UNIVERSITAT DE BARCELONA

FUNDAMENTALS OF DATA SCIENCE MASTER

Detecting incipient cognitive dysfunction in Preclinical Alzheimer's Disease subjects

---

Author:
Rachel TRIMBLE

Supervisor:
Sergio ESCALERA GUERRERO
Lorena RAMI
Pau BUCH-CARDONA

September 7, 2021

UNIVERSITAT DE BARCELONA

Abstract

Detecting incipient cognitive dysfunction in Preclinical Alzheimer's Disease subjects

by Rachel TRIMBLE

The number of people with Alzheimer's disease, a degenerative brain disorder, is projected to triple worldwide by 2060, with no current cure. There has been a paradigm shift in the diagnostic conceptualization of Alzheimer's disease (AD) based on evidence suggesting that structural and biological changes start to occur during a preclinical phase beginning decennia prior to the emergence of symptoms. However diagnostic methods for this phase are invasive and costly, thus clinicians are searching for cognitive tools for screening the population before diagnosing them.

The goal of this thesis is to support the clinicians in their search for these new cognitive tests for Preclinical AD (pre-AD) detection through machine learning. In particular to provide a tool for clinicians to validate if a test is sensitive enough to detect incipient cognitive dysfunction in pre-AD subjects.

To achieve this we first investigated multiple classifiers and ensemble methods to find a suitable one for the datasets supplied by the clinicians. We incorporate data augmentation through Synthetic Minority Oversampling Technique (SMOTE) to deal with the imbalanced nature of the dataset. We also compute the importance for each individual feature using a technique that assigns a score to these features based on how useful they were during the classification.

We found Random Forest to be the preferred choice among the tested algorithms. SMOTE proved to be a crucial step, improving both the AUC and most importantly the sensitivity. The traditional neuropsychological tests were not sensitive enough to detect incipient cognitive dysfunction in pre-AD subjects. While the new tapping tests were more sensitive. Our tool was also easily understandable for the clinicians thanks to the feature importance.

Acknowledgements

First and foremost, a huge thank you to my thesis supervisors, Sergio Escalera Guerrero, Lorena Rami and Pau Buch-Cardona, for their suggestion of this interesting topic, insightful ideas, guidance and most of all patience. I would also like to thank Adria Tort, from IDIBAPS, for his availability and promptness in helping us understand the datasets.

Additionally I'd like to extend my gratitude to the many professors in the Foundations of Data Science Masters at the UB, who taught me the skills needed for this work.

Last but not least, I would like to thank my family, in particular my husband Giuseppe, daughter Sofia and mother Madeline for their support and understanding.

Chapter 1

Introduction

1.1 Problem Statement

The pathophysiological process of Alzheimer's disease (AD) is believed to begin decennia before the diagnosis of AD. This long "preclinical" phase of AD happens in the absence of dementia, and could provide a critical opportunity for therapeutic intervention. The detection of this phase of AD is currently costly, invasive and thus not suitable for mass testing and detection. Researchers are interested in finding new tests that are sensitive enough to detect incipient cognitive dysfunction in preclinical AD patients.

Our objective is to investigate to which degree data science tools can be used to help provide insights to support clinicians in their discovery of these new cognitive tests. In particular to provide a tool that will assess which tests are important while classifying preclinical AD subjects (pre-AD) and healthy subjects to help them design new cognitive tests to detect incipient cognitive dysfunction in the preclinical group.

1.2 Similar work

There has been an increased interest in discovering a method to detect the early stages of AD. The research spans across many very different approaches, for example Hadoux et al [36] use hyperspectral imaging of the retina, Zhu et al [8] test if linguistic features can detect the early stages, Wilcockson et al [35] hypothesize that eye movement deteriorates as Alzheimer's disease progresses, with the gradual loss of the efficient control of attention and develop impairments of both inhibitory control and eye movement error-correction while Siedlecki-Wullich et al [11] found that MicroRNAs in the blood can be a predictive factor in understanding the AD stage a person is in. While some of these methods were able to detect later stages of the disease, none could conclude on a robust method for detection at the preclinical phase.

