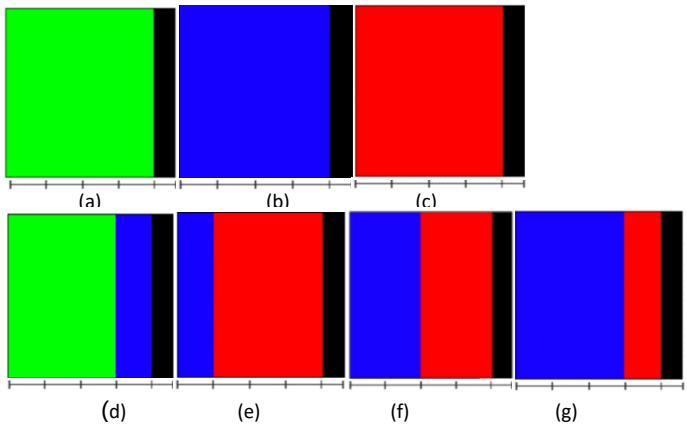


Figure 6.1. Target Images showing: (a) stable CN, (b) stable MCI, (c) stable AD, (d) CN converting to MCI at T24, (e, f, and g) are MCI that progressed to AD at time points T6, T12, and T24, respectively. The black-coded bar is added to inform on the level of uncertainty generated by the machine learning process.

6.2.3 Machine Learning Architecture

In designing the machine learning architecture shown in Figure 6.2, the standards of *stability*, *sparsity*, *interpretability*, and *accuracy* are sought with the ability to handle *missing data* and *multicollinearity* inherent to longitudinal studies [181], [210]–[217]. This architecture embeds weighing the relevance of features and means (through the black bar or what we refer to as the uncertainty region) for determining effects injected by the ML model on the visual outcome. The network is designed such that, initial layers address the intra-modality feature extraction via fully connected layers. Then, feature fusion and feature extraction for the inter-modality phase would be addressed via concatenation, a fully connected layer, tensorization with reshaping, and several transposed convolutional layers with different dilation rates. Drop-out and batch

normalization are also applied in different layers to prevent overfitting. All the kernel sizes are 3x3 and the padding style is 'same' in layer L6 in contrast to 'valid' in L8 and L9. Design details and tensor dimensions for the different layers are presented in Figure 6.2, with supplementary materials provided in the GitHub repository (https://github.com/mohaEs/ML4VisAD).

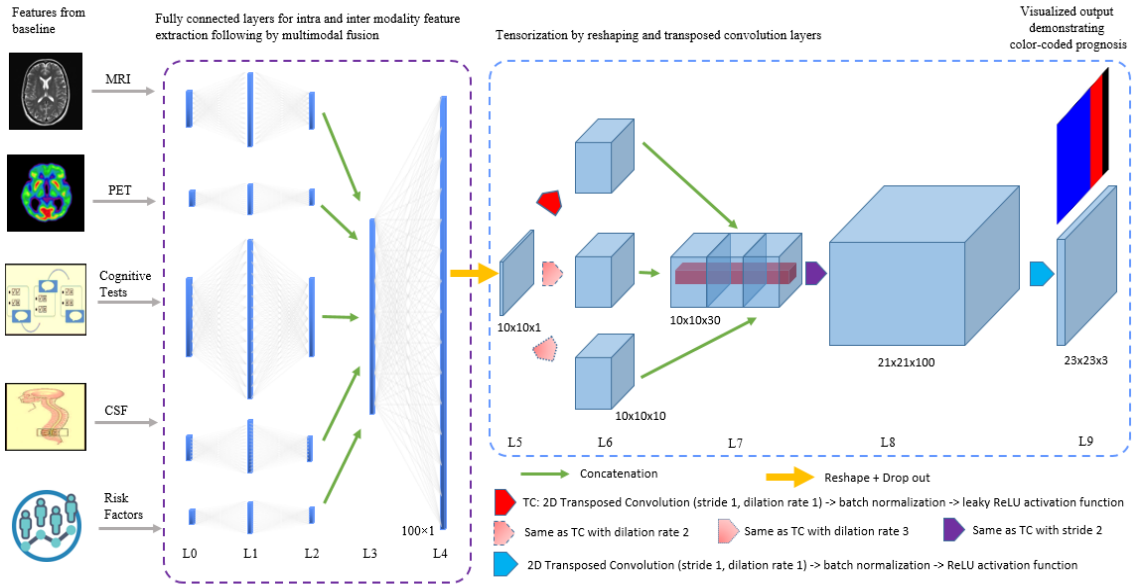


Figure 6.2. Machine learning design architecture based on convolutional neural networks with input features conform to the ADNI qt-pad-challenge and with a color-coded visual output describing disease trajectory

6.2.4 Training and Evaluation

The loss function is the Mean Absolute Error between the target image and the produced output. 10-fold cross-validation is considered in this study, and in each training session, 10 percent of data are used as a validation set (i.e. 10 times of training data split to 80/10/10 percent's as train/validation/test). The network is trained for 3000 epochs with a batch size of 500. The network for producing a larger 45x45 image size is similar to the network shown in Figure 6.2 and just the L8 layer is replicated. The network is developed through the use of Keras TensorFlow deep learning frameworks.

Using the GPU NVIDIA Geforce RTX 2080, the time it took from feeding the input to the ML model shown in Figure 6.2 to obtaining the machine learning visual outcome as a function of the image size is as follows.

Table 6.3. Processing time of machine learning model

Image Size (pixels)	Trainable parameters	Train time (sec)	Test time per subject (sec)
23x23	36,143	3000 epochs: 212.67	0.008
45x45	126,443	3000 epochs: 767.94	0.017

6.3 Results

In order to demonstrate proof of concept, different scenarios as shown in Figure 6.3 are considered to include subjects that are stable over time and subjects that transition at different time points from one state to another. These varied examples highlight the practical merits this color-coded visualization could have in facilitating diagnosis and prognosis. For each subject in the testing phase (not seen in the training phase), color-coded patterns are generated based on observed features at baseline and associated target images, demonstrating a different approach to visualizing disease progression through machine learning. Figure 6.3 provides several examples that reflect different target images and the respective visual outcomes that the ML model produces. In an effort to include different scenarios, we processed four cases for the stable cognitively normal (CN) group with the green-colored target at all four points in the first row, 4 cases of the stable mild cognitive impaired (MCI) with a blue-colored target in the second row, 4 cases who have transitioned from CN to MCI or from MCI to AD at different time points as shown respectively in the 3rd and 4th rows, 4 cases of stable AD subjects in the 5th row, and in the last row, we show 3 ambiguous cases where the target images display 3 stable cases where the ML visual outcome results in a different

disease state than the actual diagnosis of the target image. For each case, the target image is shown on the left, and the ML visual outcome is shown on the right. For added context, we also provide MMSE and CDR scores, the APOE status, and the SUVR measurements where the x-axis reflects the different brain regions for the SUVRs as annotated in Table 2, all obtained at baseline. The scores/values used for MMSE and CDR are conformed to the standards defined by ADNI. The APOE value of 0,1, or 2 specify a carrier of zero, one or two APOE e4 alleles.

For this study, three raters (M.E., S.T., and M.S.) reviewed independently all ML-generated visual outcomes for both types of classification: 3-way (CN, MCI, AD) and 4-way (CN, MCI, MCIc, AD). They were tasked to declare each visual outcome as classified correctly (X1=1), misclassified (X2=1), or inconclusive (X3=1), with Xi=1, 0 for i=1, 2, 3 with $\sum Xi$=1 for each rater. An agreement is reached when all three raters provide the same classification. The results are shown in Table 4.

Table 6.4. Classification outcomes as assessed by three raters

Classification Type	Correctly Classified	Misclassified Outcomes	Inconclusive Outcomes
3-Way (CN, MCI, AD)	0.82±0.03	0.15±0.004	0.023±0.002
5 way (CN, MCI, MCIc, AD, others)	0.68±0.05	0.29±0.01	0.023±0.002

The results obtained indicate that for most of the stable cases when using a 3-way classification (CN, MCI, AD), the machine learning model was relatively accurate with an 82% accuracy. This is exemplified in Figure 3, where cases (a), (b), (e), (k), and (q) clearly show that the ML visual is in agreement with the target image. It is observed that most of the misclassifications happen either because the ML visual outcome predicted a transition that was not reflected in the diagnosis as in cases (h), (r), and (s) or vice versa where the ML did not pick up a transition present in the target image as

in case (j) and (l), or that the ML visual outcome predicted a different stable disease state than that of the target image as in cases (u), (v) and (w). There were also cases with disagreements as to when the transition happened such as in cases (m) and (p). In fact, we observe that most of the misclassification happened for the converter cases with the conversion time points being the most contentious to resolve. It is observed that in general MMSE and CDR scores seem to guide the diagnosis reflected in the target images, while SUVRs, years of education (Edu), and APOE favor the ML-generated visual outcome in terms of the disease trajectory. When using a 5-way classification (CN, MCI, MCIc, AD, others), the classification accuracy drops to 68%. This drop in accuracy is now conceivable given all these variations in the ML visual outcomes, which as will be detailed in the discussion section, few of these cases should instead be revisited and should not necessarily be dismissed as misclassifications.

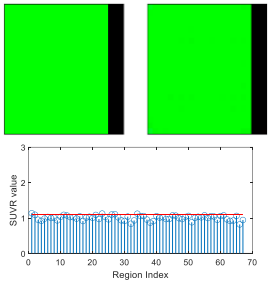
Moreover, in some of the cases, the ML model provided more sensical visual outcomes than what the diagnoses reflect in the target image. For example, case (j) shows a subject with a target image expressing an MCI converting back to CN at T24, while the ML visual outcome finds the subject to be CN with traces of MCI towards T24, suspecting that the CDR values are what contributed to the diagnoses reflected in the target image. Another such case is (v) where the target image shows a stable MCI, while the ML visual outcome places this subject as stable CN. For this case, it seems that from the high MMSE score, the low SUVR values, an APOE of 0, although the CDR is 0.5, the ML visual outcome of a stable CN seems more logical. For this case, other cognitive tests (ADAS, RVALT) may have influenced the diagnosis.

For a more meaningful assessment of disease trajectory, especially in light of the disagreements and other visual nuances and variations that the ML model generated, additional context is provided in Figure 6.5 for the challenging cases. The contextual

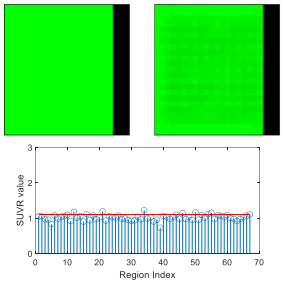
information added includes MMSE and CDR score for all four time points (T0, T6, T12, and T24), SUVR measurements for first and last time points (T0 and T24), age, sex, and years of education. As all of these challenging cases shown were misclassified by the ML model, the intent here is to use such context to deliberate on what may have led to the differences between target images and ML visual outcome. Details are provided in the Discussion Section.

Table 6.5. Brain regions for the SUVRs shown in Figure 6.5

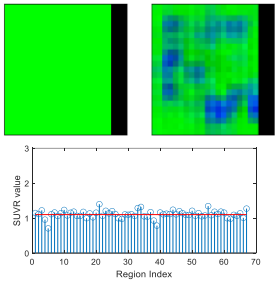
SUVR regions considered		
1)LH_CAUDALANTERIORCINGULATE 2)LH_CAUDALMIDDLEFRONTAL 3)LH_CUNEUS 4) LH_ENTORHINAL 5) LH_FRONTALPOLE 6) LH_FUSIFORM 7) LH_INFERIORPARIETAL 8) LH_INFERIORTEMPORAL 9) LH_INSULA 10) LH_ISTHMUSCINGULATE 11) LH_LATERALOCCIPITAL 12) LH_LATERALORBITOFRONTAL 13)LH_LINGUAL 14)LH_MEDIALORBITOFRONTAL 15)LH_MIDDLETEMPORAL 16)LH_PARACENTRAL 17)LH_PARAHIPPOCAMPAL 18)LH_PARSOPERCULARIS 19)LH_PARSORBITALIS 20)LH_PARSTRIANGULARIS 21)LH_PERICALCARINE 22)LH_POSTCENTRAL 23)LH_POSTERIORCINGULATE	24)LH_PRECENTRAL 25)LH_PRECUNEUS 26)LH_ROSTRALANTERIORCINGULATE 27)LH_ROSTRALMIDDLEFRONTAL 28)LH_SUPERIORFRONTAL 29)LH_SUPERIORPARIETAL 30)LH_SUPERIORTEMPORAL 31)LH_SUPRAMARGINAL 32)LH_TEMPORALPOLE 33)LH_TRANSVERSETEMPORAL 34)RH_BANKSSTS 35)RH_CAUDALANTERIORCINGULATE 36)RH_CAUDALMIDDLEFRONTAL 37)RH_CUNEUS 38)RH_ENTORHINAL 39)RH_FRONTALPOLE 40)RH_FUSIFORM 41)RH_INFERIORPARIETAL 42)RH_INFERIORTEMPORAL 43)RH_INSULA 44)RH_ISTHMUSCINGULATE 45)RH_LATERALOCCIPITAL 46)RH_LATERALORBITOFRONTAL	47)RH_LINGUAL 48)RH_MEDIALORBITOFRONTAL 49)RH_MIDDLETEMPORAL 50)RH_PARACENTRAL 51)RH_PARAHIPPOCAMPAL 52)RH_PARSOPERCULARIS 53)RH_PARSORBITALIS 54)RH_PARSTRIANGULARIS 55)RH_PERICALCARINE 56)RH_POSTCENTRAL 57)RH_POSTERIORCINGULATE 58)RH_PRECENTRAL 59)RH_PRECUNEUS 60)RH_ROSTRALANTERIORCINGULATE 61)RH_ROSTRALMIDDLEFRONTAL 62)RH_SUPERIORFRONTAL 63)RH_SUPERIORPARIETAL 64)RH_SUPERIORTEMPORAL 65)RH_SUPRAMARGINAL 66)RH_TEMPORALPOLE 67)RH_TRANSVERSETEMPORAL



