UKRAINIAN CATHOLIC UNIVERSITY

MASTER THESIS

Modeling and Prediction of Alzheimer's Disease Progression

Author:
Sevil SMAILOVA

Supervisor:
Dr. Igor KOVAL

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2021

Declaration of Authorship

I, Sevil SMAILOVA, declare that this thesis titled, "Modeling and Prediction of Alzheimer's Disease Progression" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Master of Science

Modeling and Prediction of Alzheimer's Disease Progression

by Sevil SMAILOVA

Abstract

Alzheimer's Disease is an irreversible disease that causes a decline in cognitive abilities and leads to dementia. Many efforts are applied to understand the behavior of the disease progression and foresee its future state. The metrics that assess the level of cognition are named as cognitive scores. The dynamics of cognitive scores help understand the future disease progression. However, there is a lack of understanding on what is the best benchmark for the predicted value of the cognitive score. Moreover, there could be cases when the future value of the cognitive score is not statistically different comparing to the current value.

In this work we discover those patients that by design cannot have the dynamics in their progression of cognitive scores. We justify that the dynamics of progression for Cognitively Normal patients do not change over five years. We reveal that there is no statistically significant change in progression after the 1-year follow-ups. We unified the evaluation framework of different imputation, feature selection methods and machine learning models on different time to prediction settings as well as on different patient populations.

Acknowledgements

I would like to thank my scientific advisor Igor Koval, who helped and guided me in this research work. I am thankful to Ukrainian Catholic University, Faculty of Applied Science and to Oleksii Molchanovskyi for curating the MS Data Science Program in Lviv. This work would not have been done without help of Ring scholarship program that covered my tuition fees.

Contents

List of Figures

List of Tables

List of Abbreviations

AD	**A**lzheimer's **D**isease
ADAS	**A**lzheimer's **D**isease **A**ssessment **S**cale
CN	**C**ognitive **N**ormal
MMSE	**M**ini-**M**ental **S**tate **E**xamination
MRI	**M**agnetic **R**esonance **I**maging
MCI	**M**ild **C**ognitive **I**mpairment
PET	**P**ositron **E**mission **T**omography

Chapter 1

Introduction

1.1 Domain Overview

Alzheimer's Disease (AD) is an irreversible neurodegenerative disease that causes memory, language, orientation and other cognitive disorders. 60-80% of all dementia cases belong to AD. The number of deaths from AD more than doubled for the last decade, increasing 145.2% (*Alzheimer's Association - Facts and Figures*).

AD evolves monotonously through a long period, often longer than observation at an individual scale. The pathological processes occur gradually, they develop before the first deterioration of cognitive disabilities and lead to dementia - the end stage of the disease.

Data that captures the natural history of patient visits reveals patterns of disease progression. It is a longitudinal medical data that comprises repeated measurements at multiple time points per individual. The periods between each patient's visits can vary significantly. Real-world applications and the respective data often consist of delayed observations because patients do not visit medical institutions over fixed periods. The nature of visits frequency and the overall duration of visits is heterogeneous and leads to the occurrence of missing data.

The variety of measurements helps keep track of patients' disease order. The list of such measurements has a multi-modal origin and includes cognitive test results, chemical and imaging biomarker measurements, demographic data such as age, years of education, etc.

Cognitive tests are the assessments that evaluate the severity of cognitive dysfunctions. Some tests measure several cognitive abilities, while others specialize in particular cognition. Two main tests can be classified into the first category - Mini-Mental State Exam (MMSE) and Alzheimer's Disease Assessment Scale (ADAS). While MMSE assesses orientation, attention, memory, language skills, the ADAS test, along with cognitive abilities such as memory, language and praxis, measures noncognitive behavioral dysfunctions, such as mood state and behavioral changes (Rosen, Mohs, and Davis, 1984).

The next type of measurement that assesses the physiological state of patients is biomarkers. These biomarkers are chemical and imaging indicators that reflect the disease's progression. Their values can be obtained from cerebrospinal fluid puncture (CFS), amyloid positron emission tomography (PET), magnetic resonance imaging (MRI), diffusion tensor image (DTI).

Dementia is the final observable stage that is caused by multiple pathological brain disorders (Jack et al., 2010). The conventional clinical disease stages include three phases:

1. CN - pre-symptomatic phase of cognitively normal patients with some AD pathological changes.

2. MCI - mild cognitive impairment where patients are observed with the onset of cognitive symptoms or already have cognitive disabilities but do not have dementia.

3. AD - dementia that is characterized by cognitive impairments of multiple origins.

It is essential to understand that AD is an irreversible neurodegenerative disease. The current treatment can only slow down the illness and can not cure it. Understanding the future trajectories of monotonously changing cognitions will help apply therapeutic interventions effectively on patients depending on the severity of their future progression. That is why it is crucial to know the tempo of disease progression and predict its future state on the patient level.

As described, the cognitive scores are measured by clinical assessments such as ADAS and MMSE. They are the quantitative measurements of the main types of cognition. The prediction of cognitive scores gives an understanding of the future progression of cognitions. According to the results of The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge, which summarized the outputs of 90 algorithms of the cognitive scores predictions was not satisfying. The best algorithms of 5-year predictions of ADAS-Cog13 cognitive score did not outperform the results of random guessing (constant predictions) (Marinescu et al., 2020). To our best knowledge, the previous researches did not perform better; specifically, they did not show statistically better results within the 5-year time frame.

1.2 Motivation

The current state of the research field is quite heterogeneous. Many researchers are trying to tackle specific problems, such as missing values, delayed observations, etc. Many of them apply sophisticated models/methods that make the most accurate prediction, but on the other hand, they do not take into account the presence of noise in the data. Hence, the models with the most accurate predictions predict cognitive scores for patients, which by design can not have the progression in their cognition. We call these cohorts of patients with or without progression as the scenario of interest. With this research, we aim to investigate which scenarios of interest give us room for prediction and which do not because patients do not progress over time. We also investigate how the existing approaches that work well on the scenarios without progression behave in the scenarios with more aggressive dynamics in cognition.

1.3 Thesis Structure

Chapter 2 represents an overview of the related work for the problem of predicting cognitive scores. It explains the current state of research problems in the field of missing data, delayed observations. However, we will show that they do not tackle problems with population selection, noise in the data and time to prediction.

Chapter 3 discusses the experimental pipeline that aims to investigate which scenarios of interest are subjects for prediction and which models and methods are most accurate for the current task.

Chapter 4 shows the findings of the proposed approach. It reveals scenarios of interest within which we cannot observe any progression in cognitive scores by

design and those scenarios that are subject for predictions. Additionally, we show which methods and models improve the prediction accuracy.

Chapter 5 summarizes the key results of the research and discusses the future direction of work. Here we show that there are cohorts of cognitively normal patients or patients with early symptoms of cognitive decline whose progression of cognition does not change over time. We also reveal that there is no space for improvement for short-term predictions.