Much research has also been focused on cognitive testing for early detection, as it is widely accepted that memory loss is generally the earliest cognitive change in AD. Traditional neuropsychological testing is not sensitive enough for detection of cognitive changes in the preclinical phase, as such the research is directed towards finding more demanding testing procedures. Rentz et al. [9] presented a series of demanding memory tests and found associations between a high demanding face-name associative memory test and amyloid plaques, a pathophysiological hallmark of AD. Likewise Tort-Merino et al. [4] found a highly demanding experimental associative learning test sensitive and capable of detecting subtle memory difficulties.

Other research has reported that motor dysfunction may be a sensitive marker of preclinical phases of AD [27, 6] with some literature focusing on motor speed [31]. Additionally, Mollica et al. [24] explored whether motor dysfunction could be useful for early detection, through a novel finger tapping test. In the study their findings suggest that motor dysfunction is associated with amyloid pathology and may subtly emerge during the earliest stage of the Alzheimer's continuum. Hence there are promising findings that finger tapping could be a test to detect first changes in the Pre-AD subjects.

There has also been an interest in the application of machine learning to support the search for a method for detecting the early stages of AD. Most of this research has been centralised on relatively large data sets and thus deep learning models have been applied, for example Manu Raju et al [23] utilise a transfer Learning technique on Magnetic Resonance Imaging (MRI) of the brain. However the cost of diagnosis is high, and thus not suitable for the mass-sampling that is desired.

Chapter 2

Introduction into Alzheimer's Disease

First identified by Dr. Alois Alzheimer in 1906, AD is now thought to be the most common cause of dementia. AD is a neurodegenerative disease, affecting brain function and cognitive processes related to learning and memory. [1] The disease alters synaptic connectivity through the accumulation of specific brain pathologies and the distributed neuronal connectivity in time results in the death of brain cells and as a consequence produces dementia .This results in a decline in memory, ability to articulate, problem-solve and other cognitive skills and these characteristics eventually impede the individual from performing basic everyday tasks.

With the aging population on the rise, the economic burden of AD is set to follow suit and is expected to increase rapidly with the growing percentage of this demographic in the developed world. In 2015, it was estimated that 46.8 million people are living with dementia, which is expected to double every 20 years to reach over 74 million by 2030 and AD is estimated to represent 60-80% of dementia cases [7]. By 2050, the prevalence of AD is estimated to increase to over 100 million [34] and with no cure for the debilitating and ultimately fatal disease, it is vitally important to develop new intervention tools and treatments to manage the imminent healthcare crisis.

Disease-modifying therapies that aim to treat AD patients with cognitive impairment have not demonstrated adequate efficacy in clinical trials. During the last decade research has made a paradigm shift based on evidence that suggests structural and biological changes start to occur during a preclinical phase beginning decennia prior to the appearance of clinical symptoms [3]. This preclinical stage of AD has become a major research focus as the field postulates that early intervention could provide the best prospect for finding a cure.

2.1 AD stages

In 2011, The National Institute on Aging and the Alzheimer's Association (NIA-AA) [26] revised the diagnostic criteria that was originally created in 1984 for diagnosing AD dementia and established new diagnostic guidelines for the stages of AD. The highlight of these new criteria was the introduction of biomarkers into the framework, allowing the definition of the preclinical phase of the disease [33]. The AD biomarkers are detailed in the following subsection and the diagnostic and preclinical stages are described in table below in 2.1.

Stage	Criteria
AD dementia	1. The presence of dementia, with a decline in cognition and function. 2. Insidious onset and progressive cognitive decline. 3. Impairment in two or more cognitive domains 4. Absense of other characteristics associated with other dementing disorders. 5. Increase diagnosting confidence may be suggested by the biomarker algorithm discussed in the MCI due to AD section below.
MCI due to AD	1. A change in cognition from previous levels. 2. Impairment in at least one cognitive domains. 3. Preserved independence in functional abilities. 4. No dementia. 5. Increase diagnosting confidence may be suggested by either: A. Positive Beta-Amyloid biomarker without a degeneration biomarker B. A positive degeneration biomarker without testing for Beta-Amyloid biomarkers
Preclinical AD	1. Stage 1 is marked by Beta-Amyloid 42 peptide dysregulation (reduced levels of Beta-Amyloid 42 in the cerebrospinal fluid or elevated cerebral cortical amyloid burden determinted by a PET scan) 2. Stage 2 adds synaptic/neuronal dysfunction and loss 3. Stage 3 includes charactertics of stage 1 and 2 along with subtle cognitive decline.