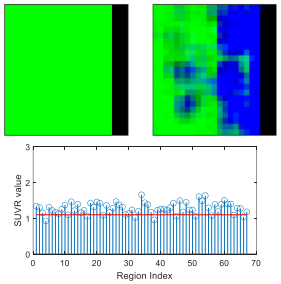
a) MMSE: 29, CDR: 0, APOE: 0



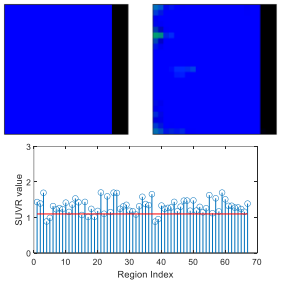
b) MMSE: 30, CDR: 0, APOE: 0



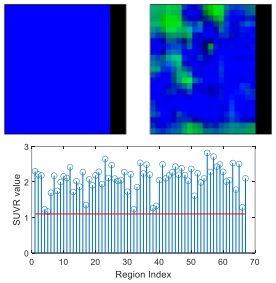
c) MMSE: 26, CDR: 0, APOE: 0



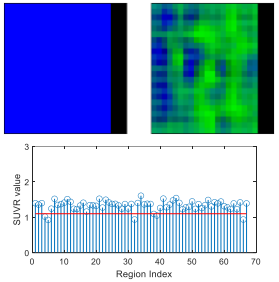
d) MMSE: 29, CDR: 0.5, APOE: 0



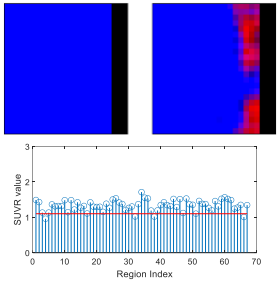
e) MMSE:24,CDR:3.5,APOE: 1



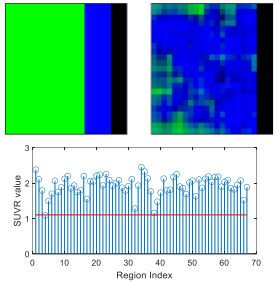
f) MMSE:28, CDR:1, APOE: 1



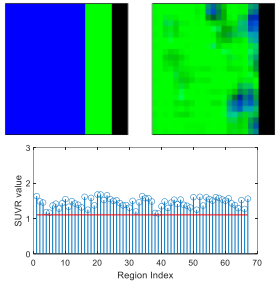
g) MMSE:29, CDR:1, APOE: 0



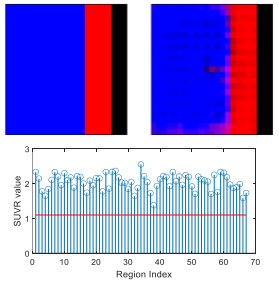
h) MMSE:29, CDR:1.5, APOE: 2



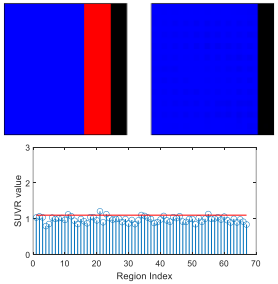
i) MMSE:28,CDR:0,APOE:0



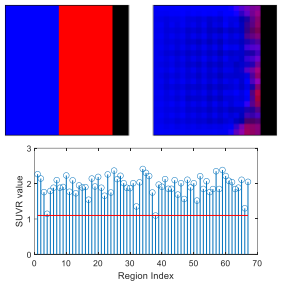
j) MMSE:29, CDR:1,APOE:0



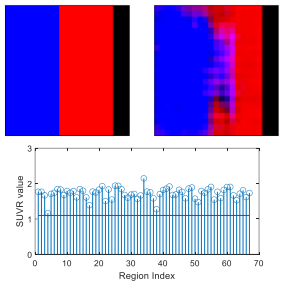
k) MMSE:27, CDR: 2.5, APOE 1



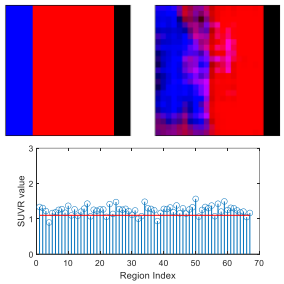
l) MMSE: 28, CDR: 5.5, APOE: 0



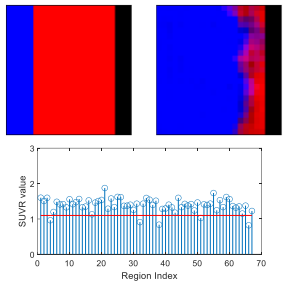
m) MMSE:26,CDR:3.5, APOE 0



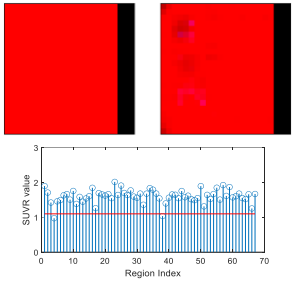
n) MMSE:28,CDR:3,APOE: 1



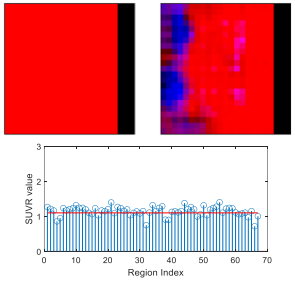
o) MMSE: 25, CDR: 4, APOE: 2



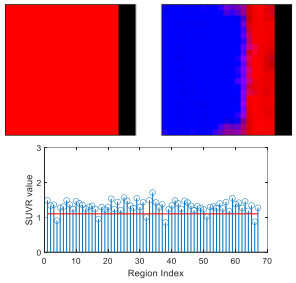
p) MMSE:28,CDR:2.5, APOE: 1



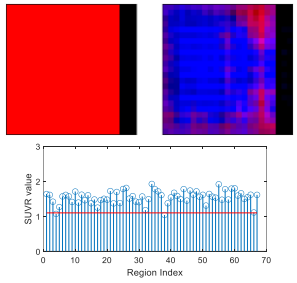
q) MMSE: 22, CDR: 5, APOE: 1



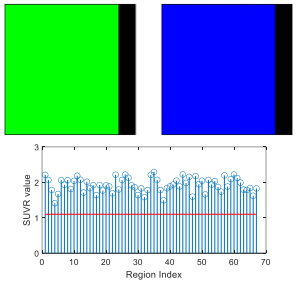
r) MMSE: 21, CDR: 6, APOE: 0



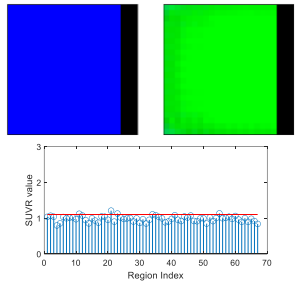
s) MMSE: 25,CDR: 5.5, APOE:2



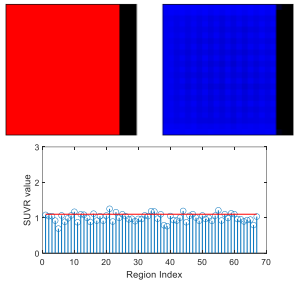
t) MMSE: 20, CDR: 5.5, APOE: 0



u) MMSE:29, CDR:1, APOE: 2



v) MMSE: 29, CDR:0.5,APOE:0v)



w) MMSE:25, CDR:3.5, APOE:1

Figure 6.3. Visualization of AD: The left and right images in each sub-figure are target and ML visual output, respectively. (a) through (d) show 4 different cases of stable CN subjects; (e) through (h) 4 different cases of stable MCI subjects; cases (i) through (p) show subjects who have transitioned either from CN to MCI or from MCI to AD at different time points; cases (q) through (t) show 4 different cases of stable AD subjects. Cases (u), (v) and (w) in the last row are challenging stable cases where the ML outcome is completely different than the target.

* The patient/record (RIDs) of the shown cases of ADNI dataset are as follow:

a) 4376, b) 4491, c) 4421, d) 4422, e) 4531, f) 2068, g) 4871, h) 4346, i) 4277, j) 4813, k) 2047, l) 4426, m) 4595, n) 4167, o) 4542, p) 4189, q) 4252, r) 4338, s) 4494, t) 4001, u) 4339, v) 4226, w) 4676

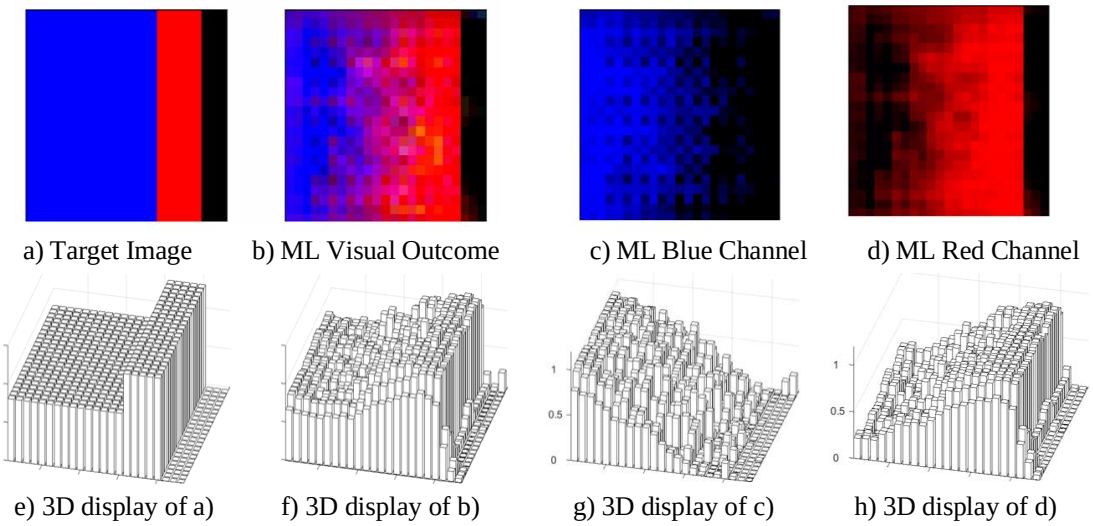


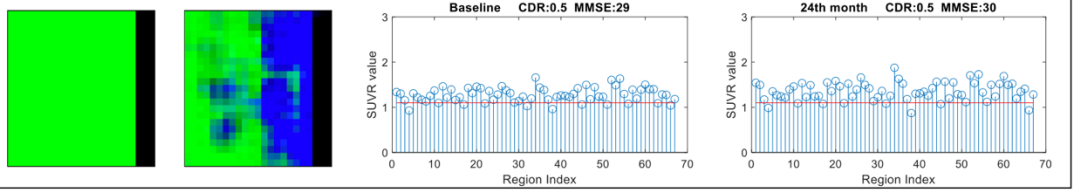
Figure 6.4. 3D Display of the RGB channels of an MCI case that transitioned to AD at T24. Note the gradual change in the ML generated displays. Also note how minimally the ML model affected the region of uncertainty (RU) in the 3D displays in f, g and h.

It should be noted that the proposed ML model is designed with the ability to display all these results in 3D as shown in Figure 6.4. For 3D visualization, the L-*a*-*b* format, which is a 3D variation of the CIE chromaticity diagram, can be used to display in 3D the RGB format without changing the contextual meaning of the outcomes reflected in the examples considered in Figures 6.3 and 6.5. In this L-*a*-*b* format, L refers to

lightness normalized from zero to 1, and *a* and *b* reflect the colors from green to red for *a* and from blue to yellow for *b*. Target and output images are shown in Fig 4a and Fig 4b, and the blue and red channels in 5c and 5d, respectively. Their respective 3D displays are shown in 4e-hh. Note the gradual change in the ML-generated visual outcomes. Observe that at T24 the ML visual outcome in 5f does indeed stabilize at the highest levels near the normalized value of 1, which is reflected in the red channel of 5h. Moreover, observe that as the blue channel reflecting the MCI state declines rapidly between T12 and T24, the red channel in 5h reflecting the AD state increases between T12 through T24 to stabilize at the maximum value of 1. Note how easy it is to ascertain the effect the ML model has on the region of uncertainty in displays (f), (g), and (h).

