APPLICATION OF CLASSICAL AND MACHINE LEARNING MODELS ON LONGITUDINAL DATA WITH BINARY RESPONSE

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

RÜMEYSA RANA ARSLAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
STATISTICS

SEPTEMBER 2023

Approval of the thesis:

APPLICATION OF CLASSICAL AND MACHINE LEARNING MODELS ON LONGITUDINAL DATA WITH BINARY RESPONSE

submitted by **RÜMEYSA RANA ARSLAN** in partial fulfillment of the requirements for the degree of **Master of Science** i**n Statistics, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Özlem İlk Dağ
Head of the Department, **Statistics** _____

Prof. Dr. Özlem İlk Dağ
Supervisor, **Statistics, METU** _____

Examining Committee Members:

Prof. Dr. Serpil Aktaş Altunay
Statistics, Hacettepe University _____

Prof. Dr. Özlem İlk Dağ
Statistics, METU _____

Assoc. Prof. Dr. Fulya Gökalp Yavuz
Statistics, METU _____

Date: 11.09.2023

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name Last name : Rümeysa Rana Arslan

Signature :

ABSTRACT

APPLICATION OF CLASSICAL AND MACHINE LEARNING MODELS ON LONGITUDINAL DATA WITH BINARY RESPONSE

Arslan, Rümeysa Rana
Master of Science, Statistics
Supervisor: Prof. Dr. Özlem İlk Dağ

September 2023, 122 pages

Alzheimer's disease (AD) is a significant global health issue that affects both individuals and society for older adults. The symptoms of the disease can be observed over time, making the structure longitudinal. Classical statistical models and machine learning algorithms can be used to analyze these datasets. This study consists of two parts: First, a real dataset is used to find the features affecting dementia status and compare the performances of models. Secondly, a simulation study based on the real dataset with a different number of subjects and an equal number of time points for each subject is conducted to apply and compare the model performances. The classical mixed models, their extended versions, and hybrid models, Boruta, GEE, GLMM, HGLM, GLMMLasso, GPBoost, GLMMTree, and HRF are used for both parts. As a result, GPBoost learns and classifies the dementia status well but overfits due to the small sample size in the dataset, and tree-based algorithms are efficient in predicting the dementia status when a new subject enters the study for the real dataset. For the simulation study, all methods have similar results, but HGLM, GPBoost and GLMMLasso algorithms have better performances regardless of the sample size and balance of the dataset.

Keywords: Longitudinal Data Analysis, Alzheimer's Disease Prediction, Machine Learning Models, Mixed Effects Models

ÖZ

İKİLİ YANITLI BOYLAMSAL VERİLERDE KLASİK VE MAKİNE ÖĞRENME MODELLERİNİN UYGULANMASI

Arslan, Rümeysa Rana
Yüksek Lisans, İstatistik
Tez Yöneticisi: Prof. Dr. Özlem İlk Dağ

Eylül 2023, 122 sayfa

Alzheimer hastalığı (AH), yaşlı bireylerde hem bireyleri hem de toplumu etkileyen önemli bir küresel sağlık sorunudur. Hastalığın belirtileri zaman içinde gözlemlenebilir, bu da veriyi uzunlamasına bir yapıya dönüştürür. Klasik istatistiksel modeller ve makine öğrenimi algoritmaları bu veri kümelerini analiz etmek için kullanılabilir. Bu çalışma iki bölümden oluşmaktadır. İlk olarak, demans durumunu etkileyen özellikleri bulmak ve modellerin performanslarını karşılaştırmak için gerçek bir veri kümesi kullanılmıştır. Ardından, gerçek veri kümesi temel alınarak farklı sayıda birim ve her bir birim için eşit sayıda zaman noktası içeren bir simülasyon çalışması yapılmış ve model performansları karşılaştırılmıştır. Klasik karma modeller, genişletilmiş versiyonları ve hibrid modeller Boruta, GEE, GLMM, HGLM, GLMMLasso, GPBoost, GLMMTree ve HRF her iki bölümde de kullanılmıştır. Sonuç olarak, GPBoost, demans durumunu iyi öğrenir ve sınıflandırır, ancak veri kümesindeki küçük örneklem boyutu nedeniyle aşırı uyum sorunu gözlemlenmiştir. Gerçek veri kümesi için yeni bir birim çalışmaya katıldığında demans durumunu tahmin etmede ağaç tabanlı algoritmalar etkilidir. Simülasyon çalışması için tüm metotlar benzer sonuçlara sahiptir ancak HGLM, GPBoost ve

GLLMMLasso algoritmaları, örneklem büyüklüğü ve veri kümesinin dengesi ne olursa olsun daha iyi performansa sahiptir.

Anahtar Kelimeler: Boylamsal Veri Analizi, Alzheimer Hastalığı Tahmini, Makine Öğrenmesi Modelleri, Karma Modeller

In the memory of my grandpa and Maya

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my supervisor Prof. Dr. Özlem İlk Dağ, for her guidance, support, patience, understanding, and dedication and for making this journey fruitful.

I am also grateful to my thesis committee members, Serpil Aktaş Altunay and Fulya Gökalp Yavuz, for their participation, valuable and constructive feedbacks, suggestions, and comments.

Special thanks are due to my beloved friends Buse Bakış, Begüm Teköz, Pelin Su Baran, my dear colleagues Serenay Çakar, Petek Aydemir, İrem Tanrıverdi, and our dear department secretary Neşe Bayram for always being there, rushing to my side whenever I needed help, both technically and emotionally, for always supporting me, being a family for me, listening to me, understanding and motivating me, giving me the courage and strength and helping me realize that despite the challenges, this journey is also enjoyable, and a very educational one. I am fortunate to have them. Another special thanks to our personal trainer Gökçe Göksel for her understanding and for always providing us with endless motivation and energy.

I am grateful for the support of my friends Mert Kaşif Ceylan, Cemre Yazıcı, Elif Gönülalan, my colleagues Ozancan Özdemir, Burcu Koca, and Onur Çamlı and my dear PASCH Alumni Association family.

My heartfelt thanks are to my grandpa, who has always believed in me and wanted to see me improving myself and doing the job I have dreamed of, to my younger sister Rengin Rüya Arslan for her hugs and love and to Maya for entering my life even for a while and making me feel the most precious emotions that I have ever felt.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ABBREVIATIONS

AD	Alzheimer's Disease
MRI	Magnetic Resonance Imaging
ADNI	Alzheimer's Disease Neuroimaging Initiative
OASIS	Open Access Series of Imaging Studies
PET	Positron Emission Tomography
CT	Computed Tomography
MCI	Mild cognitive impairment
WMH	White Matter Hyperintensity
GUI	Graphical User Inference
NiFTI	Neuroimaging Informatics Technology Initiative
MCapNet	Capsule Network
GEE	Generalized Estimating Equations
GLMM	Generalized Linear Mixed Models
HGLM	Hierarchical Generalized Linear Models
GPBoost	Gaussian Process Boosting
GLMMTree	Generalized Linear Mixed Models Tree
HRF	Historical Random Forest
LASSO	Least Absolute Shrinkage and Selection Operator
GLMMLasso	Generalized Linear Mixed Model Lasso
LR	Logistic Regression
DT	Decision Tree
RF	Random Forest
SVM	Support Vector Machines
XGBoost	Extreme Gradient Boosting
KNN	K- Nearest Neighbors
PCA	Principal Component Analysis

ANN	Artificial Neural Networks
NB	Naïve Bayes
RF GSCV	Random Forest model with Grid Search Cross Validation
LMM	Linear Mixed Models
MERF	Mixed-Effect Random Forest
RE-EM Tree	Random Effects Expectation-Maximization Tree
GD	Gradient Descent
MCAR	Missing Completely At Random
MAR	Missing At Random
MNAR	Missing Not At Random
MICE	Multiple Imputation by Chained Equations
MME	Method of Moment Estimation
MLE	Maximum Likelihood Estimation
MOB	Model-Based Recursive Partitioning
PIRLS	Penalized Iterative Least Squares
IRLS	Iterative Reweighted Least Squares
AIC	Akaike's Information Criterion
BIC	Bayesian Information Criterion
QIC	Quasi-likelihood under the Independence Model Criterion
CIC	Consistency Information Criterion
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

CHAPTER 1

INTRODUCTION

Longitudinal data is a common data type in fields such as medicine, public health, and economics, consisting of repeated measurements of individuals over time. This type of data, also called panel data or repeated measures, is multivariate due to the multiple responses collected from each individual. Unlike traditional multivariate data, longitudinal data allows for varying numbers of measurements per person. The data is hierarchical, with nested observations within subjects (Weiss, 2005). This unique characteristic allows for a more comprehensive approach to measurements and expands the dataset through temporal measurements. Compared to cross-sectional data, which offers limited observations at a single time point, longitudinal data requires fewer subjects, resulting in lower study costs. The main goal of longitudinal studies is to observe how a response changes over time and how other factors influence it. Such studies are unique as they consider subject information and provide easy interpretation analysis without restrictions. Longitudinal data studies offer several benefits that are hard to come by in other studies, such as multiple measurements of an individual often show a correlation, allowing us to predict future values based on past ones (Fitzmaurice and Ravichandran, 2008). By analyzing subject information, these studies provide a more comprehensive and less restricted interpretation, leading to more precise and powerful results. However, these studies can also present some challenges, such as difficulties in collecting data over long periods, leading to data loss, and variations in measurement time and repetition that can create bias and complicate analysis. For instance, longitudinal studies involve collecting a fixed number of repeated measurements from all participants at common time points, this makes the dataset “balanced”. However, in many real datasets, due to the duration of the studies, there are missing time points as a drawback of longitudinal studies.

Statistical models for longitudinal data incorporate two key components: a covariance model that addresses the interdependence among repeated measurements (known as the random effects part of the algorithms) and a model for the mean response (fixed effects part) and its correlation to covariates. The term "covariance" refers to the correlation between repeated measurements within an individual as well as the variation of responses across different occasions. The chosen covariance model heavily influences insights about response changes and their relationship to covariates, even though the mean response model is usually the central focus. Therefore, using the appropriate algorithms for taking the subject-specific covariances into account makes the study more effective and interpretable.

One of the areas in which longitudinal studies are used is health since it is essential to see the change over time to diagnose a disease and is generally used for early detection. Early detection is essential in health diseases such as Alzheimer's, as there is no definite treatment, but prevention is possible through early detection (Buyrukoğlu, 2021). Alzheimer's Disease (AD), first identified by Alois Alzheimer, a German physician, in 1906, is a progressive and intricate neurological disorder that affects memory, cognitive function, and daily life, and puts a heavy burden on individuals, families, and healthcare systems. It is a condition where abnormal protein deposits accumulate in the brain, leading to a breakdown in communication between nerve cells. This ultimately impairs an individual's memory, thinking, and reasoning abilities. Symptoms may be minor, such as memory lapses and difficulty recalling familiar information. However, the condition progressively worsens over time, severely impacting an individual's ability to perform daily tasks without assistance (Bhushan et al., 2018). In the studies of Alzheimer's disease, different approaches are used to understand the characteristics of the disease. These approaches include imaging methods such as MRI scans, biomarkers, and quantitative data from MRI scans. With these methods, the cognitive status of individuals could be determined. In addition to shortening the diagnosis time, the aim is to reduce human interaction by automating the diagnosis of Alzheimer's disease. Beyond reducing time, automation optimizes the diagnostic process by

reducing human interaction. Furthermore, the automation approach provides more precise results while reducing costs. For example, by analyzing MRI scans and applying predictive techniques, patients' dementia status can be better predicted. Those with early-stage Alzheimer's disease are considered to have dementia, improving overall diagnostic accuracy (Kavitha et al., 2022). Therefore, it is important to detect the features affecting Alzheimer's disease and detect the disease early and automate the detection and prediction by using modeling tools and by examining the datasets.

To pave the way for studies in the field and possible ways for early detection, some projects are providing open-access datasets, such as Alzheimer's Disease Neuroimaging Initiative (ADNI) and Open Access Series of Imaging Studies (OASIS). Most of the studies related to the field use these datasets (Buyrukoğlu, 2021; Saratxaga et al., 2021; Wang et al., 2021; Iddi et al., 2019; Leong et al., 2019; Basheer et al., 2021). The OASIS project was started by Marcus et al in 2007 to sustain free neuroimaging datasets available for researchers to have more analysis and discoveries in the field (Marcus et al., 2010). The project has three datasets about dementia, including cross-sectional, longitudinal, and image measurements. In this thesis study, the OASIS – 2 dataset, named "Longitudinal MRI Data in Nondemented and Demented Older Adults" is used. Detailed information about the dataset is given in Chapter 4.

The dataset includes both demographical and clinical measurements. The clinical measurements, Mini-Mental State Examination (MMSE), Clinical Dementia Rating (CDR), normalized Whole Brain Volume (nWBV), Atlas Scaling Factor (ASF), and estimated Total Intracranial Volume (eTIV), are obtained from MRI images and examination scores, and demographical information, age, gender, Socio-economic Status (SES), hand, education. Bhushan et al count the risk factors: age (mainly after the age of 65), genetics (genetic inheritance pattern is not clear, but the causative chromosome can mutate and cause Alzheimer's at an early age), education (less educated people are more at risk) and coexisting health problems such as heart disease, high blood pressure, high cholesterol, or depression (Bhushan et al, 2018).

They also explained the detection methods as Positron Emission Tomography (PET), Computed Tomography (CT), and Magnetic Resonance Imaging (MRI) techniques. In the OASIS-2 dataset, mainly MRI techniques are used for detection, and the variables obtained from examinations and MRI techniques:

- nWBV, normalized whole brain volume, and the risk of dementia increase when the brain shrinks.

- ASF, atlas scaling factor, a brain–volume scaling factor, is used to estimate eTIV (Buckner, et al, 2004).

- eTIV, estimated total intracranial volume related to brain size, does not change over time but affects dementia status (Jenkins et al, 2000). The effect of brain size on dementia can be seen in Figure 1.1.

- CDR, clinical dementia rating, was developed by Washington University as a global scale for denoting the stage and severity of dementia (Morris,1997).

- MMSE, mini-mental state examination, is the most reliable, well-predicted, and commonly used one to score the dementia level, and changes over time (Rodriquez, et al., 2015).

Figure 1.1 Effect of brain size in dementia (Alzheimer's Association, 2023)

This indicates that clinical measurements are important to detect the dementia status, especially the measurements about the brain size.

This study aims to discuss the methods applied to longitudinal data with a binary response and see the drawbacks of working with a dataset in which the number of subjects is small. Using these methods, finding the essential features to address dementia status, and evaluating the effect of the change in the number of subjects

and the equality of the number of time points in the dataset on model performances. The research questions addressed in this thesis are as follows

- What are the classical and machine learning methods to apply to longitudinal data with binary response and a small sample size?
- How does the performance of these methods change on such data?
- What are the main factors among clinical results and demographic information that relate to the dementia status of patients?
- What if the number of time points is balanced, how will the model performance change?
- How will the model performance change in different numbers of subjects?

To answer these questions, the Boruta algorithm is used for feature selection since it is a better identifier and an extended version of random forest. Then by using the information from Boruta and considering the relationships between variables, classical methods that consider the subject-specific classification, such as Generalized Estimating Equations (GEE), Generalized Linear Mixed Models (GLMM), Hierarchical Generalized Linear Models (HGLM), hybrid models which include mixed effects structure, Gaussian Process Boosting (GPBoost), Generalized Linear Mixed Models Tree (GLMMTree), Historical Random Forest (HRF) and Generalized Linear Mixed Model Lasso (GLMMLasso) are utilized and evaluated. The models are compared with real dataset and simulated dataset by using the performance metrics, accuracy, F1 score, sensitivity, and specificity.

This study is structured in the following manner: in the next chapter, the literature review of the studies is given in two parts, related to Alzheimer's disease and those with longitudinal data modeling algorithms. In the third chapter, the methodology includes four main parts: data preprocessing, where missing data techniques and normalization are explained; feature selection with the Boruta algorithm; modeling part, where all implied models are described in detail and model evaluation metrics. It is followed by the design of the study and dataset information in Chapter 4. This

chapter explains a detailed dataset description, data cleaning, manipulation, preprocessing, and the steps before modeling. Then, the design of the simulation study is given. Chapter 5 contains results and four main parts: feature selection, cross-validation, modeling, and simulation study. In the first three parts of Chapter 5, the application of the methods, discussed in Chapters 2 and 3, the detailed interpretation of the applied algorithms and results are identified and compared. In the fourth part, the details and model performance metrics of the simulation study are given, and the results are explained. The final chapter stands for the conclusion of the study and discussions.

CHAPTER 2

LITERATURE REVIEW

In this chapter, the literature of the studies on the detection of Alzheimer's disease and studies and applications on longitudinal data are reviewed. In the section related to the dataset, the articles related to the modeling of Alzheimer's disease and the dementia status of patients are reviewed, whereas, in the other section, the models applied to longitudinal data in the literature are discussed.

2.1 Literature Review on Studies of Alzheimer's Disease

Alzheimer's disease (AD), which is the primary cause of dementia, is a significant global health issue that affects both individuals and society (Lane et al., 2018). It generally affects older individuals, and detecting the disease early can help reduce its prevalence (Lee et al., 2019). However, almost half of the early diagnoses are incorrect. This is why many data mining methods are explored to address the early detection of Alzheimer's disease. Mainly, classification methods are used and developed in the detection of AD. In AD classification, the datasets consist of the demographic information and clinical test results of patients with MCI, AD, or nondemented patients (Buyrukoğlu, 2020). Early detection and prevention are possible by using the classification and feature selection methods on these datasets.

There are both cross-sectional and longitudinal studies in the field, which comprise the features affecting the dementia status of patients. At the same time, longitudinal studies also investigate the changes in these factors over time. There are institutions collecting information related to AD, and OASIS is one of them providing researchers with neuroimaging datasets. OASIS has three datasets collected over the years. The first dataset of OASIS is a cross-sectional study consisting of 416 subjects

categorized as very mild, moderate Alzheimer's, or without dementia, according to their CDR scores and their processed MRI data besides their demographic information (Marcus et al., 2007). The second dataset of OASIS, also used in this study, is a longitudinal study consisting of 150 patients scanned on different visits, including 373 imaging sessions (Marcus et al., 2010). Detailed information about the dataset is given in Table 4.1. Lastly, the third OASIS dataset is a longitudinal study of 1090 subjects. Still, differently in this study, MRI and PET images are included (Saratxaga et al., 2021).

Both classical approaches, machine learning algorithms and deep learning algorithms are used to classify the datasets. The classical approaches used for classification are mainly Logistic Regression (LR) for cross-sectional studies, GEE, and GLMM for longitudinal studies. In the study of Wang et al (2021), the dataset is created by 1099 participants and the collection of nine risk factors affecting dementia, their demographic information, lifestyle, and genetic risk. The aim was to predict the risk of Alzheimer's disease for older adults. LR was used to model the risk after selecting the essential features with the Least Absolute Shrinkage and Selection Operator (LASSO). The essential features to build the model are sex, age, economic status, health status, lifestyle, and genetic risk. Then, the logistic model showed that an older male with moderate economic status, good health status, smoking, drinking, and exercise habits, a regular diet, and no genetic risk have a 33.4% risk of AD (Wang et al., 2021). Brickman et al. demonstrated an analysis of a longitudinal study with six follow-ups about cerebral atrophy and severity of white matter hyperintensity obtained from MRI images in the prediction of future Alzheimer's disease status of patients. The MRI factors affecting modified mini-mental state examination are identified by using GEE models, and the result is that the bicaudate ratio has a positive relationship with mental status scores over time. Still, the baseline bicaudate ratio and white matter hyperintensity (WMH) severity have a negative relationship over time. Age, sex, and education level are also associated with poorer mental status scores. The combination of baseline atrophy and WMH severity is found to have the most substantial effect on the decrease in

mental state scores (Brickman et al., 2008). However, there were limitations that it may be a better analysis to have neuroimaging data for all participants and consider them in the analysis. The mixed model approach is seen in the study of Iddi, et al. and is followed by random forest, making the study complete in two stages. In the first stage, a joint mixed effects model and a latent time joint mixed effects model are fitted to simultaneously address the multiple markers of AD over time. Joint models are used since they consider the correlations among outcomes, and latent time joint models are for predicting long-term trajectories with short-term data. In the second stage, random forests are used to predict categorical diagnoses from predictions of continuous markers based on the first-stage model. This two-stage approach yields better diagnostic classification accuracy than separate univariate models for each marker (Iddi et al., 2019).

In Alzheimer's disease prediction and classification, mainly machine-learning approaches such as Decision Tree (DT), Random Forest (RF), Artificial Neural Networks (ANN), Naïve Bayes (NB), and Support Vector Machines (SVM), XGBoost, K-Nearest Neighbors (KNN), and Principal Component Analysis (PCA) are used. In the study of Buyrukoğlu (2020), both feature selection and machine learning algorithms are applied to identify the features affecting Alzheimer's disease and the effect of these features in predicting the early diagnosis of Alzheimer's disease. They used the ensemble feature selection algorithms to reduce the number of features and data mining algorithms, RF, ANN, LR, SVM, and NB for prediction, resulting in the RF algorithm outperforming (Buyrukoğlu, 2020). Moreover, Buyrukoğlu states the four important limitations in many ML studies: the non-usage of essential features, missing values, small sample sizes, and disregarding the model metrics like sensitivity and area under the curve. These limitations are solved in his study by identifying the important features by feature selection, imputing missing values, using a large dataset, and reporting all required metrics. This thesis considers these limitations, and feature selection, missing imputation, and reporting all possible metrics are completed. The sample size of the real dataset is small, and predictions are completed with it, but to see the effect of sample size, a simulation

study is implemented. A study conducted using the cross-sectional OASIS dataset belongs to Kavitha et al. Feature selection by using correlation coefficient and information gain is applied, and then DT, RF, SVM, XGBoost, and Voting algorithms are constructed to classify the dementia status of patients. By considering the accuracies and precisions, RF is found to be the best model to classify the dementia group, and XGBoost follows it (Kavitha et al., 2022). Another study utilizing OASIS datasets was completed by Leong and Abdullah (2019) to select the important features using Boruta, a wrapper method in feature selection. The feature selection is applied to the longitudinal OASIS-2 dataset. ASF, eTIV, nWBV, MMSE, and CDR are found to be the first five important variables in classifying the dementia status. 12 machine learning algorithms are applied by using the result of Boruta and without Boruta and the model performed better with the results of Boruta algorithm. The Random Forest model with Grid Search Cross Validation (RF GSCV) outperformed with Boruta algorithm. However, the OASIS-1 dataset is used in Graphical User Inference (GUI) prediction, and it results that the dementia status is related to the demographic, clinical, and MRI-derived information gained from patients (Leong et al., 2019).

Besides ML algorithms, DL is commonly used in Alzheimer's disease studies in both modeling numerical data and image processing. One of the studies conducted by Saratxaga et al (2021) includes machine learning and deep learning approaches implemented on a combined OASIS dataset. The proposed dataset is created by taking OASIS-1 and OASIS-2 datasets to have a large sample to avoid limitations for deep learning algorithms. They used images given in datasets besides the clinical information. The images are converted into numerical data using 16-bit Analyze 7.5 and 16-bit NiFTI1 format, and the data supplemented from images are also implemented as inputs. Different approaches are proposed, such as 2D and 3D networks, for classification models on both datasets, achieving high accuracy in detecting the presence and stage of Alzheimer's disease. The proposed approaches for classifying Alzheimer's disease have equal or better results than the state-of-the-art approaches (Saratxaga et al., 2021). The study of Basheer et al includes the

implementation of the MCapNet algorithm dataset, a DL algorithm applicable for hierarchical modeling. The models revealed that gender and age are the most influential variables in addressing dementia status. The model performs an accuracy of 92.39% which makes the model outperform other state-of-the-art ML models. The limitation of the model is that the computational complexity increases when applied to heterogeneous data (Basheer et al., 2021).

The main limitation of these studies using longitudinal AD datasets is not considering the dataset's longitudinal structure and subject-specific effects. Also, the studies mainly focus on finding the most accurate model for AD prediction. This is also the main concern of our study. The thesis proposed by Murchison (2021) emphasizes this problem. The thesis states that ML algorithms generally focus more on cross-sectional than longitudinal data and limit their predictive capacity. There are extensions in ML algorithms to be applicable in longitudinal structures by considering the time dependency and random effects. Still, they are limited and not in the same structure for all algorithms. The thesis of Murchison aims to bridge this gap by processing and comparing different longitudinal ML and generative DL methods to evaluate the utility of longitudinal ML methods in Alzheimer's disease prediction. It has resulted that the analysis method should be determined according to the researcher's objective. There is no specific best model for regression and classification of longitudinal Alzheimer's disease studies (Murchison, 2021). However, in this thesis study tree-based models, with longitudinal structure added, have better results.

To make predictions or classifications based on health data, various ML methods have been used, including Random Forest, Support Vector Machine, and sparse regression models like Lasso, and Elastic Net proposed by Knights et al (2011), Statnikov et al (2013), and Pasolli et al (2016). These methods have been proven effective in addressing the challenge of high dimensionality. However, they may not be the best option if clustered variables (i.e., a longitudinal structure) exist (Xiao et al., 2018). Therefore, besides the classical generalized linear regression modeling, our study uses ML algorithms applicable to longitudinal data to predict dementia

status in the OASIS-2 dataset. The implementation of the algorithms in the literature is discussed in Section 2.2

In the studies that which ML classification algorithms are used to classify the dementia status of patients based on OASIS datasets, the group variable is mainly taken as two, demented and nondemented. Basheer et al. (2021), Saratxaga et al. (2021), and Leong et al. (2019), directly take the converted group as demented and splitting considering the CDR score (0 is nondemented and greater than 0 is demented). In our study, the dementia group of patients is categorized considering the CDR score and in contrast to the original OASIS-2 dataset, not only the final dementia status but the dementia status in each visit is considered.