FIGURE 2.1: Alzheimer's Disease stages

2.2 Preclinical AD

AD comprises a long asymptomatic (or preclinical) phase beginning decennia prior to the emergence of the first clinical symptoms, which trigger the pathophysiological processes characteristic of the disease. This silent phase can last for years or even decades without the individual knowing they are in the phase and could potentially

progress to the later and more aggressive stages of the disease.

Clinical trials in search of a drug development targeting mild-to moderate AD have had limited success in the search for an effective treatment for AD. The preclinical stage of AD presents a window of opportunity for disease modifying therapies [15] and thus it's detection is crucial.

Historically AD has been defined as a clinical-pathological entity characterized by a particular clinical phenotype, rendering the definition of an early stage of AD ambiguous. The NIA-AA group defined a purely biological framework that excluded clinical outcomes. In this framework, preclinical AD is defined as cognitively unimpaired individuals that present Alzheimer's two pathologies, namely amyloid plaques and neurofibrillary tangles.

2.2.1 Pathophysiology of Preclinical AD

Beta-amyloid

The extracellular accumulation of β-amyloid is one of the processes distinctive to the pathological features of AD. It is a small piece of a larger protein called amyloid precursor protein (APP), in the form of sticky, starch-like plaques, in an increased manner in individuals with AD. It starts with the pieces forming small clusters called oligomers, these clusters form chains of clusters called fibrils which then form "mats" of fibrils called beta-sheets. This finally results in plaques which are clumps of beta-sheets and other substances.

The amyloid cascade hypothesis [14], the dominant model of AD pathogenesis, postulates that these stages of beta-amyloid aggregation disrupt cell-to-cell communication and activate immune cells. These immune cells trigger inflammation and eventually lead to neuronal death.

Neurofibrillary tangles (NFTs) of protein tau

The tau protein is predominantly found in neurons with one of it's many functions being to stabilize internal microtubules. The microtubules are structures that facilitate axonal transport, and thus a vital element to keep the neuron healthy.

In the presence of Alzheimer's disease, the tau detaches from these microtubules and sticks to other tau molecules which create threads that in time join to form Neurofibrillary tangles disrupting the microtubule assembly which ultimately leads to neural death [5]. Intracellular accumulation of hyperphosphorylated tau is therefore the second neuropathological hallmark of AD.

2.2.2 Detection of Preclinical AD

Early detection of AD is crucial and thus it is fundamental to develop new tools for this purpose. This will allow the detection of individuals who are at risk of the biological evolution of the disease earlier.

One method for detecting the presence of beta-amyloid and NFTs and thus detecting the preclinical phase is by performing lumbar puncture procedures and collecting the Cerebrospinal fluid (CSF) [20]. An alternative method is neuroimaging by conducting positron emission tomography (PET) on the brain [17]. While these methods have been widely used in research, they are not suitable for mass testing of the public, on subjects that seem healthy, which is essential for the detection of the early stage considering the subject does not experience symptoms. For the prior, the lumbar puncture is because the extraction of the CSF is highly invasive that introduces health risks to patients. While the latter, PET carries a high cost and has limited availability.

In light of the two aforementioned diagnostic methods being either invasive or costly, meaning that mass-scale screening of the population is not feasible, research has shifted to find a method for detecting the incipient cognitive dysfunction in pre-AD which indicates that the dementia phase is nearer. The research spans across many very different approaches, with neuropsychological assessments being one of the front runners. Cognitive and behavioral assessments through battery tests have proved successful at detecting cognitive impairment in MCI subjects but they are not sensitive enough to subtle deficits that may be present in preclinical subjects [9]. However, neuropsychologists are rising to the challenge by designing newer and more sensitive cognitive measures of Preclinical detection, and accurate computerized neuropsychological testing methods [25].

Chapter 3

Dataset

In this thesis we analyze two different datasets of subjects that have been tested for the presence of Alzheimer's Disease. The subjects are grouped into either being healthy or the stage of the Alzheimer's Disease that they are in. This categorization is achieved by taking cerebrospinal fluid through a lumbar puncture, and measuring the phosphorylated tau and Beta-amyloid values. These two aforementioned values are present in both datasets.