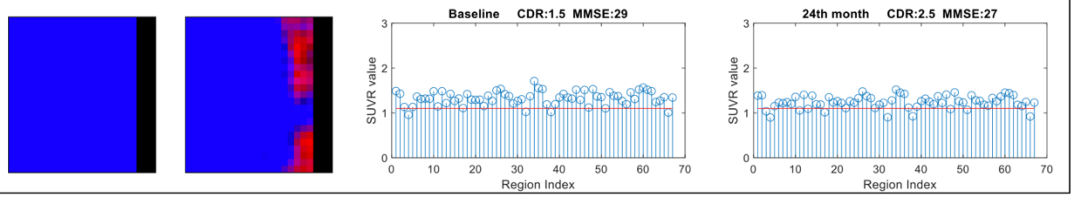
Case (d) Fig. 3

Fig #	RID	MMSE				CDR				AGE	Sex	# Edu
		T0	T6	T12	T24	T0	T6	T12	T24			
d	4422	29	30	30	30	0.5	0	0.5	0.5	70.8	Female	15



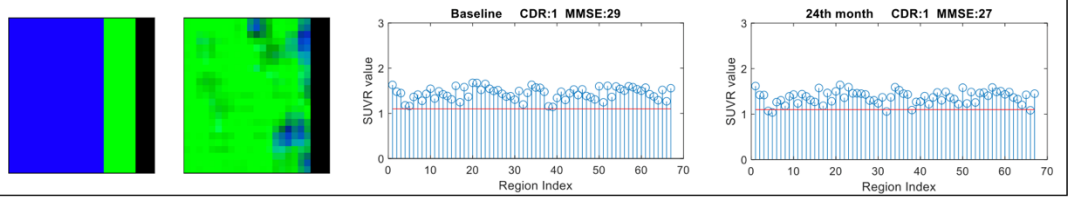
Case (h) Fig. 3

Fig #	RID	MMSE				CDR				AGE	Sex	# Edu
		T0	T6	T12	T24	T0	T6	T12	T24			
h	4346	29	29	30	27	1.5	3	2.5	2.5	71.3	Male	18



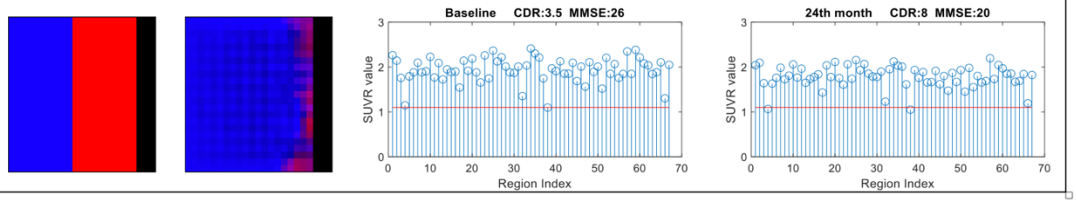
Case (j) Fig. 3

Fig #	RID	MMSE				CDR				AGE	Sex	# Edu
		T0	T6	T12	T24	T0	T6	T12	T24			
j	4813	29	29	27	27	1	1	1	1	67.6	Male	16



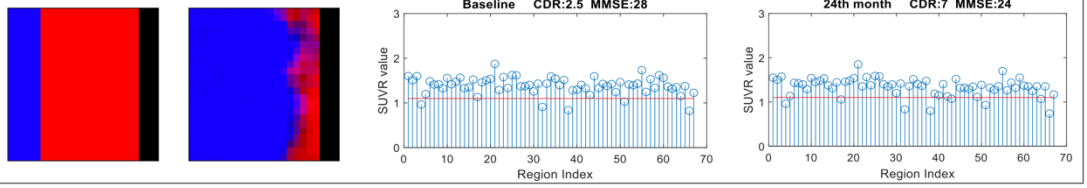
Case (m) Fig. 3

Fig #	RID	MMSE				CDR				AGE	Sex	# Edu
		T0	T6	T12	T24	T0	T6	T12	T24			
m	4595	26	27	24	20	3.5	2.5	4.5	8	76.6	Male	16



Case (p) Fig.3

Fig #	RID	MMSE				CDR				AGE	Sex	# Edu
		T0	T6	T12	T24	T0	T6	T12	T24			
p	4189	28	26	24	24	2.5	4.0	3.0	7.0	73.1	Female	15



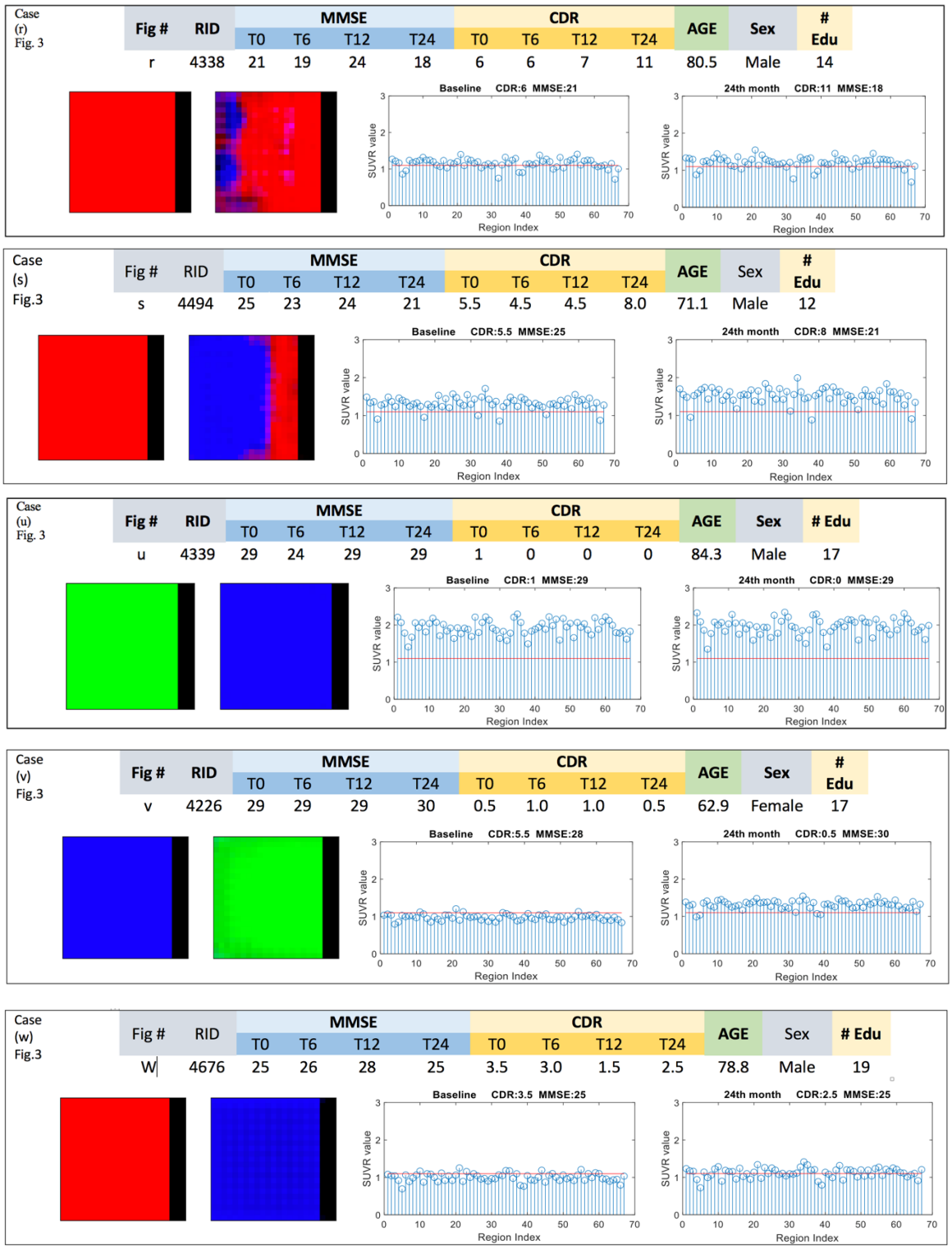


Figure 6.5. Visualization of AD Trajectory: These are complex cases from Figure 3 cases ((d), (h), (j), (m), (p), ((r), (s), (u), (v), and (w) with additional context provided, including MMSE and CDR scores for all 4 time points, Age, Sex, Years of Education, and the SUVRs for T0 and T24

6.4 Discussion

These results obtained show the need for deep reflection, especially when observing all these subtle nuances of the ML visual outcome. In this discussion of the results, recall that MMSE, CDRSB, ADAS11, ADAS13 were excluded as input measures for ML models so as not to bias the results, and if such features are displayed in the figures, it is for context only.

In the results displayed in Figure 6.3, cases a, e, k, and q clearly show that the ML visual outcomes agree fully with the target image, and these would certainly show as correctly classified as per the results in Table 4. Cases b, n, and o show that the ML visual outcome is close to the target image but could be seen as a misclassification by any one of the raters, especially if they are strict raters, and such ratings lower accuracy in classification just as it would for any automated multiclass classification algorithm.

Moreover, there are cases like (u) and (w) which are outright misclassified, but with consideration of the SUVr values and the MMSE and CDR scores, is the ML visual outcome revealing things that need re-investigating as we compare these measures that are given here for context. Also, for case v which clearly shows a misclassification, could it be that the ML model was fooled by the low SUVr values, although the CDR is rather low (varying from 0.5 to 1) and the MMSE is somewhat lower. We were truly surprised by the high SUVrs values for cases like f and i, and the ML model was affected differently for each case where the influence came more from the imaging modalities and other cognitive scores used as input for the ML model as shown earlier in Table 2. Could it be that years of education (possibly cognitive reserve) played a role in cases like d, l, and especially u?

Reflecting on Figure 6.3, the machine learning output images of cases a and b, which are identified as stable cognitively normal (CN) through their target images, suggest

faint traces of MCI in case a and more pronounced traces of MCI in case b, especially when the uncertainty bar remained unaffected. Can such outcomes lead to such diagnosis, which is an intuitive way to determine the different stages in the progression of AD, perhaps different than the EMCI and LMCI stages. But before we could validate such assertions, we need to investigate further such outcomes in terms of the level of uncertainty introduced via machine learning and in terms of delineating any effect such as the ringing effect that could be due to the convolutions and the different filters that were used.

Now consider challenging cases like d, h, l, and r. Could it be that case d transitioned to MCI mid-way as the ML visual outcome suggests? And did case h transition from MCI to AD at time point T3, or that case l did not transition to AD despite the diagnosis, and that perhaps case r had mostly MCI characteristics at baseline before it converted to AD thereafter (T6-T24).

In several of these interesting cases of Figure 6.5, we found that the higher the number of years of education the higher is the MMSE scores high but the more likely for these scores to be inconsistent through the time points, especially for stable MCI and AD and converter cases. Also, the higher the SUVR measurements are the more likely the ML visual outcome is to shift the diagnosis from CN to MCI as in case (i) and (u), and vice versa for lower SUVR values, which led to a shift in diagnosis from MCI to CN as in case (v) and from AD to MCI as in case (w).

This type of ML visual outcome foretells also why clinicians face difficulty each time they deliberate on a given patient as to which stage in the disease they may be situated. For example, it is hard to understand why the subject in case (u) in Fig. 5 had an MMSE score of 29 for T0, T12, and T24 but an MMSE score of 24 at T6. Also, the CDR score was 1 at T0 and reverted to 0 for all other subsequent time points. Although

the diagnosis is that of a stable CN, the machine learning visual outcome places this subject as stable MCI when considering all other features. Recall that the APOE for this subject is 2 at baseline and that the SUVRs is rather high. Also, could it be that the high number of years of education for this subject (17) is what led to the high MMSE scores of 29 for T0, T12, and T24, although stumbling at T6.

Such subtle nuances could imply misclassification and may be entered as such by any of the raters as shown by the results in Table 4. Remember also that all raters, in this case, have to agree to a given classification before it is recorded as such. As a consequence, the first point to be made is that this type of multiclass classification, whether it is automated or made through a rating process as was this case, does not allow for a more thorough deliberation process as these visual nuances are so hard to decipher otherwise through tabulated data or though decisional spaces showing different overlapped regions among the considered classes. Moreover, with the ML visual outcome, given the presence of the black bar that reveals the degree of uncertainty, a clear delineation can be made between a classification made with a high degree of uncertainty from an outright misclassification. Hence, it is no revelation why the more classes are considered in a multiclass classification algorithm, the less accurate are the classification results, as reflected in the results of Table 4.

Patients Record ID (RID) numbers are provided in Figures 6.3 and 6.4 and in the setup GitHub for other researchers who are interested in validating these results or would like to perform further analysis given the nuances of the ML visual outcomes that differ from their respective target images.

Through the proposed visual outcome, the processes of multiclass classification and disease projection are augmented with the ability to visualize the inner workings of the machine learning and observe what the differences between the ML visual outcome and

target image could potentially mean. In other words, the difference between them does not necessarily mean an outright misclassification, but the nuances between them should instead be revisited as to what led to such change, especially if the RU is unaffected. This last assertation is what explains why machine learning models are challenged in the ability to produce high accuracy in multiclass classification, reinforcing the fact that the more classes are considered the lower is the accuracy. This is compounded by the fact that in longitudinal studies, disease progression is depicted through the constraints of discrete and distant chronological time steps of 6 to 12 months. Moreover, ML models require large and balanced datasets, and although we used 1,123 subjects in this longitudinal study from ADNI, which is significant, it would be preferable if we had a larger and more balanced dataset in terms of the different subgroups. Moreover, the ML model needs to learn more from the different types of converter subjects in the training phase, to determine the importance of certain features that may have led to these transition phases. This is perhaps why the ML model is challenged most at these transition points yielding most of the misclassifications due to disagreements as to when such transitions may have happened.

Moreover, it is extremely important to recognize that the interrelatedness in features, along with the many variations of such multimodal features, some being temporal, others structural, functional or metabolic, genetic, demographic, and cognitive are extremely difficult to disentangle, especially when we involve thresholds as in the case for SUVR and ranges of values for MMSE a CDR. From our own experience, when you consider the ADNI data, there is an overlap in MMSE scores between CN, MCI, and even AD groups; the CDR values may resolve somewhat this overlap, but for an ML model more datasets are needed to learn more of the interplay between such features.