Chapter 2

Related Work

The current chapter reveals the related work done in the field of prediction of cognitive scores. The task of prediction of cognitive scores is related to the regression problem. The most recent researches in the domain of medical data and particularly the studies of AD examine various methods and strategies to make accurate predictions of cognitive scores and other predictive cofactors. They tackle difficulties with missing data and delayed observations (patient's visits), incorporate historical dependencies between observations. We provide an overview of the current state of the considered research field as well as highlight the proposed goal and contribution with formulation of the research direction.

2.1 Literature review

2.1.1 Different populations

The recent research papers utilize plenty of methods to predict the future dynamics of biomarkers and the level of cognition measured by clinical assessments. It can be statistical, machine learning, and deep learning models that aim to receive the most accurate prediction. It is known that there are no cognitive symptoms at the preclinical stage of the disease. Hence, the prediction precision of cognition for patients without consideration of the specifics of disease stages could be inflated.

Some approaches relate to the discovery of early changes of biomarkers of CN patients that precede the onset of cognitive symptoms (Jack et al., 2010). (Nguyen et al., 2020) attempted to broke down the patients into cohorts based on the disease stage, but the prediction was made without an analysis of dynamics within these cohorts. There are researches that do not consider different population cohorts and make predictions for the whole observed population (Ghazi et al., 2019).

2.1.2 Time to prediction

The other side of the prognosis setting is the time to prediction or, in other words, the prediction horizon. It is a setting that each researcher specifies differently. (Nguyen et al., 2020) shows results of prediction of ADAS-Cog13 and Ventricles volume 6 years ahead but does not consider those scenarios with populations that cannot have progression by design. (Ghazi et al., 2019) apply modified LSTM to predict MRI biomarkers maximum of the year ahead. (Zhang, Shen, and Initiative, 2012) consider two scenarios. The first is the prediction of ADAS and MMSE values for 24 months starting from the first month; hence the time to prediction is 2 years. In the second scenario, the authors predict the 24-th month visit based on data for the 6, 12 and 18 months patients' visits. In this case, the time to prediction is 6 months.

2.1.3 Missing data

The methods that are solving a regression problem can use either the single past observation or sequences of multiple observations. However, the difficulties come not only from catching the historical dependencies among different factors but also within the data that has to provide the long trail of records within measurements of AD patients. To make "static" predictions one-time point ahead based on the single past patient visit requires two data sequences of measurements. To introduce the dynamics into both features and target sequences, the data for analysis has to contain these observations for the required time period. To consider these observations as a sequence, the patients have to constantly visit a hospital for a long time within the same time range, for example, on a yearly basis. However, in the real world, that does not happen often. The patient's visits could be shifted in time, or even there could be no visits for specific periods. Since most machine learning models require feature-complete data, there is a need for those approaches that handle missing data.

The researchers that examine the problem of missing data apply different strategies to overcome it. Some methods impute those missing values using various imputation techniques, such as forward filling, where the missing values are imputed by the last available value; linear interpolation, where values linearly interpolated between the last and next available values and model filling, where the Machine Learning model is responsible for the value imputation (Nguyen et al., 2020). Other methods incorporate missing values as model parameters. (Ghazi et al., 2019) utilize LSTM that incorporate missing values into the neural network architecture. Authors replace missing values with zeros and apply normalization on weights of neural network that takes into account missing values in input and target. Afterward, they backpropagate zero errors for the target with missing values.

2.1.4 Models

There are plenty of papers that apply different statistical, machine learning, or deep learning models to predict the future outcome of cognitive scores or other measurements of AD. Some models make "static" predictions one-time point ahead based on a single past patient's visit. In other papers, researchers take the sequence of past visits and predict either a one-time point ahead or a sequence of future values. (**u**)tilizes SVM to predict MMSE and ADAS. Some researchers utilize the RNN to predict the future value of cognitive scores (Ghazi et al., 2019; Nguyen et al., 2020).

2.1.5 Evaluation

The important part of modeling is its evaluation. On the one hand, the model with the best performance does not have to predict better than the amount of noise in the data. On the other hand, the prediction has to be comparable to the state-of-the-art results. The researchers compete within the second category of evaluation strategy (Nguyen et al., 2020; Ghazi et al., 2019) and do not take into account the presence of noise in the data.

By design (Koval et al., 2021; Clark et al., 1999) cognitive assessments and their respective cognitive scores have measurement errors and variation in an annual score change. The patient that performs the same task within the same week could have different scores, but it will be not due to the disease progression but due to the external noise. The imaging biomarkers also contain the noisy part in their estimations (Koval et al., 2021) due to variations in the acquisition protocol, such as lightning, the patients' movements at the moment of obtaining the MRI scans, etc.

2.2 Research objectives and contribution

Considering the existing limitations discussed in motivation and literature overview the research goal of this paper is to find an ideal set-up of prediction such as the constant prediction is statistically higher then the noise in the data. This will be an indicator of changes in patient's cognitive scores progression.

The contribution of the research is formulated in a question-list below.

1. What is an ideal scenario of prediction looks like?
   1.1. Selection of patients of interest.
   1.2. Selection of time to prediction.
2. Within the ideal scenario of prediction how much accurate prediction we can receive,
   2.1. when applying different imputation techniques?
   2.2. selecting different features?
   2.3. applying different prediction methods?

Chapter 3

Methodology

In the following section we explain the steps we followed to answer two questions:

1. Which populations is meaningful to use to predict the disease progression?
2. Which methods are most accurate to make these predictions?

First, we start with data exploration. Then we discuss the ways of segregation between the populations and methods to chose the most appropriate ones. After it, we examine the methods that will be applied to improve prediction for a chosen population. In the end, we reveal an experimental protocol that explains the pipeline we followed to test our hypothesis.

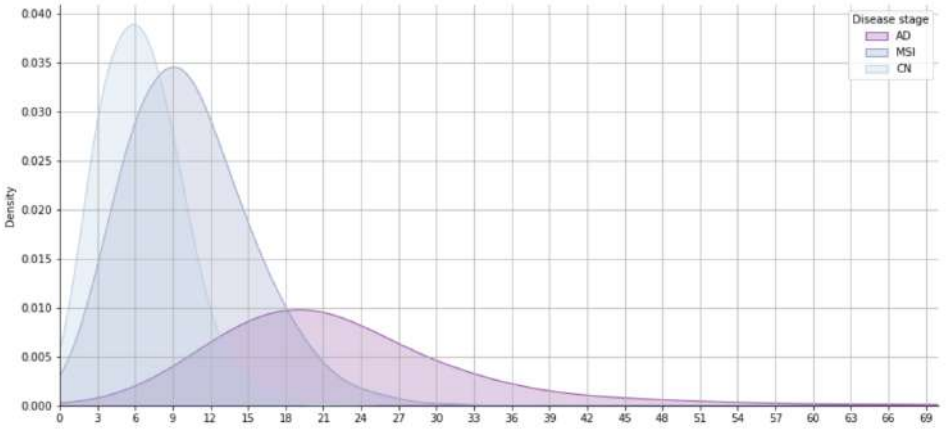
3.1 Study Data

Dataset was obtained from The Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The ADNI dataset contains the biomarkers of the following measurement subgroups: MRI, PET, CSF, diffusion tensor imaging (DTI), cognitive tests, some genetic and demographic information. Each row of the dataset represents data for one particular patient visit, and each column is a feature or a measurement. It comprises data for the 2268 patients over 13578 patient visits.