The datasets are distinct in the features that they contain, with the first dataset containing neuropsychological tests, the traditional tests used for detecting mild cognitive impairment and demographic characteristics. The second dataset also contains neuropsychological tests, although not completely the same ones, along with new motor dysfunction tests, namely tapping tests. Both datasets possess characteristics that we need to acknowledge before feeding them into any machine learning models. In the section we describe the features in the two datasets, summarize their characteristics and finally mention the data preprocesses applied.

3.1 Features

3.1.1 Dataset 1 - Battery tests

Subjects were divided into four groups according to their stage in Alzheimer's Disease: 80 control subjects, 25 Pre-AD, prodromal/MCI 43 and 30 AD patients. All subjects underwent a neuropsychological battery and their age and gender was also noted in the dataset.

Neuropsychological battery test

We tested 22 of the battery tests in our model, in particular;

1. buschke_AL: Free and cued selective reminding test free learning score
2. buschke_AT: Free and cued selective reminding test total learning score
3. buschke_RDL: Free and cued selective reminding test delayed free recall score

4. buschke_RDT: Free and cued selective reminding test Delayed total recall score
5. tam: Memory alteration test
6. vis_cerad: visual memory test
7. paisajes_tot: visual memory test
8. bnt: Boston Naming Test score
9. flu_anim: category fluency task animals in one minute
10. compren: comprehension test
11. ideom: cognition test
12. prax_cerad: cognition test
13. tdp: cognition test
14. VOSP_num: Visual Object and Space Perception battery number location subtest
15. VOSP_letras: Visual Object and Space Perception battery incomplete letters subtest
16. tmtA: Trail Making Test Form A
17. tmtB: Trail Making Test Form B
18. fas_total: fluency test
19. Stroop_lect: Stroop test word reading subtest
20. Stroop_color: Stroop test color reading subtest
21. Stroop_I: Stroop test word-color reading subtest
22. Clave_num: speed test

3.1.2 Dataset 2 - Battery tests and Tapping Features

Subjects were divided into three groups according to their stage in Alzheimer's Disease: 37 control subjects, 20 Pre-AD and 15 AD patients. Control and Pre-AD subjects underwent a neuropsychological battery and all subjects were given a finger tapping task. The finger tapping task, based on a novel version [32], to measure the tapping speed and intrasubject variability. Below we mention the features in the battery and tapping tests that we utilized. The dataset also contains the age, gender and years of education for each subject.

Neuropsychological battery test

We tested 13 of the battery tests with our model, in particular;

1. buschke_AL: Free and cued selective reminding test free learning score
2. buschke_AT: Free and cued selective reminding test total learning score
3. buschke_RDL: Free and cued selective reminding test delayed free recall score
4. buschke_RDT: Free and cued selective reminding test Delayed total recall score
5. bnt: Boston Naming Test score
6. flu_anim: category fluency task animals in one minute
7. VOSP_num: Visual Object and Space Perception battery number location subtest
8. VOSP_letras: Visual Object and Space Perception battery incomplete letters subtest
9. tmtA: Trail Making Test Form A
10. tmtB: Trail Making Test Form B
11. Stroop_lect: Stroop test word reading subtest
12. Stroop_color: Stroop test color reading subtest
13. Stroop_I: Stroop test word-color reading subtest

Tapping Features

The tapping test included six different blocks of 10s each, for each block the subjects were instructed to tap as many times possible. The features used to measure the motor dysfunction from the tapping test were the *tapping speed* and *intrasubject variability*. These variables showed a strong relationship with the variables used to form the grouping, phosphorylated tau and Beta-amyloid values, during the data analysis we performed.

It should be noted that we engineered other features from the tapping dataset, but none showed a strong correlation with phosphorylated tau and Beta-amyloid values. The tapping rate is computed simply as the number of taps made over all of the blocks, while the tapping intrasubject variability is computed by dividing the subject's standard deviation by their mean (SD/mean). Where a higher intrasubject variability indicates greater inconsistent performance across trials.

3.2 Dataset characteristics

Small with large dimensions

The first dataset contains 178 samples and 1,443 variables while the second contains 72 samples and 51 variables. Small datasets can be difficult to model as they are accompanied by difficulties such as overfitting. This happens when we increase the dimensionality of a model without increasing the number of training samples, resulting in the feature space becoming more sparse and leading the classifier to overfit easily. As a first step, to model the small datasets, we reduced their dimensionality. We removed features that were irrelevant for our analysis, and worked with the clinicians that created the datasets to gain domain knowledge to reduce the dimensionality further.