Chapter 7 Conclusion

This research endeavor explored new machine learning techniques and designs that focused on extracting relevant features that define the prodromal states of Alzheimer's disease and the normal control state to evolve into determining the best approaches for multimodal multiclass classification and prediction. Investigations were carried out when these two tasks are either performed separately or performed simultaneously through multitasking.

When designing a profile-specific regression model for disease progression of Alzheimer's using longitudinal data, we presented a profile-specific SVM-RBF regression framework based on L1 feature selection and MLP classification. This framework came in support of predicting cognitive scores for AD at multiple future time steps using multimodal data. In order to understand the relationship between the predictive AD features and the temporal variations in cognitive scores, an MLP classification model has been adopted to segregate patients based on the disease severity using baseline data. For each baseline pattern of AD stages, separate sets of temporal regression kernels are trained to increase the longitudinal prediction accuracy. We investigated the prediction performance of our proposed method with other generic linear and nonlinear regression models such as Ridge, Lasso, Elastic Net, and Gradient Boosting regression. This SVM-RBF regression framework showed that utilizing the nonlinear RBF kernel built on the MLP classifiers on sets of discriminated AD features yields more accurate predictions comparing to fitting a single regression model for all classes of subjects.

This initial SVM-RBF framework stimulated interest in investigating the merits of using recurrent neural networks for the prediction modeling of Alzheimer's disease

using longitudinal data. For tracking the progression of the AD at multiple future intervals and gauging the merits and gradual effects of any potential treatment plan in longitudinal AD studies, this approach aimed to apply Recurrent Neural Networks to the ADNI dataset. Three historical time points from subjects in three categories of CN, MCI, and AD were selected to form a feature space. Then, the model is trained on 75% of the data to predict three future MMSE scores and diagnosis labels of the subjects with two different variations of RNN (LSTM and GRU). This approach showed that employing *L1* feature extraction prior to application of the RNNs leads to higher performance in both regression and classification models in comparison to other state-of-the-art algorithms, which can be observed from the results provided in Tables 3 and 4.

To focus on a single cognitive test score, namely the MMSE test score, used among other tests for labeling subjects at baseline, a distributed multitask multimodal approach was then developed for the prediction of MMSE cognitive test scores in a longitudinal study as means to gauge disease progression. Predicting MMSE over time, through multimodal longitudinal data, could augment our prospects for analyzing the interplay between the different multimodal features used in the input space in relation to the predicted MMSE scores. Such a prediction model could also be used to ascertain the effectiveness of treatment or therapeutic protocol by comparing taken MMSE tests against predicted scores by the model, allowing at the same time to observe the conversion rate in the different stages of individuals who are at risk of developing AD. A novel distributed multitask multimodal framework is introduced for predicting cognitive measures in the progression of Alzheimer's disease even when burdened with the missing data challenge. The model is capable of handling size discrepancy between the number of observations belonging to different time points and assuming different

recording modalities. The proposed approach also has the potential to directly consider the inherent temporal sparsity patterns of different modalities and their relative correlation strength. This provides flexibility in utilizing complementary information from multimodal data. Furthermore, the model also terminates the propagation of potential error from one modality to another which may have originated from corrupted data. The experimental results proved that this method can effectively predict the progression of Alzheimer's disease over a period of four years in terms of the predicted MMSE scores based solely on neuroimaging features (MRI and PET), cognitive tests that excluded those used for labeling the subjects or found to be highly correlated with the MMSE test in order to avoid any bias, cerebrospinal fluid (CSF) and other risk factors associated with age, gender, years of education, and the APOE gene. While the proposed approach mitigates the consequence of the negative correlation between various modalities, there could still be unrelated information between different tasks within a single modality. Future studies using longitudinal data may be able to improve the performance of these prediction algorithms. The general approach described for predicting progression used in this study, as expressed in Figure 4.2, could be extended not only to other longitudinal studies involving other neurological disorders but could also be used for the prediction of other cognitive scores such as ADAS11 and RAVLT to assess the singular merits of such cognitive scores and how related and correlated they may be to the MMSE test.

At a deeper level of this research endeavor, a tensorized multitask deep learning network for progression prediction and multiclass classification of Alzheimer’s disease was developed to determine if an integrated machine learning framework can provide optimal multiclass classification accuracy as well as prediction results with minimal root mean square error deviation. This aspect of the research investigates whether the

same set of features that optimizes one task would also optimize the results of the second task in one undertaking. In this line of thought, a novel neural network structure with multitask learning, modality fusion, kernelization, and tensorization was developed to predict and classify the different stages of Alzheimer's disease in a multiclass population. Using the features from baseline, this newly developed network is shown to predict the cognitive status (through the MMSE scores) of the patients in a 24-month longitudinal study involving the AD/MCI-C/MCI-NC/CN groups (taking into consideration the converter (C) and non-converter groups (NC) in the MCI category). Multitask learning has been explored to enhance prediction performance by incorporating the common relationship or interrelatedness between the regression and classification tasks. Furthermore, the power of modality fusion, kernelization, and tensorization have also been investigated to efficiently extract important features hidden in the lower-dimensional feature space without being distracted by those deemed as irrelevant. Empirical evaluations on the longitudinal multimodal ADNI dataset were conducted to evaluate the network performance. The results reveal that the proposed KTMnet framework not only predicts the cognitive scores with relatively high accuracy but can also enhance the multiclass classification accuracy for early-stage diagnosis and prognosis of the MCI conversion group. It is emphasized here that although we are aware of the overlap that exists in the MMSE scores in between subject groups, making the prediction of MMSE scores difficult, we still removed from consideration in the training phase the predictive biomarkers of ADAS13, MoCA, and CDR, which are found to be highly correlated to MMSE. Their inclusion otherwise would have favored the proposed machine learning design and would have increased the accuracy for both prediction and multiclass classification.

Intrigued by the difficulty for consolidating the tasks of multiclass classification and prediction and the opacity of the black box problem associated with machine learning, the last of this research endeavor envisioned the use of a unique color-coded visualization system with a fully integrated machine learning model for the enhanced diagnosis and prognosis of Alzheimer's disease. This system was designed to generate a visual image that portrays AD trajectory in a 2-year longitudinal study using baseline features only. Target images are created using different colors to define each stage of the disease at the 4 observation time points (T0, T6, T12, and T24), with T0 being the baseline timepoint. A unique characteristic of this model is that it is trained with known target images with color-coded diagnoses at all 4 time points to then generate a visual output that predicts disease trajectory based on baseline features only. This research could also lead to new insights as to the gradual changes that happen in between transition phases as a function of the input feature space considered. Three-way (CN, MCI, and AD) and 5-way classifications (CN, MCI, MCIc, AD and others) are considered. Since only baseline features are used as input, this design is amenable to both cross-sectional and longitudinal studies. An interesting aspect of this design is the inclusion of a black-color coded bar to the target image defined as the region of uncertainty (RU) as means to evaluate any effect this model could inject onto the results. The motive here is that a reliable model would have minimal or no effect on this black-color coded bar. Moreover, although the results are displayed mainly in 2D images, the ML model could display these results in 3D as reflected in Figure 6.5. As can be observed with the 3D displays we can gauge better the gradual transitions in the converter cases and could see clearly the effects the ML model has on the region of uncertainty. Future research in this line of research could look into defining the uncertainty region that would delineate the impact the calculations in the machine

learning have on this region but at the same time, one needs to determine if this region of uncertainty is also affected by the nature of the features selected and how well such features define the state at each transition phase of the disease. These are many of the challenges we will continue to address, which are also viewed as part of the limitations and improvements that need to be made. These include:

- Revisiting the region of uncertainty so we can define the effect of the ML model on the region of uncertainty (RU) (i.e., the black color-coded bar) itself and be able to remove such effect on the results part of the ML visual image through all time points. This is akin to finding the transfer function (H) of this effect (that may have been introduced through the different convolutions, filtering, and dilations performed by the ML model) and then perform image restoration with the image I convolved with H^{-1} in the Fourier domain.
- Since all of our input features that were fed into the ML model were from the baseline, we wish we had a balance of samples between CN, MCI, and AD. As it stands, and from the availability of data, we had nearly 4 times more MCIs than AD and twice as many MCIs than CNs, which may have skewed the training process.
- Although we dealt with 1,123 subjects, since this is a machine learning model, its efficacy is driven by the availability of many folds this number if the training phase is to capture all the nuances that distinguish the different groups in the dataset.
- Most importantly, for added interpretability, serious consideration needs to be given to the network design in terms of the inner workings of the initial layers to determine optimally, how intra-modality features could be extracted via fully connected layers and then combined in the inter-modality phase as we contend

with multicollinearity and relatedness in between these many features. In other words, we need to define better how these tensors are encoding our multimodal data to combine time, space, different numerical ranges of the cognitive scores.

- Also, beyond the necessity for the region of uncertainty, we need to take a closer look at any noise effect that may have been introduced through the different convolutions, filtering, and dilations performed by the ML model.
- As for the number of classes to be considered for such machine learning designs, our data necessitated the use of three colors (R, G, B) for the three classes (CN, MCI, AD), but if there is a need to use additional classes such as EMCI and LMCI or any other type of subgroups, we could add more colors, and augment the primary colors currently used (R, G, B) with the secondary colors of yellow, cyan and magenta (Y, C, M) to include such additional labels.

Although many of these challenges are defined in context to the visualization system described in Chapter 6, many of these ideas apply to the development of the other machine learning systems for both multiclass classification and prediction, especially if more insight is gained into the black box effect of machine learning.

REFERENCES

[1] Alzheimer Association, "2016 Alzheimer's Disease Facts and Figures," Alzheimer's & Dementia 2016, vol. 12, no. 4, pp. 1–80, 2016, doi: 10.1016/j.jalz.2016.03.001.

[2] B. M. Jedynak et al., "A computational neurodegenerative disease progression score: Method and results with the Alzheimer's disease neuroimaging initiative cohort," NeuroImage, vol. 63, no. 3, pp. 1478–1486, Nov. 2012, doi: 10.1016/j.neuroimage.2012.07.059.

[3] T. Nimmy John, S. D. Puthankattil, R. Menon, T. N. John, S. D. Puthankattil, and R. Menon, "Analysis of long range dependence in the EEG signals of Alzheimer patients," Cognitive Neurodynamics, vol. 12, no. 2, pp. 183–199, 2018, doi: 10.1007/s11571-017-9467-8.

[4] S. S. Poil, W. de Haan, W. M. van der Flier, H. D. Mansvelder, P. Scheltens, and K. Linkenkaer-Hansen, "Integrative EEG biomarkers predict progression to Alzheimer's disease at the MCI stage," Frontiers in Aging Neuroscience, vol. 5, no. OCT, pp. 1–12, 2013, doi: 10.3389/fnagi.2013.00058.

[5] C. M. Stonnington, C. Chu, S. Klöppel, C. R. Jack, J. Ashburner, and R. S. J. Frackowiak, "Predicting clinical scores from magnetic resonance scans in Alzheimer's disease," NeuroImage, vol. 51, no. 4, pp. 1405–1413, 2010, doi: 10.1016/j.neuroimage.2010.03.051.

[6] Z. Lao, D. Shen, Z. Xue, B. Karacali, S. M. Resnick, and C. Davatzikos, "Morphological classification of brains via high-dimensional shape transformations and machine learning methods," NeuroImage, vol. 21, no. 1, pp. 46–57, 2004, doi: 10.1016/j.neuroimage.2003.09.027.

[7] B. Magnin et al., "Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI," Neuroradiology, vol. 51, no. 2, pp. 73–83, 2009, doi: 10.1007/s00234-008-0463-x.

[8] L. Sørensen et al., "Early detection of Alzheimer's disease using MRI hippocampal texture," Human Brain Mapping, vol. 37, no. 3, pp. 1148–1161, 2016, doi: 10.1002/hbm.23091.

[9] M. H. Azmi, M. I. Saripan, A. J. Nordin, F. F. Ahmad Saad, S. A. Abdul Aziz, and W. A. Wan Adnan, "18F-FDG PET brain images as features for Alzheimer classification," Radiation Physics and Chemistry, vol. 137, pp. 135–143, 2017, doi: 10.1016/j.radphyschem.2016.08.028.

[10] G. E. Alexander, K. Chen, P. Pietrini, S. I. Rapoport, and E. M. Reiman, "Longitudinal PET evaluation of cerebral metabolic decline in dementia: A potential outcome measure in Alzheimer's disease treatment studies," American Journal of Psychiatry, vol. 159, no. 5, pp. 738–745, 2002, doi: 10.1176/appi.ajp.159.5.738.

[11] S. M. Landau et al., "Amyloid Deposition, Hypometabolism, and Longitudinal Cognitive Decline," Ann Neurol, vol. 72, no. 4, pp. 578–586, 2012, doi: 10.1002/ana.23650.Amyloid.

[12] F. D. G. Pet, "Early detection of Alzheimer's disease using PiB and FDG PET," pp. 117–122, 2015, doi: 10.1016/j.nbd.2014.05.001.Early.

[13] E. Trushina, T. Dutta, X. M. T. Persson, M. M. Mielke, and R. C. Petersen, "Identification of Altered Metabolic Pathways in Plasma and CSF in Mild Cognitive Impairment and Alzheimer's Disease Using Metabolomics," PLoS ONE, vol. 8, no. 5, 2013, doi: 10.1371/journal.pone.0063644.