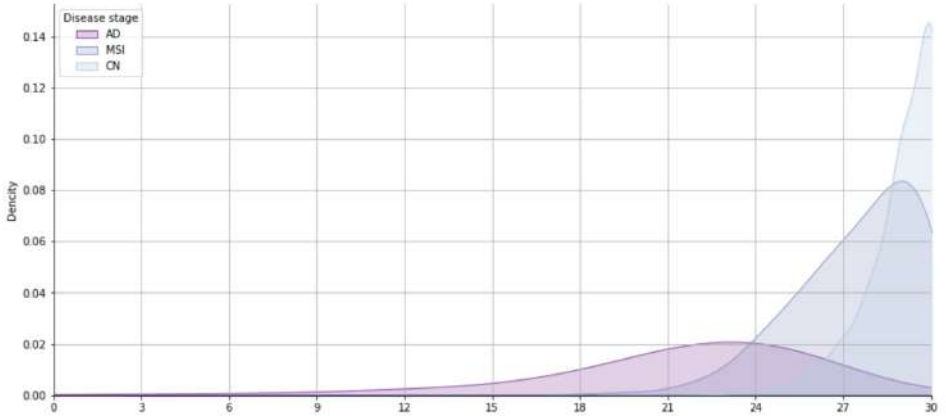
The cognitive scores of ADAS and MMSE give an understanding of the quantifiable dynamics of disease progression. They are measured by clinical experts. The tests include questions on which patient answers and gain scores. Each test has its own scale. MMSE assesses the level of cognitive dysfunction at a scale from 0 to 30 points, where 30 points is an indication that the patient doesn't have cognitive disabilities. The ADAS estimates patients at a scale from 0 to 70 points, where 0 points is an indication that a person does not have cognitive dysfunctions. The distribution of MMSE and ADAS (Figure 3.1) shows that distributions of cognitive scores on different disease stages overlap. Patients with MMSE values between 27 to 30 can be classified as CN or MCI or even AD.

Cognitive score	Disease Stage	Mean	STD
ADAS	CN	6.417	3.199
	MCI	10.329	5.080
	AD	22.183	9.762
MMSE	CN	29.042	1.216
	MCI	27.499	2.250
	AD	21.602	4.661

TABLE (3.1) Descriptive statistics for ADAS and MMSE.



(A) ADAS.



(B) MMSE.

FIGURE (3.1) Distribution of ADAS and MMSE within disease stages.

3.1.1 Data preprocessing

Since the absolute values of morphological features of the brain vary due to different patients' head sizes, it is a standard practice to normalize them by intracranial volume (ICV) (Sargolzaei et al., 2015; Voevodskaya et al., 2014). Hence, the MRI and DTI volumetric biomarkers were normalized by the corresponding patient's ICV. Then data were normalized within the min-max scaler algorithm.

Before running any machine learning model, some columns were removed from the feature space as their inclusion could have led to data leakage: columns indicating the disease stage and age at the first patient visit.

3.2 Scenario of interest

The original distribution of patient visits is shown in Figure 3.3 within the first light purple bars. The second darker bar shows the amount of resampled data. One way to take data as an input is to use the first patient's visit as a feature value and each next visit as a target. However, in this case, we are tied to the patients' first visit. Suppose the patient has visited the hospital 5 consecutive years. In that case, there will be one observation with a one-year follow-up starting from the first patient's visit, and we will not consider cases with 1-year follow-ups, starting the second and subsequent patient's visits. To avoid such scenarios, we resampled the patients'

visits. We took each patient's visit and corresponding subsequent next patient's visit (1-year or n-years ahead) as a target no matter if the first observation coincides with the first visit to the hospital or not (Figure 3.2). After the resampling process, we increased the amount of input data that is shown as the second bar in Figure 3.3.

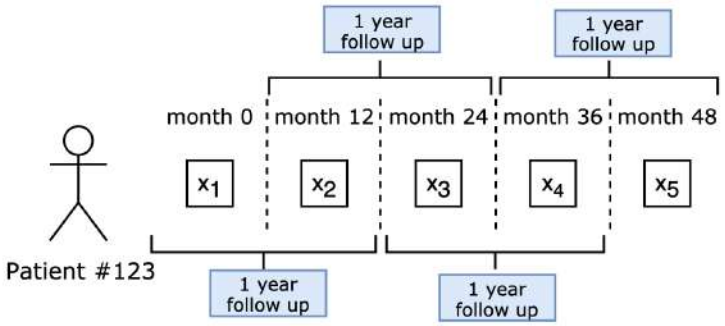


FIGURE (3.2) Generating one year follow-up observations where $x_1, ..., x_5$ are the measurements of AD.

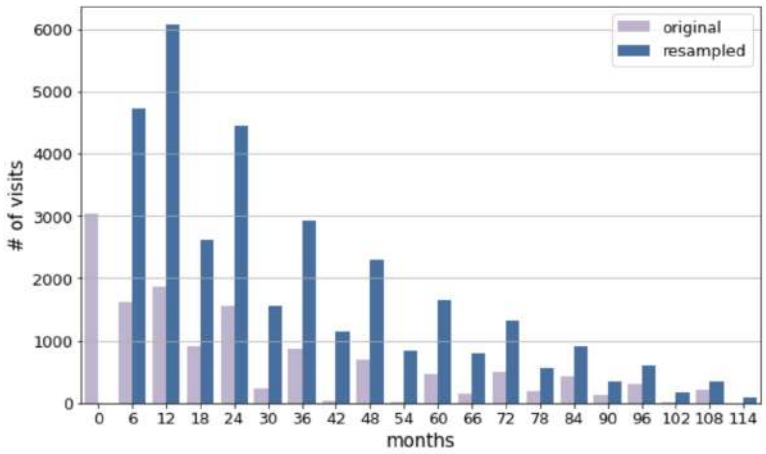


FIGURE (3.3) Distribution of patients' visits during 9 years follow-ups.

We introduce the upper and lower boundaries for the models' error estimations to select the right setting for prediction. The upper boundary represents the baseline model error, the lower bound shows the distribution of noise. The baseline error is the error of the model with constant prediction. This constant prediction is not the trainable model; it is a simple adjustment for which the prediction for the target variable is obtained by increasing its value by 1. The prediction setting is considered meaningful if the distributions of noise and constant prediction error are statistically different. If distributions are not statistically different, then the model catches the noise. That means that we cannot predict better than the baseline model since it is already the best possible prediction and this best prediction is noise itself. The distribution of noise is the change in target value within the six months follow-up. Such distribution is called test-retest error.

Time to prediction is a crucial setting that helps to understand which point in a future is meaningful to predict and then how far we can predict accurately. As we show, there are setting for "short time" predictions where the distributions of test-retest and constant prediction do not not statistically differ.