Imbalanced

The two datasets are imbalanced with a smaller proportion of subjects in the Pre-AD group, representing 31-35% of the entire datasets. This imbalance property, that is common to many real healthcare datasets, makes classification a challenging task. Most classification models are not able to deal with imbalanced datasets, they require the dataset to be sufficiently balanced for the model to learn from. As outlined in the subsequent sections, we oversampled the minority class to balance both of the datasets.

Missing Values

Both datasets contain many samples with missing feature values, which is very common in these types of medical datasets. There are two types of missing data present in the datasets:

1. Missing completely at random (MCAR): The missing data are unrelated to the observation being studied or the other variables in the data set.
2. Missing at random (MAR): The fact that data are missing can be predicted from the other variables in the study, but not from the missing data themselves.

As we outline in the subsequent subsection, we handle these missing values through inputting.

3.3 Dataset preprocessing

3.3.1 Standardization

While scaling of the data is not a necessary step for random forest, we decided to scale the data before applying the imputing method since it uses a distance based measure (KNN). This is because we do not want the imputing to be affected by the magnitude of the features, which without scaling would be biased towards features with higher magnitude.

We scale the data through normalization. We chose this method over standardization since standardization assumes that your observations fit a Gaussian distribution with a well-behaved mean and standard deviation which is not the case for our datasets. While you can still standardize data that don't meet these expectations, reliable results are not guaranteed.

The data is normalized using the scikit-learn object MinMaxScaler, for which the a value is normalized as follows:

$$y = (x - min)/(max - min)$$

3.3.2 Imputing

As mentioned we have many missing values of the types MCAR and MAR. To handle these missing values we decided to impute them. Imputation is the process of substituting the missing values in the dataset. Much research has concluded that k-Nearest Neighbor Imputation (KNN Imputation) is the superior imputation choice [21, 10]. We tested this method along with replacing the values with the mean, median, most frequent value and a constant (zero), and witnessed the same results as the aforementioned papers, indicating that KNN Imputation is the superior method.

With KNN imputation, a new sample is imputed for those missing by finding the samples in the training set that are closest to it, using euclidean distance, and averaging these nearby points to fill in the value [28]. If the variable is a categorical one, the most frequent category is taken.

3.3.3 One-hot encoding

When we have categorical data, that is, data that takes only a limited number of values, are plugged into a machine learning models in Python without being encoded first we most likely cause an error. Thus coding category data is necessary for most machine learning models.

One of the most commonly used types of encoding is one-hot encoding. It is primarily used when there are not too many categories for the particular variable. It creates new binary columns that indicate the presence of each category in the variable.

Tree-based models, such as Random Forest don't usually work well with one-hot encoding when there are many levels. However in our case we do not have any categorical variables with more than two levels, for example "Gender" and thus one-hot encoding can be used for our datasets.

Chapter 4

Proposed Method

Considering the intention of this thesis is to classify two subjects, control and Pre-AD, we utilize classification models. To choose the most appropriate model we took a number of aspects into consideration, including the characteristic of data and the respective model, along with the performance of the classification model. From this analysis we deemed that Random Forest was the suitable model to classify the data. We use sklearn library to implement random forest and fine tune the model to be robust against overfitting. We compute the importance of all features to provide explainability of the model. Furthermore, considering the imbalanced nature of the data we propose a data augmentation for the minority class, namely Synthetic Minority Oversampling Technique (SMOTE).

In the preceding subsections we give a light introduction into Random Forest and expand on our motivations for the model choice. Following this we introduce the concept of SMOTE and finally explain the method of validation for the model.

4.1 Random Forest

Devised by L. Breiman in the early 2000s [19], Random Forest still remains one of the most popular machine learning ensemble methods, due to its adaptability to a wide range of prediction problems and having few parameters to fine tune. Additionally it is recognized for its accuracy, along with its capacity to handle small sample sizes and high-dimensional feature spaces.