[14] M. A. Colijn and G. T. Grossberg, "Amyloid and Tau Biomarkers in Subjective Cognitive Impairment," Journal of Alzheimer's Disease, vol. 47, no. 1, pp. 1–8, 2015, doi: 10.3233/JAD-150180.

[15] L. M. Shaw et al., "Cerebrospinal fluid biomarker signature in alzheimer's disease neuroimaging initiative subjects," Annals of Neurology, vol. 65, no. 4, pp. 403–413, 2009, doi: 10.1002/ana.21610.

[16] D. M. Michaelson, "APOE ε4: The most prevalent yet understudied risk factor for Alzheimer's disease," Alzheimer's and Dementia, vol. 10, no. 6, pp. 861–868, 2014, doi: 10.1016/j.jalz.2014.06.015.

[17] J. A. Rogers et al., "Combining patient-level and summary-level data for Alzheimer's disease modeling and simulation: A beta regression meta-analysis," Journal of Pharmacokinetics and Pharmacodynamics, vol. 39, no. 5, pp. 479–498, 2012, doi: 10.1007/s10928-012-9263-3.

[18] G. Chen et al., "A learning deficit related to age and b-amyloid plaques in a mouse model of Alzheimer's disease," Nature, vol. 408, no. 1998, pp. 975–979, 2000, doi: 10.1038/35050103.

[19] D. Mungas, B. R. Reed, W. G. Ellis, and W. J. Jagust, "The effects of age on rate of progression of Alzheimer disease and dementia with associated cerebrovascular disease.," Archives of neurology, vol. 58, no. 8, pp. 1243–7, 2001, doi: 10.1001/archneur.58.8.1243.

[20] L. a Farrer et al., "Effects of Age, Sex, and Ethnicity on the Association Between Apolipoprotein E Genotype and Alzheimer Disease," The Journal of the American Medical Association, vol. 278, no. 16, pp. 1349–1356, 1997, doi: 10.1001/jama.1997.03550160069041.

[21] A. E. H. Corder et al., "Gene Dose of Apolipoprotein E Type 4 Allele and the Risk of Alzheimer ' s Disease in Late Onset Families Published by : American Association for the Advancement of Science Stable URL : http:// www.jstor.org/stable/2882127," vol.261, no. 5123, pp. 921–923, 2008.

[22] X.-A. Bi, Q. Shu, Q. Sun, and Q. Xu, "Random support vector machine cluster analysis of resting-state fMRI in Alzheimer's disease," PLoS ONE, vol. 13, no. 3, pp. 1–17, 2018, doi: 10.1371/journal.pone.0194479.

[23] G. B. Frisoni et al., "The topography of grey matter involvement in early and late onset Alzheimer's disease," Brain, vol. 130, no. 3, pp. 720–730, 2007, doi: 10.1093/brain/awl377.

[24] S. Duchesne, A. Caroli, C. Geroldi, D. L. Collins, and G. B. Frisoni, "Relating one-year cognitive change in mild cognitive impairment to baseline MRI features," NeuroImage, vol. 47, no. 4, pp. 1363–1370, 2009, doi: 10.1016/ j.neuroimage.2009.04.023.

[25] Y. Li et al., "Discriminant analysis of longitudinal cortical thickness changes in Alzheimer's disease using dynamic and network features," Neurobiology of Aging, vol.33, no. 2, pp. 1–29, 2012, doi: 10.1016/ j.neurobiolaging.2010.11.008.

[26] K. Buerger et al., “CSF tau protein phosphorylated at threonine-231 correlates with cognitive decline in MCI subjects.,” Neurology, vol. 59, pp. 627–629, 2002.

[27] C. R. Jack et al., “Longitudinal tau PET in ageing and Alzheimer’s disease,” Brain, no. May, 2018, doi: 10.1093/brain/awy059.

[28] M. J. De Leon et al., “Longitudinal CSF and MRI biomarkers improve the diagnosis of mild cognitive impairment,” Neurobiology of Aging, vol. 27, no. 3, pp. 394–401, 2006, doi: 10.1016/j.neurobiolaging.2005.07.003.

[29] T. Tong, K. Gray, Q. Gao, L. Chen, and D. Rueckert, “Multi-modal classification of Alzheimer’s disease using nonlinear graph fusion,” Pattern Recognition, vol. 63, no. May 2016, pp. 171–181, 2017, doi: 10.1016/j.patcog.2016.10.009.

[30] K. Ritter, J. Schumacher, M. Weygandt, R. Buchert, C. Allefeld, and J. D. Haynes, “Multimodal prediction of conversion to Alzheimer’s disease based onincomplete biomarkers,” Alzheimer’s and Dementia: Diagnosis, Assessment and Disease Monitoring, vol. 1, no. 2, pp. 206–215, 2015, doi: 10.1016/j.dadm.2015.01.006.

[31] E. Westman, J. S. Muehlboeck, and A. Simmons, “Combining MRI and CSF measures for classification of Alzheimer’s disease and prediction of mild cognitive impairment conversion,” NeuroImage, vol. 62, no. 1, pp. 229–238, 2012, doi: 10.1016/j.neuroimage.2012.04.056.

[32] A. Daoqiang Zhanga, Yaping Wanga, b, Luping Zhoua, Hong Yuana, Dinggang Shena and A. D. N. Initiative1, “Multimodal Classification of Alzheimer’s Disease and Mild Cognitive Impairment,” NeuroImage, vol. 55, no. 3, pp. 856–867, 2011, doi: 10.1016/j.neuroimage.2011.01.008.Multimodal.

[33] M. F. Mendez, “Early-Onset Alzheimer Disease,” Neurologic Clinics, vol. 35, no. 2. pp. 263–281, 2017. doi: 10.1016/j.ncl.2017.01.005.

[34] A. L. Pierce, S. S. Bullain, and C. H. Kawas, “Late-Onset Alzheimer Disease,” Neurologic Clinics of NA, vol. 35, no. 2, pp. 283–293, 2017, doi: 10.1016/j.ncl.2017.01.006.

[35] B. A. Lawlor, T. M. Ryan, J. Schmeidler, R. C. Mohs, and K. L. Davis, “Clinical symptoms associated with age at onset in Alzheimer's disease.,” Am J Psychiatry, vol. 151, no. 11, pp. 1646–1649, 1994.

[36] M. S. Wolfe, “Prospects and Challenges for Alzheimer Therapeutics,” in Developing Therapeutics for Alzheimer’s Disease: Progress and Challenges, 2016, pp. 605–637. doi: 10.1016/B978-0-12-802173-6.00023-X.

[37] R. S. Doody, V. Pavlik, P. Massman, S. Rountree, E. Darby, and W. Chan, “Predicting progression of Alzheimer ’ s disease,” Alzheimer’s research & therapy, p. 77030, 2010, doi: 10.1186/alzrt38.

[38] W. M. Van Der Flier and P. Scheltens, “Alzheimer disease: Hippocampal volume loss and Alzheimer disease progression,” Nature Reviews Neurology, vol. 5, no. 7. pp. 361–362, 2009. doi: 10.1038/nrneurol.2009.94.

[39] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, and J. Tohka, “Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects,” NeuroImage, vol. 104, pp. 398–412, 2015, doi: 10.1016/j.neuroimage.2014.10.002.

[40] R. E. Curiel et al., "Semantic Intrusions and Failure to Recover From Semantic Interference in Mild Cognitive Impairment: Relationship to Amyloid and Cortical Thickness," Current Alzheimer Research, vol. 15, no. 9, pp. 848–855, 2018, doi: 10.2174/1567205015666180427122746.

[41] G. Lizarraga et al., "A neuroimaging web services interface as a cyber physical system for medical imaging and data management in brain research: Design study," Journal of Medical Internet Research, vol. 20, no. 4, pp. 1–17, 2018, doi: 10.2196/medinform.9063.

[42] C. Li, D. A. Loewenstein, R. Duara, M. Cabrerizo, W. Barker, and M. Adjouadi, "The relationship of brain amyloid load and APOE status to regional cortical thinning and cognition in the ADNI cohort," Journal of Alzheimer's Disease, vol. 59, no. 4, pp. 1269–1282, 2017, doi: 10.3233/JAD-170286.

[43] D. A. Loewenstein et al., "Recovery from Proactive Semantic Interference in Mild Cognitive Impairment and Normal Aging: Relationship to Atrophy in Brain Regions Vulnerable to Alzheimer's Disease," Journal of Alzheimer's Disease, vol. 56, no. 3, pp.1119–1126, 2017, doi: 10.3233/JAD-160881.

[44] S. Sargolzaei et al., "Estimating Intracranial Volume in Brain Research: An Evaluation of Methods," Neuroinformatics, vol. 13, no. 4, pp. 427–441, 2015, doi: 10.1007/s12021-015-9266-5.

[45] R. Duara et al., "Insights into cognitive aging and Alzheimer's disease using amyloid PET and structural MRI scans," Clinical and Translational Imaging, vol. 3, no. 1. pp. 65–74, 2015. doi: 10.1007/s40336-015-0110-6.

[46] S. Minhas, A. Khanum, F. Riaz, S. Khan, and A. Alvi, "Predicting Progression from Mild Cognitive Impairment to Alzheimer's Disease using Autoregressive Modelling of Longitudinal and Multimodal Biomarkers," IEEE Journal of Biomedical and Health Informatics, pp. 1–1, 2017, doi: 10.1109/JBHI.2017.2703918.

[47] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A Multi-Task Learning Formulation for Predicting Disease Progression Categories and Subject Descriptors".

[48] L. Huang, Y. Jin, Y. Gao, K.-H. Thung, and D. Shen, "Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest," Neurobiology of Aging, vol. 46, pp. 180–191, 2016, doi: 10.1016/j.neurobiolaging.2016.07.005.

[49] E. Moradi, I. Hallikainen, T. Hänninen, and J. Tohka, "Rey's Auditory Verbal Learning Test scores can be predicted from whole brain MRI in Alzheimer's disease," NeuroImage: Clinical, vol. 13, pp. 415–427, 2017, doi: 10.1016/j.nicl.2016.12.011.

[50] R. S. Eldholm et al., "Progression of Alzheimer's disease: A longitudinal study in Norwegian memory clinics," Journal of Alzheimer's Disease, 2018, doi: 10.3233/JAD-170436.

[51] H.-I. Suk, S.-W. Lee, and D. Shen, "Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis," Brain Structure and Function, vol. 221, no. 5, pp.2569–2587, 2016, doi: 10.1007/s00429-015-1059-y.

[52] P. Cao, X. Liu, J. Yang, D. Zhao, M. Huang, and O. Zaiane, "ℓ2,1−ℓ1regularized nonlinear multi-task representation learning based cognitive performance prediction of

Alzheimer's disease," Pattern Recognition, vol. 79, pp. 195–215, 2018, doi: 10.1016/j.patcog.2018.01.028.

[53] Alzheimer Association, "2016 Alzheimer's Disease Facts and Figures," Alzheimer's & Dementia 2016, vol. 12, no. 4, pp. 1–80, 2016, doi: 10.1016/j.jalz.2016.03.001.

[54] M. Aghili, S. Tabarestani, M. Adjouadi, and E. Adeli, "Predictive Modeling of Longitudinal Data for Alzheimer's Disease Diagnosis Using RNNs," in PRedictive Intelligence in MEdicine, 2018, pp. 112–119.

[55] Y. Fan, S. M. Resnick, X. Wu, and C. Davatzikos, "Structural and functional biomarkers of prodromal Alzheimer's disease: A high-dimensional pattern classification study," NeuroImage, 2008, doi: 10.1016/j.neuroimage.2008.02.043.

[56] M. Adjouadi, "Denoising of ultrasound images affected by combined speckle and Gaussian noise," IET Image Processing, Sep. 2018.

[57] R. Cuingnet et al., "Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database," NeuroImage, vol. 56, no. 2, pp. 766–781, 2011, doi: 10.1016/j.neuroimage.2010.06.013.

[58] L. Nie, L. Zhang, L. Meng, X. Song, X. Chang, and X. Li, "Modeling Disease Progression via Multisource Multitask Learners: A Case Study with Alzheimer's Disease," IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 7, pp. 1508–1519, 2017, doi: 10.1109/TNNLS.2016.2520964.

[59] C. Hinrichs, V. Singh, L. Mukherjee, G. Xu, M. K. Chung, and S. C. Johnson, "Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset," NeuroImage, vol. 48, no. 1, pp. 138–149, 2009, doi: 10.1016/j.neuroimage.2009.05.056.

[60] Y. Gao et al., "MCI identification by joint learning on multiple MRI data," 2015. doi: 10.1007/978-3-319-24571-3_10.

[61] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via multi-task learning," NeuroImage, vol. 78, pp. 233–248, 2013, doi: 10.1016/j.neuroimage.2013.03.073.

[62] L. Huang, Y. Jin, Y. Gao, K.-H. Thung, and D. Shen, "Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest," Neurobiology of Aging, vol. 46, pp. 180–191, 2016, doi: 10.1016/j.neurobiolaging.2016.07.005.

[63] G. Gavidia-Bovadilla, S. Kanaan-Izquierdo, M. Mataroa-Serrat, and A. Perera-Lluna, "Early prediction of Alzheimer's disease using null longitudinal model-based classifiers," PLoS ONE, 2017, doi: 10.1371/journal.pone.0168011.

[64] and A. M. Tabarestani S., Aghili M., Shojaie M., Freytes C., "Profile-Specific Regression Model for Progression Prediction of Alzheimer's Disease Using Longitudinal Data," 2018.