To chose the population of interest, we select a cohort of patients from the general distribution of patients based on some filters. To filter population, we use the disease stage or value of cognitive score used to differentiate between patients' disease stages. Different populations of interest have different progression of cognitive scores. We will show which of them do not have the signs of progression and those which progression has to be tackled further.

3.3 Prediction methods

3.3.1 Feature Selection

We tried different approaches to select the features with the highest prediction power:

1. Based on TADPOLE challenge.
2. Correlation based.
3. Boruta feature selection.

The first cohort of features was taken from the recommendations from a TADPOLE challenge and consist of 14 measurements. The correlation-based method took features that the most correlated within the targets.

The Boruta feature selection method is a wrapper for a Random Forest algorithm. First, the original dataset was augmented with its shuffled copy, so each original column received its shuffled copy. Then Random Forrest Classifier is trained on the augmented dataset. Then, the original feature is considered as important if its importance value is higher than the importance of the most important augmented feature (Kursa and Rudnicki, 2011).

The list of features selected in each method could be found in Table 3.2.

Feature Selection Type	Cognitive Scores	MRI	PET	CFS	Other
TADPOLE based	CDRSB, ADAS, MMSE, RAVLT	Hippocampus, WholeBrain, Entorhinal and MidTemp volumes	FDG, AV45	tau and amyloid-beta levels	APOE4, AGE
Correlation based	CDRSB, ADAS, MMSE, RAVLT, LM	Hippocampus and Entorhinal volumes	FDG, AV1451	tau level	APOE4, AGE, years of education
Boruta feature selection	ADAS, MMSE, BNT	TEMPORALPOLE, INSULA, INFTEMPORAL, PARAHIP		ANIM	APOE4, GDS, NPIQ, NPI

TABLE (3.2) The list of selected features by corresponding method.

3.3.2 Imputation techniques

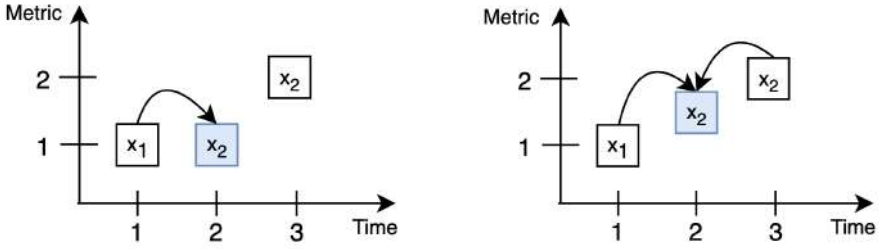
The following imputation methods was used to fill missing values:

1. Forward-filling
2. Linear interpolation

3. MMSE_TOT based

The forward filling imputation technique imputes the missing value based on the value of the last time point with available data. The linear interpolation technique takes the values for the last and next time points with available data and linearly fill the missing values 3.4.

The MMSE-based imputation method takes the MMSE value for the missing feature and imputes the average feature value of other patients with the same MMSE score. If we look for patients within the same value for MMSE with a patient with missing features, we can end up within a small number of such patients. Instead of choosing the patients within the exact value of MMSE, we calculate the minimal distribution around MMSE and impute the average feature value among the patients whose MMSE falls into that distribution. The minimal distribution was calculated as the measure of noise for MMSE, specifically the distribution of test-retest error for the six-month follow-up.



(A) Forward-filling imputation. (B) Linear interpolation.

FIGURE (3.4) Imputation techniques schema. Each schema impute missing value x_2.

3.3.3 Machine Learning models

The regression problem we were solving is a prediction of a future value for cognitive score given one time point. The following classical Machine Learning models were used to predict cognitive scores:

1. Linear Regression
2. Elastic Net
3. Random Fores Regression
4. Support Vector Regression

3.3.4 Validation techniques

To train and validate models, the dataset was divided into train and test to have different patients.

The nonparametric statistical Mann-Whitney U Test was used to verify whether the obtained error distribution is statistically different from the test-retest error. With this test, we check the hypothesis that the two distributions come from the same general distribution or have the same median (Milenović, 2011).

3.4 Experimental protocol

To understand the impact of all listed methods on prediction accuracy for different scenarios, we have to broke down the future work into pipelines.

We start experiments within the model selection step of the pipeline. We apply the forward filling technique on this step because the ML models accept feature complete data and chose the features recommended by the TADPOLE. When the most accurate model for a given scenario was chosen, we study other imputation techniques and feature selection methods to choose the ones that increase prediction accuracy.

The steps of the entire pipeline of experiments are described below:

1. TADPOLE features + forward filling imputation technique + ML model selection.
2. The best model from Pipeline 1 + imputation techniques.
3. The best output from Pipeline 2 + feature selection.

The schema of pipeline is described on the pipelines described in Figure 3.5.

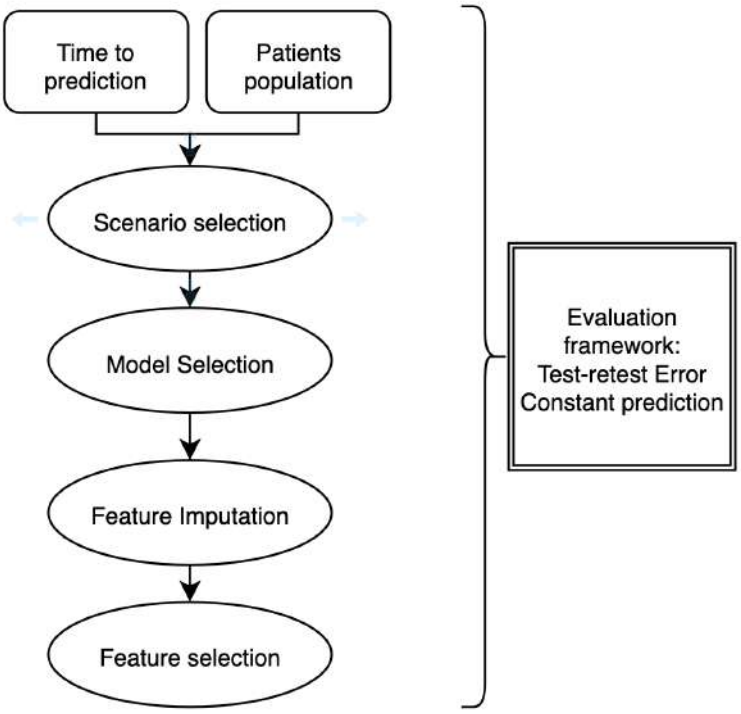


FIGURE (3.5) Experimental pipeline

Chapter 4

Results and evaluation

4.1 Scenario of interest

To understand which populations do not have the change in the dynamics of cognitive scores, we have to investigate the differences in change of cognitive scores per different cohorts of patients. We take a cohort of patients with specified criteria and perform a statistical test that checks if this cohort's distribution changes over the years. The Algorithm 1 explains how we calculate the significance of distributions change over time. Since we resampled the ADNI dataset (as described in Figure 3.3), the baseline value in an Algorithm 1 represents the first visit from two consequent patient's visits with the difference of six months. The difference between baseline values and six-month follow-up is named test-retest error. Hence, the statistical significance between test-retest error and i-year ($i = [1, ..., n]$) follow-ups change in the progression is calculated for three consequent patient visits. The difference between first and second (six-month after the first visit) visits is a test-retest error. The difference between the first and third (i-years after the first visit) visits is the change in progression after i-year from the patient's first visit.