Random Forest is an ensemble model based on decision trees that follows the bagging technique. The term bagging stands for bootstrapping and aggregating: rather than building one single predictor, a single decision tree in this case, the method creates many decision trees from random samples of data points drawn with replacement, known as bootstrapping. Once the forest of trees has been built, the aggregation happens where a sample is classified by taking the majority vote among all of the trees in the forest. The technique of bagging, with its multiple decision trees, helps in reducing the variance in the data, and increasing the number of trees will reduce the variance of the estimator. Random Forests also apply a method referred to as random subspace projection [18]. This process randomly selects subsets of features used in each data sample, preventing overfitting on features that are powerful predictors for the target class.

4.1.1 Random Forest algorithm

1. Ntree bootstrap samples are drawn from the data.
2. For each sample, a decision tree is built and a prediction is estimated.
3. Predict new data by aggregating the predictions of the ntree trees (majority votes).
4. The final prediction is the most voted prediction result.

Algorithm 1 Pseudo code for the random forest algorithm

To generate c classifiers:
for $i = 1 \rightarrow c$ **do**
    Randomly sample the training data D with replacement to produce D_i
    Create a root node, N_i containing D_i
    Call BuildTree(N_i)
end for
BuildTree(N):
if N contains instances of only one class **then** *return*
else
    Randomly select x ▷ of the possible splitting features in N
    Select the feature F with the highest information gain to split on
    Create f child nodes of $N, N_1, ..., N_f$, where F has f possible values $(F_1, ..., F_f)$
    for $i = 1 \rightarrow f$ **do**
        Set the contents of $N_i \rightarrow D_i$, where D_i is all instances in N that match F_i
        Call BuildTree(N_i)
    end for
end if

FIGURE 4.1: Pseudo code for Random Forest algorithm

4.1.2 Motivation for Random Forest

High dimensional data

Random forests work well with high dimensional data sets since it works with subsets of the data. This is an important characteristic for our purpose considering clinicians use many tests to assess the cognition state of subjects.

Feature Importance

Estimates of each feature's importance during the classification are computed by measuring how much it decreases the impurity at each split in the node, where the higher the decrease the greater the importance. This feature of random forest is crucial to demonstrate to the clinicians the importance of each feature (test), firstly to confirm that how the model gives importance to the features is aligned with their domain knowledge and secondly to provide a means of assessing the importance of newly introduced features.

Overfitting

They are able to cope with overfitting due to the power of averaging [2]. Thus when there is a robust number of decision trees in a random forest, overfitting is unlikely,

but not impossible, since the averaging of uncorrelated trees lowers the overall variance and prediction error. We fine tune the parameters of the model to avoid overfitting.

Unexcelled in performance among tested algorithms

We tested multiple classifiers and ensemble methods on the datasets and found random forest to be a high performer in terms of accuracy and area under the curve (AUC). A table of this performance can be found in the annex.

Robust to outliers

Outliers are essentially binned making random forest more robust to outliers compared to other models such as logistic regression.

Multicollinearity

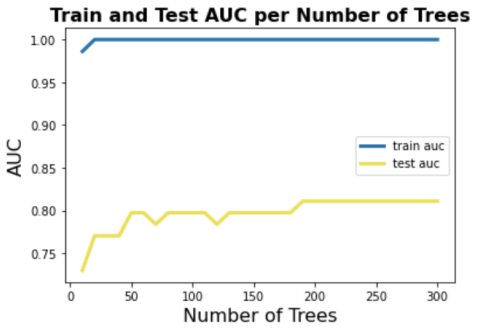
This occurs when features are strongly dependent on each other. If multicollinearity is present in the dataset, it is difficult to determine the importance of a feature. While some models are sensitive to multicollinearity, like logistic regression, this is less of a problem for random forest. This is because random forests select features one at a time and even if there are multiple features that are equally good, random forest can simply choose one of them at random.

4.1.3 Overfitting

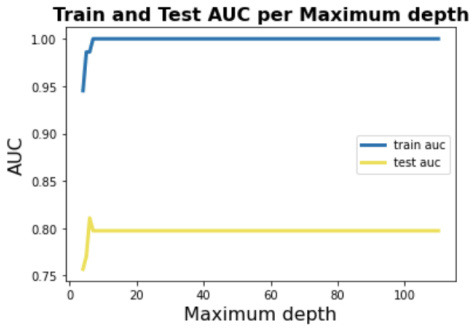
Overfitting occurs when a model learns the details and noise of the training dataset to the extent that it negatively impacts the performance of the model on unseen data. There has been a misconception on the definition of overfitting, with some believing that a 100% accuracy in the training set indicates overfitting. The confusion arises from mixing overfitting as a phenomenon with its indicators. A simple, but yet more accurate definition of overfitting is when a model is no longer as accurate as we want it to be on the data important to us.