2.2 Literature Review on Studies of the Algorithms Applied in Longitudinal Studies

In this section, the algorithms applicable for analyzing longitudinal data are discussed. The algorithms progress from simple to complex, starting with the most common classical regression models, then the extended versions of statistical models, and lastly, ML algorithms with random effect integrations.

Longitudinal studies are important for examining and analyzing changes in the target variable over time and very common in health studies. Because of correlations between repeated measures, heterogeneity, and missing responses, longitudinal studies should use models appropriate for the data structure. These models typically include algorithms that analyze the dataset's fixed and subject-based effects. By using the longitudinal continuous response dataset from a study of pulmonary function decline and the effects of smoking, linear mixed-effects models are used by Fitzmaurice and Ravichandran (2008). The study results show that the fitted models are appropriate for the data type and emphasize a valid effect of smoking and time on forced expiratory volume as a risk factor for lung diseases. (Fitzmaurice and Ravichandran, 2008). Hadgu and Koch (1999) proposed another longitudinal study

in clinical trials, considered a dental clinical trial in which each subject's outcome was measured two times. The study used an extension of GLM, Generalized Estimating Equations (GEE), since the model is advantageous in flexible error and correlation structures, handling different covariates and imbalances in subject-specific measurements. The study aimed to compare the impact of the three mouth rinses in interrupting the development of dental plaque. It was concluded that two experimental rinses are more influential than the control rinses (Hadgu and Koch, 1999).

The hybrid models combine GLMMs and several feature selections of classification algorithms. The combination of LASSO and GLMM was used in the study of Muslim et al. (2021) in analyzing monthly rainfall data, a longitudinal dataset. The study compared three methods, LASSO, GLMM, and GLMMLasso. When the number of covariates was small, then GLMM seemed to be a good way, but still, the correlation between covariates was a limitation of GLMM since they have caused multicollinearity and biased results. By implementing LASSO, the L1 penalty term was added, and feature selection and shrinkage were applied, then this multicollinearity problem in GLMM was overcome. The rainfall data is high-dimensional and applying GLMMLasso resulted in a consistent analysis without losing information. As results also showed, applying the GLMMLasso method for various lambda parameters outperformed applying GLMM and LASSO separately. Another rainfall study in which GLMMLasso has been used was conducted by Muslim and Hayati (2021), and they compared the GLMMLasso model with the normally distributed response and the GLMM model with the Gamma distributed response. Again, it was concluded that the GLMMLasso model has better performance metrics than the GLMM model (Hayati and Muslim, 2021). Another hybrid method that combines the Gaussian process, mixed effects models, and boosting is GPBoost. The method can be applied to datasets where a within-subject correlation is accounted for, and the grouping variable is modeled as random effects, as stated in the study of Zhou et al. (2022). The study compared the GPBoost prediction model and seven other regression models in which the within-subject

correlations were not considered, such as linear regression, linear SVM, quadratic SVM, Gaussian SVM, DT, RF, and XGBoost. At first, all 10 covariates in the dataset were included in the GPBoost model. A reduced model was constructed with five important features, and the reduced GPBoost prediction model outperformed others (Zhou et al., 2022). In medical studies, because of the longitudinal structure of many datasets, GPBoost is an optimal algorithm. In the simulation study conducted by Turgal (2021) on the evaluation of mixed-effect machine learning algorithms, LMM, MERF, and GPBoost were applied. Their performances were compared based on simulated datasets with different sample and subject sizes. The study shows that GPBoost performed better than LMM and MERF in model performance and computational time (Turgal, 2021).

Tree-based regression models are statistical models used to predict continuous and binary responses. Recently, hybrid methods have been developed to combine the random effects structure and tree-based modeling. GLMMTree is one of those hybrid methods developed by Fokkema et al. (2018) to identify the treatment-subgroup interactions and consider the clustered structure of a dataset. The algorithm applies model-based recursive partitioning to detect interactions and a GLMM to estimate random-effects parameters (Fokkema et al., 2018). This algorithm was used in a microbiome study conducted by Xiao et al. (2018) to capture the dense signals, which are community-level changes where many species that rely on each other contribute to the outcome. The algorithm identified the microbiome similarity to predict outcomes while incorporating other covariates. Simulation studies and real data analysis were applied and GLMMTree outperformed existing methods in clustered dense signal scenarios (Xiao et al., 2018). In the study of Çakar and Yavuz in neuroscience, to predict the HbO value, besides LMM, hybrid methods are used, such as tree-based ones (RE-EM Tree and GLMM Tree), and GPBoost with and without using the nested structure of the dataset. The LMM algortihm indicates the significance of the relation between HbO values with mean response time and n-back condition level, although the the machine learning algorithms did not capture it. Among the machine learning algortihms GLMM Tree is evaluated to be the better

algorithm in predicting the HbO value when a new subject entres the data (Çakar and Yavuz, 2023). Different optimizers in GPBoost are trained and the default optimizer, Gradient Descent(GD) is found to have a better performance. Another hybrid tree-based algorithm is Historical Random Forest (HRF), a random forest algorithm for classification and regression considering subject and time effects. Hu and Szymczak (2023) stated in their review of random forest algorithms applied to longitudinal data that the HRF approach combines the past and present observations of variables to create a model for the conditional mean of the response variable (Hu and Szymczak, 2023).

The classical regression models and machine learning algorithms applicable to binary response longitudinal data are utilized in our thesis study. The application and results are discussed in Chapter 5.

CHAPTER 3

METHODOLOGY

In Chapter 2, the studies based on Alzheimer's disease and longitudinal studies are discussed. It is seen that the main aim of all studies about Alzheimer's disease is detecting the disease level or status of patients by using statistical, machine learning, and deep learning models for supplying early detection and, thereby, prevention. Also, the features influencing the disease are identified. Moreover, many statistical and machine learning methods are applied to longitudinal datasets considering the within-subject correlations. The methods discussed in this chapter are evaluated for Alzheimer's disease specifically.

In this chapter, the background information on the methods is given. The chapter consists of four main parts, processes applied before modeling to prepare the dataset for modeling and evaluation, feature selection with Boruta, modeling longitudinal data, and model performance metrics for evaluating the applied methods.

3.1 Data Preprocessing

Data preprocessing is a major step before starting the analysis (García et al., 2015). The quality and unbiasedness of the analysis depend on the data quality. In real-life, raw data consists of irrelevant and unnecessary variables, missing values, noise, inconsistent parts, and outliers. To increase the efficiency of data and so as analysis, preprocessing is important (Alasadi and Bhaya, 2017).

This section explains missing data applications, including missing data mechanisms and imputation methods, scaling, and cross-validation methods.

3.1.1 Missing data applications

It is common to encounter missing cases in statistical applications for several reasons. Especially, in longitudinal studies, it happens more due to the long duration of studies compared to cross-sectional ones. The percentage of missing values around 15-20% can be considered usual in longitudinal studies (İlk, 2004).

3.1.1.1 Missing data mechanisms in longitudinal data

The reasons for missing values can be manual data entry mistakes, equipment errors, incorrect measurements, or subject-dependent causes (García et al., 2015). In longitudinal studies, commonly subject-dependent causes are considered. Subjects may drop out of the study or take a break from the study (they continue later), there may be drawbacks of the study that affect the subject. The subject cannot continue, subjects may be bored with the study and unwilling to continue participating or get disturbed because of the content of the study (İlk, 2004). These cause missing patterns in the data.

Missing data mechanisms are proposed by Little and Rubin as

- Missing completely at random (MCAR),
- Missing at random (MAR),
- Missing not at random (MCAR).

To formulate the missing mechanisms, let the response variable be Y_i, and covariates be X_i. $Y_i = \left(Y_{i1}, Y_{i2}, \dots, Y_{in_i}\right)'$ denotes the repeated measurements for the i-th subject where $i = 1,2,3, \dots, N$ and Y_i is partitioned as $Y_i = (Y_i^o, Y_i^m)'$ where Y_i^o denotes the observed values and Y_i^m denotes the missing values in case that Y_i has some missing values. Similarly, $X_i = \left(X_{i1}, X_{i2}, \dots, X_{in_i}\right)'$ denotes the covariates for the i-th subject. The missing values of the response are denoted by an indicator $R_i = \left(R_{i1}, R_{i2}, \dots, R_{in_i}\right)'$ where $R_i = \begin{cases} 1, & Y_i \; is \; observed \\ 0, & Y_i \; is \; measured \end{cases}$.

Missing Completely at Random

In the MCAR case, the probability of a missing response is unrelated to observed and unobserved values. The mechanism is formulated as $f(R_i\,|Y_i^o, Y_i^m, X_i) = f(R_i)$, which means the distribution of the response variable does not depend on the observed and missing values of Y and covariates.

Missing At Random

In the MAR case, the probability of a missing response is unrelated to unobserved values but related to observed values, which means the missingness may be because of a side effect of the study. The side effect is important but not the objective of the analysis. The mechanism is formulated as $f(R_i\,|Y_i^o, Y_i^m, X_i) = f(R_i|Y_i^o, X_i)$, which means the distribution of the response variable depends on the observed responses and covariates.

Missing Not at Random

In the MNAR case, the probability of a missing response may be related to observed and unobserved values. This is also known as non-ignorable missing since the absence of an observation may be because of the effects of the study itself. The mechanism is formulated as $f(R_i\,|Y_i^o, Y_i^m, X_i) = f(R_i|Y_i^o, Y_i^m, X_i)$, which means the distribution of the response variable depends on the missing and observed responses, and covariates.

To decide on the missing mechanism of data, the percentage and pattern of missing can be checked, and the relationship between the variable having missing values and other variables can be evaluated. It is important to select the appropriate solution to eliminate missing values. If the mechanism is MCAR and MAR, the missingness can be ignored, which means the missing values can be predicted from the existing data, and there is no need to apply special modeling, any application is appropriate. However, if it is MNAR, then the missingness cannot be predicted from the existing data and it should be modeled by special algorithms.

3.1.1.2 Missing data techniques

3.1.1.2.1 Complete case analysis

Complete case analysis means working with a dataset with no missing values, which happens when the missing values are directly omitted. There are both advantages and disadvantages to this method, such as there is no need to use any special methods or apply any special analysis to deal with missing values, but it will be a serious loss of information. It may be efficient in MCAR cases and results in an unbiased estimation, but if it is not the case, complete data does not show the characteristics of the whole dataset and it will cause biased estimates (Nakai and Ke, 2011).

3.1.1.2.2 Single imputation

Single imputation replaces the missing values with only one observation based on the remaining complete part of the dataset. It is practical but may also cause biased estimates in the analysis, ruin the structure of the dataset so as the analysis, and cannot show the sampling variability accurately (Nakai and Ke, 2011). There are four main ways to apply single imputation, as the last observation carried forward, mean imputations, hot-deck imputation, and imputation using expectation-maximization.

Last Observation Carried Forward

It is a commonly used method that implements the last observed value of that variable for each subject separately, especially in the pharmaceutical industry. But this method changes the summary statistics of the variables, such as mean, covariance, and variance structures, which may cause biased estimates.

Mean Imputation

Mean imputation substitutes mean values of observed values in missing values. It has three methods: subject mean imputation, occasional mean imputation, and

conditional mean imputation. In subject mean imputation, the missing value is replaced by the mean, median, or mode of the column that the observation is missing for that specific subject. For instance, in a longitudinal study with five time points and the fourth time point is missing, the mean/median/mode of that variable for that subject is taken and imputed in the fourth visit. Occasion mean imputation is similar to the subject mean but replaces the missing observation with the mean, median, or mode of that time point, which means considering the example, the mean/median/mode of the fourth time point of all subjects. Conditional mean imputation uses regression to impute the missing value. Specifically, all subjects or time points with missing values are removed, then a regression model is fitted, and using the model equation, the missing value is calculated and imputed (Little and Rubin, 2022).

Hot-deck Imputation

In hot-deck imputation, the missing values are replaced with similar values in the sample. It is generally used in survey studies. The replaced values do not ruin the distribution of the variable, but it is hard to find the values having similar characteristics to the missing value (Nakai and Ke, 2011).

Imputation by using Expectation – Maximization Algorithm

The method uses an expectation-maximization algorithm which iteratively finds parameters that maximize log-likelihood and benefits from the relationship between missing values and unknown parameters of the model based on data. However, its convergence rate may be very slow when the percentage of missing values is high (Nakai and Ke, 2011).

3.1.1.2.3 Multiple imputation

Single imputations are easy to apply but lack of accurate estimates and decrease variability, which is important for the reasonability of parameter estimates and standard errors (Patrician, 2002). Multiple imputation is recently the most popular

method in missing data imputation (Nakai and Ke, 2011). Rubin developed multiple imputations in 1978 to impute the predictive distribution of the missing values based on the observed values, and this distribution naturally relies on a model and demonstrates how the conclusions are affected by reasonable model selections (Rubin, 1978). Multiple imputation estimates parameters several times by using regression and these estimates are combined to have effective imputations and reasonable standard errors (Daniel, 2003). The multiple imputation method has three phases: filling in the missing data x times to obtain x complete datasets, analyzing these x complete datasets using standard procedures, and combining these results from the x complete datasets for inference (Nakai and Ke, 2011).

Recently, a multiple imputation method, Multiple Imputation by Chained Equations (MICE) has been used and the process has four main steps:

1. Initial imputation: every missing value in the dataset is initially imputed with a basic estimation such as mean imputation and these imputations are thought as "placeholders".
2. Reset and prepare: The provisional imputations for a specific variable are then reverted to missing.
3. Regression-based estimation: the known values from the imputed variable in step 2 are used as the dependent variables in the regression model and this model involves other variables as well, which could encompass all or only a subset of the variables in the dataset. The other variables behave like independent variables in the regression process.
4. Substitute missing values: The vacant values for the imputed variable (in step 2) are then substituted with imputations obtained from the regression model.
5. Cycle through variables: Steps 2-4 are repeated sequentially for each variable, including missing data, and this process is called a "cycle". At the end of each cycle, all missing values are imputed with predictions stemming from the regressions.
6. Repeat and update: Steps 2-4 are iterated numerous times, with imputations being refreshed at each cycle (Azur et al., 2011)

Figure 3.1 shows the scheme of the MICE application in the R programming language. Firstly, the “mice” package is applied to the data frame with missing values and several imputed datasets are obtained. Then, by the “with” function, the best-imputed dataset is selected, by checking the structure of imputed and existing values in the dataset and building regression models to see whether the imputed values reflect accurate estimates. After that, the results are obtained for each imputed dataset, and by using the “pool” function, the imputed dataset is implemented in the existing dataset (Zhang, 2016).

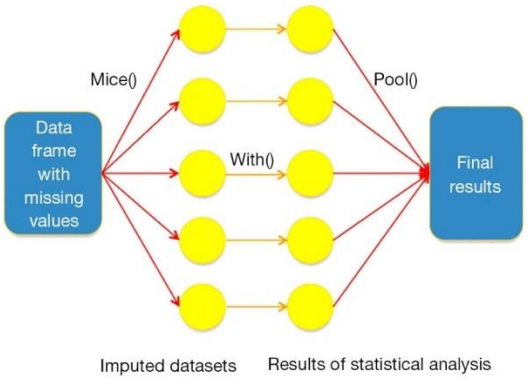

Figure 3.1 Illustration of the process of MICE package (Zhang, 2016)

3.1.2 Normalization

Normalization is a crucial pre-processing stage in data mining. Its purpose is to manipulate data by scaling it up or down, making them on the same scale before it is used in further stages. The two important normalization techniques are Min-Max and Z-score normalization (Patro and Sahu, 2015).

Min-Max Normalization is a technique that shifts and rescales the numeric variables into a pre-defined boundary. It is sensitive to outliers.

Patro and Sahu demonstrate the formula of min-max normalization as

$$A' = \frac{A - minimum\ value\ of\ A}{maximum\ value\ of\ A - minimum\ value\ of\ A} * (D - C) + C \qquad (1)$$

where [C, D] are the elements of the predefined boundary if done and A is the original data (Patro and Sahu, 2015).

Z-score normalization centers the data around the mean and can be used for normally distributed or unknown data. It is less sensitive to outliers and keeps the relationship between variables.

Patro and Sahu demonstrate the formula of Z-score normalization as $v_i' = \frac{v_i - \bar{E}}{std(E)}$ where v_i is the value of row E in i-th column, $std(E) = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(v_i - \bar{E})^2}$ and $\bar{E} = \frac{1}{n}\sum_{i=1}^{n} v_i$ or directly the mean value (Patro and Sahu, 2015).

3.2 Feature selection with Boruta

Feature selection is an important step in predictive modeling since working with all variables makes the analysis impractical, and computationally slow, and having irrelevant variables in the analysis can affect the visibility of important variables and may change the analysis pattern (Stanczyk and Jain, 2015). Feature selection approaches are classified as filters, wrappers, and embedded methods (Stanczyk and Jain, 2015).

Boruta is a feature selection algorithm, a wrapper built around the random forest. It selects a variable based on the random forest algorithm without user-defined parameters. The name Boruta is the name of the god of the forest in Slavic mythology (Kursa and Rudnicki, 2010). By using an iterative process, it eliminates the less relevant variables based on statistical tests. Boruta separates variables into three classes: confirmed, tentative, and rejected (Stanczyk et al., 2015).

In random forest, firstly, the algorithm classifies the objects, and each tree contributes its votes to classifying objects not involved in its criterion. Trees vote for

the correct class, and the significance of variables is measured by comparing the original and permuted systems. Variable importance is determined by averaging these measures across trees or calculating Z-score.

These importance scores alone cannot identify meaningful relationships between variables and decisions. Breiman assumes that the importance is normally distributed since the correlation between trees is low, and therefore, Z-score can be used but the assumption is wrong, and another algorithm is required to detect the truly important variables (Breiman, 2001; Kursa et al., 2010). The algorithm Boruta is developed to serve this purpose. It implements an extra layer of randomness that generates a randomized duplicate of the system, combines it with the original, and is employed to build a classifier for the augmented setup. Evaluating the importance of a variable in the original system involves comparing it to the importance of randomized variables. Only the variables with importance scores exceeding the randomized variables are considered significant (Kursa et al., 2010).

In Boruta, the Z-score is used as an importance measure, but it requires an external benchmark to differentiate between the importance of significant features and random fluctuations. It is addressed by extending the information system with purposely designed random attributes, called “shadows”. The values of shadows are achieved by shuffling the original attributes across objects. Classification is conducted, and the importance of shadow attributes serves as a reference for discerning genuine importance. In Boruta, adding extra randomness results in a better insight into which attributes are important, confirmed, or not, rejected (Kursa and Rudnicki, 2010).

3.3 Modelling longitudinal data

3.3.1 Marginal models with binary response

Marginal models are commonly used in longitudinal data analysis. The method is developed to examine hypotheses concerning the effects of factors on response variables with binary and exponential family distributions, collected over time within subjects. It extends generalized linear models and enables regression analysis for non-normally distributed responses (Ballinger, 2004). Marginal models are explained in the thesis study of Asar (2012) as an expansion of GLMs to be applicable in longitudinal data. They allow for direct regression of covariate impacts on the mean response, without the need to account for random effects or past responses. This dissociates the regression parameter interpretation from within-person associations, hence the term "marginal models" (Fitzmaurice et al., 2004). While longitudinal data analysis often focuses on the correlation among recurring outcomes, acknowledging covariate effects and within-subject associations is crucial for accurate deductions. Marginal models prove advantageous in epidemiological and health investigations that aim to contrast subsets based on average population responses (Fitzmaurice et al., 2004).

Marginal models are formulated as

$$g(\mu_{it}) = X_{it}\beta \qquad (2)$$

where $\mu_{it} = E(Y_{it}|X_{i1}, \ldots, X_{it})$ represents the mean response which depends on the covariates X_{it} for the i-th subject at time t. It is also assumed as $E(Y_{it}|X_{i1}, \ldots, X_{it}) = E(Y_{it}|X_{it})$ meaning that the mean outcome at time t is the mean of the outcome for the i-th subject at time t related with the covariates of the same subject at time t is taken. $g(.)$ is the link function used for making the effect of covariates on the mean response linear, and β is the vector for regression parameters having dimensions $p \times 1$.

The diagram of the marginal models is given in Figure 3.2. In the figure Y_{it}, $i = 1,2, \ldots N$ represents the response and the response values are obtained by applying β parameters to the covariates X_{it}. The α parameter stands for the within-subject association and shows the relation between the time points for each subject. The exchangeable correlation structure is assumed in this figure.

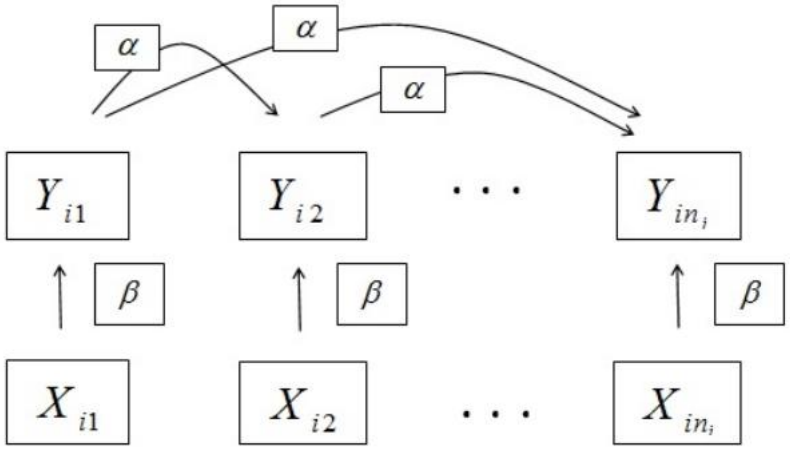

Figure 3.2 Diagram of the marginal models (Asar, 2012)

For binary outcomes {0,1}, the illustration of the marginal models is

$$logitP(Y_{it} = 1|X_{it}) = \beta_0 + \beta_1 * X_{1it} + \beta_2 * X_{2it} + \cdots \quad (3)$$

The formula can be extended according to the number of covariates. The $logit$ function here defines the logarithm of odds of $Y = 1$ compared to $Y = 0$. The interpretation of the parameters is in the form of odds ratio, and the interpretations are not subject-based, but covariate based which means if X_1 is gender, the interpretation of gender is between males and females.

3.3.1.1 Generalized estimating equations

The generalized estimating equations technique was developed by Liang and Zeger in 1986. In GEEs, a variance function is employed, which entails a transformation matrix determined by the observed mean. The variance function aids in computing

parameter variances, allowing for non-constant variances linked to specific functions of the mean (Ballinger, 2004). GEEs can be used to model the longitudinal data without any assumptions in the response distributions and considers within-subject dependence by assigning a working correlation matrix using the variance-covariance matrix of the response in different time points. It is essential that GEE is not a model but a parameter estimation technique of marginal models (Molenberghs and Kenward, 2010).

Liang and Zeger (1986) proposed an extension of the quasi-likelihood algorithm to be used for both discrete and continuous responses for longitudinal data modeling. In the quasi-likelihood approach, it is not required to specify the entire distribution of the response. It has two main assumptions: the relationship between the mean of the response and covariates is specified, and the relationship between the mean and variance of the response is specified.

The demonstration of the system of quasi-likelihood is

$$U(\beta) = \sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \beta} v_i^{-1} (Y_i - \mu_i) = 0 \tag{4}$$

Here, the number of subjects is given by N and $\frac{\partial \mu_i}{\partial \beta}$ represents the first derivative of the mean of the response with respect to the regression parameters. The parameter estimates of quasi-likelihood are found by solving the system (4) using iteratively re-weighted least squares method (Gentle et al., 2004).

The relationship between the response and covariates in GEE approach is the same with the marginal model defined, in the equation (2). The relationship between mean and variance of the response is

$$v_{it} = h(\mu_{it})/\phi \tag{5}$$

where $v_{it} = \text{Var}(Y_{it} \mid X_{it})$, $h(.)$ is the diagonal matrix of known variance functions and ϕ is given as the dispersion parameter (Karadağ and Aktaş, 2018).

Also, the covariance structure of the responses, $Y_i = (Y_1, \dots, Y_{n_i})$, in longitudinal data should be identified. The working correlation matrix, V_i, of responses is given in formula (6).

$$V_i = \frac{A_i^{1/2} R_i(\alpha) A_i^{1/2}}{\phi} \tag{6}$$

in which, A_i represents a $n_i \times n_i$ diagonal matrix with $h(\mu_{it})$, $R_i(\alpha)$ represents a $n_i \times n_i$ working correlation matrix which deals with the relation between repeated measures, and ϕ represents a scale parameter which is taken as a nuisance parameter and should be estimated or known.