[65] S. Tsao et al., "Feature selective temporal prediction of Alzheimer's disease progression using hippocampus surface morphometry," Brain and Behavior, vol. 7, no. 7, pp. 1–11, 2017, doi: 10.1002/brb3.733.

[66] M. Nguyen, N. Sun, D. C. Alexander, J. Feng, and B. T. Thomas Yeo, “Modeling Alzheimer’s disease progression using deep recurrent neural networks,” 2018. doi: 10.1109/PRNI.2018.8423955.

[67] R. Cui, M. Liu, and G. Li, “Longitudinal analysis for Alzheimer’s disease diagnosis using RNN,” Proceedings - International Symposium on Biomedical Imaging, vol. 2018-April, no. Isbi, pp. 1398–1401, 2018, doi: 10.1109/ISBI.2018.8363833.

[68] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, “Using recurrent neural network models for early detection of heart failure onset,” Journal of the American Medical Informatics Association, vol. 24, no. 2, pp. 361–370, 2017, doi: 10.1093/jamia/ocw112.

[69] T. Wang, R. G. Qiu, and M. Yu, “Predictive Modeling of the Progression of Alzheimer ’ s Disease with Recurrent Neural Networks,” Scientific Reports, pp. 1–12, 2018, doi: 10.1038/s41598-018-27337-w.

[70] M. Aghili, S. Tabarestani, M. Adjouadi, and E. Adeli, “Predictive Modeling of Longitudinal Data for Alzheimer’s Disease Diagnosis Using RNNs,” in PRedictive Intelligence in MEdicine, 2018, pp. 112–119.

[71] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, 1997, doi: 10.1162/neco.1997.9.8.1735.

[72] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, “Learning Phrase Representations using {RNN} Encoder-Decoder for Statistical Machine Translation,” CoRR, vol. abs/1406.1078, 2014.

[73] D. Zhang, Daoqiuang; Shen, “Multi modal multi task learning for joint prediction of multiple regression and classification variables in Alzheimer’s disease,” Neuroimage, vol. 59, no. 2, pp. 895–907, 2013, doi: 10.1016/j.neuroimage.2011.09.069.Multi-Modal.

[74] H. Wang et al., “Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance,” in Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 557–562. doi: 10.1109/ICCV.2011.6126288.

[75] M. C. Tierney, J. P. Szalai, and W. G. Snow, “Prediction of probable Alzheimer ’ s disease in memory-impaired patients :,” Neurology, vol. 46, pp. 661–665, 1996, doi: 10.1212/WNL.46.3.661.

[76] D. Zhang and D. Shen, “Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers,” PLoS ONE, vol. 7, no. 3, 2012, doi: 10.1371/journal.pone.0033182.

[77] W. Izquierdo et al., “PREDICTING COGNITIVE TEST SCORES IN ALZHEIMER’S PATIENTS USING MULTIMODAL LONGITUDINAL DATA,” Alzheimer’s & Dementia: The Journal of the Alzheimer’s Association, vol. 13, no. 7, pp. P796–P797, Jul. 2017, doi: 10.1016/j.jalz.2017.06.1078.

[78] R. Caruana, “Multitask Learning,” Machine Learning, vol. 28, no. 1, pp. 41–75, 1997, doi: 10.1023/A:1007379606734.

[79] D. Dong, H. Wu, W. He, D. Yu, and H. Wang, “Multi-Task Learning for Multiple Language Translation,” Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural

Language Processing (Volume 1: Long Papers), pp. 1723–1732, 2015, doi: 10.3115/v1/P15-1166.

[80] K. Greenlaw, E. Szefer, J. Graham, M. Lesperance, and F. S. Nathoo, "A Bayesian group sparse multi-task regression model for imaging genetics," Bioinformatics (Oxford, England), vol. 33, no. 16, pp. 2513–2522, 2017, doi: 10.1093/bioinformatics/btx215.

[81] T. Evgeniou, C. Micchelli, and M. Pontil, "Learning multiple tasks with kernel methods," Jmlr, vol. 6, pp. 615–637, 2005.

[82] Y. Zhang and D.-Y. Yeung, "A Convex Formulation for Learning Task Relationships in Multi-Task Learning," 2012.

[83] C. Widmer, C. M. Org, and G. Rätsch, "Multitask Learning in Computational Biology," Conference Proceedings, vol. 27, pp. 207–216, 2012.

[84] J. Bi et al., "An Improved Multi-task Learning Approach with Applications in Medical Diagnosis," Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases-Part I, pp. 117–132, 2008.

[85] Y. Xue, X. Liao, L. Carin, B. Krishnapuram, and B. K. Com, "Multi-Task Learning for Classification with Dirichlet Process Priors," Journal of Machine Learning Research, vol. 8, pp. 35–63, 2007.

[86] Z. Yang, J. Rong, and H. H. Steven C., "Exclusive Lasso for Multi-task Feature Selection," Aistats, vol. 9, pp. 988–995, 2010.

[87] X. Zhu, H. Il Suk, L. Wang, S. W. Lee, and D. Shen, "A novel relational regularization feature selection method for joint regression and classification in AD diagnosis," Medical Image Analysis, vol. 38, pp. 205–214, 2017, doi: 10.1016/j.media.2015.10.008.

[88] Y. Zhang and D. Y. Yeung, "Multi-Task Learning in Heterogeneous Feature Spaces," Aaai, vol. 1, p. 1, 2011.

[89] Y. Li, X. Tian, T. Liu, and D. Tao, "On Better Exploring and Exploiting Task Relationships in Multitask Learning: Joint Model and Feature Learning," IEEE Transactions on Neural Networks and Learning Systems, 2017. doi: 10.1109/TNNLS.2017.2690683.

[90] K. Weinberger, A. Dasgupta, J. Attenberg, J. Langford, and A. Smola, "Feature Hashing for Large Scale Multitask Learning," no. Icml, 2009, doi: 10.1145/1553374.1553516.

[91] A. Kumar and H. Daume, "Learning Task Grouping and Overlap in Multi-task Learning," 2012.

[92] B. Bakker and T. Heskes, "Task Clustering and Gating for BayesianMultitask Learning," Journal of Machine Learning Research, vol. 1, no. 1, pp. 83–99, 2003, doi: 10.1162/153244304322765658.

[93] D. Zhang, Daoqiuang; Shen, "Multi modal multi task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease," Neuroimage, vol. 59, no. 2, pp. 895–907, 2013, doi: 10.1016/j.neuroimage.2011.09.069.Multi-Modal.

[94] S. Emrani, A. McGuirk, and W. Xiao, "Prognosis and Diagnosis of Parkinson's Disease Using Multi-Task Learning," Proceedings of the 23rd ACM SIGKDD International

Conference on Knowledge Discovery and Data Mining - KDD '17, pp. 1457–1466, 2017, doi: 10.1145/3097983.3098065.

[95] L. Nie, L. Zhang, L. Meng, X. Song, X. Chang, and X. Li, "Modeling Disease Progression via Multisource Multitask Learners: A Case Study with Alzheimer's Disease," IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 7, pp. 1508–1519, 2017, doi: 10.1109/TNNLS.2016.2520964.

[96] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12, p. 1095, 2012, doi: 10.1145/2339530.2339702.

[97] H.-I. Suk, S.-W. Lee, and D. Shen, "Deep ensemble learning of sparse regression models for brain disease diagnosis," Medical Image Analysis, vol. 37, pp. 101–113, 2017, doi: 10.1016/j.media.2017.01.008.

[98] S. Emrani, N. Carolina, A. Mcguirk, N. Carolina, and N. Carolina, "Prognosis and Diagnosis of Parkinson ' s Disease Using Multi-Task Learning," pp. 1457–1466, 2017, doi: 10.1145/3097983.3098065.

[99] B. Jie, D. Zhang, B. Cheng, and D. Shen, "Manifold regularized multitask feature learning for multimodality disease classification," Human Brain Mapping, vol. 36, no. 2, pp. 489–507, 2015, doi: 10.1002/hbm.22642.

[100] P. Cao, X. Shan, D. Zhao, M. Huang, and O. Zaiane, "Sparse shared structure based multi-task learning for MRI based cognitive performance prediction of Alzheimer's disease," Pattern Recognition, vol. 72, pp. 219–235, 2017, doi: 10.1016/j.patcog.2017.07.018.

[101] H. Wang et al., "High-Order Multi-Task Feature Learning to Identify Longitudinal Phenotypic Markers for Alzheimer's Disease Progression Prediction," p. 9.

[102] P. Cao, X. Liu, J. Yang, D. Zhao, M. Huang, and O. Zaiane, "ℓ2,1−ℓ1 multi-task representation learning based cognitive performance prediction of Alzheimer's disease," Pattern Recognition, vol. 79, pp. 195–215, 2018, doi: 10.1016/j.patcog.2018.01.028.

[103] B. Jie, M. Liu, J. Liu, D. Zhang, and D. Shen, "Temporally Constrained Group Sparse Learning for Longitudinal Data Analysis in Alzheimer's Disease," IEEE Transactions on Biomedical Engineering, vol. 64, no. 1, pp. 238–249, Jan. 2017, doi: 10.1109/TBME.2016.2553663.

[104] B. Cheng et al., "Multimodal manifold-regularized transfer learning for MCI conversion prediction," Brain Imaging and Behavior, vol. 9, no. 4, pp. 1805–1817, 2015, doi: 10.1007/s11682-015-9356-x.

[105] Y. Zhu, X. Zhu, M. Kim, D. Shen, and G. Wu, "Early Diagnosis of Alzheimer's Disease by Joint Feature Selection and Classification on Temporally Structured Support Vector Machine," in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, vol. 9900, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds. Cham: Springer International Publishing, 2016, pp. 264–272. doi: 10.1007/978-3-319-46720-7_31.

[106] J. Zhou, J. Chen, and J. Ye, "User's Manual MALSAR: Multi-tAsk Learning via StructurAl Regularization," Arizona State University, 2012.

[107] X. Zhu, H.-I. Suk, S.-W. Lee, and D. Shen, “Subspace Regularized Sparse Multitask Learning for Multiclass Neurodegenerative Disease Identification,” IEEE Transactions on Biomedical Engineering, vol. 63, no. 3, pp. 607–618, Mar. 2016, doi: 10.1109/TBME.2015.2466616.

[108] X. Zhu, H. Li, and Y. Fan, “Parameter-Free Centralized Multi-Task Learning for Characterizing Developmental Sex Differences in Resting State Functional Connectivity,” p. 8.

[109] J. Friedman, “Greedy Function Approximation : A Gradient Boosting Machine Author (s): Jerome H . Friedman Source : The Annals of Statistics , Vol . 29 , No . 5 (Oct ., 2001), pp . 1189-1232 Published by : Institute of Mathematical Statistics Stable URL : http://www,” The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/009053606000000795.

[110] J. O. Ogutu, H. P. Piepho, and T. Schulz-Streeck, “A comparison of random forests, boosting and support vector machines for genomic selection,” BMC Proceedings, vol. 5, no. SUPPL. 3, p. S11, 2011, doi: 10.1186/1753-6561-5-S3-S11.

[111] A. Anoop, P. K. Singh, R. S. Jacob, and S. K. Maji, “CSF Biomarkers for Alzheimer’s Disease Diagnosis,” International Journal of Alzheimer’s Disease, vol. 2010, pp. 1–12, 2010, doi: 10.4061/2010/606802.

[112] D. P. Hanger, B. H. Anderton, and W. Noble, “Tau phosphorylation: the therapeutic challenge for neurodegenerative disease,” Trends in Molecular Medicine, vol. 15, no. 3, pp. 112–119, Mar. 2009, doi: 10.1016/j.molmed.2009.01.003.

[113] W. Noble, D. P. Hanger, C. C. J. Miller, and S. Lovestone, “The Importance of Tau Phosphorylation for Neurodegenerative Diseases,” Frontiers in Neurology, vol. 4, 2013, doi: 10.3389/fneur.2013.00083.

[114] A. Bussy et al., “Effect of apolipoprotein E4 on clinical, neuroimaging, and biomarker measures in noncarrier participants in the Dominantly Inherited Alzheimer Network,” Neurobiology of Aging, vol. 75, pp. 42–50, Mar. 2019, doi: 10.1016/j.neurobiolaging.2018.10.011.

[115] Y. Stern, “Cognitive reserve in ageing and Alzheimer’s disease,” The Lancet Neurology, vol. 11, no. 11, pp. 1006–1012, Nov. 2012, doi: 10.1016/S1474-4422(12)70191-6.

[116] R. L. Buckner, “Memory and executive function in aging and AD: Multiple factors that cause decline and reserve factors that compensate,” p. 14.

[117] M. Sugiyama and M. Krauledat, “Covariate Shift Adaptation by Importance Weighted Cross Validation,” p. 21.

[118] “2018 Alzheimer’s disease facts and figures,” Alzheimer’s & Dementia, vol. 14, no. 3, pp. 367–429, Mar. 2018, doi: 10.1016/j.jalz.2018.02.001.

[119] E. Pellegrini et al., “Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review,” Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring, vol. 44, no. August, pp. 1–17, 2018, doi: 10.1016/J.DADM.2018.07.004.

[120] R. J. Perrin, A. M. Fagan, and D. M. Holtzman, “Multimodal techniques for diagnosis and prognosis of Alzheimer’s disease,” Nature, vol. 461, no. 7266, pp. 916–922, Oct. 2009, doi: 10.1038/nature08538.