Algorithm 1 Progression change comparison

```
test_retest_distribution = 6_month_followup - baseline_values
historical_followups = [1, ..., n]
u-statistics = Mann-Whitney-test
for i in historical_followups do
    followup_change_distribution = historical_followups[i] - baseline_values
    u-statistics(test_retest_distribution, followup_change_distribution)
```

Suppose the difference between test-retest and change in progression for a longer period is not statistically significant according to the Mann–Whitney U test. In that case, this scenario is considered to be a scenario without progression. That means that there is no statistically significant difference in the progression of cognitive scores for these patients within years.

The summary of the progression of cognitive scores (Table 4.1) shows that the MMSE progression for CN patients at baseline is not statistically significant up until five years. On the other hand, MCI and AD patients have a significant change in progression within only one year of follow-up.

The crucial part to remember when selecting the scenario of interest, including the population and time to prediction, is to pick meaningful scenarios to predict and those that help identify the first changes in cognition. When we answer this question, we can move on to the methods that could improve the prediction accuracy of the future value of the cognitive score. According to the results of Table 4.1 it is reasonable to predict the MMSE value for CN at least six years ahead and ADAS at

Disease stage	1 year	2 years	3 years	4 years	5 years	6 years
MMSE						
CN	1.036 (p=0.417)	0.993 (p=0.436)	1.053 (p=0.252)	1.118 (p=0.126)	1.181 (p=0.075)	1.291 (p=0.035)
MCI	1.769 (p=3.79E-03)	2.309 (p=1.55E-10)	2.830 (p=4.20E-16)	3.321 (p=1.99E-16)	3.803 (p=5.34E-19)	3.821 (p=1.46E-13)
AD	3.113 (p=3.45E-06)	4.825 (p=2.08E-14)	6.343 (p=8.66E-05)	4.733 (p=5.16E-03)	5.462 (p=1.79E-03)	4.333 (p=6.94E-03)
ADAS						
CN	2.281 (p=0.045)	2.218 (p=0.237)	2.501 (p=0.137)	2.619 (p=0.0002)	2.951 (p=2.4E-06)	3.627 (p=1.4E-18)
MCI	3.237 (p=8.9E-03)	4.149 (p=1.0E-08)	5.262 (p=1.2E-14)	6.443 (p=2.3E-16)	7.701 (p=3.3E-21)	7.870 (p=3.9E-22)
AD	5.462 (p=2.2E-04)	9.510 (p=1.0E-16)	12.766 (p=4.1E-06)	11.436 (p=7.4E-03)	12.273 (p=9.9E-03)	13.619 (p=1.1E-03)

TABLE (4.1) The MAE value of MMSE and ADAS distributions of change over time. The changes in a progression that are not statistically significant marked in gray.

least four years ahead. However, here we have to consider the volume of data we are working with. Figure 3.3 shows that the amount of patients visits decrees within time. Hence the relatively small amount of data can affect the model performance to generalize well.

There is an important omission within an approach of patient segregation based on the disease stage. The MCI cohort is a linking stage between CN and AD. It includes patients with early symptoms of cognitive disabilities and those with more severe cognitive impairments. If the problem is about identifying patients with early changes and preventing the future decline in cognition, it is reasonable to predict those patients' cognition with early symptoms. Hence, we have to break down the MCI cohort into smaller sub-cohorts. Further, we divide patients and form new cohorts based on disease stage and MMSE and ADAS values (Table 4.2). The statistical significance of the difference was calculated for cohorts with more than 50 observations. Since patients at baseline with CN do not show the statistically significant change in distribution during the first five years, we do not take this cohort into the future analysis. Since current research aims to investigate the early progression of decline in cognitive abilities, we do not study patients at the baseline with AD.

Comparing results on the deeper level of granularity of patients cohorts, we can see no MMSE and ADAS dynamics progression for one-year follow-ups. The sub-cohorts with less than 50 observations that were not taken into account considered as the "tails" of cognitive scores distributions. These tails could be the reason why the one-year follow-up observation in Table 4.1 shows statistically significant progress comparing to the sub-cohorts from the Table 4.2. The interesting dynamic is shown for ADAS sub-cohorts for CN patients at baseline. The CN patients at baseline with ADAS less than 6 show a change in progression at the fourth year of medical examinations. However, the CN patients with ADAS at baseline from 6 to 12 do not show changes over six years of observations. The MCI patients with ADAS less than 6 points have a stagnation period at the fourth year follow-up. The MCI patients with

Disease stage	Cognitive score range	1 year	2 years	3 years	4 years	5 years	6 years
MMSE							
MCI	(27; 30]	1.408 (p=0.38)	1.817 (p=1.7E-02)	2.093 (p=3.4E-03)	2.523 (p=3.4E-04)	2.963 (p=1.3E-07)	2.464 (p=3.6E-04)
	(23; 27]	2.178 (p=0.300)	2.964 (p=2.9E-04)	4.071 (p=2.7E-08)	4.626 (p=2.0E-11)	4.789 (p=1.3E-07)	6.339 (p=6.7E-07)
ADAS							
CN	(0:6]	1.905 (p=0.26)	1.812 (p=0.35)	1.902 (p=0.31)	2.642 (p=7.5E-04)	3.191 (p=4.0E-04)	4.042 (p=8.4E-11)
	(6:12]	2.533 (p=0.36)	2.662 (p=0.29)	2.972 (p=0.38)	2.541 (p=0.29)	2.503 (p=0.4)	2.872 (p=0.21)
MCI	(0:6]	2.343 (p=0.35)	2.738 (p=0.062)	2.701 (p=0.044)	2.607 (p=0.14)	3.552 (p=0.005)	4.233 (p=0.001)
	(6:12]	2.941 (p=0.24)	3.772 (p=6.5E-04)	4.905 (p=9.0E-08)	6.212 (p=1.5E-08)	7.572 (p=2.8E-10)	6.782 (4.1E-08)
	(12:20]	4.041 (p=0.062)	5.712 (p=2.4E-07)	7.707 (p=7.8E-10)	10.369 (p=3.9E-11)	11.831 (p=2.1E-09)	14.026 (p=2.5E-07)

TABLE (4.2) The MAE value of MMSE and ADAS distributions of change over time. The changes in a progression that are not statistically significant marked in gray.

ADAS between 6 and 12 and between 12 and 20 show the change in cognition starting the second year from the baseline observation.