The working correlation matrix does not have to be correctly specified. Even if it is not correctly specified, the parameter estimates and variances are similar, but it is important to increase the efficiency. To obtain the best working correlation, the relations between repeated measures are used. The common correlation structures are given, for a longitudinal dataset with four time points for each subject, in a 4×4 matrix format as

Simple:
$$\begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ & \sigma^2 & 0 & 0 \\ & & \sigma^2 & 0 \\ & & & \sigma^2 \end{bmatrix}$$

Unstructured:
$$\begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ & \sigma_{22}^2 & \sigma_{23} & \sigma_{24} \\ & & \sigma_{33}^2 & \sigma_{34} \\ & & & \sigma_{44}^2 \end{bmatrix}$$

Compound symmetry (exchangeable):
$$\begin{bmatrix} \sigma^2 + \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ & \sigma^2 + \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ & & \sigma^2 + \sigma_1^2 & \sigma_1^2 \\ & & & \sigma^2 + \sigma_1^2 \end{bmatrix}$$

Toeplitz:

$$\sigma^2\begin{bmatrix} 1 & \sigma_1 & \sigma_2 & \sigma_3 \\ & 1 & \sigma_1 & \sigma_2 \\ & & 1 & \sigma_1 \\ & & & 1 \end{bmatrix}$$

AR(1):

$$\sigma^2\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ & 1 & \rho & \rho^2 \\ & & 1 & \rho \\ & & & 1 \end{bmatrix}$$

The parameter estimates, $\hat{\beta}, \hat{\alpha}, \hat{\phi}$, can be found by deriving the extension of quasi-likelihood:

$$\sum_{i=1}^{N} D_{it}^T V_i^{-1}(Y_{it} - \mu_{it}) = 0 \tag{7}$$

where, $D_{it} = \partial\mu_{it}/\,\partial\beta, \mu'_{it} = (\mu_{i1}, \dots, \mu_{in_i})$.

The equation can be solved by an iterative algorithm which uses two-stage estimation: firstly, using a modified Fisher scoring algorithm, given the estimates of α and ϕ, the estimates of β are estimated in each iteration. Second, given the estimates of β the parameters α and ϕ are found by the Method of Moment Estimation (MME) by using the Standardized Pearson Residuals (Asar, 2012).

3.3.1.2 Implementation in R

In R programming language, "geepack" package can be used to fit GEE models. The package is developed by Højsgaard et al (2006). In the package, "geeglm" function builds the marginal model with generalized estimating equations. Besides the model formula, the id variable, family link and correlation structure are assigned. In order to decide on the correlation structure, three main options can be applied: the relations between repeated measurements can be considered, the model can be fitted under the unstructured covariance assumption and the working correlation of that model can be checked, and QIC or CIC can be used for comparing models with different

correlation structures (the detailed information is in Section 3.4.1.2). The control parameters can be added to increase the efficiency of the model and decrease the computational time by using the "geese.control" function. The parameters to be controlled are:

- epsilon: used for positive convergence tolerance. If the absolute value of the difference in parameter estimate is smaller than this value, the iteration converges. The default value is 0.0001.
- maxit: the maximum number of interactions in fisher scoring. The default is 25.

3.3.2 Generalized Linear Mixed Effect Models (GLMM)

Generalized linear mixed models are an extended version of both generalized linear models and linear mixed models in terms of considering any type of responses and random effects caused by within-subject correlations, respectively. They are widespread in longitudinal studies. They enable to obtain subject-based model parameters, which means heterogeneity among subjects.

3.3.2.1 Random Effect Models

The general structure of the random effect model is

$$Y_{it} \sim_{ind} \text{Bernoulli}(p_{it}), \tag{8}$$

$$\log\left(\frac{p_{it}}{1-p_{it}}\right) = X_{it}\beta + Z_{it}b_i, \tag{9}$$

where $b_i \sim_{iid} \text{N}(0, G)$ and it is usually assumed as $G = \sigma_0^2$, Z_{it} is the subset of covariates, and b_i is the random effect coefficients for each subject.

The conditional variance of Y_{it} given the random effects vector b_i is

$$\text{Var}(Y_{it} \mid b_i) = v\{E(Y_{it} \mid b_i)\}\phi, \tag{10}$$

where $v(\cdot)$ is the variance function, $E(Y_{it} \mid b_i)$ is the conditional mean function. Furthermore, for given the random effects b_i, Y_{it}s are assumed to be independent.

The specifications for the binary response case:

- Y_{it}, the response vector with a Bernoulli distribution given the random effects variance

$$\mathrm{Var}(Y_{it} \mid b_i) = E(Y_{it} \mid b_i)\{1 - E(Y_{it} \mid b_i)\} \tag{11}$$

- The conditional mean is defined as

$$\eta_{it} = X'_{it}\beta + Z'_{it}b_i = X'_{it}\beta + b_i, \tag{12}$$

when $Z_{it} = 1$ for all $i = 1, \dots, N$, and $t = 1, \dots, n_i$, with

$$\log\left\{\frac{\Pr(Y_{it} = 1 \mid b_i)}{\Pr(Y_{it} = 0 \mid b_i)}\right\} = \eta_{it} = X'_{it}\beta + b_i. \tag{13}$$

- A single random effect b_i is distributed normally as $b_i \sim_{iid} \mathrm{N}(0, G)$ (Fitzmaurice et al.,2012).

3.3.2.2 Estimation Techniques

Unlike the marginal models which do not completely determine the joint distribution of the response vector through the specification of marginal means, variances, and pairwise associations, GLMMs fully specify the joint distributions of both the response vector and the random effects vector, allowing estimation and inference based on the likelihood function.

The overview of the maximum likelihood (ML) estimation process for the fixed effects (β), the covariance parameters of random effects (G), and the prediction of random effects is explained by Fitzmaurice et al (2012) as a three-part specification.

$$f(Y_i, b_i) = f(Y_i \mid b_i)f(b_i), \tag{14}$$

is the joint probability of response and random effects where

$$f(Y_i \mid b_i) = f(Y_{i1} \mid b_i) f(Y_{i2} \mid b_i) \cdots f\left(Y_{in_i} \mid b_i\right) \tag{15}$$

under the assumption of "conditional independence". Also, it is assumed that $f\left(Y_{ij} \mid b_i\right)$ has an exponential family distribution and b_i has a multivariate normal distribution, $b_i \sim_{iid} \mathrm{N}(0, G)$.

The inference about β and G is related to the likelihood function (16). The random effects b_i are unobserved and the likelihood function is derived by performing integration over or averaging across the distribution of the unobserved random effects.

$$L(\beta, \phi, G) = \prod_{i=1}^{N} \int f(Y_i \mid b_i) f(b_i) db_i, \tag{16}$$

The marginal likelihood incorporates an integral, which signifies the process of averaging across the distribution of the random effects and by that, the resulting function becomes independently solely on β, ϕ, G. In other words, the reliance of the marginal likelihood is on the covariance of b_i while excluding consideration of the unobserved b_i.

The ML estimates for β, ϕ, G are essentially the values of these parameters that maximize the likelihood function. For the solution, numerical integration methods are required. For instance, Gaussian quadrature approximate the integral within the marginal likelihood by a weighted sum:

$$L(\beta, \phi, G) \approx \prod_{i=1}^{N} \sum_{k=1}^{K} f(Y_i \mid b_i = v_k) w_k, \tag{17}$$

where the predefined quadrature points (the weights, w_k, and the evaluation points, v_k are selected to yield a precise numerical approximation).

The number of quadrature points, denoted as K, can be adjusted as necessary. Utilizing more quadrature points enhances the accuracy of the approximation. Nevertheless, with an increase in quadrature points, there is a simultaneous rise in computational demands, particularly when the number of random effects grows exponentially. Generally, the computational time is minimal compared to the time spent on gathering longitudinal data. Thus, Fitzmaurice et al. (2012) propose augmenting the number of quadrature points until the indications of stability in parameter estimates and standard errors emerge.

The random effect for any subject is predicted by given the estimates of

$$\hat{b}_i = E\left(b_i \mid Y_i; \hat{\beta}, \hat{\phi}, \hat{G}\right), \tag{18}$$

where the ML estimates of β, ϕ, G are given as $\hat{\beta}, \hat{\phi}, \hat{G}$.

3.3.2.3 Implementation in R

GLMM models can be constructed in R programming language by function "lme4". The function is created by Bates et al. (2014) and the last version is released in 2023.

Besides the model formula, in which the subject based random intercept and random slopes can be added by "(1 | Subject)" or "(Visit | Subject)", respectively. The family of the response is set. To improve the model performance, and decrease the computational time, optimizers can be added by "glmerControl" function. The possible optimizers are

- bobyqa, uses minqa package and derivative-free.
- Nelder_Mead, default for glmer, a.k.a. simplex method and gradient-free.
- nlminbwrap, uses nlminb package.
- nmkbw, uses Nelder Mead algorithm.
- nloptwrap, default for lmer.
- optimx.L-BFGS-B, uses optim and nlminb packages.
- nloptwrap.NLOPT_LN_NELDERMEAD,

- nloptwrap.NLOPT_LN_BOBYQA.

Out of these optimizers, bobyqa, Nelder_Mead, and nlminbwrap are built-in in the package. Also, special optimizers in "optimx" package, and special extensions of bobyqa and Nelder Mead can be used.

3.3.3 Hierarchical Generalized Linear Models (HGLM)

Hierarchical generalized linear models are an extended version of generalized linear models where the hierarchical or cluster structure of the dataset is accounted which means HGLMs can consider subject-based effects. In HGLM, random effects may have conjugate exponential family distributions such as Gaussian, Gamma, Beta, or inverse-Gamma where the random effect distributions of GLMs and GLMMs are mainly Gaussian (Rönnegård et al., 2010). Also, a dispersion model can be added to the models which estimates the dispersion parameter as a function of the covariates and considers the variance of the response among dependencies between subjects and captures heteroscedasticity in the dataset. By considering the dispersion, even if a dispersion model is not specified, the model gives more accurate and efficient parameter estimates.

Consider y as the response and u as the unobserved random effects. The hierarchical model is

$$y \mid u \sim f_m(\mu, \phi), \qquad u \sim f_d(\psi, \lambda) \tag{19}$$

where f_m and f_d represent specified distributions for the mean and dispersion parts of the model. The dispersion term ϕ can be linked to a linear predictor $X_d\beta_d$ given a link function $g_d(.)$ with $g_d(\phi) = X_d\beta_d$. Details of the algorithm can be found in Lee et al. (2006).

3.3.3.1 Implementation in R

The main parameters in the model are assigned as a formula for the fixed effects, a formula for the random effects, the dataset, the family of the response, and the family of the random effects. Also, the "disp" parameter can add a dispersion model. The mean and dispersion model summaries are given in the output. The mean model is the main part of the model and includes the summary of the fixed effect estimates and random effects estimates. The dispersion model, if assigned, gives the model estimates of the dispersion term, and dispersion parameter for the mean model and the random effects. Plotting the model gives the plots of the residuals in both mean and dispersion models, deviances, and hat values of the model.

Function	Description	Reference
`Beta`	Extended usage of the Beta family	
`CAR`	Conditional autoregressive family	
`GLM.MME`	Internal IWLS estimation for `hglm()`	Lee et al. (2006)
`hglm`	Fitting hierarchical generalized linear models	Lee and Nelder (1996)
`hglm2`	`hglm` model specified in `lmer` single equation style	
`inverse.gamma`	Extended usage of the inverse-Gamma family	
`SAR`	Simultaneous autoregressive family	
- Utilities -		
`logLik`	Print log-likelihoods while `calc.like = TRUE`	
`lrt`	Print likelihood-ratio test while `calc.like = TRUE`	
`plot`	Plot individual deviances and hatvalues for the fitted `hglm` objects	
`print`	Produce basic statistics from `hglm` estimation in a simplified way	
`summary`	Produce standard summary statistics for the fitted `hglm` objects	

Figure 3.3 Functions in hglm package (Rönnegård et al, 2010)

Model name	$y\|u$ distribution	Link $g(\mu)$	u distribution	Link $v(u)$
Linear mixed model	Gaussian	identity	Gaussian	identity
Binomial conjugate	Binomial	logit	Beta	logit
Binomial GLMM	Binomial	logit	Gaussian	identity
Binomial frailty	Binomial	comp-log-log	Gamma	log
Poisson GLMM	Poisson	log	Gaussian	identity
Poisson conjugate	Poisson	log	Gamma	log
Gamma GLMM	Gamma	log	Gaussian	identity
Gamma conjugate	Gamma	inverse	Inverse-Gamma	inverse
Gamma-Gamma	Gamma	log	Gamma	log

Figure 3.4 Distributions and link functions applicable with hglm()(Rönnegård et al, 2010)

Model name	hglm() code: family =	hglm() code: rand.family =
Linear mixed model	gaussian(link = identity)	gaussian(link = identity)
Beta-Binomial	binomial(link = logit)	Beta(link = logit)
Binomial GLMM	binomial(link = logit)	gaussian(link = identity)
Binomial frailty	binomial(link = cloglog)	Gamma(link = log)
Poisson GLMM	poisson(link = log)	gaussian(link = identity)
Poisson frailty	poisson(link = log)	Gamma(link = log)
Gamma GLMM	Gamma(link = log)	gaussian(link = identity)
Gamma conjugate	Gamma(link = inverse)	inverse.gamma(link = inverse)
Gamma-Gamma	Gamma(link = log)	Gamma(link = log)

Figure 3.5 hglm() codes for models (Rönnegård et al, 2010)

The functions and their descriptions are given in Figure 3.4 The "hglm" function is used to build the model and utilities are the functions to obtain information from model output.

In Figure 3.5 and Figure 3.6, the distributions, corresponding link functions, and codes are given. Since hierarchical generalized linear models can deal with different distributions of the response and random effects, these distributions can be implemented to the models by using appropriate family names, random family names, and link functions.

3.3.4 Gaussian Process Boosting (GPBoost)

Boosting is a ML algorithm introduced by Freud and Schapire in 1996 which is recognized for frequently attaining superior predictive accuracy. It is the most potent ready-to-use nonlinear learning algorithm for various application challenges (Johnson and Zhang, 2013). It also works well in feature selection. In boosting and many other supervised ML methods, a function connects predictor variables to an outcome variable. However, any residual correlation not captured by the predictor function is usually overlooked. This approach is often impractical with spatial and spatiotemporal data.

Gaussian processes, proposed by Williams and Rasmussen (2006), stand as versatile function models devoid of strict parametric constraints, leading to cutting-edge

predictive precision and enabling the formulation of probabilistic forecasts. They are used in diverse domains such as non-parametric regression, modeling time series, spatial, and spatiotemporal data, replicating extensive computer experiments, refining costly black-box function optimization, and fine-tuning parameters within machine learning models. Moreover, mixed-effect models involving grouped or clustered random effects have gained extensive traction across various scientific fields. These models are precious for addressing data characterized by a structured grouping pattern.

In Gaussian process and mixed effects models, the prior mean is usually assumed to be zero or a linear function of predictor variables. Then, the structured variance that remains unaccounted for is modeled through a zero-mean Gaussian process and/or a grouped random effects model. However, in some cases, these assumptions may not be valid, and relaxing them could improve prediction accuracy (Fuglstad et al., 2015; Schmidt and Guttorp, 2020).

The Gaussian process boosting algorithm, developed by Sigrist in 2022, is an algorithm that combines tree boosting, and the Gaussian process to be a solution for the limitations of boosting and make use of the advantages of the algorithm itself. In the algorithm, the Gaussian process stands for applicability, incorporating the complex properties of data into the model, and boosting is a powerful machine learning algorithm that creates ensemble decision trees and improves the model's performance. The algorithm relaxes the assumptions of linearity or zero prior mean in the Gaussian process and mixed effects models, and the independence assumption in boosting. The method employs a mixed effects model where the predictor function is a non-parametric function of predictor variables. The random effects involve combinations of Gaussian processes and grouped random effects. The non-linear predictor function is modeled using an ensemble of base learners, such as regression trees, learned through boosting. The covariance structure parameters of the random effects are jointly estimated with the predictor function (Sigrist, 2022).

3.3.4.1 Combining Gaussian Process and Mixed Effects Models with Boosting

The combination of the Gaussian process and mixed-effects models with the boosting algorithm is briefly explained in the thesis study of Çakar (2022). As the mixed-effects model given shows the relationship between fixed effects and the response

$$\mathrm{Y_{it}} = X_{it}\beta + Z_{it}b_i + \varepsilon_{it}, \tag{20}$$

$$\varepsilon \sim N(0, \sigma^2 I_n) \tag{21}$$

where $Y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ denotes the response variable, X_{it} is the covariate denotation and β is the vector for estimates of the fixed effects, $b \in \mathbb{R}^m$ stands for the random effects with covariance matrix $\Sigma \in \mathbb{R}^{m \times m}$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top \in \mathbb{R}^n$ denotes the error term. The number of observations is given by n, the random effects' dimension is denoted by m, the number of fixed effects (the number of parameters) is denoted by p, and the group-level random effects to observations is given as Z matrix.

The likelihood function is optimized,

$$p(y \mid F, \theta) = \int p(y \mid b, F, \theta) p(b \mid \theta) db, \tag{22}$$

where

$$p(b \mid \theta) \quad = \exp\left(-\frac{1}{2} b^T \Sigma^{-1} b\right) |\Sigma|^{-1/2} (2\pi)^{-m/2} \tag{23}$$

$$p(y \mid b, F, \theta) \quad = \exp\left(-\frac{1}{2\sigma^2} (y - F - Zb)^T (y - F - Zb)\right) (2\pi\sigma^2)^{-n/2}. \tag{24}$$

Here, $F = F(X)$ and $\theta \in \Theta \subset \mathbb{R}^q$ stands for all variance and covariance parameters. The function " $F(\cdot)$ " denotes the function F evaluated at X. Also, it is assumed that $\theta_1 = \sigma^2$.

By using GPBoost model, the joint minimizer can be obtained combining the Gaussian process and mixed models. The details of the algorithm can be found in the main article of the algorithm written by Sigrist (2022).

3.3.4.2 Implementation in R

In the "gpboost" package in R programming language, which was developed by Sigrist in 2022, there are two main models. One is for the Gaussian process implementing the mixed effects model and helping to see the significance of covariates, and the other one is for boosting, which only has the importance of features, not estimates. However, the parameters of the first model can be assigned to the second model.

The "fitGPModel" function is used to fit the GP model, and the covariates and response are assigned as different matrices. Also, the grouping variable is added to implement the random effects structure to the model, and the family of likelihood is set. To obtain a computationally fast model with higher performance, it is assigned additional parameters:

- maxit, used to set the maximum number of iterations for the optimization algorithm. The default value is 1000.
- optimizer_coef, optimizers used for estimating the regression coefficients. The options are:
    - gradient_descent, used with gradient descent covariance optimizer and appropriate for likelihoods other than Gaussian.
    - wls, used with gradient descent covariance optimizer and appropriate for Gaussian data.
    - nelder_mead, used with nelder_mead covariance optimizer.

    - bfgs, used with bfgs covariance optimizer.
    - adam, used with adam covariance optimizer.
- delta_rel_conv, used to set the convergence tolerance and makes the algorithm stop when the relative change is below this value. The default is 0.000001.
- optimizer_cov, the optimizers used for estimating the covariance parameters. The options are:
    - Gradient_descent, the default optimizer
    - Fisher_scoring,
    - Nelder_mead
    - Bfgs
    - adam

These parameters can be assigned in the boosting model as well. To fit the boosting model, firstly the optimizer parameters and parameters for grid search are set. The required hyperparameters for the grid search to be tuned for obtaining the optimal boosting algorithm are

- learning_rate, important for boosting and the lower values are used to improve the model performance but in that case the iteration number should be high. The default is 0.1.
- min_data_in_leaf, used to set the minimum number of samples per leaf. The default value is 20.
- max_depth, used as the maximum depth of a tree and it has no limit.
- num_leaves, used to assign the number of leaves in a tree. The default is 31.
- max_bin, used to set the maximum number of bins to aggregate feature values. The default is 255.

The hyperparameters given are the ones used in this thesis study. Detailed information can be found in the main article of the algorithm written by Sigrist (2022).

To tune these hyperparameters, “gpb.grid.search.tune.parameters” function is used. The dataset is converted to a gp dataset format by “gpb.Dataset” function. Using the tuning result, best hyperparameters are obtained and the boosting model is fitted by “gpb.train” function (Sigrist, 2022).

Finally, variable importance results are obtained in cover, gain, and frequency. The cover values show the proportion of observations associated with a feature, indicating its significance. On the other hand, frequency provides a more straightforward evaluation of gain which simply tallies how often the feature is utilized across all constructed trees (Chen et al, 2018).

3.3.5 Generalized Linear Mixed Model Trees (GLMMTree)

Fokkema & Zeileis (2019) proposed generalized linear mixed-effects model trees to detect treatment subgroup interactions in clustered datasets. GLMM can interpret the factors affecting the response but cannot directly indicate what needs to be done in the decision-making process. In cases where a definitive conclusion cannot be reached, recursive partitioning defines the relationship between the response and covariates through binary decision trees rather than a mathematical model formula. Due to its ease of implementation and interpretation, it is widely used in clinical studies and psychotherapy. Decision trees have a non-parametric nature. They do not require any assumptions, such as linear relationships or the distribution of residuals. Also, they allow for identifying numerous potential predictor variables, surpassing the number of observations.

The GLMMTree algorithm is a specialized form of GLMTree, the model-based recursive partitioning algorithm proposed by Zeileis et al. in 2008. It consists of nodes that contain an intercept and subgroup-specific GLMs, which include the effects of predictive variables. Subgroups are defined by segmentation variables that are not considered model predictors. The GLMM tree algorithm is an updated version of the GLM tree. It considers the dependence between observations caused

by the repeated measurement structure of longitudinal datasets and accounts for the random effects (Fokkema et al., 2018).

3.3.5.1 Model-based recursive partitioning

The GLMM tree algorithm is a flexible framework for subgroup detection that considers the clustered structure of datasets, building on model-based recursive partitioning (MOB). The idea behind MOB is partitioning the data based on the additional covariates and finding better fits in each partition when a single global parametric model may not be well-fitted for the data.

The MOB algorithm checks parameter stability using additional covariates (partitioning variables, denoted by U) by an iterative cycle:

1. A parametric model is fitted.
2. The parametric instability of partitioning variables is statistically tested. A test statistic is calculated for each potential partitioning variable to quantify parameter instability. A p-value is then calculated for each partitioning variable based on the known distribution of these test statistics under the null hypothesis of parameter stability.
3. When instability is observed, the dataset is split concerning the variable with the highest instability. If any partitioning variable has a p-value below α, step (3) creates two subsets using Uk^*, the variable with the minimal p-value in step (2). The loss function is minimized separately in each subgroup, and the split point that yields the minimum sum of loss functions is selected.
4. The steps are repeated in each subgroup until the null hypothesis of parameter stability cannot be rejected, or the subsets are too small.

The partition in the MOB algorithm is the decision tree. If GLM is used, then it is a GLMTree with fixed effects regression model in each j-th terminal node:

$$g(\mu_{it}) = X_i^{\intercal}\beta_j \tag{25}$$

3.3.5.2 Including random effects

For the longitudinal datasets with clustered observations, GLMM is used, and this mixed-effects structure is added to the decision tree model. The equation of GLM is updated as (26) by adding random effects by $Z_i^\intercal b$.

$$g(\mu_{it}) = X_i^\intercal \beta + Z_i^\intercal b \tag{26}$$

Therefore, the GLMMTree algorithm has global and local parts in which random effects terms, all observations, and locally estimated fixed effects terms are included to be well-fit for the dataset.

After adding the random effects part, the MOB algorithm creates a decision tree based on the GLMM model in each j-th terminal node:

$$g(\mu_{it}) = X_i^\intercal \beta_j + Z_i^\intercal b \tag{27}$$

It is possible to identify variations in intercepts between terminal nodes and detect variations in slopes, such as treatment impacts, by using GLMMTree to approximate the fixed effects part of the GLMM. Also, it is possible to use GLMMTree for all binary, count, or continuous responses.

The steps of the GLMMTree algorithm:

0. Setting r and all values $\hat{b}_{(r)}$ to 0.
1. Increasing r by one and estimating a GLMTree by using an offset, $Z_i^\intercal \hat{b}_{(r-1)}$
2. Building the mixed-effects model $g(\mu_{it}) = X_i^\intercal \beta_j + Z_i^\intercal b$ by assigning the terminal node $j(r)$ from the estimated GLMTree. Extract posterior predictions $\hat{b}_{(r)}$ from the estimated model.
3. Applying the first and second steps until convergence.

The algorithm starts with b set to 0 and re-estimates the GLM tree in the first step and the fixed and random effects parameters in the second step at each iteration. Random effects are estimated globally, and fixed effects are estimated locally in

partition cells. The algorithm converges until the log-likelihood criterion of the mixed effects model does not change between iterations.

The GLMM tree algorithm considers a random intercept for cluster indicator, treatment indicator, treatment outcome, and potential partitioning variables (Fokkema et al., 2021).

3.3.5.3 Implementing in R

The GLMMTree models can be constructed in R programming language using the "glmmtree" package created by Fokkema and Zeileis. The package uses "partykit" package (Hothorn & Zeileis, 2015) to find the partition and the "lme4" package (Bates et al.,2015) to estimate the mixed-effects model. The "glmmtree" package consists of modeling functions for continuous and binary responses. The classification structure is set by entering the response, covariates, id, random effects part, and family of the response. Then the tree plots, random effects, and model predictions can be obtained (Fokkema & Zeileis, 2019).

The important hyperparameters of the model are

- minsize, the smallest size for the node and generally taken as 0.10*nrow(data).
- maxdepth, the number of nodes in the longest path from the root node to the farthest leaf node.
- maxit, the maximum number of iterations used in estimation the tree.
- abstol, the convergence criterion and used for model estimation. If the log-likelihood difference of model's random effects from two iterations is smaller than that value, the model converges.