[121] T. Tong, K. Gray, Q. Gao, L. Chen, and D. Rueckert, “Multi-modal classification of Alzheimer’s disease using nonlinear graph fusion,” Pattern Recognition, vol. 63, no. October 2016, pp. 171–181, 2017, doi: 10.1016/j.patcog.2016.10.009.

[122] X. Wang, X. Zhen, Q. Li, D. Shen, and H. Huang, “Cognitive Assessment Prediction in Alzheimer ’ s Disease by Multi-Layer Multi-Target Regression,” pp. 285–294, 2018.

[123] H. Wang, F. Nie, H. Huang, J. Yan, and S. Kim, “High-Order Multi-Task Feature Learning to Identify Longitudinal Phenotypic Markers for Alzheimer’s Disease Progression Prediction.,” Nips, pp. 1–9, 2012.

[124] R. Wei, C. Li, N. Fogelson, and L. Li, “Prediction of conversion from mild cognitive impairment to Alzheimer’s Disease using MRI and structural network features,” Frontiers in aging neuroscience, vol. 8, p. 76, 2016.

[125] W. Huang, X. Li, X. Li, G. Kang, Y. Han, and N. Shu, “Combined Support Vector Machine Classifier and Brain Structural Network Features for the Individual Classification of Amnestic Mild Cognitive Impairment and Subjective Cognitive Decline Patients,” Frontiers in aging neuroscience, vol. 13, 2021.

[126] W. Lin et al., “Predicting Alzheimer’s disease conversion from mild cognitive impairment using an extreme learning machine-based grading method with multimodal data,” Frontiers in aging neuroscience, vol. 12, p. 77, 2020.

[127] A. Tolonen et al., “Data-driven differential diagnosis of dementia using multiclass disease state index classifier,” Frontiers in aging neuroscience, vol. 10, p. 111, 2018.

[128] Y. Gupta, J.-I. Kim, B. C. Kim, and G.-R. Kwon, “Classification and graphical analysis of Alzheimer’s disease and its prodromal stage using multimodal features from structural, diffusion, and functional neuroimaging data and the APOE genotype,” Frontiers in aging neuroscience, vol. 12, p. 238, 2020.

[129] S. Sarraf and G. Tofighi, “Classification of Alzheimer’s Disease using fMRI Data and Deep Learning Convolutional Neural Networks,” pp. 8–12, 2016, doi: arXiv:1607.06583.

[130] C. K. Fisher, A. M. Smith, and J. R. Walsh, “Using deep learning for comprehensive , personalized forecasting of Alzheimer ’ s Disease progression,” pp. 1–36, 2018.

[131] J. Zhang, Q. Li, R. J. Caselli, J. Ye, and Y. Wang, “Multi-task Dictionary Learning based Convolutional Neural Network for Computer aided Diagnosis with Longitudinal Images,” 2017.

[132] D. Lu, K. Popuri, G. W. Ding, R. Balachandar, and M. F. Beg, “Multiscale deep neural network based analysis of FDG-PET images for the early diagnosis of Alzheimer’s disease,” Medical Image Analysis, vol. 46, pp. 26–34, 2018, doi: 10.1016/j.media.2018.02.002.

[133] H. Choi and K. H. Jin, “Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging,” Behavioural Brain Research, vol. 344, no. February, pp. 103–109, 2018, doi: 10.1016/j.bbr.2018.02.017.

[134] N. Amoroso et al., “Deep learning reveals Alzheimer’s disease onset in MCI subjects: Results from an international challenge,” Journal of Neuroscience Methods, vol. 302, pp. 3–9, 2018, doi: 10.1016/j.jneumeth.2017.12.011.

[135] T. Jo, K. Nho, and A. J. Saykin, “Deep learning in Alzheimer’s disease: diagnostic classification and prognostic prediction using neuroimaging data,” Frontiers in aging neuroscience, vol. 11, p. 220, 2019.

[136] the Alzheimer’s Disease Neuroimaging Initiative, M. Liu, D. Cheng, K. Wang, and Y. Wang, “Multi-Modality Cascaded Convolutional Neural Networks for Alzheimer’s Disease Diagnosis,” Neuroinformatics, vol. 16, no. 3–4, pp. 295–308, Oct. 2018, doi: 10.1007/s12021-018-9370-4.

[137] H.-I. Suk and D. Shen, “Deep Learning-Based Feature Representation for AD/MCI Classification,” in Advanced Information Systems Engineering, vol. 7908, C. Salinesi, M. C. Norrie, and Ó. Pastor, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 583–590. doi: 10.1007/978-3-642-40763-5_72.

[138] S. Liu et al., “Multimodal Neuroimaging Feature Learning for Multiclass Diagnosis of Alzheimer’s Disease,” IEEE Transactions on Biomedical Engineering, vol. 62, no. 4, pp. 1132–1140, Apr. 2015, doi: 10.1109/TBME.2014.2372011.

[139] J. Liu, S. Shang, K. Zheng, and J. R. Wen, “Multi-view ensemble learning for dementia diagnosis from neuroimaging: An artificial neural network approach,” Neurocomputing, vol. 195, pp. 112–116, 2016, doi: 10.1016/j.neucom.2015.09.119.

[140] Dept. of Information and comm., Engineering Chosun University, D. Jha, and G.-R. Kwon, “Alzheimer’s Disease Detection Using Sparse Autoencoder, Scale Conjugate Gradient and Softmax Output Layer with Fine Tuning,” International Journal of Machine Learning and Computing, vol. 7, no. 1, pp. 13–17, 2017, doi: 10.18178/ijmlc.2017.7.1.612.

[141] T. Wang, R. G. Qiu, and M. Yu, “Predictive Modeling of the Progression of Alzheimer ’ s Disease with Recurrent Neural Networks,” Scientific Reports, pp. 1–12, 2018, doi: 10.1038/s41598-018-27337-w.

[142] M. Liu, D. Cheng, W. Yan, and Alzheimer’s Disease Neuroimaging Initiative, “Classification of Alzheimer’s Disease by Combination of Convolutional and Recurrent Neural Networks Using FDG-PET Images,” Frontiers in Neuroinformatics, vol. 12, Jun. 2018, doi: 10.3389/fninf.2018.00035.

[143] L. Kang, J. Jiang, J. Huang, and T. Zhang, “Identifying early mild cognitive impairment by multi-modality mri-based deep learning,” Frontiers in aging neuroscience, vol. 12, p. 206, 2020.

[144] M. Liu, J. Zhang, E. Adeli, and D. Shen, “Joint Classification and Regression via Deep Multi-Task Multi-Channel Learning for Alzheimer’s Disease Diagnosis,” IEEE Transactions on Biomedical Engineering, vol. PP, no. c, pp. 1–1, 2018, doi: 10.1109/TBME.2018.2869989.

[145] X. Zhu, H.-I. Suk, S.-W. Lee, and D. Shen, “Canonical feature selection for joint regression and multi-class identification in Alzheimer’s disease diagnosis,” Brain Imaging and Behavior, vol. 10, no. 3, pp. 818–828, Sep. 2016, doi: 10.1007/s11682-015-9430-4.

[146] J. Shi, X. Zheng, Y. Li, Q. Zhang, and S. Ying, "Multimodal Neuroimaging Feature Learning With Multimodal Stacked Deep Polynomial Networks for Diagnosis of Alzheimer's Disease," IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 1, pp. 173–183, Jan. 2018, doi: 10.1109/JBHI.2017.2655720.

[147] X. Zhen, M. Yu, X. He, and S. Li, "Multi-Target Regression via Robust Low-Rank Learning," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 2, pp. 497–504, 2018, doi: 10.1109/TPAMI.2017.2688363.

[148] D. Zhang, Daoqiuang; Shen, "Multi modal multi task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease," Neuroimage, vol. 59, no. 2, pp. 895–907, 2013, doi: 10.1016/j.neuroimage.2011.09.069.Multi-Modal.

[149] X. Zhu, H. Il Suk, and D. Shen, "A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis," NeuroImage, vol. 100, pp. 91–105, 2014, doi: 10.1016/j.neuroimage.2014.05.078.

[150] S.-H. Wang, P. Phillips, Y. Sui, B. Liu, M. Yang, and H. Cheng, "Classification of Alzheimer's Disease Based on Eight-Layer Convolutional Neural Network with Leaky Rectified Linear Unit and Max Pooling," Journal of Medical Systems, vol. 42, no. 5, p. 85, 2018, doi: 10.1007/s10916-018-0932-7.

[151] J. Fei, N. Zhao, Y. Shi, Y. Feng, and Z. Wang, "Compressor performance prediction using a novel feed-forward neural network based on Gaussian kernel function," Advances in Mechanical Engineering, vol. 8, no. 1, p. 1687814016628396, 2016.

[152] A. Srisuphab and J. L. Mitrpanont, "Gaussian kernel approximation algorithm for feedforward neural network design," Applied mathematics and computation, vol. 215, no. 7, pp. 2686–2693, 2009.

[153] O. Debals and L. De Lathauwer, "Stochastic and deterministic tensorization for blind signal separation," in International Conference on Latent Variable Analysis and Signal Separation, 2015, pp. 3–13.

[154] A. Novikov, D. Podoprikhin, A. Osokin, and D. Vetrov, "Tensorizing neural networks," arXiv preprint arXiv:1509.06569, 2015.

[155] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv:1412.6980 [cs], Dec. 2014, Accessed: Jan. 14, 2019. [Online]. Available: http://arxiv.org/abs/1412.6980

[156] F. Chollet and others, Keras. 2015.

[157] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), 2016, pp. 265–283. Accessed: Jan. 14, 2019. [Online]. Available: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf

[158] "Tensorizing Generative Adversarial Nets," GroundAI. https://www.groundai.com/project/tensorizing-generative-adversarial-nets/ (accessed Jan. 14, 2019).

[159] B. Lei, F. Jiang, S. Chen, D. Ni, and T. Wang, "Longitudinal analysis for disease progression via simultaneous multi-relational temporal-fused learning," Frontiers in Aging Neuroscience, vol. 9, no. MAR, pp. 1–17, 2017, doi: 10.3389/fnagi.2017.00006.

[160] X. Zhu, H. Il Suk, S. W. Lee, and D. Shen, “Subspace Regularized Sparse Multitask Learning for Multiclass Neurodegenerative Disease Identification,” IEEE Transactions on Biomedical Engineering, vol. 63, no. 3, pp. 607–618, 2016, doi: 10.1109/TBME.2015.2466616.

[161] S. Spasov, L. Passamonti, A. Duggento, P. Liò, and N. Toschi, “A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer’s disease,” NeuroImage, vol. 189, pp. 276–287, Apr. 2019, doi: 10.1016/j.neuroimage.2019.01.031.

[162] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, and J. Tohka, “Machine learning framework for early MRI-based Alzheimer’s conversion prediction in MCI subjects,” NeuroImage, vol. 104, pp. 398–412, Jan. 2015, doi: 10.1016/j.neuroimage.2014.10.002.

[163] K. Liu, K. Chen, L. Yao, and X. Guo, “Prediction of Mild Cognitive Impairment Conversion Using a Combination of Independent Component Analysis and the Cox Model,” Frontiers in Human Neuroscience, vol. 11, Feb. 2017, doi: 10.3389/fnhum.2017.00033.

[164] S. H. Hojjati, A. Ebrahimzadeh, A. Khazaee, and A. Babajani-Feremi, “Predicting conversion from MCI to AD using resting-state fMRI, graph theoretical approach and SVM,” Journal of Neuroscience Methods, vol. 282, pp. 69–80, Apr. 2017, doi: 10.1016/j.jneumeth.2017.03.006.

[165] S. Natarajan et al., “Relational learning helps in three-way classification of Alzheimer patients from structural magnetic resonance images of the brain,” International Journal of Machine Learning and Cybernetics, vol. 5, no. 5, pp. 659–669, Oct. 2014, doi: 10.1007/s13042-013-0161-9.

[166] H. Il Suk, S. W. Lee, and D. Shen, “Deep sparse multi-task learning for feature selection in Alzheimer’s disease diagnosis,” Brain Structure and Function, vol. 221, no. 5, pp. 2569–2587, 2016, doi: 10.1007/s00429-015-1059-y.

[167] M. Liu, D. Cheng, K. Wang, and Y. Wang, “Multi-Modality Cascaded Convolutional Neural Networks for Alzheimer’s Disease Diagnosis,” Neuroinformatics, pp. 1–14, 2018, doi: 10.1007/s12021-018-9370-4.

[168] G. Lizarraga et al., “A neuroimaging web services interface as a cyber physical system for medical imaging and data management in brain research: Design study,” JMIR medical informatics, vol. 6, no. 2, p. e9063, 2018.

[169] Q. Li, “Overview of data visualization,” in Embodying Data, Springer, 2020, pp. 17–47.

[170] K. Seo et al., “Visualizing Alzheimer’s disease progression in low dimensional manifolds,” Heliyon, vol. 5, no. 8, p. e02216, 2019.

[171] A. E. Blanken et al., “Distilling heterogeneity of mild cognitive impairment in the National Alzheimer Coordinating Center Database using latent profile analysis,” JAMA network open, vol. 3, no. 3, pp. e200413–e200413, 2020.

[172] X. Liu, D. Tosun, M. W. Weiner, N. Schuff, A. D. N. Initiative, and others, “Locally linear embedding (LLE) for MRI based Alzheimer’s disease classification,” Neuroimage, vol. 83, pp. 148–157, 2013.