4.1.1 Constant prediction

The meaningful prediction settings are the ones that allow predicting better than the constant prediction and higher than the test-retest error. The results observed in Table 4.2 can be compared to the distribution plots at Appendix A.1. Here we can see how the MMSE and ADAS progression develops over time and in which scenarios test-retest error is the same as the constant predictions. The visual representation of the same error distributions is presented at Figure 4.1. It represents the model's performance, Constant Prediction and Test-Retest error of ADAS 6-month prediction. The distributions of Constant Predictions (marked as CP+1, CP+2, CP+3) have smaller variance than Test-Retest error distribution (marked as TR). Mann Whitney U test also indicated that distributions of errors do not statistically differ from each other. That means that there is no space for prediction improvement since the error of the baseline model, such as Constant Prediction, is already the best predictor.

In all cases, with yearly predictions, the constant prediction is statistically the same as the test-retest error. As time passed, we can see that the baseline model becomes less accurate, and here we have a space to improve the accuracy of the predictions. These settings are the predictions starting from 2 or 3 till 6 years ahead. As a result, we will go with further analysis with the MCI sub-cohorts for MMSE and ADAS cognitive scores 2-5 years prediction ahead.

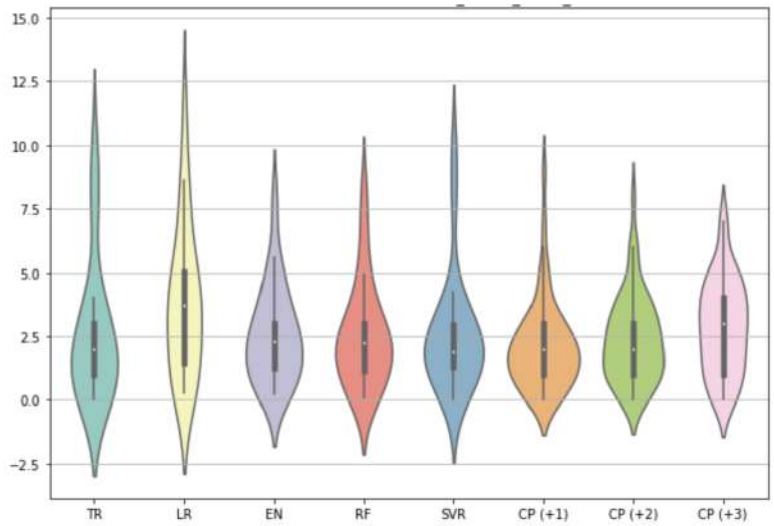


FIGURE (4.1) Distribution of error for 6 month ADAS prediction. The TR stands for Test-Retest error, LR - Linear Regression model error, EN - Elastic Net, RF - Random Forest, SVR - Support Vector Regression, CP+1 - error for Constant prediction with added 1 point to the ADAS value, CP+2 and CP+3 with added 2 and 3 points respectively

4.2 Methods comparison

This section aims to test the different methods to predict the cognitive scores with selected scenarios. Each subsection follows the Experimental Protocol design and represents the studied step of the entire pipeline.

4.2.1 ML model selection

We start with the model selection at the first iteration for improvement of prediction accuracy. We freeze other setups such as imputation or feature selection techniques and examine the performance of the specified models. We use the next set of models to predict the values of cognitive scores: Linear Regression (LR), Elastic Net (EN), Random Forest Regression (RF), and Support Vector Regression (SVR). Here we apply forward filling techniques to impute missing data and use the set of features recommended by the TADPOLE. The output of Table 4.3 shows the best MAE for the current scenario among all tested models. The model parameters were optimized using the GridSearch algorithm. The grid of parameters for each model is shown in Appendix **??**.

The error distribution for 2-years predictions for all scenarios shows good results comparing to the test-retest error. The scenario for MCI patients with ADAS less than 6 points starts to overfit the data on 2 and 3 years of predictions. The reason for such performance for the 2-year prediction is clear - the initial setup of the scenario shows that there is no progression with 2-years follow-ups (Table 4.2), hence the model starts to overfit.

The 5 scenarios out of 7 for the 4-years predictions show that there is room for improvement for prediction accuracy, as well as for the 6 scenarios for the 5 years predictions and 5 scenarios for the 6 years prediction.

Disease stage	Cognitive score range	2 years	3 years	4 years	5 years	6 years
MMSE						
MCI	(27; 30]	[RF] 1.42 (p=0.385)	[EN] 1.39 (p=0.108)	[FR] 1.71 (p=0.005)	[EN] 1.84 (p=0.010)	[FR] 1.55 (p=0.045)
	(23; 27]	[RF]$^{+}$ 1.92 (p=0.385)	[EN] 3.15 (p=0.005)	[FR] 3.30 (p=0.013)	[FR] 2.53 (p=0.183)	[FR] 3.99 (p=0.022)
	all patients	[EN] 1.53 (p=0.230)	[RF] 1.81 (p=0.001)	[RF] 2.21 (p=0.000)	[RF] 2.43 (p=0.000)	[RF] 2.90 (p=0.000)
ADAS						
MCI	(0:6]	[RF]$^{+}$ 1.90 (p=0.054)	[RF]$^{+}$ 2.06 (p=0.151)	[RF] 2.45 (p=0.436)	[RF] 2.01 (p=0.106)	[RF] 3.33 (p=0.018)
	(6:12]	[RF] 2.85 (p=0.093)	[EN]* 4.35 (p=0.000)	[RF] 4.37 (p=0.000)	[RF] 4.13 (p=0.001)	[RF] 5.32 (p=0.000)
	(12:20]	[RF] 4.00 (p=0.074)	[RF] 4.89 (p=0.033)	[RF] 7.85 (p=0.000)	[SVR] 6.72 (p=0.001)	[RF]* 12.30 (p=0.005)
	all patients	[FR] 2.82 (p=0.366)	[RF] 3.17 (p=0.235)	[RF] 3.95 (p=0.010)	[RF] 4.48 (p=0.000)	[RF] 5.72 (p=0.000)

TABLE (4.3) The MAE value of best models for MMSE and ADAS prediction errors and p-value for statistical significance test between predicted error and test-retest error. The distribution of prediction error that is statistically different from the distribution of test-retest error is marked in gray. Those MAE that are higher than Constant Predictions have a star sign and MAE that are lower than test-retest error have a plus sign.

Random Forest showed the best results for most scenarios; hence, this model will be used in further analysis. Those scenarios that started to overfit will not be considered in the subsequent iterations.

4.2.2 Imputation technique selection

With this iteration, we test different imputation techniques such as Forward-Filling (FF), Linear Interpolation (LN), and imputation based on MMSE value (CB).

As it is shown in Table 4.5 in most cases, the Forward Filling method delivered the best results. There were only a few scenarios when Linear Interpolation or MMSE-based Imputation outperformed the first technique. There are two scenarios where the error distribution becomes closer to the noise distribution: the 4-year MMSE prediction scenario for the cohort of MCI patients with MMSE values between 27 and 30 and the 3-year prediction for MCI patients with MMSE value between 23 and 27. The MMSE-based imputation technique improved the prediction performance only in 3 scenarios out of 33. For Linear interpolation, the amount of improved scenarios is 7. As a result, there is a slight improvement in patients' subcohorts accuracy and smaller improvement for the cohorts of higher granularity, such as whole MCI patients groups for MMSE and ADAS. Since the Forward-Filling imputation technique gave the best results, it will be used in the next step of the pipeline.