Tuning the hyperparameters enhances the performance of the model.

3.3.6 Historical Random Forests (HRF)

The algorithm Historical Random Forest was proposed by Sexton and Laake in 2018 as a historical observations-added version of random forests. It considers the covariates' history, which makes the algorithm applicable to longitudinal data.

3.3.6.1 The random forest algorithm

Random Forest (RF) is an ensemble learning method for classification, regression, and other tasks that operate by constructing a multitude of decision trees. It is a decision tree ensemble where each tree is built from a bootstrapped sample of the training dataset. To split a node, the algorithm aims to minimize the impurity of response values, measured by the Gini index for categorical variables or by variance for continuous ones. The final estimate for the observation in a Random Forest is based on majority voting or averaging the results of all decision trees in the forest. The final estimate for the observation in a Random Forest is based on majority voting or averaging the results of all decision trees in the forest. However, the RF algorithm can suffer from issues when directly used for longitudinal data analysis.

For clustered data, subjects have several observations, and these correlated observations can be included in bootstrap samples from different trees, correlated or homogeneous trees can be obtained, which leads to worse prediction performances. Because of the high similarity between observations of the same subject, the estimated prediction errors are very low, and it causes biased results. Therefore, some extensions of the standard random forest algorithm should be developed to be applicable in longitudinal studies (Hu and Szymczak, 2023).

3.3.6.2 Historical RF

The extensions of random forest for clustered data, averaging, and subject-level bootstrapping are developed. Averaging takes the averages of repeated measurements in each subject and applies random forest to them, but decreasing the number of observations badly affects the model prediction performance. Therefore, Karpievity et al. (2009) developed subject-level bootstrapping, which includes all observations from the subject in a bootstrap sample. Then in 2011, Adler et al, developed a two-stage bootstrapping, which selects a subject and takes its observations, and then for each selected subject, training samples are selected randomly from all observations. Both methods have performed better than the standard RF but they ignore the random effects obtained from the subject and time effect (Hajjem et al., 2014). Therefore, the time effect is considered, and the historical random forest algorithm was developed by Saxton and Laake in 2018.

In the training data, $\{Y_{it}, t_{it}, X_{it}\}$, $i = 1,2,3, \dots N$ denotes the number of subjects and $t = 1,2,3, \dots, n_i$, denotes the number of observations in each subject where Y_{it}, t_{it}, X_{it} stand for response, time and covariates respectively. The response Y_{it} is modeled using both time and covariates of the i-th subject in all time points. The algorithm consists of time–varying covariates part and concurrent covariates part. For concurrent covariates, a standard random forest model is used by considering its current values when building each tree. However, current values are also considered for time-varying (historical) covariates besides historical information. In time-varying covariates, the historical information is summarized by a function. The function for the i-th subject at t-th time counts past observations of response and covariates measures within the maximum of η_1 units of time before the t-th time point and smaller than η_2. The summary function:

$$S_C(\eta; \bar{H}_{ita}) = \sum_{t_{ia} \in [t_{ia} - \eta_1, t_{ia})} I(H_{ila} < \eta_2), \tag{28}$$

where $\bar{H}_{it} = \{H_{il} = (y_{il}, x_{il}) : t_{il} < t_{it}\}$ is the past observations of the i-th subject at time t and $\bar{H}_{ita}$ is the a-th component. This summation results in a single number for each observation and variable for a fixed value of $\eta = (\eta_1, \eta_2)$. Each summary function also has another version where an upper bound limits the time interval, $t_{il} \in [t_{it} - \eta_1, t_{it} - \eta_3)$. Partitioning at a node is applied by using the covariates with the smallest Gini-impurity for categorical responses. For time-varying covariates, the optimal cut-off point determination includes optimizing the summary function parameters, significantly increasing computing expenses for high dimensional datasets. Therefore, Sexton and Laake added an extra level of randomization to reduce the effects of parameter optimization in the summary function. They used a subsample of observations from the specified time interval to optimize the breakpoint and adopted a subject-level resampling strategy in RF construction. Besides being the solution in cases where the previous methods are insufficient, this algorithm keeps the entire observation history of a subject (Hu and Szymczak, 2023).

3.3.6.3 Implementing in R

In 2017, Sexton created the "htree" package for historical tree ensembles for the R programming language, and in 2018, he updated the package. The package includes historical random forest and tree boosting for longitudinal data and these models' prediction, partial dependence, quantile, and variable importance functions. The models are created using covariates, response, time, and id. The algorithm has hyperparameters that will improve the model's performance by being tuned.

- mtry, the number of covariates in each split.
- nsplit, the number of splits in trees.
- nsample, the number of delta values, which are minimized in the determination of the split, in the sample.
- ntree, the number of ensembled trees.

After creating the random forests, obtaining variable importance is essential and the algorithm uses Z-score to reveal the variable importance of the covariates (Sexton, 2017).

3.3.7 Generalized Linear Mixed Models combined with Least Absolute Shrinkage and Selection Operator (GLMMLasso)

A combined model called GLMM-Lasso is proposed by Schelldorfer, Meier and Bühlmann (2014) to handle independence and multicollinearity issues by using both GLMM and Lasso penalty ($L1$) on fixed effects. The GLMM-Lasso model reduces complexity for high dimensional data with many predictors/independent variables and it is used for variable selection and shrinkage (Hayati and Muslim, 2021).

3.3.7.1 Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO was proposed by Tibshirani (1996) as an approach for variable selection by using penalization. It uses an L1 penalty on regression model coefficients and it shrinks the coefficients to zero and some directly to zero according to their weights on predicting the response. The goal is to maximize the likelihood of the model while limiting the L1-norm of the parameter vector β. The Lasso estimate,

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}}\, l(\beta)\,, \qquad \|\beta\|_1 \leq s, \qquad s \geq 0 \tag{29}$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}}[l(\beta) - \lambda\,\|\beta\|_1]\,, \qquad \lambda \geq 0 \tag{30}$$

The denotion $\|.\|$ stands for the L_1 – norm and λ and s are the parameters to be tuned to obtain the optimal estimate.

3.3.7.2 The GLMM Formulation

Let the index of the groping variables be denoted as $i = 1,2, \dots N$ and $t = 1,2, \dots n_i$ denotes the t-th response in i-th group and lastly $n = \sum_{i=1}^{N} \mathrm{n_i}$ be the total number of observations.

Let X, Z, Y, and b be the fixed-effects, $N \times q$ random-effects design matrices, N dimensional random response vector, and q-dimensional vector of random effects, respectively. y of Y is observed but b is unobserved and the specification of the GLMM involving the unconditional distribution of b and the conditional distribution of $Y \mid b$:

1. $Y_i \mid b$ are independent for $i = 1, \dots, N$
2. $Y_i \mid b$ has an exponential family distribution with density

$$\exp\{\phi^{-1}(y_i\xi_i - b(\xi_i)) + c(y_i, \phi)\}, \tag{31}$$

where $b(.)$ and $c(.,.)$ are known functions, ϕ is the dispersion parameter and ξ_i is related with the conditional mean $\mu_i := E[Y_i \mid b]$, that is, $\xi_i = \xi_i(\mu_i)$.

3. μ is the conditional mean vector depending on b through the established link function g and the linear predictor $\eta = X\beta + Zb$, where $\eta = g(\mu)$ on a component wise basis. In this context, β is the unknown p-dimensional parameter vector, known as fixed effects, and b is the unknown q-dimensional vector of random effects.
4. The random vector b follows a multivariate normal distribution, $b \sim N_q(0, \Sigma_\theta)$, where the covariance matrix Σ_θ is defined by the unknown parameter vector $\theta \in \mathbb{R}^d$. It is assumed that Σ_θ is positive semidefinite, meaning that, $\Sigma_\theta \geq 0$. The value of d which shows dimensionality is usually small, $d \leq 10$.

The Σ_θ can be written as $\Sigma_\theta = \Lambda_\theta \Lambda_\theta^T$ in terms of its Cholesky decomposition, and the unobserved random variable $\mathcal{U}$ can be identified as $b := \Lambda_\theta \mathcal{U}$ where $\mathcal{U} \sim N_q(0, 1_q)$. Then the linear predictor η is $\eta = X\beta + Z\Lambda_\theta u$. The parameters β, θ, and ϕ are estimated using the ML method and the random effects u are predicted.

3.3.7.3 The Likelihood Function

By using the notation of the density function, the likelihood function of GLMM is

$$
\begin{aligned}
L(\beta, \theta, \phi) = & \int_{\mathbb{R}^+} \prod_{i=1}^{N} \left[\exp\{\phi^{-1}(y_i \xi_i(\beta, \theta) - b(\xi_i(\beta, \theta))) + c(y_i, \phi)\}\right] \\
& \times \frac{1}{(2\pi)^{q/2}} \exp\left\{-\frac{1}{2} \| u \|_2^2\right\} du \\
= & \frac{1}{(2\pi)^{q/2}} \int_{\mathbb{R}^q} \exp\left\{\sum_{i=1}^{N} \left(\frac{y_i \xi_i(\beta, \theta) - b(\xi_i(\beta, \theta))}{\phi} + c(y_i, \phi)\right)\right. \\
& \left. - \frac{1}{2} \| u \|_2^2\right\} du
\end{aligned}
\tag{32}
$$

The integral in (32) cannot be solved directly and numerical approximations are required (Jiang, 2007; Morlenbergs and Verbeke, 2005).

3.3.7.4 The GLMMLasso Estimator

In the high dimensional datasets, the number of covariates p is higher than the number of observations N, and the β_0 values (actual underlying fixed-effects vector) of many covariates are not zero, which makes those covariates influential in modeling the response. The Lasso type approach is applied to make the β_0 values of unimportant coefficients zero and select the critical features, and L_1-penalty for the fixed-effects vector β is added to the likelihood function. The objective function in (33) is considered

$$Q_\lambda(\beta, \theta, \phi) = -2\log L(\beta, \theta, \phi) + \lambda \parallel \beta \parallel_1, \quad (33)$$

where $\lambda \geq 0$ is a regularization parameter.

In order to estimate β (fixed-effect parameter), θ (covariance parameter), and ϕ (dispersion parameter) by the formula of GLMMLasso estimator

$$(\hat{\beta}, \hat{\theta}, \hat{\phi}) := \underset{\beta,\theta,\phi}{\arg\min}\, Q_\lambda(\beta, \theta, \phi). \quad (34)$$

3.3.7.5 Implementation in R

In R programming language, "glmmLasso" package is used to apply the algorithm which is developed by Groll and Tutz in 2014 to be used in variable selection for GLMMs by using ℓ_1 penalty. The package includes datasets a modeling function and a control function. In the modeling function, both fix and random parts of the model are assigned and the family is stated. Lambda value is required to be given and the optimal lambda value is obtained by tuning. Generally, loops are used by trying several lambda values in the models and finding the optimal one by checking the model performances. The hyperparameters that can change and make the model performance better are:

- nue, the weakness of the learner and generally takes value between o and 1
- steps, the number of iterations
- epsilon, convergence speed controller. The default is 0.0001.

3.4 Model Evaluation

3.4.1 Information Criterions

3.4.1.1 AIC and BIC

Akaike's information criterion (AIC) and Bayesian information criterion (BIC) are commonly used to evaluate model performances in longitudinal studies. Both provide a way of choosing and selecting the best models. The basic idea is to evaluate the model's trade-off between fit and simplicity. AIC and BIC evaluate the model based on its complexity and log-likelihood (Kuha, 2004).

AIC is used to evaluate the predictive ability and fit of models. Lower AIC values indicate better model performance. AIC considers the model's quality and complexity (such as the number of parameters). When the AIC value is calculated, less complex models and more data may be appropriate.

BIC is also used for model selection and, like AIC, evaluates the trade-off between fit and complexity. However, BIC is more model selective and gives more weight to more data. Therefore, there may be cases where BIC evaluates more complex models unfavorably than AIC.

The difference between the two criteria depends on which factors are emphasized more in model selection. AIC emphasizes predictive ability and the model fit more, whereas BIC favors more data and lower dimensional models (with fewer parameters).

Both AIC and BIC are tools to guide model selection but may be preferred depending on the context and objective for which they are used. When choosing a good model, it is important to consider both AIC and BIC values and evaluate your model against these criteria (Kuha, 2004).

3.4.1.2 QIC and CIC

In GEE models, AIC and BIC cannot be used since they are calculated by using likelihood estimate. However, GEE has its own estimation technique. Therefore, Quasi-likelihood is developed to interpret the results of the GEE model (Gosho et al., 2011).

- QIC (Quasi-likelihood under the independence model criterion) is used to evaluate the model's compatibility. The lower the QIC value, the better the model fits the dataset. QIC value is a function of likelihood probability and the number of covariates.
- QICu (Unbiased Quasi-likelihood under the independence model criterion) is similar to QIC but unbiased. It differs from QIC in selecting a reliable model for small sample sizes.
- QIC likelihood is a different type of QIC and is used in model selection by considering the compatibility and complexity of the model.
- CIC (Consistency Information Criterion) is important to balance model fit and complexity. A lower CIC value indicates better data fit and fewer overfitting issues.
- QICC (Corrected Quasi-likelihood under the independence model criterion) is an updated and regularized version of QIC used to deal with the overfitting problem.

Both QIC and CIC are used to select the working correlation and the wellness of the model fit in GEE models (Ballinger, 2004).

3.4.2 Model performance metrics

Evaluating the model performance is important to select the best-estimated model that fits the data well. Confusion matrix is a tool used to evaluate the performance of a classification model. A 2x2 confusion matrix (given in Table 3.1) with actual class and predicted class values, is used to calculate the model performance metrics in

binary classification cases. It shows how correctly or incorrectly the model classifies different classes. To obtain two classes, the estimated probabilities are calculated for each model and predictions are decided by setting a threshold value. If the probability is smaller than that value, it is classified as zero; if bigger, it is one. In Table 3.1, four different values are given: true positive (TP), false positive (FP), true negative (TN) and false negative (FN).

Table 3.1 Confusion matrix for model performance

		Actual class	
		Positive =1	Negative =0
Predicted class	True =1	True Positive (TP)	False Positive (FP)
	False =0	False Negative (FN)	True Negative (TN)

True Positive (TP): Represents cases where samples that are actually positive are correctly predicted to be positive.

False Positive (FP): Represents cases where samples that are actually negative are incorrectly predicted as positive.

True Negative (TN): Represents cases where samples that are negative are correctly predicted as negative.

False Negative (FN): Represents cases where samples that are positive are incorrectly predicted as negative (Vujović, 2021).

Niels Bohr says that "The opposite of a correct statement is a false statement. But the opposite of a profound truth may well be another profound truth. " Therefore, finding all cells of the contingency table in Table .. may give a significant result.

The main performance metrics are:

- Accuracy
- Sensitivity
- Specificity
- F1 Score

Accuracy represents the ratio of correct predictions of a classification model to the total data points.

$$Accuracy = \frac{TN + TP}{TN + TP + FP + FN} \tag{35}$$

It can be a useful performance measure in some cases, but there are some factors that should be considered. Accuracy may be inadequate if the dataset has a class imbalance, i.e., the number of data points from different classes is very different. In this case, other performance measures (e.g., F1 score, sensitivity, specificity, etc.) can provide a complete picture (Buyrukoğlu, 2020). Furthermore, accuracy may be insufficient for some special cases, such as medical diagnostics, false positives or false negatives can have serious consequences for living beings. Therefore, it is important to consider different performance metrics depending on the requirements and characteristics of the analysis.

Sensitivity and specificity are two important metrics used to evaluate the performance of medical tests or anomaly detection systems. They provide insight into how well a model performs for different aspects of classification.

Sensitivity (a.k.a. True Positive Rate or Recall) measures the ability of the model to accurately identify positive instances from among all true positive instances. It focuses on minimizing false negatives when positive samples are mistakenly predicted as negatives. It mainly answers the question: “Out of all positive cases, how many of them did the model correctly detect?”

$$Sensitivity = \frac{TP}{TP + FN} \tag{36}$$

Specificity measures the ability of the model to accurately identify negative samples among all true negative samples. It focuses on minimizing false negatives when negative samples are falsely detected as positives. It answers the question: "How many of all true negative cases did the model identify as negative?"

$$Specificity = \frac{TN}{TN + FP} \tag{37}$$

The F1 score is particularly important for measuring performance on imbalanced datasets (when one class is much more numerous than another). The F1 score provides a balanced assessment of the model's predictions, considering both precision and sensitivity and a balanced view of how the model predicts classes.

$$Precision = \frac{TP}{TP + FP} \tag{38}$$

$$F1 - Score = \frac{2 * Precision * Sensitivity}{Precision + Sensitivity} \tag{39}$$

CHAPTER 4

DESIGN OF THE STUDY AND DATASET INFORMATION

The propose of this study is to compare the performances of classical and machine learning approaches in the classification of Alzheimer's disease dataset. In this section, the practical application of the research, and the simulation study to see the effect of sample size are demonstrated.

4.1 Design of the Study

In longitudinal studies, although there are limited methods because of the subject-specific and repeated structure of the observations, they have been evolving over time. The classical methods like generalized estimating equations and random effects models are not only the most popular and known as commonly used techniques but also many applications that combine these methods in a more complex manner have emerged. Moreover, various machine learning methods are being continuously developed to be available for longitudinal data structures. Recently, shrinkage methods have also been adapted to be applicable to longitudinal data structures.

The aim of this study is to explore and compare the proper methods for analyzing a longitudinal data with a binary response, small sample size and irregular time points. The objective is to assess the performance of these methods and observe how the disadvantages, which may be sample size and irregularity of the time points, of this dataset affect the performance metrics of the models. Therefore, the study is divided into two main parts as the analysis with real dataset and simulation. In the first part, methods are compared using a real dataset called "Longitudinal MRI Data in Nondemented and Demented Older Adults" while in the second part a simulation study is designed to investigate the impact of the dataset on the performance metrics of the models.

4.2 Data Description and Pre-processing

4.2.1 Data Description

The Open Access Series of Imaging Studies (OASIS) is a project with the primary goal of providing the scientific community with free access to neuroimaging data sets of the brain. In the initial study, the OASIS project included a cross-sectional MRI data set comprising over 400 individuals across the adult lifespan, including both demented and nondemented individuals (Marcus et al., 2007). Building on this foundation, a longitudinal sample of MRI data specifically focusing on older adults, with and without Alzheimer's disease is introduced (Marcus et al., 2009).

The dataset called “Longitudinal MRI Data in Nondemented and Demented Older Adults” is the second dataset of the OASIS study based on the MRI data on the dementia status of older adults and it consists of demographics of the patients, clinical results, and derived anatomic volumes.

The dataset includes socio-demographic variables such as gender, education, age and socioeconomic status, and clinical predictor variables which are measured as a result of clinical tests, and some of them are directly related to the brain images since the brain size is an important factor in the detection of dementia (Fulton, et.al.,2019).

Table 4.1 gives the detailed information about the descriptions, classes and definitions about the dataset and Table 4.2 gives the summary statistics of the categorical variables. There are 373 longitudinal MRI data collected from 150 subjects each measured 5 times at most. The group variable stands for the dementia status categorized as demented converted and nondemented and each category is determined according to the Clinical Dementia Rating (CDR) score of participants which is one of the clinical results of the study. Clinical dementia rating is a rating method used in longitudinal studies and clinical trials to determine the stages of Alzheimer’s disease by interviewing the patients in six cognitive categories as memory, orientation, judgement, problem solving, community affairs, home and

hobbies, and personal care on a scale that 0 = none, 0.5 = very mild AD, 1= mild AD, 2= moderate and 3= severe. In categorizing dementia status by using CDR, zero means no dementia and the other ones can be categorized as dementia but as very mild, mild, moderate and severe respectively.

Table 4.1 Information about the variables

Variable Name	**Variable Description**	**Class of the Variable**	**Details**
Subject ID	ID number of subjects	Numeric	150 subjects, 373 observations
MRI ID	MRI number of subjects	Numeric	5 MRI's
Group	Dementia group of subjects	Categorical	Demented Converted Nondemented
Visit	Visit number of subjects	Categorical	5 Visits
MR Delay	MR Delay of subjects	Numeric	Ranges from 0 to 2639
M/F	Gender of Subjects	Categorical	Female Male
Hand	Handedness	Categorical	Right Left
Age	Age of subjects	Numeric	Ranges from 60 to 98
EDUC	Years of education of subjects	Numeric	Ranges from 6 to 23
SES	Socioeconomic status of subjects	Categorical	1– Highest 5 – Lowest
MMSE	Mini-mental state examination scores	Numeric	0 – Worst 30 – Best
CDR	Clinical dementia rating	Categorical	0 – No Dementia 0.5 – Very mild AD 1 – Mild AD 2 – Moderate AD
eTIV	Estimated total intracranial volume	Numeric	Ranges from 1106 to 2004
nWBV	Normalized whole brain volume	Numeric	Ranges from 0.644 to 0.837
ASF	Atlas scaling factor	Numeric	Ranges from 0.876 to 1.587

According to these categories, dementia group of the participants are classed by considering the CDR scores. When the CDR score is 0 in the first and the last visit, it is recorded as nondemented for all visits even if the CDR score is other than zero in the middle visits. Further, for the ones having the CDR score 0.5, 1 or 2, it means being demented for all visits. Lastly, the ones having CDR as zero in the initial visit and converted into demented in the following visits are categorized as converted. Demented group is seen 146 times where nondemented group and converted group are seen 190 and 37 times, respectively. In the dataset related to the study, 88 females and 62 males aged between 60 and 98 are included since the dataset is for older adults. All participants are right-handed. Education levels of the participants are given yearly and ranged from 6 to 23. Socioeconomic status of the participants is ranked from 1 as highest to 5 as lowest. While 33 participants are stating that they have the highest socioeconomic status, 3 of them are stating they have the lowest. MMSE, the other clinical result, shows the mini mental status of the participants. It is a numeric variable ranged from 0 to 30 meaning worst and best, respectively. Derived anatomic volumes are measured by Estimated Total Intracranial Volume (eTIV) in the scale of mm^3, Atlas Scaling Factor (ASF) and Normalized Whole Brain Volume (nWBV) which are ranged from 1106 to 2004, 0.644 to 0.837, and 0.876 to 1.587, respectively (Marcus et al., 2009).

Table 4.2 Summary statistics of categorical variables

Variable Name	**Categories**
Group	Demented: 146, Converted: 37, Nondemented: 190
Visit	1: 150, 2: 144, 3: 58, 4: 15, 5: 6
M/F	Female: 88, Male: 62
Hand	Right: 150, Left: 0
SES	1: 33, 2: 42, 3: 34, 4: 30, 5: 3
CDR	0: 206, 0.5: 123, 1: 41, 2: 3

MRI ID and Visit variables state the number of sequential visits. Therefore, their bar plots show the same frequencies in Figure 4.1. This Figure is the bar plot showing the frequency of categorical variables by the dementia group. In all visits, nondemented participants seemed to be more than the other ones and the frequency of nondemented participants increases in the later visits. Females are more than males in terms of being nondemented, which means males are seen to be facing dementia more than females. In terms of socioeconomic status, there are more nondemented participants who have stated that their socioeconomic status is high (one, two or three) whereas there are more demented participants stating that their socioeconomic status is low (five) with only three patients. The categorization of the CDR scores of the participants proves the logic behind the relationship between CDR scores and dementia status. The ones having zero CDR are mainly nondemented and the ones having other CDR scores are mainly demented.

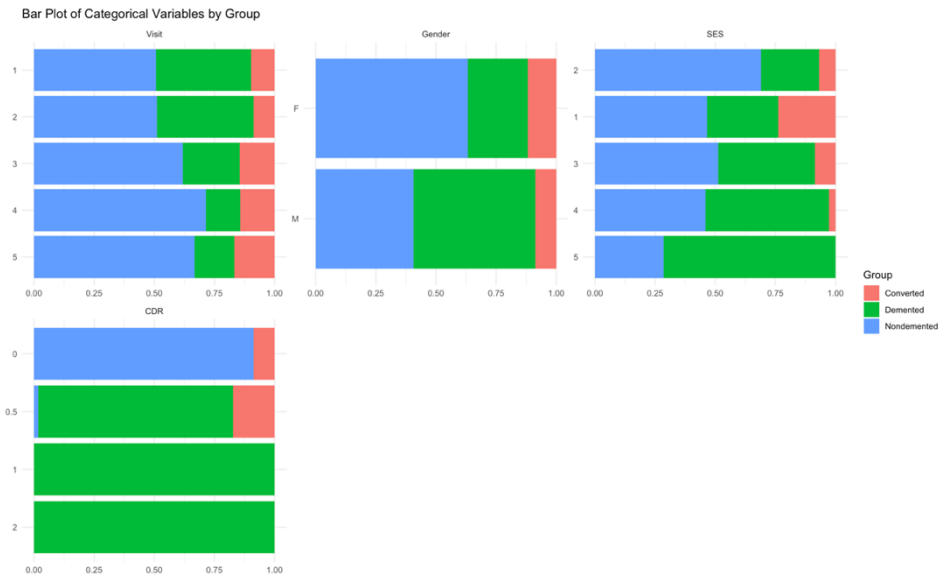

Figure 4.1 Bar plot of categorical variables by group

In Figure 4.2, box plots of numeric variables by group are seen. The median age of the nondemented ones seem slightly higher compared to the demented ones, and converted participants are older than the demented and nondemented ones. Also, the range of the age of the nondemented ones is wider. There is an outlier out of demented participants having age around 98. The age variable is important to state

that the patients with dementia have a lower survival rate (Kavitha et al.,2022). However, age is not a determining variable in the OASIS-2 dataset, since all participants are old. The estimated total intracranial volume (eTIV) is the same for the groups demented and nondemented, but converted ones have lower eTIV value. There seem some outliers in the demented group having high eTIV values. The ASF variable is a one-parameter scaling factor that is used to compare the eTIV considering the differences in human anatomy. The median of the ASF is like eTIV for nondemented and demented participants. Values are same for the groups of demented and nondemented. Importantly, for the converted group the values of eTIV exhibit an inverse relationship in comparison to the other groups. The ASF range for the nondemented participants is also wider. There seem differences in the medians of education levels. Nondemented and converted participants are seemed to be more educated than the demented ones. Again, the education range of nondemented participants is wider. There seem differences in the medians of education levels. Nondemented and converted participants are seemed to be more educated than the demented ones. Again, the education range of nondemented participants is wider. The MMSE variable stands for the results of an examination of mini mental state. It is a reliable and validated 30-point examination and used widely to be helpful in detecting the dementia level of patients with mild cognitive impairment (Fulton, et.al., 2019). The median values of MMSE exhibit a high variability among the variables within the groups demented and nondemented. However, the median of nondemented and converted groups are the same. The range in the demented group is wider than the nondemented and converted groups. The higher the values of MMSE the lower the level of dementia status of the patients (Gluhm, et al.,2013). There are outliers in the lower MMSE values in all groups. MR Delay shows the time between visits in days, does not change much between groups, but the ranges differ and there are high values in the demented group as outliers. The median of normalized whole brain volume is higher in the nondemented group, and the demented group has the lowest median since the disease influences the brain tissues

directly and results in a shrink in the brain tissues. Among all variables, education, and MMSE seem to be changed within demented and nondemented groups.