[173] S. Gerber et al., “Manifold modeling for brain population analysis,” Medical image analysis, vol. 14, no. 5, pp. 643–653, 2010.

[174] D. Berron, D. van Westen, R. Ossenkoppele, O. Strandberg, and O. Hansson, “Medial temporal lobe connectivity and its associations with cognition in early Alzheimer’s disease,” Brain, vol. 143, no. 4, pp. 1233–1248, 2020.

[175] T. Montez et al., “Altered temporal correlations in parietal alpha and prefrontal theta oscillations in early-stage Alzheimer disease,” Proceedings of the National Academy of Sciences, vol. 106, no. 5, pp. 1614–1619, 2009.

[176] R. F. Buckley et al., “Functional network integrity presages cognitive decline in preclinical Alzheimer disease,” Neurology, vol. 89, no. 1, pp. 29–37, 2017.

[177] J. K. Wisch et al., “Resting state functional connectivity signature differentiates cognitively normal from individuals who convert to symptomatic Alzheimer’s disease,” Journal of Alzheimer’s Disease, vol. 74, no. 4, pp. 1085–1095, 2020.

[178] D. Toddenroth, T. Ganslandt, I. Castellanos, H.-U. Prokosch, and T. Bürkle, “Employing heat maps to mine associations in structured routine care data,” Artificial intelligence in medicine, vol. 60, no. 2, pp. 79–88, 2014.

[179] P. Klemm et al., “3D regression heat map analysis of population study data,” IEEE transactions on visualization and computer graphics, vol. 22, no. 1, pp. 81–90, 2015.

[180] S. Qiu et al., “Development and validation of an interpretable deep learning framework for Alzheimer’s disease classification,” Brain, vol. 143, no. 6, pp. 1920–1933, 2020.

[181] I. Jelistratova, S. J. Teipel, and M. J. Grothe, “Longitudinal validity of PET-based staging of regional amyloid deposition,” Human brain mapping, vol. 41, no. 15, pp. 4219–4231, 2020.

[182] R. Ossenkoppele et al., “Prevalence of amyloid PET positivity in dementia syndromes: a meta-analysis,” Jama, vol. 313, no. 19, pp. 1939–1950, 2015.

[183] S. M. Landau et al., “Amyloid-β imaging with Pittsburgh compound B and florbetapir: comparing radiotracers and quantification methods,” Journal of Nuclear Medicine, vol. 54, no. 1, pp. 70–77, 2013.

[184] M. J. Grothe et al., “In vivo staging of regional amyloid deposition,” Neurology, vol. 89, no. 20, pp. 2031–2038, 2017.

[185] P. Parbo et al., “Brain inflammation accompanies amyloid in the majority of mild cognitive impairment cases due to Alzheimer’s disease,” Brain, vol. 140, no. 7, pp. 2002–2011, 2017.

[186] M. Lynch, “New Alzheimer’s association report reveals sharp increases in Alzheimer’s prevalence, deaths, cost of care,” Alzheimer’s & Dementia, 2018.

[187] L. Einav, A. Finkelstein, S. Mullainathan, and Z. Obermeyer, “Predictive modeling of US health care spending in late life,” Science, vol. 360, no. 6396, pp. 1462–1465, 2018.

[188] L. Mucke, “Alzheimer’s disease,” Nature, vol. 461, no. 7266, pp. 895–897, 2009.

[189] P. N. Young et al., “Imaging biomarkers in neurodegeneration: current and future practices,” Alzheimer’s research & therapy, vol. 12, no. 1, pp. 1–17, 2020.

[190] C. R. Jack Jr et al., “Tracking pathophysiological processes in Alzheimer’s disease: an updated hypothetical model of dynamic biomarkers,” The Lancet Neurology, vol. 12, no. 2, pp. 207–216, 2013.

[191] D. A. Loewenstein et al., “Utilizing semantic intrusions to identify amyloid positivity in mild cognitive impairment,” Neurology, vol. 91, no. 10, pp. e976–e984, 2018.

[192] D. J. Selkoe, “Early network dysfunction in Alzheimer’s disease,” Science, vol. 365, no. 6453, pp. 540–541, 2019.

[193] F. Jessen et al., “The characterisation of subjective cognitive decline,” The Lancet Neurology, vol. 19, no. 3, pp. 271–278, 2020.

[194] A. C. van Loenhoud et al., “Cognitive reserve and clinical progression in Alzheimer disease: A paradoxical relationship,” Neurology, vol. 93, no. 4, pp. e334–e346, 2019.

[195] R. Nortley et al., “Amyloid β oligomers constrict human capillaries in Alzheimer’s disease via signaling to pericytes,” Science, vol. 365, no. 6450, 2019.

[196] M. Schöll and A. Maass, “Does early cognitive decline require the presence of both tau and amyloid-β?,” Brain, vol. 143, no. 1, pp. 10–13, 2020.

[197] S. N. Lockhart et al., “Amyloid and tau PET demonstrate region-specific associations in normal older people,” Neuroimage, vol. 150, pp. 191–199, 2017.

[198] A. Montagne et al., “APOE4 leads to blood–brain barrier dysfunction predicting cognitive decline,” Nature, vol. 581, no. 7806, pp. 71–76, 2020.

[199] T. Li et al., “APOE ε4 and cognitive reserve effects on the functional network in the Alzheimer’s disease spectrum,” Brain imaging and behavior, vol. 15, no. 2, pp. 758–771, 2021.

[200] J. Therriault et al., “Association of apolipoprotein E ε4 with medial temporal tau independent of amyloid-β,” JAMA neurology, vol. 77, no. 4, pp. 470–479, 2020.

[201] B. Olsson et al., “CSF and blood biomarkers for the diagnosis of Alzheimer’s disease: a systematic review and meta-analysis,” The Lancet Neurology, vol. 15, no. 7, pp. 673–684, 2016.

[202] P. Lewczuk et al., “Cerebrospinal Fluid Aβ 42/40 Corresponds Better than Aβ 42 to Amyloid PET in Alzheimer’s Disease,” Journal of Alzheimer’s Disease, vol. 55, no. 2, pp. 813–822, 2017.

[203] M. Sabbagh et al., “Effects of a combined transcranial magnetic stimulation (TMS) and cognitive training intervention in patients with Alzheimer’s disease,” Alzheimer’s & Dementia, 2019.

[204] S. M. Romanella et al., “Sleep, Noninvasive Brain Stimulation, and the Aging Brain: Challenges and Opportunities,” Ageing research reviews, vol. 61, p. 101067, 2020.

[205] N. Geifman, R. E. Kennedy, L. S. Schneider, I. Buchan, and R. D. Brinton, “Data-driven identification of endophenotypes of Alzheimer’s disease progression: implications for

clinical trials and therapeutic interventions," Alzheimer's research & therapy, vol. 10, no. 1, pp. 1–7, 2018.

[206] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT press, 2016.

[207] C. K. Fisher, A. M. Smith, and J. R. Walsh, "Machine learning for comprehensive forecasting of Alzheimer's Disease progression," Scientific reports, vol. 9, no. 1, pp. 1–14, 2019.

[208] E. Pellegrini et al., "Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review," Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, vol. 10, pp. 519–535, 2018.

[209] A. Ezzati et al., "Optimizing machine learning methods to improve predictive models of Alzheimer's disease," Journal of Alzheimer's Disease, vol. 71, no. 3, pp. 1027–1036, 2019.

[210] S. Tabarestani et al., "A distributed multitask multimodal approach for the prediction of Alzheimer's disease in a longitudinal study," NeuroImage, vol. 206, p. 116317, 2020.

[211] M. Huang, W. Yang, Q. Feng, and W. Chen, "Longitudinal measurement and hierarchical classification framework for the prediction of Alzheimer's disease," Scientific reports, vol. 7, no. 1, pp. 1–13, 2017.

[212] L. M. Aksman et al., "Modeling longitudinal imaging biomarkers with parametric Bayesian multi-task learning," Human brain mapping, vol. 40, no. 13, pp. 3982–4000, 2019.

[213] L. Huang et al., "Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest," Neurobiology of aging, vol. 46, pp. 180–191, 2016.

[214] K. Chiotis et al., "Longitudinal changes of tau PET imaging in relation to hypometabolism in prodromal and Alzheimer's disease dementia," Molecular psychiatry, vol. 23, no. 7, pp. 1666–1673, 2018.

[215] M. Bilgel, J. L. Prince, D. F. Wong, S. M. Resnick, and B. M. Jedynak, "A multivariate nonlinear mixed effects model for longitudinal image analysis: Application to amyloid imaging," Neuroimage, vol. 134, pp. 658–670, 2016.

[216] L. Lu, H. Wang, S. Elbeleidy, and F. Nie, "Predicting cognitive declines using longitudinally enriched representations for imaging biomarkers," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4827–4836.

[217] W. Jung, E. Jun, H.-I. Suk, A. D. N. Initiative, and others, "Deep recurrent model for individualized prediction of Alzheimer's disease progression," NeuroImage, vol. 237, p. 118143, 2021.

VITA

SOLALE TABARESATNI

2003 - 2007	B. Sc., Electrical Engineering, Azad University, Iran
2009 - 2012	M. Sc., Electrical Engineering, Shahid Beheshti University, Iran
2016 - 2020	M. Sc., Electrical and Computer Engineering, Florida International University
June 2020 – Aug 2020	Data Scientist Intern, 3M
May 2021 – Aug 2021	Machine learning Engineer, Tesla

SELECTED PUBLICATIONS

Tabarestani, S., Aghili, M., Eslami, M., Cabrerizo, M., Barreto, A., Rishe, N., Curiel, R.E., Loewenstein, D., Duara, R. and Adjouadi, M., 2020. A distributed multitask multimodal approach for the prediction of Alzheimer's disease in a longitudinal study. NeuroImage, 206, p.116317.

Eslami, M., Tabarestani, S., Albarqouni, S., Adeli, E., Navab, N. and Adjouadi, M., 2020. Image-to-images translation for multi-task organ segmentation and bone suppression in chest x-ray radiography. IEEE transactions on medical imaging, 39(7), pp.2553-2565.

Mafi, M., Tabarestani, S., Cabrerizo, M., Barreto, A. and Adjouadi, M., 2018. Denoising of ultrasound images affected by combined speckle and Gaussian noise. IET Image Processing, 12(12), pp.2346-2351.

Shojaie, M., Tabarestani, S., Cabrerizo, M., DeKosky, M. T., Vaillancourt, D. E., Loewenstein, D., Duara, R., and Adjouadi, M, 2021 "PET Imaging of Tau Pathology and Amyloid-β, and MRI for Alzheimer's Disease Feature Fusion and Multimodal Classification", accepted in Journal of Alzheimer's Disease.

Aghili, M., Tabarestani, S. and Adjouadi, M., 2021, "Addressing the Missing Data Challenge in Multi-Modal Datasets for the Diagnosis of Alzheimer's Disease", accepted in Journal of Neuroscience Methods.

Aghili, M., Tabarestani, S., Freytes, C., Shojaie, M., Cabrerizo, M., Barreto, A., Rishe, N., Curiel, R.E., Loewenstein, D., Duara, R. and Adjouadi, M., 2019. Prediction Modeling of Alzheimer's Disease and Its Prodromal Stages from Multimodal Data with Missing Values. International Journal of Medical and Health Sciences, 13(2), pp.36-40.

Eslami, M., Tabarestani, S. and Adjouadi, M., 2021. Feasibility Assessment of Multitasking in MRI Neuroimaging Analysis: Tissue Segmentation, Cross-Modality Conversion and Bias correction. arXiv preprint arXiv:2105.14986.

Eslami, M., Tabarestani, S. and Adjouadi, M., 2020, April. Joint Low Dose CT Denoising And Kidney Segmentation. In 2020 IEEE 17th International Symposium on Biomedical Imaging Workshops (ISBI Workshops) (pp. 1-4). IEEE.

Tabarestani, S., Aghili, M., Shojaie, M., Freytes, C., Cabrerizo, M., Barreto, A., Rishe, N., Curiel, R.E., Loewenstein, D., Duara, R. and Adjouadi, M., 2019, May. Longitudinal prediction modeling of alzheimer disease using recurrent neural networks. In 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI) (pp. 1-4). IEEE.

Tabarestani, S., Aghili, M., Shojaie, M., Freytes, C. and Adjouadi, M., 2018, December. Profile-Specific Regression Model for Progression Prediction of Alzheimer's Disease Using Longitudinal Data. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1353-1357). IEEE.

Eslami, M., Karami, M., Tabarestani, S., Torkamani-Azar, F., Eslami, S. and Meinel, C., 2018, November. SignCol: Open-Source Software for Collecting Sign Language Gestures. In 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS) (pp. 365-369). IEEE.

Aghili, M., Tabarestani, S., Adjouadi, M. and Adeli, E., 2018, September. Predictive modeling of longitudinal data for Alzheimer's Disease Diagnosis Using RNNs. In International Workshop on PRedictive Intelligence in MEdicine (pp. 112-119). Springer, Cham.

Tabarestani, S., Eslami, M. and Torkamni-Azar, F., 2015, November. Painting style classification in persian miniatures. In 2015 9th Iranian Conference on Machine Vision and Image Processing (MVIP) (pp. 209-213). IEEE.

Aliasgari, M., Birjandtalab, J., Fakhraie, S.M. and Tabarestani, S., 2012, November. Deep out-of-band radiation reduction by using joint filterbank and cancellation carriers in cognitive radios. In 6th International Symposium on Telecommunications (IST) (pp. 271-276). IEEE.