Disease stage	Cognitive score range	2 years	3 years	4 years	5 years	6 years
MMSE						
MCI	(27; 30]	[CB] 1.41 (p=0.122)	[FF] 1.57 (p=0.104)	[LN] 1.62 (p=0.062)	[FF] 1.97 (p=0.003)	[LN] 1.76 (p=0.048)
	(23; 27]	[LN] 1.95 (p=0.433)	[LN] 2.24 (p=0.149)	[FF] 2.56 (p=0.037)	[FF] 2.53 (p=0.183)	[FF] 3.22 (p=0.004)
	all patients	[FF] 1.79 (p=0.002)	[FF] 2.01 (p=0.000)	[LN] 1.96 (p=0.011)	[CB] 2.29 (p=0.001)	[LN] 2.04 (p=0.004)
ADAS						
MCI	(0:6]	-	-	[FF] 2.45 (p=0.436)	[FF] 2.01 (p=0.106)	[FF] 3.33 (p=0.018)
	(6:12]	[FF] 2.85 (p=0.093)	[FF] 4.16 (p=0.000)	[FF] 4.37 (p=0.000)	[FF] 4.13 (p=0.001)	[FF] 5.32 (p=0.000)
	(12:20]	[FF] 4.00 (p=0.074)	[FF] 4.89 (p=0.033)	[FF] 7.85 (p=0.000)	[CB] 6.68 (p=0.000)	[LN] 7.26 (p=0.004)
	all patients	[FF] 2.82 (p=0.336)	[FF] 3.17 (p=0.235)	[FF] 3.95 (p=0.010)	[LN] 3.91 (p=0.001)	[CB] 4.44 (p=0.003)

TABLE (4.4) The MAE value of model with best imputation technique for MMSE and ADAS prediction errors and p-value for statistical significance test between predicted error and test-retest error. The distribution of prediction error that is statistically different from the distribution of test-retest error is marked in gray.

4.2.3 Feature selection

This section represents the different feature selection techniques. The first group of features was taken from the TADPOLE challenge's recommendation. The second group of features is the result of the correlation analysis, and the last group of features was obtained using Boruta algorithm. The list of features chosen in each feature selection type is represented in Table 3.2. Since the set of features recommended by TADPOLE and features selected using correlation analysis almost overlap, we merged these features into a single feature group "Core" (CR), and used instead of feature set obtained based on correlation analysis.

There are a few scenarios where features based on Boruta algorithm outperformed TADPOLE features, but overall the best-performed feature set was the TADPOLE one.

Disease stage	Cognitive score range	2 years	3 years	4 years	5 years	6 years
MMSE						
MCI	(27; 30]	[TP] 1.33 (p=0.030)	[TP] 1.57 (p=0.104)	[TP] 1.78 (p=0.062)	[TP] 1.97 (p=0.003)	[TP] 1.76 (p=0.048)
	(23; 27]	[BR] 2.06 (p=0.123)	[TP] 2.24 (p=0.149)	[TP] 2.56 (p=0.037)	[TP] 2.53 (p=0.183)	[FF] 3.22 (p=0.004)
	all patients	[TP] 1.79 (p=0.002)	[TP] 1.81 (p=0.001)	[TP] 2.21 (p=0.000)	[TP] 2.43 (p=0.000)	[BR] 2.82 (p=0.000)
ADAS						
MCI	(0:6]	-	-	[TP] 2.45 (p=0.436)	[TP] 2.01 (p=0.106)	[BR] 3.64 (p=0.068)
	(6:12]	[TP] 2.85 (p=0.093)	[BR] 3.50 (p=0.086)	[TP] 4.37 (p=0.000)	[TP] 4.13 (p=0.001)	[TP] 5.32 (p=0.000)
	(12:20]	[TP] 4.00 (p=0.074)	[TP] 4.89 (p=0.033)	[TP] 7.85 (p=0.000)	[TP] 6.24 (p=0.000)	[BR] 8.21 (p=0.000)
	all patients	[TP] 2.82 (p=0.336)	[TP] 3.17 (p=0.235)	[TP] 3.95 (p=0.010)	[TP] 4.48 (p=0.000)	[TP] 5.72 (p=0.000)

TABLE (4.5) The MAE value of model with best feature selection technique for MMSE and ADAS prediction errors and p-value for statistical significance test between prediction error and test-retest error. The distribution of prediction error that is statistically different from the distribution of test-retest error is marked in gray.

Chapter 5

Conclusions and Future work

In our work, we examined the different scenarios for the prediction of cognitive scores. We showed that there are scenarios that, by design, could not progress with time. These scenarios include cognitively normal patients at baseline, 1-year predictions, and a scenario with MSI patients with the first demonstration of cognitive dysfunctions. By predicting the values of cognitive scores for these cohorts, we do not learn the progression of the cognitive scores. On the other hand, the long-time predictions are harder to follow. The models that show good results on short-term prediction in terms of accuracy cannot perform well on long-term predictions.

We introduced an evaluation framework that helps to identify those scenarios without progression. By comparing the distribution of progression change of each cognitive score, we describe the dynamics of whether cognitive decline can be observed with a given cohort of patients or not. The constant prediction, which is a simple adjustment and is calculated by adding 1 point to the target, can be considered the best model for those scenarios without progression. This constant prediction does not statistically differ from noise.

Finally, we test a list of techniques and models to produce a future value of cognitive scores in the scope of the proposed evaluation framework. We apply different machine learning models, imputation techniques and techniques for feature selection. As a result, we received a group of scenarios (MSI with ADAS < 6, 4-6 years predictions) where the distribution of prediction error is statistically the same as the distribution of test-retest error (noise), but the absolute value of MAE still not so close to the MAE value of test-retest. For other MSI settings, the results are worse.

5.1 Future work

To understand the long-term dynamics better the more sophisticated methods have to be used. As for imputation techniques, the Machine Learning models or Deep Learning models could potentially help as well as the incorporation of missing values into DL networks architecture.

The potential to grasp the long-term dependencies has the Recurrent Neural Networks. However, the analysis has to be done carefully considering that there is a limited amount of data that is needed to train DL models well. The further the prediction has to be made, the less data is left to make these predictions (Figure 3.3).