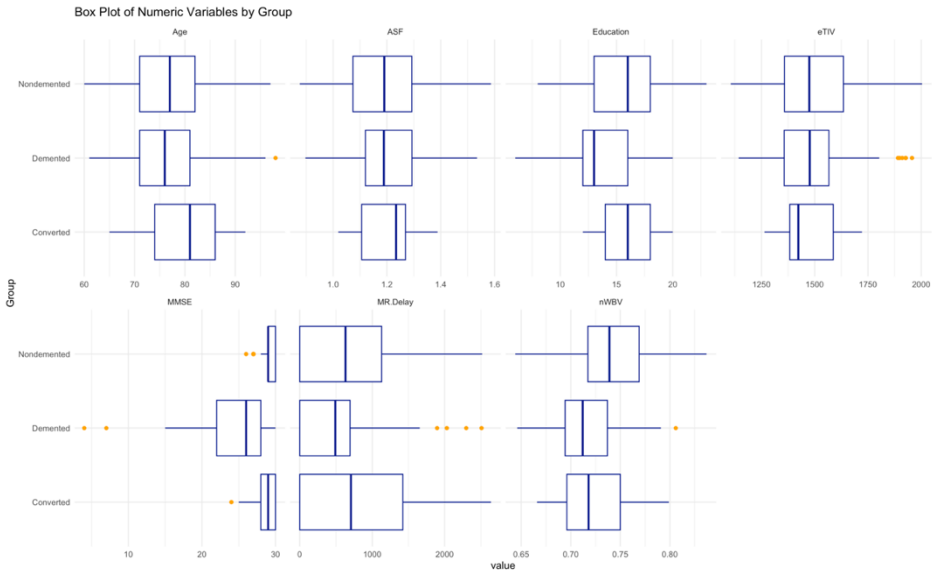

Figure 4.2 Box plot of numeric variables by group

4.2.2 Data Pre-processing

In this section, the steps taken to prepare the dataset for analysis are explained. To start analyzing and classifying the dataset, it is required to clean and prepare the dataset. To clean the dataset, firstly the variables that describe the same condition and the ones having no apparent impact on the analysis are examined. Missing values are identified and imputed, and the correlation structure is checked to ensure data readiness for analysis.

Data Cleaning

As seen in Table 4.1, all the participants are right-handed. Therefore, there is no need to add the hand variable to the analysis. Also, MRI ID and Visit, both show the same condition, the visit count of the patient. Therefore, the MRI ID variable is excluded.

The response variable of the study is planned to be group variable. However, since the structure of the dataset is longitudinal, but group variable does not show a longitudinal structure, the group variable is the same for all visits in each subject, a new response is created by using CDR variable. Considering the CDR scores, zero is directly taken as nondemented; the values greater than zero are taken as demented since even 0.5 shows a very mild dementia, it is a potential beginning of Alzheimer's disease. After inducing the categories in the response into two, the relationship between variables and group seen more clearly. The group variable excluded and replaced with response.

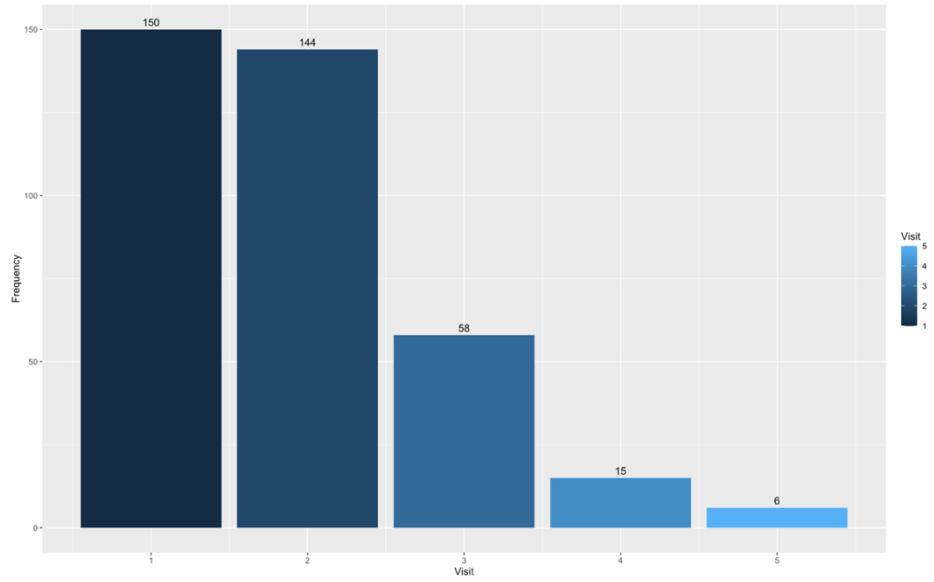

Figure 4.3 Bar plot of visit

The unbalanced structure of the visit variable is explored, and it is also seen in Figure 4.3, the frequencies in the fourth and fifth visits are considerably low compared to the other visits and it causes nonreliable results in determining the relations between variables and in modeling the response. Therefore, the fourth and fifth visits are eliminated, and analysis are conducted with three visits.

The relationship of the variables with visits is also examined in Figure 4.4 and Figure 4.5. It is important to see the effect of visits in other variables in order to plan the structure of the longitudinal analysis. In Figure 4.4, the relationship between visits

and categorical variables is given. Looking at the visits of the participants, there does not seem to be a noticeable difference for each visit in the gender, socioeconomic status, and dementia group (response variable). In CDR scores, two is seen only in the second visit, and there are minor differences such as in the frequency of zero. Moreover, nondemented group, is slightly higher in the third visit but this is because of the nondemented ones continuing the study are more than the demented ones.

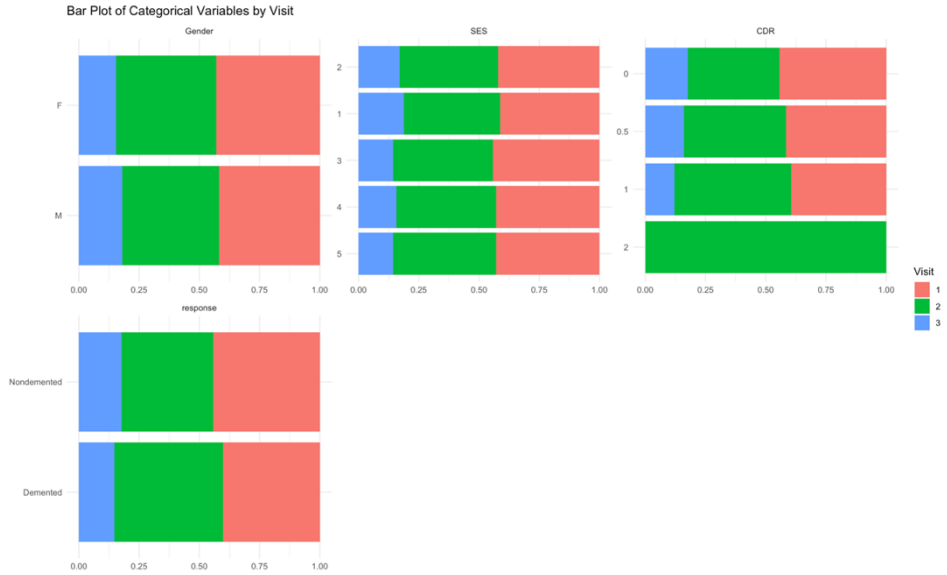

Figure 4.4 Bar plot of categorical variables by visit

Figure 4.5 shows the relationships between numeric variables and visit. The change in age and MR delay are clearly seen since age increases in each visit and MR delay is the delay of the participants according to the initial visit. Since MMSE is directly affecting the dementia group, the effect of visit in MMSE is also considerable since the demented ones are seen more in the second visit and MMSE scores are lower in the second visit as well. There seems a minor decrease in the brain volumes as the study progress.

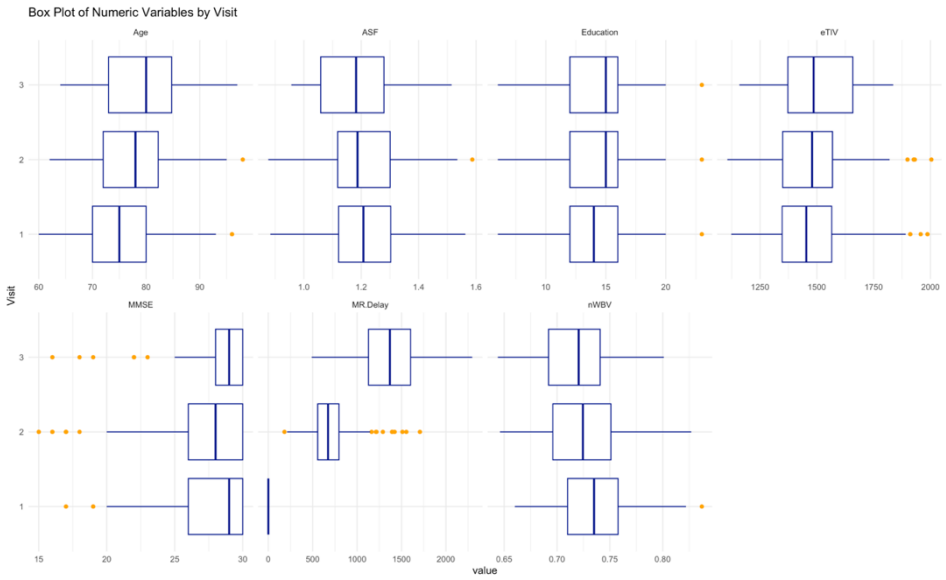

Figure 4.5 Box plot of continuous variables by visit

After cleaning the directly inter-related variables, the final structure of dataset is seen in Figure 4.6 and Figure 4.7. There are more demented participants in the second visit.

Figure 4.6 Bar plot of categorical variables by group

Male participants are still more likely to be demented than females. In terms of socioeconomic status of participants, the relation between SES and dementia group

is clearly seen that as while socioeconomic status level of participants is getting worse, the frequency of facing with dementia increases. Likewise with the results before cleaning, education, MMSE and nWBV values are considerably changing according to the dementia group of patients.

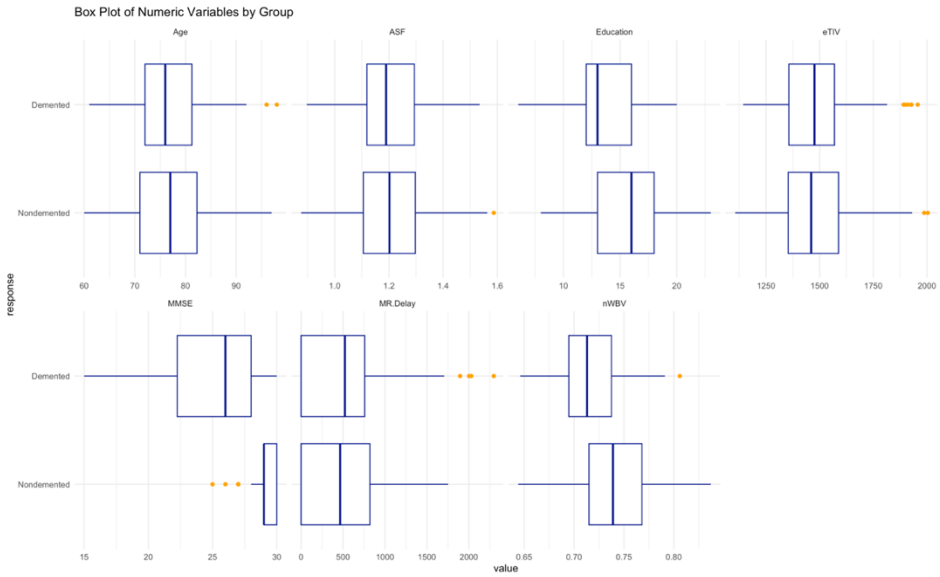

Figure 4.7 Box plot of numeric variables by group

Missing Imputation

Last step of the pre-processing is detecting the missing information, extract the missingness pattern and impute the missing values. In the dataset, there are missing values in MMSE and SES variables. As seen in Figure 4.8 the number of missing values correspond to approximately 0.05% of the observations in MMSE variable and around 5% of the observations in SES variable. Since the missing values do not have a specific pattern and a specific relation with other variables, the missing pattern is assumed to be MCAR or MAR. In order to detect the missing pattern, Little's MCAR test is applied (Little, 1988). The result of the test shows that the missing mechanism is not missing completely at random. Therefore, it is assumed to be missing at random. A more reliable imputation method is planned since there are several variable classes in the dataset and a multiple imputation method, MICE is

applied. After imputation, the relationship between imputed variables and response do not change and that is also a proof of the assumed missing pattern.

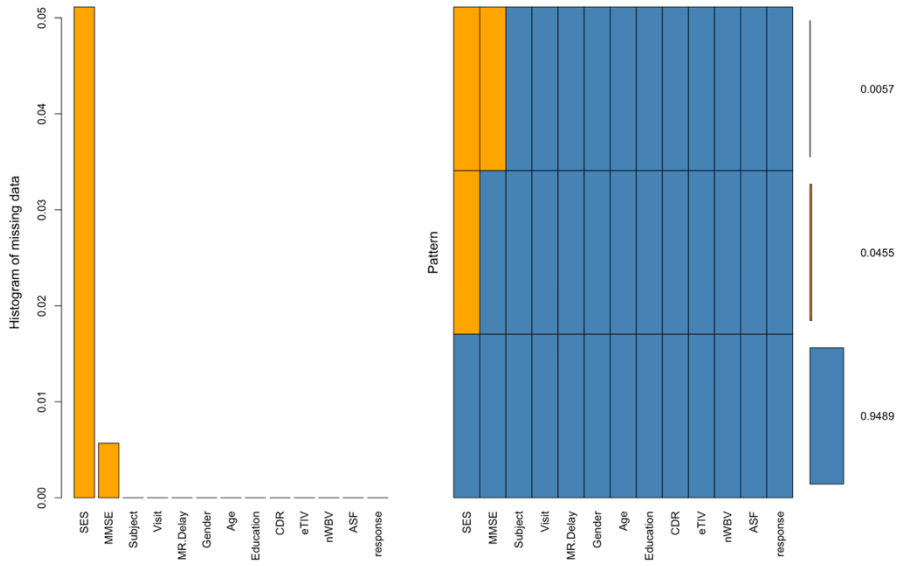

Figure 4.8 Missing patterns

Correlation Structure

Checking the correlation is very important before starting the modelling since it shows the relations between variables and gives an insight about the content of the models. The correlation structure of the ready-to-analyze data is checked, and the correlation plot is given in Figure 4.9. Visit seems to have a weak relationship with many of the variables. eTIV variable has a strong negative correlation with ASF, since ASF is used to estimate eTIV values, and these may cause a multicollinearity in models. It is not surprising to see that CDR and response are highly positively correlated, since response is derived from CDR. So as MR delay and visit since MR delay directly increases when the visit count increases.

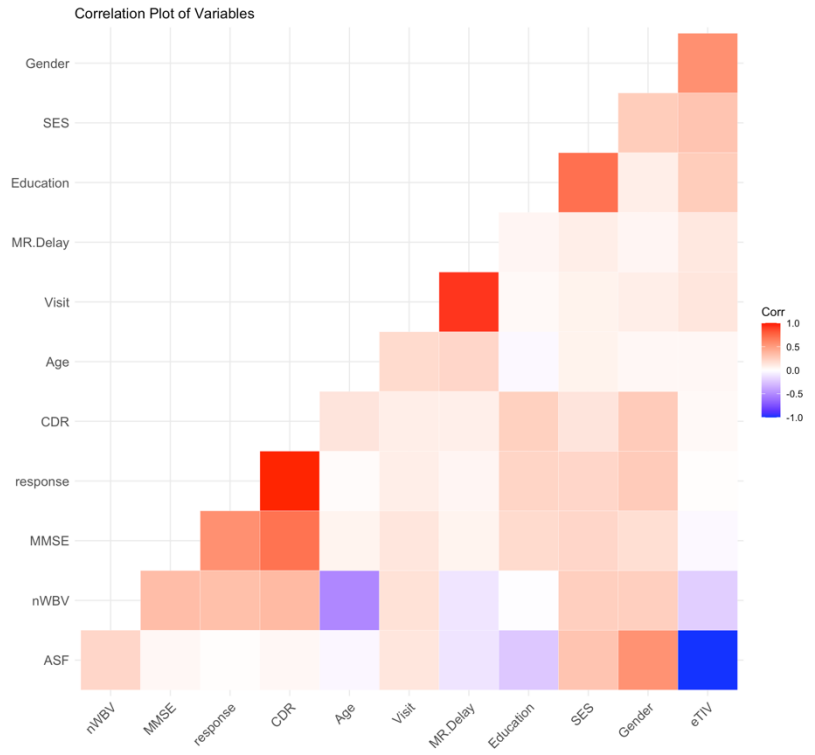

Figure 4.9 Correlation plot

Education and SES variables are also seemed to be considerably positively correlated since the better socioeconomic status of patients the better educated they are or vice versa. nWBV and age seems to have a negative correlation which means brain volume is getting smaller while people are getting older. The correlation between CDR and MMSE is high and positive, and this is reflected to the correlation between MMSE and response. The ones having high MMSE score are meant to have a higher probability to face with dementia. Considering the correlations, the variables which are highly correlated give idea about the factors effecting the dementia level.

4.3 Simulation Study

A simulation study is designed to see the effect of the sample size and the balance in time points in longitudinal data modelling. Four different number of subjects are simulated, and all models are trained in those subjects, also balanced and unbalanced time points are compared. In creating the simulations, characteristics of the original

dataset is used. In this section, the usage of the original dataset, the aim and structure of the simulation are explained.

First, the long-formed dataset is converted into wide-form to see the relations and correlations between time points. The structure of all variables in the dataset are evaluated. Ranges, distributions, mean and median values of numeric variables and frequencies of categorical variables are checked. Since generating numbers from multivariate normal distribution is the most common way, in order to use normal distribution, the relationship of numeric variables with normal distribution is checked, and mainly boxcox and log transformations are applied to convert each numeric variable to normal. Then the correlation between variables is saved to be used in simulating the new dataset. By using the "mvrnorm" function in R Studio, and assigning the mean and correlation matrix, the base of the dataset is simulated. After that, the wide-formed simulated dataset is converted into long-form, and back transformations and required arrangements are applied by considering the structure of the variables in the original dataset by using "faux" library in R Studio. Response and gender are converted to binary by assigning their classification probabilities, means and standard deviations. CDR and SES are converted to a likert form by assigning their minimum and maximum values, means, standard deviations and the mean and standard deviation of the normally simulated variables. Education is converted to uniform, again by using its minimum and maximum values, mean and standard deviation. nWBV and ASF are converted to normal by entering their minimum and maximum values, means, standard deviations and the mean and standard deviation of the normally simulated variables. Finally, age, MMSE and eTIV are generated by using their information in the first visit since they have a visit-related structure and the correlation with time must have been considered. As a last step, the structure of the simulated dataset is checked, and its characteristics are found very similar to the original dataset.

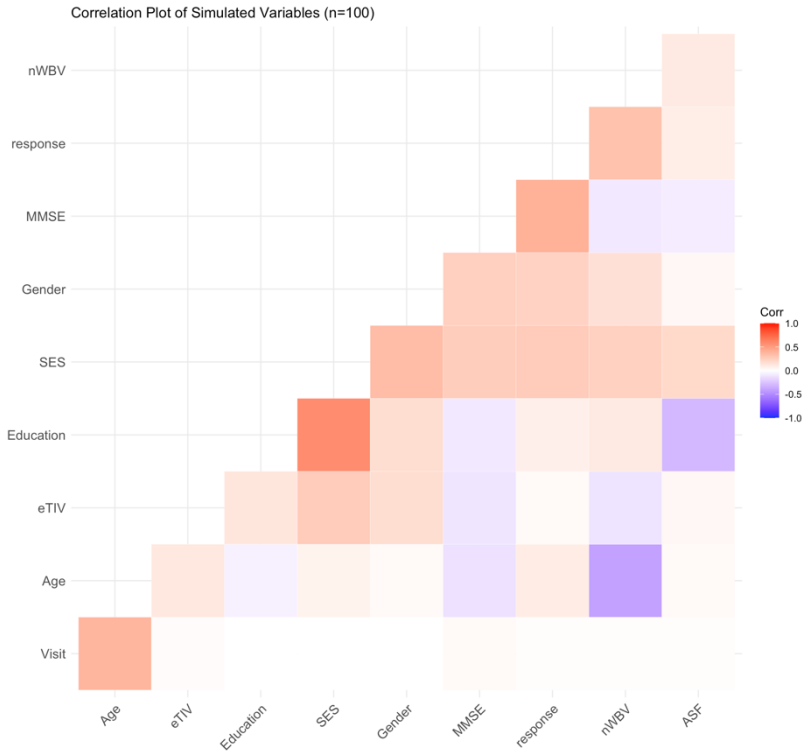

Figure 4.10 Correlation plot of simulated variables (N=100)

After simulating the dataset, the cross validation is applied as 80% train and 20% test. The classification models, GEE, GLMER, HGLM, GPBoost, GLMMTree, HRF and GLMMLasso are fitted and the model metrics, specifically, accuracy, F1 score, sensitivity and specificity are saved. The number of subjects is changed as 100, 150, 250 and 500 to see the effect of the sample size on model performances. In balanced simulated dataset, there were three time points for each subject, and the effect of having a balanced dataset is also evaluated. In Figure 4.10, the correlation structure of the simulated dataset can be seen. The correlation structure is not directly the same since this is a balanced structure and sample size differs, but it is very similar to the original dataset.

The unbalanced dataset is also simulated in the same structure, but all time points were not equal. The unbalanced structure is created by using again the original dataset. The frequency of missing time points in the original dataset are detected and the same frequency is applied to the simulated dataset. The number of subjects is designed to be 100, 150, 250 and 500 and the effects of both the sample size and

equality of time points are checked by constructing the same models as the original dataset and comparing the model performances.

In both balanced and unbalanced structures, the independent variables are kept the same and selected considering the models constructed with original dataset. The common significant variables are selected as MMSE, gender and education. For all simulation scenarios, the iteration is set to be 1000 and the model performance metrics are saved in Excel. Then their means are calculated and reported.

CHAPTER 5

RESULTS

In this chapter, the results of the analysis are given and model performances are reported and compared. Firstly the models constructed with real dataset is given, and the selected features and metrics are compared. Then the model performance metrics of simulation study are given and compared. In order to compare the models, accuracy, F1 score, sensitivity and spesificity are used.

5.1 Feature selection with Boruta

In the study, firstly, Boruta is used. To apply Boruta, the "Boruta" package in R is used. The algorithm gives an importance plot to show the importance of the confirmed attributes. This confirmation status is decided by shadows which work like a threshold to confirm a variable or not. The importance plot of the first Boruta model is given in Figure 5.1. The blue box plots show shadows, the red one, between *shadowMin* and *shadowMean,* is the rejected variable, and the green ones are the confirmed variables. It is seen that MMSE has a significant impact on dementia status, and it is followed by nWBV, education, gender, eTIV, ASF, SES, and age. The "Visit" variable is selected as unimportant according to Boruta.

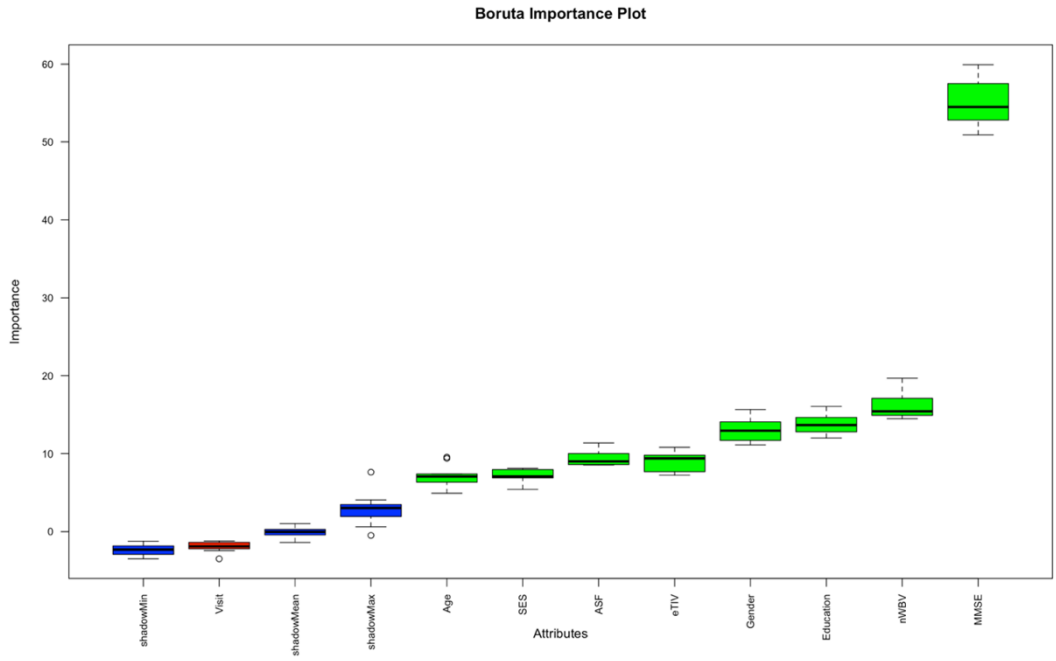

Figure 5.1 Boruta importance plot

A second model is fitted without the visit variable and after the elimination of the visit, all variables are confirmed which means all variables are relevant to the response. The second Boruta algorithm results after 10 iterations in 0.663 seconds.

In the following parts of the modeling, the visit variable is one of the first variables to be eliminated.

5.2 Cross Validation

In the beginning of the modelling part, the dataset is separated into two as 80% train and 20% test in order to make the models learn by using train set and check their performances by an unseen test set. Table 5.1 shows the number of observations, subjects and covariates in train and test sets. The division ratio is determined by trying 0.9 – 0.1, 0.8 – 0.2 and 0.7 – 0.3 as train and test sets for all models and calculating the model performance metrics. Because the sample size is small, giving a smaller percentage to the test set caused overfitting and biased results since the

model has plenty of observations to learn from but a limited number of observations to test the model performance.

Table 5.1 Cross validation structure

	Train Set	Test Set
Number of Observations	278	74
Number of Subjects	120	30
Number of Covariates	11	11

On the other hand, when the train percentage is reduced, the number of observations is not enough for model to learn. Therefore, the optimal division ratio is determined as 0.8 for train and 0.2 for test. All models are constructed by using train set and the model performance metrics are calculated by using the test set.

5.3 Modelling

The real life "Longitudinal MRI Data in Nondemented and Demented Older Adults" dataset is used to build the models in this part. First, the Boruta algorithm is used to see the importance of the features. After Boruta, the dataset is split into test and train, and both sets are scaled by standardization. The first part of data used to train GEE, GLMM, HGLM, GPBoost, GLMERTree, HRF, and GLMMLasso.

5.3.1 Generalized Estimating Equations (GEE)

The GEE model is constructed by using the "geepack" package in R programming. The algorithm requires a correlation structure. For the real dataset, the correlation structure resembles an exchangeable structure. Therefore, the models are constructed using an exchangeable correlation structure.

In the function, there is a model, data, id, family, and correlation structure. Firstly, the model is constructed to see the relations between response and all other predictors

except the subject variable. The subject variable is assigned as id, the family is chosen to be binomial, and the correlation structure is assigned as exchangeable. The multicollinearity between variables is checked by the variance inflation factor(vif) function under the library "car", and it is seen that the high correlation between ASF and eTIV resulted in multicollinearity. Reduced models are constructed by considering the result of Boruta and the correlations between response and independent variables.

Table 5.2 Information criteria comparison of GEE models

	QIC	QICu	CIC	QICC	parameters
Fit 1	188	183	16.63	193	14
Fit 2	188	183	15.61	193	13
Fit 3	**183**	**176**	9.32	**184**	6
Fit 4	201	197	7.10	202	5
Fit 5	200	196	**5.74**	200	4
Fit 6	203	200	18.87	207	12

Six models are trained, and the best one is selected as the one which relates the response with education, gender, and MMSE. The QIC values of the models are compared and reported in Table 5.2. The estimated marginal model summary is given in Table 5.4.

Table 5.2 gives the QIC, QICu, CIC, and QICC values of all trained models. In terms of QIC values, the third model seems to have lower QIC values, which indicates that the third model is a better fit to the dataset and has less overfitting. However, the CIC value is lower for the fifth model, and this means that the fifth model is both well-fitted to the data and has lower complexity. Additionally, examining the performance metrics of the third and fifth models (Table 5.3), it is observed that the performance of the fifth model is slightly better in test metrics, while the third model is better in train metrics. Also, the computational time of the third model is slightly smaller. Overall, the third model is selected as the best estimated marginal model.

In order to enhance the model, several values of the control parameters, specifically, positive convergence tolerance (epsilon) and the number of maximum iterations (maxiter), are tested on the third model. Results show that the control parameters do not influence model performance metrics and QIC values. They only change the computational time. Therefore, the final model is selected with epsilon = 0.01 and maxit=10 since the model is built in 0.012 seconds.

Table 5.3 Model performance metrics and computational time of third and fifth models

	Train				
	Accuracy	F1 Score	Sensitivity	Specificity	Time
Fit 3	**0.848**	**0.834**	0.823	**0.870**	**0.029**
Fit 5	0.832	0.809	**0.829**	0.835	0.032
	Test				
	Accuracy	F1 Score	Sensitivity	Specificity	
Fit 3	0.787	0.785	0.737	0.843	
Fit 5	**0.796**	**0.892**	**0.750**	**0.846**	

The estimated model is

$$logit(\hat{P}(Y_{it}=1)) = 0.002 - 0.815Age_{it} - 0.831Education_{it} - 1.236nWBV_{it} + 1.131Gender_{it} - 2.227MMSE_{it} \quad (40)$$

$$\hat{P}(Y_{it}=1) = \frac{e^{0.002-0.815Age_{it}-0.831Education_{it}-1.236nWBV_{it}+1.131Gender_{it}-2.227MMSE_{it}}}{1+e^{0.002-0.815Age_{it}-0.831Education_{it}-1.236nWBV_{it}+1.131Gender_{it}-2.227MMSE_{it}}} \quad (41)$$

The estimates are in logit scale, and to interpret them, exponentiation is used to obtain the odds ratio. The odds ratio of the estimates is given in Table 5.4. For instance, the odds ratio of education is 0.436, meaning one-unit education increase results in the patient being approximately 0.436 times less likely to have dementia. Similarly, for every one-unit increase in age, nWBV and MMSE, the participant is respectively 0.443,0.290 and 0.108 times less likely to have dementia. Female is selected as the reference group, and the odds ratio is calculated as

$\frac{P(demented|Male)}{P(nondemented|Male)} / \frac{P(demented|Female)}{P(nondemented|Female)} = 3.101$. The odds ratio is interpreted as males are 3.101 times more likely to be demented compared to females. Since the estimates are higher than the standard errors, the results seem to be reliable.

Table 5.4 GEE model summary

	Estimate	Standard error	P-value	Odds Ratio
Intercept	0.002	0.308	0.994	1.002
Age	-0.815	0.290	0.005*	0.443
Education	-0.831	0.265	0.002*	0.436
nWBV	-1.236	0.299	<0.001*	0.290
Gender (M)	1.131	0.491	0.021	3.101
MMSE	-2.227	0.518	<0.001*	0.108

"*" indicates the significant variables

In Table 5.5, the estimated alpha indicates the relationship between the repeated measures. The estimated value is 0.465, which means there seems a positive moderate relationship between repeated measurements. In terms of the significance, $Z\ score = \frac{estimate}{standard\ error} = \frac{0.436}{0.092} = 4.74$, which is higher than the z-score in 0.05 significance level. It means, the correlation parameter for the repeated measures is significant.

Table 5.5 Estimated correlation parameters of GEE model

	Estimate	Standard error
alpha	0.436	0.092

5.3.2 Generalized Linear Mixed Model (GLMM)

In this study, generalized linear mixed models are used to see the relationship between the dementia status of participants and the predictors related to dementia by building several models and trying the optimizers. In building the models, the results of the Boruta and the relationship between variables have a big role. Five models are trained, and the best estimated model is selected according to AIC, BIC, and model performance metrics. In building GLMMs, the "lme4" package in the R programming language is used.

The first model is constructed as default with all variables by assigning the dementia group(response) as the outcome; visit, gender, age, education, SES, MMSE, eTIV, nWBV, and ASF as predictors, and subject as the random effect. Family is assigned as binomial. Because of the high correlation between ASF and eTIV, multicollinearity is observed. Therefore, the variables are eliminated step by step by also considering their significance. The best-estimated model is selected as the one with covariates gender, age, education, MMSE, and nWBV with random intercept only.

Later, considering the possible changes of dementia status over visits, visit variable is included as a random slope. In the data exploration part, it is observed that there is not an obvious change in dementia status and in other covariates over visits and the random slope estimates also confirmed this. Adding other variables as random slope gives the same result. Therefore, the final model is decided to be the one without random slope, and it indicates that modeling only subject-specific relationships is beneficial to classify the dementia status.

In order to enhance the model, control parameters are added to the model. Optimizers improve the model by finding the parameter values that minimize the negative loglikelihood function of the model and thereby increasing the model's goodness-of-fit. The type of optimizer can affect the convergence speed and accuracy of the model. Also, an appropriate optimization method helps to achieve more reliable

parameter estimates. The default optimizer of the GLMM under the lme4 package is "bobyqa". The built-in optimizers are "Nelder Mead", "bobyqa", "optimx", "nmkbw", "nloptwrap" and "nlminwrap" but there are commonly used optimizers in "optimx" package such as "L-BFGS-B" and there are options in nloptwrap such as "NLOPT_LN_NELDERMEAD" and "NLOPT_LN_BOBYQA" (Bates, et.al.,2023). These possible optimizers are tried and the one giving smaller AIC and BIC values, higher model performance metrics are planned to be selected. However, all optimizers give similar values and metrics. Therefore, their working speed is calculated, and "nlminbwrap" is found to be the fastest one and selected as the best optimizer.

The estimated model for the first subject in the train set is given in equations (42) and (43),

$$logit\left(\hat{P}(Y_{1t}=1)\right) = -0.572 + 3.951Gender_{1t} - 2.467Age_{1t} - 2.524Education_{1t} - 5.530MMSE_{1t} - 4.347nWBV_{1t} + 0.253 \quad (42)$$

$$\hat{P}(Y_{1t}=1) = \frac{e^{-0.572+3.951Gender_{1t}-2.467Age_{1t}-2.524Education_{1t}-5.530MMSE_{1t}-4.347nWBV_{1t}+0.253}}{1+e^{-0.572+3.951Gender_{1t}-2.467Age_{1t}-2.524Education_{1t}-5.530MMSE_{1t}-4.347nWBV_{1t}+0.253}} \quad (43)$$

The random effect behavior is given in Table 5.6. The variance indicates the variability between subjects across dementia groups, not the variance between subjects within each group. So, 32.7% of variability within dementia groups is explained by subject-based differences which is an adequate percentage to say that using a mixed model and considering the subject differences as random intercept are effective for classifying dementia group for OASIS 2 dataset.

Table 5.6 Random effects

Random effect	Variance	Standard deviation
Subject	32.7	5.72

The summary of the model and odds ratio of the estimates are given in Table 5.7 and seen in Figure 5.2. For instance, the estimate of gender is 3.951. The odds ratio of having dementia among two gender groups is calculated by the formula, $\frac{P(demented|Male)}{P(nondemented|Male)} / \frac{P(demented|Female)}{P(nondemented|Female)} = 51.968$ which means males are 51.968 times more likely to face with dementia compared to females.

Table 5.7 GLMM summary

	Estimate	Standard error	P-value	Odds Ratio
Intercept	-0.572	1.048	0.585	0.564
Gender (M)	3.951	1.997	0.048*	51.968
Age	-2.467	1.186	0.037*	0.085
Education	-2.524	1.115	0.024*	0.080
MMSE	-5.530	1.470	<0.001*	0.004
nWBV	-4.347	1.759	0.013*	0.013

"*" indicates the significant variables

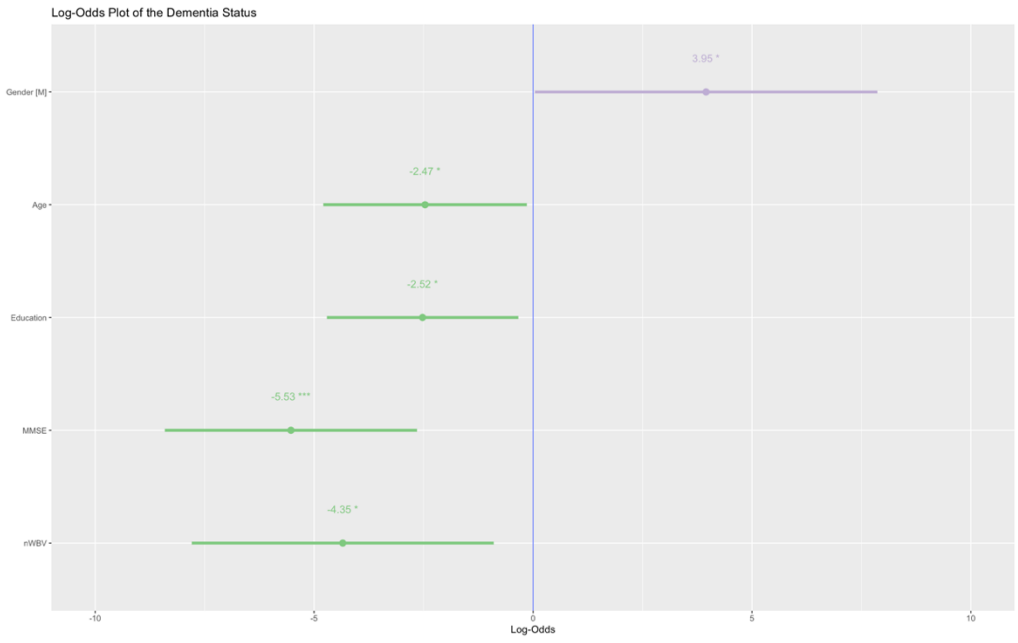

Figure 5.2 Log-odds plot of dementia status

The estimate of age is -2.467. When it is exponentiated, the odds ratio showing the effect of one unit change in age is 0.085, meaning that a one-unit increase in age results in the patient being 0.085 times less likely to be in the demented group. Similarly, the odds ratios of education, MMSE, and nWBV represent that a one-unit increase in education, MMSE and nWBV results in the patient having dementia are respectively 0.80, 0.004 and 0.013 times less likely.

Since the estimates are higher than the standard errors, the results seem to be reliable. Also, the overdispersion of the model is checked and no overdispersion is detected.

5.3.3 Hierarchical Generalized Linear Model (HGLM)

In order to build hierarchical generalized linear models, the "hglm" package in the R programming language is used.

In the study, several HGLMs are constructed initially using all variables in the dataset and the ones disrupting the model are eliminated one by one. The elimination process is performed considering the results of Boruta, multicollinearity between variables, standard errors of the variables in the model output (since high standard error results in unreliable estimates), and the significance of the variables. In the final model reached, variables Age, Gender, Education, nWBV and MMSE are found to be significant. The family is assigned as binomial in all models, and only a random intercept is added.

On the one hand, the distribution of the random effects is kept in the default format, as Gaussian, and the model giving the highest model performance metrics is saved. On the other hand, the distribution of random effects is changed to Beta and Gamma, by also changing the family link (as given in Table 3.6) as required. However, in both cases, the model did not converge. Therefore, it is continued with Gaussian as random effect distribution.

Moreover, since HGLM has an extension that deals with dispersion, a dispersion model is added to see the variance estimates of covariates, but the models with the

dispersion model added also did not converge. Therefore, the dispersion model could not be added but the dispersion parameters are recorded in the estimated model. Also, different values for convergence criteria "conv", and the maximum number of iterations "maxit" are tried and the model performance metrics did not change. Hereby, the estimated model is selected with the parameters having lower computational time and without the dispersion model added.

The estimated model for the first subject in the train set:

$$logit\left(\hat{P}(Y_{1t}=1)\right) = 1.366 - 1.004Age_{1t} - 1.132Gender_{1t} - 0.776Education_{1t} - 3.367MMSE_{1t} - 1.312nWBV_{1t} + 1.760 \quad (44)$$

$$\hat{P}(Y_{1t}=1) = \frac{e^{1.366-1.004Age_{1t}-1.132Gender_{1t}-0.776Education_{1t}-3.367MMSE_{1t}-1.312nWBV_{1t}+1.760}}{1+e^{1.366-1.004Age_{1t}-1.132Gender_{1t}-0.776Education_{1t}-3.367MMSE_{1t}-1.312nWBV_{1t}+1.760}} \quad (45)$$

The summary of the model and odds ratio of estimates are given in Table 5.8. The mean model structure is a usual GLMM model structure, interpreted like GLMM. Having a binary response, the estimates are in logit scale, and the odds ratio is used to interpret them. The odds ratio of education estimate is 0.460, meaning that one unit increase in education variable results in the patient being 0.460 times less likely to have dementia. Similarly, the odds ratios of age, MMSE, and nWBV means a one-unit increase in those variables causes the patient to be respectively 0.366, 0.034, and 0.269 times less likely to be in the demented group.

Table 5.8 HGLM mean model summary

	Estimate	Standard error	P-value	Odds Ratio
Intercept	1.366	0.312	<0.001*	3.921
Age	-1.0043	0.242	<0.001*	0.366
Gender 1	-1.1320	0.341	0.001*	0.322
Education	-0.776	0.202	<0.001*	0.460
MMSE	-3.367	0.466	<0.001*	0.034
nWBV	-1.312	0.242	<0.001*	0.269

"*" indicates the significant variables

In this model, male (0) is the reference group. The odds ratio of having dementia among two gender groups is $\frac{P(demented|Female)}{P(nondemented|Female)} / \frac{P(demented|Male)}{P(nondemented|Male)} = 0.322$ which means females are 0.322 times less likely to face with dementia compared to males.

Furthermore, as seen in Table 5.9, the low value of dispersion parameter in the mean model proves the presence of homoscedasticity in the model and it indicates that the model is well-classifying the data and there seems no overdispersion problem. Therefore, it is not necessary to add a dispersion model.

Table 5.9 Dispersion model outputs

Dispersion parameter for the mean model:	0.663
Dispersion parameter for the random effects:	0.018

In HGLMs, the dispersion parameter for random effects is also calculated and it is used to estimate the variance of the random effects. The dispersion parameter for random effects is important to qualify the variability between subjects and it is found as 0.018, which means there seems not any outlier observations among the subjects.

Finally, since the hglm package does not contain a "predict" function, a prediction function is created manually by using the dataset and the model's fixed effects and random effects. To obtain predictions on the test set, random effects are taken as zero by taking it similar to the random effects in glmm predictions on the test set.

5.3.4 Gaussian Process Boosting (GPBOOST)

The algorithm is applied using the "gpboost" package in the R programming language. In the package, there are two main models, one is for Gaussian Process (GP) implemented mixed effects modeling and helps to see the significance of covariates, and the other one is for boosting, which only has the importance of features, not the estimates. However, the parameters of the first model can be assigned to the second model. Both models are performed, and the results are saved.

The GP model, combined with the mixed effects model, by using the "fitGPModel" function, is built by assigning the group data as subject, likelihood as Bernoulli logit, and the response and all covariates except the subject. Since gpboost can deal with multicollinearities, none of the variables are eliminated. Then, the options for better performance are checked, and by trying several optimizer coefficients, training speed, and accuracy are planned to be improved. The optimizers, Gradient Descent, Nelder Mead, bfgs, and adam, are tried and Gradient Descent is selected as the one giving the lowest AIC and BIC values and highest model performance metrics. The maximum number of iterations is set to 1000000 to increase the convergence rate and reach optimal parameter values. Also, the convergence tolerance, putting a convergence threshold, works as a stopper of the algorithm if the relative change in the parameters is below the convergence tolerance value (Sigrist, et.al.,2023). The default value is generally 0.000001, but 0.01 gives the highest model performance metrics. Therefore, the model with the best model performance metrics and lowest AIC and BIC values is recorded as the model with 1000000 as the maximum iterations, gradient descent as the optimizer, and 0.01 as the convergence tolerance.

The model, after tuning the parameters, is selected as the first GP model, and it is recorded as GPBoost with optimizers in Table 5.10.

The estimated model for the first subject in the train set:

$$logit\left(\hat{P}(Y_{i1}=1)\right) = 1.845 + 0.074Visit_{1t} - 1.109Age_{1t} - 2.045Education_{1t} + 0.292SES_{1t} - 4.824MMSE_{1t} + 0.145eTIV_{1t} - 2.918nWBV_{1t} - 0.189ASF_{1t} - 3.545Gender_{1t} + 0.566 \equiv m* \quad (46)$$

$$\hat{P}(Y_{1t}=1) = \frac{e^{m*}}{1+e^{m*}} \quad (47)$$

In the model, education, MMSE, gender, and nWBV seem to be significant.

Table 5.10 GP Model with optimizers summary

	Estimate	Standard error	P-value	Odds Ratio
Intercept	1.845	2.713	0.497	6.326
Visit	0.074	0.546	0.893	1.076
Age	-1.109	0.927	0.231	0.329
Education	-2.045	1.052	0.052*	0.129
SES	0.292	0.877	0.739	1.339
MMSE	-4.824	1.282	<0.001*	0.008
eTIV	0.145	4.583	0.975	1.156
nWBV	-2.918	0.962	0.002*	0.054
ASF	-0.189	4.464	0.966	0.827
Gender	-3.545	1.805	0.049*	0.029
Covariance parameter for random effects:				
Subject	24.8			

"*" indicates the significant variables

In the model output (Table 5.10), the covariance parameter for the random effects is 24.8, which means subject-based differences explain 24.8% of variability within dementia groups, there seems an effect of subject-specific modeling in classifying the response and therefore, it is meaningful to add a random effect part.

Interpretation of the significant covariates is that a one-unit change in education, MMSE, nWBV and the change in Gender (from female to male) result is the patient respectively 0.129, 0.008, 0.054 and 0.029 times less likely to be demented.

The optimizers are used to build the boosting model with hyperparameters. The model is again constructed using all covariates and setting a grouping variable. To tune the hyperparameters for obtaining the model with the best performance, grid search is applied by using the function "gpb.grid.search.tune.parameters". The best hyperparameters are given in Table 5.11.

The variable importance plot after boosting is given in Figure 5.3. The model gives the most important six variables in the classification of the dementia group. As seen, MMSE has the highest score in terms of gain, cover and frequency and the frequency of nWBV is the same with MMSE. In gain and cover plots, MMSE is followed by nWBV. Surprisingly, in these plots, visit seems an important variable to classify dementia status. However, in model results and in the feature selection of Boruta, it is not significant and eliminated.

Table 5.11 Hyperparameters of the model

Hyperparameters	Obtained values
Learning rate	0.0001
Minimum data in leaf	10
Maximum depth	5
Number of leaves	106421
Maximum bin	255

Table 5.12 Model performance metrics of GP models

	Train				Test			
	Accuracy	F1	Sensitivity	Specificity	Accuracy	F1	Sensitivity	Specificity
GP Model optimizers	0.975	0.973	0.956	0.992	0.731	0.724	0.691	0.774
GP Model hyperparameters	0.971	0.968	0.964	0.977	0.463	0.633	0.463	NA

The model performance metrics obtained by GP models are given in Table 5.12. It is seen that the model with optimizers is a good fit, train metrics are high and test metrics are considerably high, there is not a sign of overfitting. However, the boosting model overfits due to the small number of observations in the test set. This indicates that boosting works well in high-dimensional data but may be overfitting in small samples. The specificity is NA since in the test predictions, no zero value is obtained which means there is no subject assigned to be demented.