Appendix A

Methods

A.1 ML models parameters

Here is the list of parameters for models that were fitted to Grid Search algorithm.
Elastic Net:

- alpha: [0.5, 1, 2]
- l1_ratio: [0.3, 0.5, 0.7]

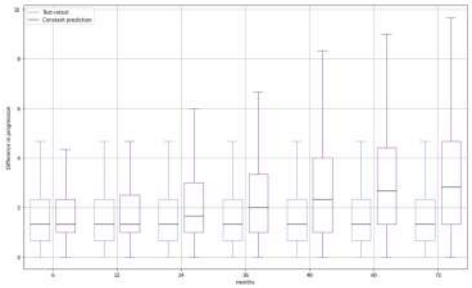
Random Forest Regressor:

- n_estimators: [20, 60, 100]
- max_depth: [40, 70, 100]
- min_samples_split: [2, 5, 10]
- min_samples_leaf: [2]

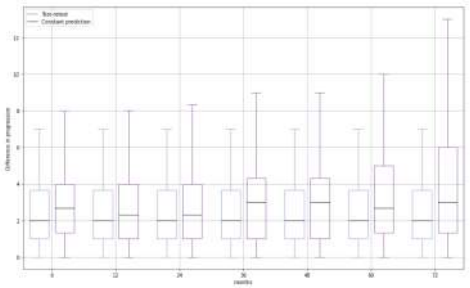
Support Vector Regression:

- kernel: ['linear', 'rbf', 'sigmoid']
- C: [1, 10]
- coef0: [0.01, 10]
- max_iter: [500]

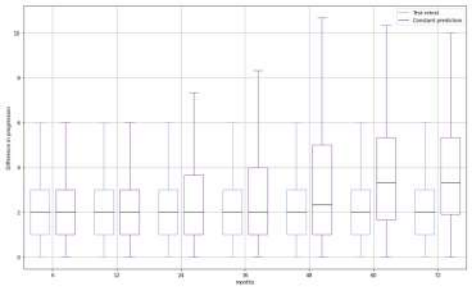
A.2 Figures



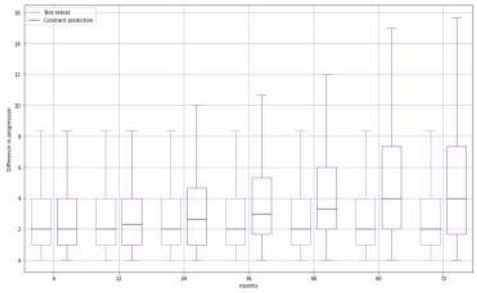
(A) ADAS between 0 and 6, CN.



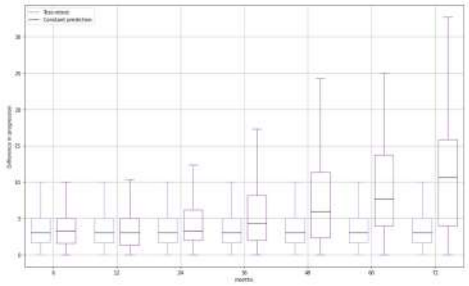
(B) ADAS between 6 and 12, CN.



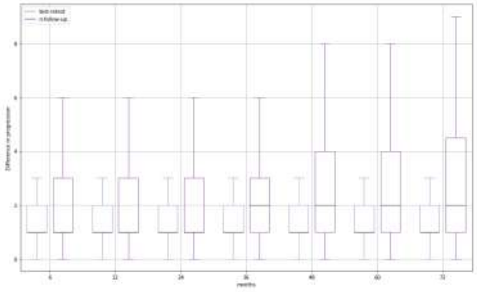
(C) ADAS between 6 and 12, MSI.



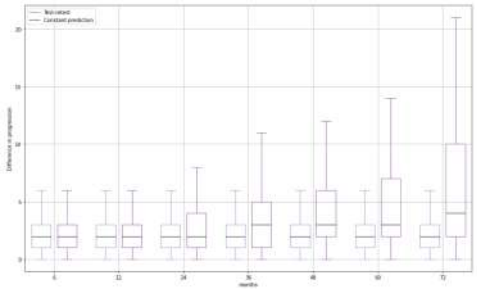
(D) ADAS between 6 and 12, MSI.



(E) ADAS between 12 and 20, MSI.



(F) MMSE between 27 and 30, MSI.



(G) MMSE between 23 and 27, MSI.

FIGURE (A.1) Distribution of Test-Retest and Constant prediction error for ADAS and MMSE.

Bibliography

Alzheimer's Association - Facts and Figures. Accessed: 2020-05-02. URL: https://www.alz.org/alzheimers-dementia/facts-figures.

Clark, Christopher M. et al. (July 1999). "Variability in annual Mini-Mental State Examination Score in patients with probable Alzheimer disease: a clinical perspective of data from the consortium to establish a registry for Alzheimer's disease". In: *Archives of Neurology* 56.7, pp. 857–862. ISSN: 0003-9942. DOI: 10.1001/archneur.56.7.857.

Ghazi, Mostafa Mehdipour et al. (2019). "Training recurrent neural networks robust to incomplete data: application to Alzheimer's disease progression modeling." In: *Medical Image Analysis* 53, pp. 39–46. DOI: 10.1016/J.MEDIA.2019.01.004.

Jack, Clifford R et al. (2010). "Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade". In: *Lancet Neurology* 9.1, pp. 119–128. DOI: 10.1016/S1474-4422(09)70299-6.

Koval, Igor et al. (2021). "AD Course Map charts Alzheimer's disease progression." In: *Scientific Reports* 11.1, pp. 8020–8020. DOI: 10.1038/S41598-021-87434-1.

Kursa, Miron B. and Witold R. Rudnicki (2011). "The all relevant feature selection using Random Forest". In: *arXiv preprint arXiv:1106.5112*. URL: https://arxiv.org/pdf/1106.5112.pdf.

Marinescu, Razvan V. et al. (2020). "The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) challenge: results after 1 year follow-up". In: *arXiv e-prints*, p. 35. URL: https://arxiv.org/pdf/2002.03419.pdf.

Milenović, Živorad (2011). "Application of Mann-Whitney U Test in Research of Professional Training of primary school teachers". In: *Metodički obzori : časopis za odgojno-obrazovnu teoriju i praksu* 6.11, pp. 73–79. DOI: 10.32728/MO.06.1.2011.06.

Nguyen, Minh et al. (2020). "Predicting Alzheimer's disease progression using deep recurrent neural networks." In: *NeuroImage* 222, p. 117203. DOI: 10.1016/J.NEUROIMAGE.2020.117203.

Rosen, Wilma G., Richard C. Mohs, and Kenneth L. Davis (1984). "A new rating scale for Alzheimer's disease." In: *American Journal of Psychiatry* 141.11, pp. 1356–1364. DOI: 10.1176/ajp.141.11.1356.

Sargolzaei, Saman et al. (2015). "A practical guideline for intracranial volume estimation in patients with Alzheimer's disease". In: *BMC Bioinformatics* 16.7, pp. 1–10. DOI: 10.1186/1471-2105-16-S7-S8.

Voevodskaya, Olga et al. (2014). "The effects of intracranial volume adjustment approaches on multiple regional MRI volumes in healthy aging and Alzheimer's disease". In: *Frontiers in Aging Neuroscience* 6, p. 264. ISSN: 1663-4365. DOI: 10.3389/fnagi.2014.00264.

Zhang, Daoqiang, Dinggang Shen, and Alzheimer's Disease Neuroimaging Initiative (2012). "Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers". In: *PLOS ONE* 7.3. DOI: 10.1371/JOURNAL.PONE.0033182.