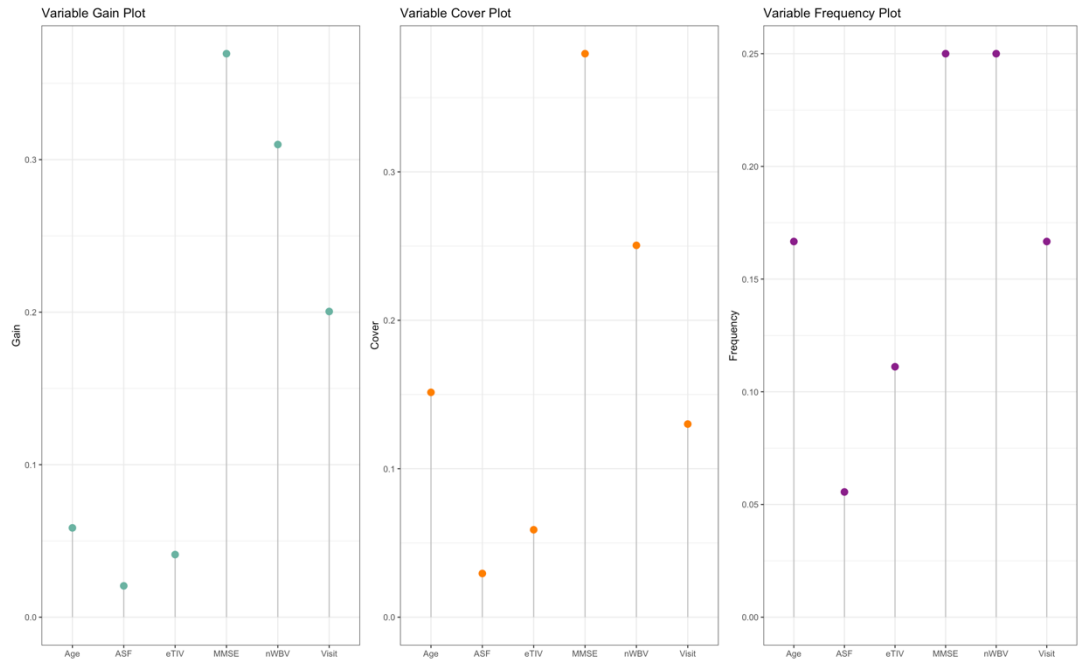

Figure 5.3 Variable importance plots of GP Boosting

5.3.5 Generalized Linear Mixed Model Tree (GLMMTree)

In the study, GLMMTree is applied by using the "glmertree" package in the R programming language. The package combines mixed and tree-based modeling. Therefore, a clustering variable (Subject) is assigned to add a random effects part to the model. Also, since the outcome is binary, family parameter is set to be binomial. The initial tree consists of 5 inner nodes and 6 terminal nodes. Then, to improve the performance of the model and reduce the computational time, the hyperparameters, maxdepth, maxit, and abstol are tuned and alpha and minsize are added to the model. The alpha parameter is set to be 0.05 to be able to compare p values with it and the default formula for minsize, 0.10*nrow(data) is applied. The hyperparameters giving the best performance metrics are planned to be selected but all give the same accuracies. Therefore, the ones having the least computational time are selected as maxdepth =10, maxit=100, and abstol=0.00001. After tuning, the number of inner nodes is reduced to 3 and the number of terminal nodes is reduced to 4. The model gives a predicted random effects plot of subjects (Figure 5.5) built by the mixed effects part of the model and the model tree (Figure 5.4).

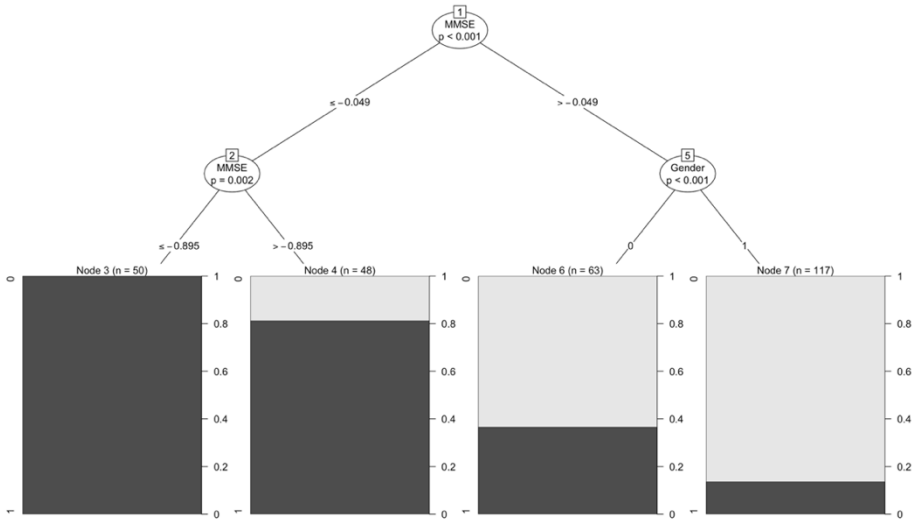

Figure 5.4 Model tree

In the model tree, the splitting variables and their corresponding p-values used to identify the significance of the variables are reported in every inner node. The significance level is set to be 0.05; therefore, these p-values are compared with 0.05 (Fokkema et al.,2018).

Although many trees are fitted with several variables, the most important variables come up as MMSE and gender to address the dementia status. Since the MMSE values are in the standardized form, they are converted into real MMSE values in interpretation as -0.049 means 27.2 and -0.895 means 24.4. The ones having smaller than or equal to 27.2 MMSE value continue with node 2, and if they have smaller than or equal to 24.4 MMSE value, they are in node 3. In this node, there are 50 observations and all of them are detected as demented. The ones having MMSE values between 24.4 and 27.2 end up with node 4 where there are 48 observations and almost 80% of them are detected as demented. On the other hand, the ones having greater than 27.2 MMSE value continue with node 5 and get split according to their gender. The males (i.e. Gender=0) having greater than 27.2 MMSE value end up in node 6 where there are 63 observations and approximately 38% of them are detected as demented. Otherwise, the females end up in node 7 where 117 observations are in, and 18% of them are detected as demented. Like the other models, while the MMSE value increases, subjects are more likely to be nondemented and males are more likely to be grouped as demented.

In the predicted random effects plot of the subjects, random intercepts are given. Each vertical line corresponds to one subject and the expected dementia group score for many subjects has different ranges. This indicates higher variability between subjects. The subjects having positive values tend to have higher expected dementia group scores whereas the ones on the left part of zero tend to have lower scores. Zero means the subjects tend to have average expected dementia group scores.

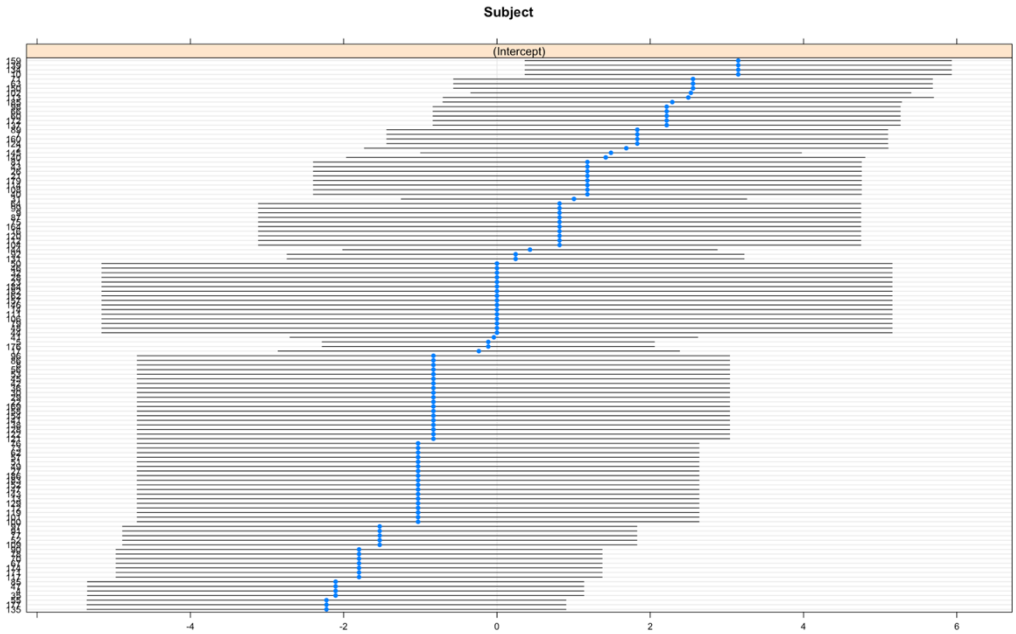

Figure 5.5 Predicted random effects plot of subjects

5.3.6 Historical Random Forest (HRF)

In the study, HRF is applied by using the “htree” package in the R programming language. The algorithm combines random forest with subject-based classification by adding id and time parameters, and parameters for historical(vh) and concurrent(vc) predictors. The id and time parameters are set to be subject and visit respectively. Historical predictors are the ones changing with time and assigned as age, MMSE, eTIV, nWBV, and ASF, whereas the concurrent predictors are education, SES, and gender. The historical parameter is set to be TRUE to add the historical structure to the model. The initial model has the default hyperparameters as mtry=3, nsamp=5, and ntrees=100. The hyperparameters are tuned to increase the model performance and the model having the highest performance is with hyperparameters nsamp = 4, ntrees=100, and mtry=3.

The model performance metrics are calculated by taking all trees into account and calculating the mean of each model with a tree since considering all trees makes the result more accurate.

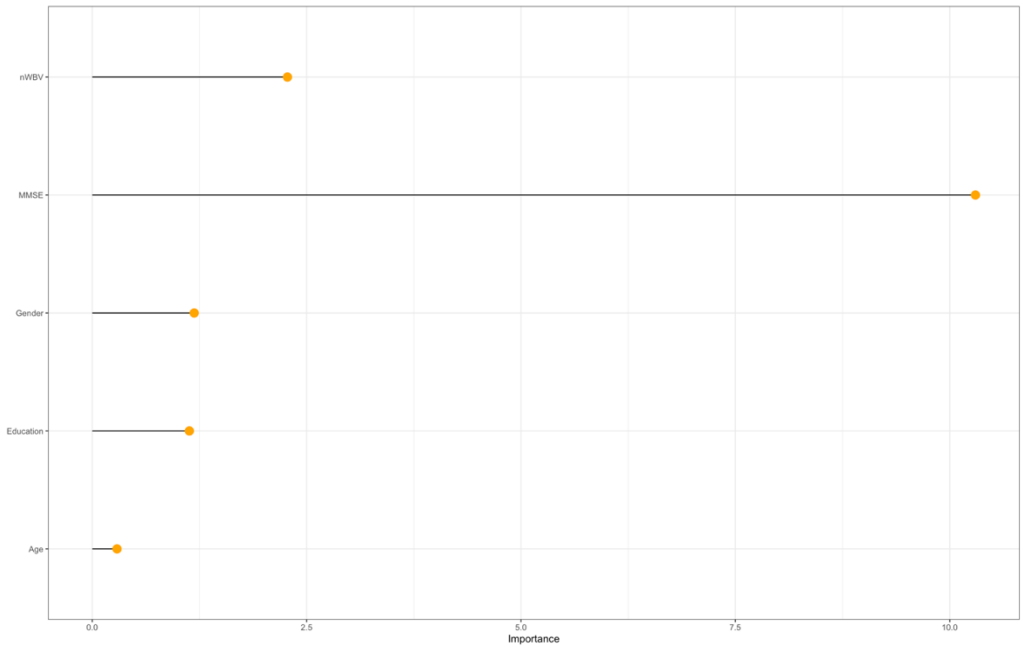

Figure 5.6 Variable importance plot of HRF

The importance of the variables is given in Figure 5.6. The MMSE variable seems to have the highest importance in classifying the dementia status and it is followed by nWBV, gender, education, and age.

5.3.7 Generalized Linear Mixed Model Least Absolute Shrinkage and Selection Operator (GLMMLasso)

GLMMLasso is available in R programming language in the “glmmLasso” package which is used in the study. In the function, there are both fixed and random model parts. In the fixed part the usual model with the response as the outcome and all variables except the subject as covariates are set whereas in the random part, the subject is assigned as the random effect. The family of the outcome is taken as “binomial (link = "logit")”. To choose the optimal lambda value a for loop is written and the model metrics for each tried lambda value are found the same. All possible lambda values are tried one by one and the one giving the most optimal parameter estimates is chosen as 2.2. Then the control parameters nue, steps, and epsilon are

tuned, and the final best-estimated model is selected as nue=1.3, steps = 300, and epsilon = 0.0001.

The estimated model for the first subject in the train set:

$$logit\left(\hat{P}(Y_{1t}=1)\right) = 2.283 - 0.305Visit2_{1t} - 0.479Visit3_{1t} - 1.046Gender_{1t} \quad (48)$$
$$- 1.021Age_{1t} - 1.315Education_{1t} - 1.343SES2_{1t} - 0.345SES3_{1t}$$
$$- 1.227SES4_{1t} - 4.467SES5_{1t} - 3.384MMSE_{1t} + 1.063eTIV_{1t}$$
$$- 1.393nWBV_{1t} + 1.070ASF_{1t} + 0.158 \equiv m*$$

$$\hat{P}(Y_{1t}=1) = \frac{e^{m*}}{1+e^{m*}} \quad (49)$$

Table 5.13 Estimated model summary of GLMMLasso

	Estimate	Standard error	P-value	Odds Ratio
Intercept	2.283	0.318	<0.001*	9.807
Visit 2	-0.305	0.436	0.484	0.737
Visit 3	-0.479	0.717	0.505	0.620
Gender (F)	-1.046	0.562	0.063*	0.351
Age	-1.021	0.326	0.002*	0.360
Education	-1.315	0.413	0.001*	0.268
SES 2	-1.343	0.523	0.010*	0.261
SES 3	-0.345	0.612	0.573	0.708
SES 4	-1.227	1.041	0.239	0.293
SES 5	-4.467	2.277	0.049*	0.012
MMSE	-3.384	0.625	<0.001*	0.034
eTIV	1.063	1.393	0.445	2.894
nWBV	-1.393	0.338	<0.001*	0.248
ASF	1.070	1.403	0.446	2.915

"*" indicates the significant variables

The variables gender, age, education, SES, MMSE, and nWBV are found to have a significant effect on classifying dementia status. The interpretation is that

- Females are 0.351 times less likely,
- One unit increase in age results in the patients being 0.360 times less likely,
- One unit increase in education results in the patients being 0.268 times less likely,
- Socioeconomic status = 1, highest, is taken as the reference group and compared to the ones having the highest socioeconomic status the ones having the second SES are 0.261 times less likely, and the ones having the fifth SES are 0.012 times less likely,
- One unit increase in MMSE results in the patients being 0.034 times less likely, and
- One unit increase in nWBV results in the patients being 0.248 times less likely to be in the demented group.

5.3.8 Comparison of the Methods in Real Data

Classical and machine learning approaches are applied to OASIS – 2 datasets to classify the dementia status of patients. Since the study is longitudinal, the algorithms having mixed modeling or cluster-based modeling approaches are applied. The model performances are measured by accuracy, F1 score, sensitivity, and specificity. Detailed information about these metrics is given in Section 3.5.2. The metrics calculated on the train set are given in Table 5.14, whereas those calculated on the test set are in Table 5.15. The computational times of the final models are also included in Table 5.14. The model performance metrics on the test set are essential to see the classification performance of the models when new data comes in. Therefore, the higher the model performance metrics on the test set, the better the model fits.

Table 5.14 Model performance metrics on train set and computational time

	GEE	GLMM	HGLM	GPModel optimizers	GPModel Boosting	GLMMTree	HRF	GLMMLasso
Accuracy	0.848	0.857	0.844	**0.975**	0.971	0.827	0.816	0.881
F1	0.834	0.840	0.824	**0.973**	0.968	0.788	0.798	0.869
Sensitivity	0.823	0.844	0.840	0.956	**0.964**	0.908	0.78	0.865
Specificity	0.870	0.867	0.948	**0.992**	0.977	0.783	0.847	0.895
Time	0.029	0.137	0.495	**0.008**	0.014	0.266	0.138	0.802

The model performance metrics on the train set show that the models fit the dataset at a similar rate. Only GPBoost models outperform the other models with higher model metrics. Also, the computational time of GPBoost models is very low compared to the others. This indicates that GP models learn better and work faster. However, the model performance metrics on the test set show that predictions of GP models on the test set are not as good as the ones on the train set, even overfitting is observed.

Table 5.15 Model performance metrics on test set

	GEE	GLMM	HGLM	GPModel optimizers	GPModel Boosting	GLMMTree	HRF	GLMMLasso
Accuracy	0.787	0.750	0.796	0.731	0.463	**0.851**	**0.899**	0.769
F1	0.785	0.738	0.784	0.724	0.633	**0.814**	**0.889**	0.762
Sensitivity	0.737	0.769	0.703	0.691	0.463	**0.889**	**0.966**	0.727
Specificity	0.843	0.821	0.838	0.774	NA	**0.830**	**0.85**	0.811

The tree-based models seem to have better metrics on test set, especially HRF outperforms. This indicates that in a small-sampled longitudinal data analysis, ensembled methods are better to classify the binary response and interpret the importance of the variables.

In Table 5.16, the features mainly related to the dementia status are given based on each model. It is seen that MMSE and nWBV are common in all models, except GLMMTree. Mainly age, education and gender are obtained as the important

features since they are significant in most of the algorithms to explain the dementia status. The model results prove the analysis in Chapter 4, that MMSE and nWBV are the most important factors affecting dementia, since they are directly related to the behaviors and situations of patients and brain volume respectively, the increase in both decreases the rate that the patient have dementia. Also, the age, education and gender effects dementia as seen in Chapter 4. The older patients or the less educated ones are more likely to be demented. Also, males are more likely to be demented. The analysis shows that to address the dementia status, both clinical results and demographical information are important.

Table 5.16 Important features based on models

GEE	GLMM	HGLM	GPModel optimizers	GPModel Boosting	GLMMTree	HRF	GLMMLasso
MMSE nWBV Age Education Gender	MMSE nWBV Age Education Gender	MMSE nWBV Age Education Gender	MMSE nWBV Education Gender	MMSE nWBV Visit Age eTIV ASF	MMSE Gender	MMSE nWBV Age Gender Education	MMSE nWBV Age Education Gender SES

5.4 Simulation Study

The simulation study explained in Chapter 4.3 is designed. Firstly, the datasets are simulated separately, and the models are constructed with the covariates commonly seen in Table 5.16, as MMSE, nWBV, education, and gender. For each data type, a simulation function is created that iterates a "for" loop n times. Then, the number of iterations is set to be 1,000 and sample size is changed as desired. In the tables given in Chapter 5.4.1, the model performance metrics of the models constructed using the train and test set of simulated balanced and unbalanced data are given. In each table, the metrics with train and test set of one algorithm is given.

5.4.1 Comparison of the Methods in Simulation Study

In Tables 5.17-5.23, model performance metrics of GEE, GLMM, HGLM, GPBoost, GLMMTree, HRF, and GLMMLasso fitted with balanced dataset are given respectively. Also, in Table 5.24, comparison of the models with balanced dataset when N=500 is given. Tables 5.25-5.31 provide the performance metrics of the same models trained with simulated unbalanced data and Table 5.32 gives the comparison of the models with ubalanced dataset when N=500.

Table 5.17 Model performance metrics of GEE in simulated balanced data

	Generalized Estimating Equations			
Train Set	Accuracy	F1 Score	Sensitivity	Specificity
N=100	**0.724**	**0.660**	0.678	**0.756**
N=150	0.722	0.654	**0.684**	0.748
N=250	0.713	0.652	0.665	0.746
N=500	0.715	0.655	0.671	0.745
Test Set				
N=100	0.674	0.601	0.629	0.722
N=150	0.693	0.623	0.659	0.729
N=250	0.693	0.629	0.645	0.736
N=500	**0.707**	**0.646**	**0.664**	**0.741**

When the sample sizes were changed, slight differences are seen between the models' performance metrics, but in general, the values are very close to each other. The highest values in the train (and test) data are marked in bold in the tables. In GEE (Table 5.17), the model learns slightly better when the number of subjects is 100, but when an unseen data is faced, model with the high sample size gives better metrics.

Table 5.18 Model performance metrics of GLMM in simulated balanced data

	Generalized Linear Mixed Models			
Train Set	Accuracy	F1 Score	Sensitivity	Specificity
N=100	0.723	**0.636**	0.698	**0.737**
N=150	**0.725**	0.634	**0.716**	0.731
N=250	0.714	0.625	0.694	0.725
N=500	0.715	0.629	0.698	0.724
Test Set				
N=100	0.678	0.582	0.653	0.709
N=150	0.697	0.602	0.690	0.714
N=250	0.695	0.604	0.673	0.717
N=500	**0.709**	**0.622**	**0.692**	**0.721**

Similarly, in GLMM (Table 5.18), the model learns better when the number of subjects is 100 and 150 (when N=150, sensitivity is higher compared to the other metrics) but when predictions are obtained with the test set, model with the high sample size gives better metrics.

Table 5.19 Model performance metrics of HGLM in simulated balanced data

	Hierarchical Generalized Linear Models			
Train Set	Accuracy	F1 Score	Sensitivity	Specificity
N=100	0.874	0.846	0.866	0.883
N=150	0.886	0.864	0.868	**0.903**
N=250	0.887	0.863	0.875	0.897
N=500	**0.892**	**0.871**	**0.882**	0.900
Test Set				
N=100	0.672	0.567	0.651	0.702
N=150	0.692	**0.599**	0.682	**0.715**
N=250	0.686	0.584	0.662	0.705
N=500	**0.697**	**0.599**	**0.683**	0.707

However, in HGLM the metrics of the models (Table 5.19) with both train and test predictions are better in higher number of subjects (N=500). This corresponds to our aim to see that the higher the sample size, the better the model performs.

Table 5.20 Model performance metrics of GPBoost in simulated balanced data

	Gaussian Process Boosting			
Train Set	Accuracy	F1 Score	Sensitivity	Specificity
N=100	0.890	0.870	0.859	0.913
N=150	0.893	0.874	0.862	0.918
N=250	**0.896**	0.877	**0.867**	0.919
N=500	0.895	**0.878**	0.863	**0.920**
Test Set				
N=100	0.673	0.596	0.631	0.719
N=150	0.693	0.617	**0.665**	0.725
N=250	0.693	0.624	0.647	0.732
N=500	**0.706**	**0.642**	**0.665**	**0.737**

Table 5.21 Model performance metrics of GLMMTree in simulated balanced data

	Generalized Linear Mixed Models Tree			
Train Set	Accuracy	F1 Score	Sensitivity	Specificity
N=100	0.715	0.646	0.703	**0.768**
N=150	0.722	0.630	**0.729**	0.743
N=250	0.718	0.632	0.711	0.744
N=500	**0.725**	**0.653**	0.704	0.747
Test Set				
N=100	0.636	0.516	0.641	0.686
N=150	0.664	0.523	0.702	0.679
N=250	0.671	0.522	0.699	0.686
N=500	**0.687**	**0.557**	**0.715**	**0.695**

Table 5.20 indicates that the GPBoost models with both N=250 and N=500 sample sizes are good in learning the model, when an unseen data comes, higher sample size is better to predict it.

Table 5.22 Model performance metrics of HRF in simulated balanced data

	Historical Random Forest			
Train Set	Accuracy	F1 Score	Sensitivity	Specificity
N=100	**0.791**	**0.755**	0.744	**0.832**
N=150	0.780	0.737	**0.745**	0.810
N=250	0.767	0.725	0.720	0.806
N=500	0.758	0.717	0.712	0.795
Test Set				
N=100	0.637	0.623	0.571	0.756
N=150	0.652	0.630	0.588	0.755
N=250	0.650	0.636	0.578	0.767
N=500	**0.662**	**0.658**	**0.589**	**0.787**

Table 5.23 Model performance metrics of GLMMLasso in simulated balanced data

	Generalized Linear Mixed Models Lasso			
Train Set	Accuracy	F1 Score	Sensitivity	Specificity
N=100	0.880	0.857	0.854	0.900
N=150	0.872	0.847	0.845	0.894
N=250	**0.886**	**0.865**	**0.861**	**0.905**
N=500	0.818	0.779	0.791	0.836
Test Set				
N=100	0.682	0.601	0.644	0.721
N=150	0.694	0.514	0.670	0.720
N=250	0.699	0.623	0.663	0.729
N=500	**0.711**	**0.640**	**0.678**	**0.735**

Similar to HGLM, GLMMTree (Table 5.21) gives better metrics when N=500 which indicates that with higher sample size, GLMMTree learns and identifies the category of the unseen data better. Even though both are tree-based algorithms, higher metrics of GLMMTree and HRF differs in different sample sizes. In contrast to GLMMTree, HRF (Table 5.22) learns better with a small number of subjects but performs better in predicting the unseen data in a large sample size.

Lastly, it is given in Table 5.23 that GLMMLasso behaves like the previously explained models in terms of test set predictions but it performance better in learning with N=250.

In learning, predictions with the train set, HGLM, GPBoost and GLMMLasso have better metrics with N=500, N=250 and N=250 respectively. In predictions with the test set, GEE, GLMM, GPBoost and GLMMLasso have higher metrics with sample size n=500.

Table 5.24 Model performance metrics for test set with balanced data when N=500

	Accuracy	F1 Score	Sensitivity	Specificity
GEE	**0.707**	**0.646**	**0.664**	**0.741**
GLMM	**0,709**	**0,622**	**0,692**	**0,721**
HGLM	**0,697**	**0,599**	**0,683**	**0,707**
GPBoost	**0,706**	**0,642**	**0,665**	**0,737**
GLMMTree	0,687	0,557	**0,715**	0,695
HRF	0,662	**0,658**	0,589	**0,787**
GLMMLasso	**0,711**	**0,640**	**0,678**	**0,735**

In overall, HGLM, GPBoost and GLMMLasso are found to be appropriate models to train a balanced, longitudinal dataset since they have very similar and higher performances compared to the other models in train predictions. In test predictions, as given in Table 5.24 when N=500, the metrics of all models are generally very close to each other, but still, the accuracies of the models other than GLMMTree and

HRF are higher than them. However, GLMMTree has higher sensitivity, while HRF has higher F1 score and specificity than other models. Considering both train and test accuracies, for such dataset, GLMMLasso, GPBoost and HGLM are appropriate algorithms.

Looking at the performances of the models fitted with balanced data, it can be said that there are very small (1%-5%) differences between the metrics as the sample size changes.

Table 5.25 Model performance metrics of GEE in simulated unbalanced data

	Generalized Estimating Equations			
Train Set	Accuracy	F1 Score	Sensitivity	Specificity
N=100	0.718	**0.656**	0.672	**0.752**
N=150	**0.724**	0.649	**0.680**	**0.752**
N=250	0.706	0.637	0.657	0.739
N=500	0.715	0.651	0.674	0.742
Test Set				
N=100	0.676	0.611	0.633	0.727
N=150	0.701	0.626	**0.669**	0.734
N=250	0.693	0.622	0.647	0.731
N=500	**0.706**	**0.641**	0.664	**0.737**

Table 5.25 indicates that GEE predicts the train set better with 150 subjects but in terms of the predictions with the test set, higher sample size works better. The metrics of GLMM (Table 5.26) are the same as GEE in terms of sample sizes.

Parallel to the HGLM metrics with balanced data, the higher the sample size, the better the HGLM predicts in both train and test set, seen in Table 5.27.

Table 5.26 Model performance metrics of GLMM in simulated unbalanced data

	Generalized Linear Mixed Models			
Train Set	Accuracy	F1 Score	Sensitivity	Specificity
N=100	0.722	**0.635**	0.705	**0.733**
N=150	**0.727**	0.623	**0.723**	0.732
N=250	0.709	0.607	0.692	0.718
N=500	0.713	0.622	0.703	0.719
Test Set				
N=100	0.681	0.588	0.659	0.710
N=150	0.705	0.601	0.709	**0.716**
N=250	0.696	0.592	0.681	0.711
N=500	**0.706**	**0.613**	**0.693**	**0.716**

Table 5.27 Model performance metrics of HGLM in simulated unbalanced data

	Hierarchical Generalized Linear Models			
Train Set	Accuracy	F1 Score	Sensitivity	Specificity
N=100	0.888	0.860	0.879	0.897
N=150	0.907	0.888	0.889	0.923
N=250	0.904	0.882	**0.902**	0.907
N=500	**0.913**	**0.897**	0.899	**0.923**
Test Set				
N=100	0.668	0.565	0.651	0.698
N=150	0.699	0.600	**0.689**	**0.717**
N=250	0.690	0.577	0.672	0.702
N=500	**0.703**	**0.610**	0.686	0.715

Further, in Table 5.28, GPBoost has higher metrics with 150 subjects in learning the dataset but it predicts better with 500 subjects.

Table 5.28 Model performance metrics of GPBoost in simulated unbalanced data

	Gaussian Process Boosting			
Train Set	Accuracy	F1 Score	Sensitivity	Specificity
N=100	0.903	0.887	0.873	0.929
N=150	**0.915**	**0.899**	0.882	**0.941**
N=250	0.913	0.897	0.882	0.937
N=500	0.913	0.898	**0.886**	0.934
Test Set				
N=100	0.673	0.601	0.636	0.717
N=150	0.702	0.617	**0.679**	0.728
N=250	0.694	0.614	0.655	0.725
N=500	**0.705**	**0.633**	0.669	**0.731**

Table 5.29 Model performance metrics of GLMMTree in simulated unbalanced data

	Generalized Linear Mixed Models Tree			
Train Set	Accuracy	F1 Score	Sensitivity	Specificity
N=100	0.719	0.634	0.724	**0.763**
N=150	**0.728**	0.612	**0.751**	0.743
N=250	0.715	0.626	0.704	0.749
N=500	0.719	**0.646**	0.699	0.746
Test Set				
N=100	0.655	0.516	0.691	0.688
N=150	0.671	0.499	**0.727**	0.679
N=250	0.671	0.511	0.700	0.684
N=500	**0.683**	**0.545**	0.714	**0.691**

Table 5.30 Model performance metrics of HRF in simulated unbalanced data

	Historical Random Forest			
Train Set	Accuracy	F1 Score	Sensitivity	Specificity
N=100	**0.792**	**0.753**	**0.751**	**0.834**
N=150	0.782	0.733	0.747	0.812
N=250	0.768	0.729	0.714	0.814
N=500	0.757	0.715	0.710	0.795
Test Set				
N=100	0.637	0.623	0.571	0.756
N=150	0.652	0.630	0.588	0.755
N=250	0.650	0.636	0.578	0.767
N=500	**0.662**	**0.658**	**0.589**	**0.787**

GLMERTree (Table 5.29) has similar results with GPBoost, it learns better with 150 subjects and predicts the unseen data better with 500 subjects. In contrast to the previous two models, HRF (Table 5.30) performs well with a sample size of 100, however, it also performs better in test predictions when the sample size is increased to 500. The distinguishing factor of the GLMMLasso is that its metrics in learning are better at 250 subjects. Similarly, its test prediction metrics are better at a higher sample size (Table 5.31).

In a manner similar to the metrics with the balanced dataset, with the unbalanced dataset, predictions with the train set, HGLM, GPBoost, and GLMMLasso have better metrics with regardless of the sample size since there are only minor differences for different sample sizes.

Table 5.31 Model performance metrics of GLMMLasso in simulated unbalanced data

	Generalized Linear Mixed Models Lasso			
Train Set	Accuracy	F1 Score	Sensitivity	Specificity
N=100	0.865	0.835	0.839	0.885
N=150	0.850	0.815	0.825	0.867
N=250	**0.876**	**0.849**	**0.851**	**0.894**
N=500	0.838	0.804	0.815	0.855
Test Set				
N=100	0.666	0.592	0.632	0.704
N=150	0.703	0.625	**0.680**	0.728
N=250	0.699	0.619	0.664	0.726
N=500	**0.706**	**0.633**	0.673	**0.730**

Table 5.32 Model performance metrics for test set with unbalanced data when N=500

	Accuracy	F1 Score	Sensitivity	Specificity
GEE	**0.706**	**0.641**	**0.664**	**0.737**
GLMM	**0.706**	**0.613**	**0.693**	**0.716**
HGLM	**0.703**	**0.610**	**0.686**	**0.715**
GPBoost	**0.705**	**0.633**	**0.669**	**0.731**
GLMMTree	0.683	0.545	**0.714**	0.691
HRF	0.657	**0.648**	0.587	**0.774**
GLMMLasso	**0.706**	**0.633**	**0.673**	**0.730**

In Table 5.32, predictions with the unseen data when N=500 are given. GEE, GLMM, HGLM, GPBoost and GLMMLasso have higher metrics. However, HRF has the highest F1 score and specificity, and GLMMTree has the highest sensitivity in unbalanced data as well. Considering both train and test accuracies, for a

longitudinal dataset with unbalanced time points, GLMMLasso, GPBoost and HGLM are again found to be appropriate algorithms.

To sum up, before starting the simulation, it was aimed to show that when the sample size increases, the model performance also increases. Also, the models are assumed to work and fit well in the balanced dataset. However, simulation results show that, in both balanced and unbalanced data, common models such as, GPBoost and GLMMLasso have good metrics. However, both train and test predictions do not have a specific pattern for sample size since all the values are very close to each other (the only pattern is that for both data types, all models are slightly more successful in test predictions when the number of subjects is 500). This means that the data being balanced and with high number of subjects do not make any difference in the model that works well.

CHAPTER 6

DISCUSSION AND CONCLUSION

Longitudinal studies are mainly used in health. This study aims to find the proper algorithms to model a longitudinal study with a binary response and small sample size, and find the important features related to the response variable. Therefore, the OASIS-2 dataset on predicting Alzheimer's disease status of older patients is used. Then, a simulation study is conducted to see the effect of the challenges in the real dataset. The dataset has 150 subjects in total with a maximum of five visits each. The clinical results from MRI sessions, examination results about detecting dementia status, and demographical information of patients are given. In the dataset, there are redundant columns with the same observations for each participant and show the same results as other variables. They are eliminated since they would not show any different results for each subject when added to the model. Also, the time points are reduced because very few patients are attending the last two visits and this causes convergence problems and biased results in explaining the effects of covariates to the response. The cleaning and preprocessing are conducted to obtain more precise results in finding the important variables to address the dementia status and more accurate modeling performance.

The simulation study is designed to have a similar structure to the real dataset. Specifically, two types of datasets are created with an equal number of time points in each subject.Four different subject sizes are trained in both datasets to see the effect of sample size and balance of the time points in the model performances.

Classical and machine learning techniques are utilized first on the OASIS-2 dataset to classify patients' dementia status, then the same models are applied to the simulated datasets by setting the common important variables obtained in the models of the real dataset. Due to the study being longitudinal, algorithms considering the

subject-based covariance and random effects variability are applied. This means the selected models are mainly a version of GLMM. Model performances are measured using accuracy, F1 score, sensitivity, and specificity, and the results are interpreted by using the odds ratio. The models' performance metrics on the test set are crucial to evaluate classification performance with new data. Therefore, before modeling the datasets are split into two as train and test, and the performances in both sets are evaluated. The models with higher performance metrics on the test set are better.

Before starting the modeling, a feature selection is applied by the Boruta algorithm and the confirmed features are found as MMSE, nWBV, education, gender, eTIV, ASF, SES, and age in order of importance, and the visit variable is rejected. The results are almost consistent with the study of Leong and Abdullah (2019), where in their study, the Boruta algorithm gives ASF, eTIV, nWBV, MMSE, and CDR variables as primary features in the OASIS-2 dataset. The difference may be because of our preprocessing and creating a new response by using CDR and eliminating CDR in modeling. The results of Boruta are considered in variable elimination in modeling.

Research and the visualizations in Chapter 4 show that the increase in MMSE, nWBV, and education has numerically adverse effects on dementia status. Also, an increase in age is assumed to show more probability that the patient is demented, and males are more likely to be demented. The results of the marginal model (GEE), random effects model (GLMM), and hierarchical GLM are similar regarding significant variables such as MMSE, nWBV, education, gender, and age. The effect of MMSE, nWBV, education, and age is numerically negative on the response. Moreover, in GPBoost and GLMMLasso, the effects of the common significant variables MMSE, nWBV, education, and gender are numerically negative. Additionally, age and SES seem significant in GLMMLasso with numerically negative effects on the dementia status. These results indicate that

- the increase in MMSE score means the patient is less likely to be demented,

- the increase in normalized whole brain volume indicates a less probability to be demented,
- increase in socioeconomic status results in the patient having less probability to be demented,
- males are more likely to be demented,
- age has a negative impact on being demented,
- the increase in education years results in the patient having less probability of being demented.

Surprisingly, in the literature, females are more likely to be demented. For instance, almost 66% of Alzheimer's cases are females, according to the Alzheimer's Association (2018). But this data set provides a reverse relationship. Similarly, age has a negative effect on dementia in the literature, which means the increase in age results in the patient being more likely to be demented (Tucker and Stern, 2011). However, in this study, age seems to have a negative effect on dementia status, such as an increase in age results in an increase in the probability of being nondemented. The article by Gardner et al. (2013) states that a slower decline is possible at a higher age at onset. In our study, the reversed relationship might be because of the scope of the OASIS-2 dataset. The dataset is only for elderly people, which means the ages of the people in the study are already old. Among these patients, the effect of age on dementia status is not identified, even though the variable is significant in most models. Raket (2020) also highlighted a similar finding (Raket, 2020).

Interpreting the model performances on the real dataset, GPBoost models outperform others with 97% accuracy in classifying the dementia status in the train set, and they have faster computational times as 0.008 and 0.014 seconds. However, they perform worse on the test set, even the GPModel with boosting indicates overfitting with 46% accuracy in test set. Tree-based models, especially HRF, perform better on test sets with 89% accuracy for small-sampled longitudinal data analysis. While GPBoost performs better than other methods in learning the real dataset, HRF works better for an upcoming new subject. The essential features to address the dementia status are deduced as MMSE, nWBV, age, education and gender of the patients. The analysis

indicates that clinical results and demographic information are vital in addressing dementia status, and hybrid models are efficient in modeling longitudinal data with binary response.

Prior to the simulation, the objective was to demonstrate that model performance increases with sample size. It was also assumed that the models perform well on balanced datasets. However, simulation results show no specific pattern about sample sizes and the equality in the number of time points. Regardless of balance of the dataset and the number of subjects, HGLM, GPBoost, and GLMMLasso algorithms are appropriate for such simulated datasets. The test prediction accuracies of HGLM, GPBoost, and GLMMLasso for balanced dataset with N=500 are 69.7%, 70.6%, and 71.1%, respectively; for unbalanced dataset, they are 70.3%, 70.5%, and 70.6% respectively.

The study has some limitations, mainly because of the dataset and design of the study considering the structure of the dataset. The OASIS-2 dataset is the one with a small sample size and irregularity in time points. Therefore, it is hard to find proper models and obtain remarkable results. In the literature, no such study uses only the OASIS-2 dataset for analysis. This was the motivation for this study and the reason why this dataset was used. Because in real life, especially in health studies, not all the datasets are high dimensional, and have a balanced structure. This thesis aims to identify the algorithms applicable to a longitudinal dataset with binary response, small sample size, and irregular time points and extract the important features in the classification of the response.

As a second limitation, some contradictions are observed in the results of real dataset and simulation. HRF and GLMMTree are the best algorithms in the analysis with real dataset but their performances in simulation studies are worse than others. This is because of the design of the simulation. Hyperparameters and control parameters could not be tuned in the simulation, since, in each iteration, different parameters would be the best ones. Besides being a time-consuming procedure, finding and applying the best parameters for each 1000 iteration are also impossible. However,

they are tuned with the real dataset; the performances are better than the simulation, and the results differ from the simulation.

As a last limitation, because of the small number of people in the study, the number of observations in test and validation are very small when the dataset is split into three as train, validation and test. It causes overfitting and convergence problems in many of the models. Therefore, adding the validation set is abandoned; tuning is done with the train set. This solved the convergence and overfitting problems. The overfitting problem is seen only in the GPBoost model, and this is because hyperparameter tuning is completed by using the train set, and train prediction metrics are very high, but test prediction metrics are worse.

In future studies, OASIS-3 dataset can be used for the classification of dementia status, and the same models can be applied. Also, based on longitudinal data with binary response, the implemented algorithms can be improved by integrating the random effects into state-of-the-art algorithms such as SVM, KNN, XGBoost, Ridge, and Elastic Net. Also, the OASIS-2 dataset can be analyzed by using deep learning algorithms.

REFERENCES

Alasadi, S. A., & Bhaya, W. S. (2017). Review of data preprocessing techniques indata mining. Journal of Engineering and Applied Sciences, 12(16), 4102-4107.

Alzheimer's Association. (2018). 2018 Alzheimer's disease facts and figures. Alzheimers Dement. 14, 367–429.

Alzheimer's Association. (2023.). What Is Alzheimer's? Brain Tour Part 2. URL: https://www.alz.org/alzheimers-dementia/what-is-alzheimers/brain_tour_part_2

Arevalo-Rodriguez I, Smailagic N, Roqué i Figuls M, Ciapponi A, Sanchez-Perez E, Giannakou A, Pedraza OL, Bonfill Cosp X, Cullum S. Mini-Mental State Examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI). Cochrane Database of Systematic Reviews 2015, Issue 3. Art. No.: CD010783.

Asar, Ö. (2012). On multivariate longitudinal binary data models and their applications in forecasting (Master's thesis, Middle East Technical University).

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? International Journal of Methods in Psychiatric Research, 20(1), 40-49.

Ballinger, G. A. (2004). Using generalized estimating equations for longitudinal data analysis. Organizational research methods, 7(2), 127-150.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Bhushan, I., Kour, M., Kour, G., Gupta, S., Sharma, S., & Yadav, A. (2018). Alzheimer's disease: Causes & treatment–A review. *Ann Biotechnol*, *1*(1), 1002.

Brickman AM, Honig LS, Scarmeas N, et al. Measuring Cerebral Atrophy and White Matter Hyperintensity Burden to Predict the Rate of Cognitive Decline in Alzheimer Disease. Arch Neurol. 2008;65(9):1202–1208.

Buyrukoğlu, S. (2021). Early detection of Alzheimer's disease using data mining: Comparison of ensemble feature selection approaches. Konya Journal of Engineering Sciences, 9(1), 50-61.

Caruana EJ, Roman M, Hernández Sánchez J, Solli P. (2015). Longitudinal studies. J Thorac Dis, 7(11), E537-E540.

Cakar, S., & Yavuz, F. G. (2023). Hybrid statistical and machine learning modeling of cognitive neuroscience data. Journal of Applied Statistics.

Çakar, S. (2022). Longitudinal data analysis with statistical and machine learning methods in neuroscience. (Master's thesis, Middle East Technical University).

Feng Zhou, Areen Alsaid, Mike Blommer, Reates Curry, Radhakrishnan Swaminathan, Dev Kochhar, Walter Talamonti & Louis Tijerina (2022) Predicting Driver Fatigue in Monotonous Automated Driving with Explanation using GPBoost and SHAP, International Journal of Human–Computer Interaction, 38:8, 719-729,

Fitzmaurice, G. M., & Ravichandran, C. (2008). A primer in longitudinal data analysis. Circulation, 118(19), 2005-2010.

Fokkema, M., Smits, N., Zeileis, A. et al. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. Behav Res, 50, 2016–2034.

Fokkema, M., Zeileis, A., & Fokkema, M. M. (2019). Package 'glmertree'.

Fulton, L. V., Dolezel, D., Harrop, J., Yan, Y., & Fulton, C. P. (2019). Classification of Alzheimer's Disease with and without Imagery Using Gradient Boosted Machines and ResNet-50. Brain Sciences, 9(9), 212.

García, S., Luengo, J., & Herrera, F. (2015). Data preprocessing in data mining (Vol. 72, pp. 59-139). Cham, Switzerland: Springer International Publishing.

Gardner, R. C., Valcour, V., & Yaffe, K. (2013). Dementia in the oldest old: A multi-factorial and growing public health issue. Alzheimers Res. Ther., 5, 27.

Gentle, J. E., Haerdle, W., & Mori, Y. (2004). Handbook of Computational Statistics. Heidelberg: Springer - Verlag.

Groll, A., & Tutz, G. (2014). Variable selection for generalized linear mixed models by L1-penalized estimation. Statistics and Computing, 24, 137-154.

Hayati, M., & Muslim, A. (2021). Generalized linear mixed model and lasso regularization for statistical downscaling. Enthuastic Int. J. Stat. Data Sci, 1(1), 36-53.

Hayati, Marwah & Muslim, Agus. (2021). Generalized Linear Mixed Model and Lasso Regularization for Statistical Downscaling. Enthusiastic: International Journal of Applied Statistics and Data Science. 1. 36-52. 10.20885/enthusiastic.vol1.iss1.art6.

Højsgaard, S., Halekoh, U., & Yan, J. (2005). The R Package geepack for Generalized Estimating Equations. Journal of Statistical Software, 15(2), 1–11.

Hu, J., & Szymczak, S. (2023). A review on longitudinal data analysis with random forest. Briefings in Bioinformatics, 24(2), bbad002.

Iddi, S., Li, D., Aisen, P. S., et al. (2019). Predicting the course of Alzheimer's progression. Brain Inf., 6, 6.

Ilk, O. (2004). Exploratory multivariate longitudinal data analysis and models for multivariate longitudinal binary data. Iowa State University.

İnan, G. (2009). A simulation study on the comparison of methods for the analysis of longitudinal count data (Master's thesis, Middle East Technical University).

Jenkins R, Fox NC, Rossor AM, Harvey RJ, Rossor MN. (2000). Intracranial Volume and Alzheimer Disease: Evidence Against the Cerebral Reserve Hypothesis. Arch Neurol. 57(2), 220–224.

Jürg Schelldorfer, Lukas Meier & Peter Bühlmann (2014) GLMMLasso: An Algorithm for High-Dimensional Generalized Linear Mixed Models Using ℓ1-Penalization, Journal of Computational and Graphical Statistics, 23:2, 460-477.

Karadağ, Ö., & Altunay, S. A. (2018). Generalized estimating equations for genetic association studies of multi-correlated longitudinal family data. *Gazi University Journal of Science*, *31*(1), 273-280.

Karpievitch YV, Hill EG, Leclerc AP, et al. (2009). An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of RF++. PLoS One, 4(9), e7087.

Kavitha C, Mani V, Srividhya SR, Khalaf OI, & Tavera Romero CA (2022) Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models. Front. Public Health 10:853294.

Kuha, J. (2004). AIC and BIC: Comparisons of Assumptions and Performance. Sociological Methods & Research, 33(2), 188–229.

Kursa, M. B., Jankowski, A., & Rudnicki, W. R. (2010). Boruta–a system for feature selection. Fundamenta Informaticae, 101(4), 271-285.

Kursa, M. B., Jankowski, A., & Rudnicki, W. R. (2010). Boruta–a system for feature selection. *Fundamenta Informaticae*, *101*(4), 271-285.

Lane, C. A., Hardy, J., & Schott, J. M. (2018). Alzheimer's disease. European Journal of Neurology, 25(1), 59-70.

Lee Kuok Leong and Azian Azamimi Abdullah (2019). J. Phys.: Conf. Ser. 1372 012065

Lee, G., Nho, K., Kang, B., Sohn, K., & Kim, D. (2019). Predicting Alzheimer's disease progression using multi-modal deep learning approach. Scientific Reports, 9(1), 1-12.

Little, R. J., & Rubin, D. B. (2019). Statistical analysis with missing data (Vol. 793). John Wiley & Sons.

Marcus DS, Fotenos AF, Csernansky JG, Morris JC, Buckner RL. (2010). Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. Journal of cognitive neuroscience, 22(12), 2677-2684.

Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL. (2007). Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle-aged, nondemented, and demented older adults. Journal of cognitive neuroscience, 19(9), 1498-1507.

McCullagh, P., & Nelder, J. A. (1989). Generalized Linear Models, London: Chapman & Hall.

Molenberghs, G., & Kenward, M. G. (2010). Semi-parametric marginal models for hierarchical data and their corresponding full models. Computational Statistics & Data Analysis, 54(2), 585-597.

Morchid, M., Karita, Y., Mellebeek, B., Dugast, C., and Besacier, L. (2017). Multilingual Automatic Speech Recognition: Challenges, Opportunities, and Future Directions. ACM Trans. Speech Lang. Process. 14, 4, 2:1–2:18.

Morris JC. (1997). Clinical dementia rating: a reliable and valid diagnostic and staging measure for dementia of the Alzheimer type. Int Psychogeriatr. 9 Suppl 1, 173-6.

Murchison, C. F. (2021). *Applications of Longitudinal Machine Learning Methods in Multi-Study Alzheimer's Disease Datasets* (Doctoral dissertation, The University of Alabama at Birmingham).

Nakai, M., & Ke, W. (2011). Review of the methods for handling missing data in longitudinal data analysis. International Journal of Mathematical Analysis, 5(1), 1-13.

Newman, D. A. (2003). Longitudinal Modeling with Randomly and Systematically Missing Data: A Simulation of Ad Hoc, Maximum Likelihood, and Multiple Imputation Techniques. Organizational Research Methods, 6(3), 328–362.

Patrician, P.A. (2002). Multiple imputation for missing data. Res. Nurs. Health, 25(1), 76-84.

Patro, S. G., & Sahu, K. K. (2015). Normalization: A Preprocessing Stage. ArXiv. /abs/1503.06462

Raket LL (2020) Statistical Disease Progression Modeling in Alzheimer Disease. Front. Big Data 3:24.

Roderick J. A. Little (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values, Journal of the American Statistical Association, 83:404, 1198-1202.

Rönnegård, L., Shen, X., Alam, M. (2010). Hglm: A package for fitting hierarchical generalized linear models. The R Journal, 2(2), 20-28.

Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. Proceedings of the survey research methods section of the American Statistical Association, 1, 20-34.

Saratxaga, C. L., Moya, I., Picón, A., Acosta, M., Moreno-Fernandez-de-Leceta, A., Garrote, E., Bereciartua-Perez, A. (2021). MRI Deep Learning-Based Solution for Alzheimer's Disease Prediction. J. Pers. Med., 11(9), 902.

Sexton J. htree: historical tree ensembles for longitudinal data. R package version 0.1.1, 2017.

S. Basheer, S. Bhatia and S. B. Sakri, (2021). Computational Modeling of Dementia Prediction Using Deep Neural Network: Analysis on OASIS Dataset. IEEE Access, 9, 42449-42462.

Stańczyk, U., & Jain, L. C. (2015). Feature selection for data and pattern recognition: An introduction (pp. 1-7). Springer Berlin Heidelberg.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B, 58, 267–288.

Tucker, A. M., and Stern, Y. (2011). Cognitive reserve in aging. Curr. Alzheimer Res. 8, 354–360.

Turgal, E. (2022). Evaluation of Traditional Machine Learning and Mixed Effect Machine Learning Model Performances by Simulation Study. Türkiye Klinikleri Biyoistatistik, 14(3), 147-157.

Vujović, Ž. (2021). Classification model evaluation metrics. International Journal of Advanced Computer Science and Applications, 12(6), 599-606.

Wang, L., Li, P., Hou, M., Zhang, X., Cao, X., & Li, H. (2021). Construction of a risk prediction model for Alzheimer's disease in the elderly population. BMC Neurology, 21.

Weiss, R. E. (2005). Modeling longitudinal data (Vol. 1). New York: Springer.

Zeger, S. L., Liang, K. Y., & Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. Biometrics, 1049-1060.

Zhang, Z. (2016). Multiple imputation with multivariate imputation by chained equation (MICE) package. Annals of Translational Medicine, 4(2).