Overfitting happens when adding additional complexity to a model results in the test error increasing. To ensure our model was not overfitting, we measured the model's performance (AUC) on the training dataset and the model's performance for unseen data through LOOCV. We do this for three of the model's parameters that are mostly associated with overfitting, in particular, the number of trees, the maximum depth and the minimum number of samples per leaf. We take a range of values for each of the parameters and plot the training and test AUC for each of the parameters values in the defined range. If overfitting is present we expect to see a divergence between the training and test AUC.

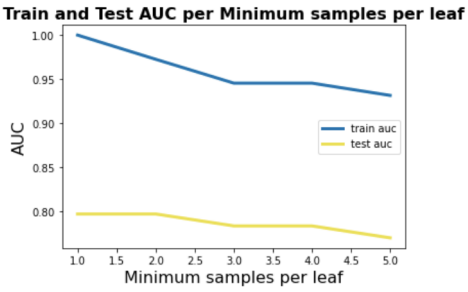
Below are the graphs of this analysis per parameter. From these we see that neither the number of trees nor the maximum depth paramaters are an issue. However it is evident that when the parameter for the minimum samples per leaf is set at 1 overfitting takes place. We avoid this by setting it equal to 2.



(A) Number of trees



(B) Maximum depth



(C) Minimum samples per leaf

FIGURE 4.2: Graphs for checking overfitting

4.2 Handling Imbalanced data

Imbalanced data is common in medical datasets, where the classes in a dataset have a highly unequal number of samples. Traditional classification methods, including Random Forest, perform poorly on minority class examples as they assume the data is balanced and their misclassification cost during training is equal. However the misclassification cost of patient samples is higher than that of healthy subjects [37]. It is therefore crucial to improve the identification of patients without disturbing the classification of healthy subjects.

There are two primary approaches to deal with imbalanced data while using the random forest model, namely cost-sensitive learning and sampling. The prior, cost-sensitive approach, works by assigning different weights to the classes, where a higher weight is assigned to the minority class and thus assigning a higher misclassification cost. This helps reduce the biases towards the majority class. The latter approach, sampling, is achieved by either undersampling the majority class or oversampling the minority class. While undersampling is an efficient strategy, it throws away many potentially useful data and considering our datasets are relatively small, oversampling is more suitable.

In terms of oversampling there are many techniques that can improve the model performance . We chose SMOTE [30], a widely used technique that has been shown to be an optimal choice [12].

4.2.1 SMOTE

We use SMOTE from the imblearn library with the default parameters. This technique synthesises new minority instances from existing minority instances. First a random instance is taken from the minority class, then k of the nearest neighbors are identified (in our case k=5). One of these neighbours is selected at random and a new synthetic instance is created at a randomly selected point between the two instances in feature space [29]. The technique is effective since the new synthetic instances from the minority class are relatively close in feature space to existing instances from the minority class.

4.3 Validation Method

4.3.1 Performance Metrics

To validate the performance of our model we utilized three performance metrics.

Area under the curve (AUC)

Commonly used in binary classification, it measures the ability of a classifier to distinguish between classes, the higher the AUC, the better the model is at distinguishing between the classes.

Sensitivity

Measures how often the model correctly identifies people who have the condition (Pre-AD). This is a key metric since the cost of misclassifying a subject with Pre-AD is higher to that of misclassifying a healthy subject.

Specificity

Measures how often the model correctly identifies people who do not have the condition (Pre-AD). This metric is not as important as the sensitivity since the cost of misclassifying a healthy subject is lower to that of misclassifying a subject with Pre-AD .

4.3.2 Cross Validation

Cross-validation, or k-fold cross-validation [22], is a model evaluation method better than residuals that estimates the performance of a model on previously unseen data. Under cross-validation the dataset is divided into k subsets, and the model is trained k times. Each time, one of the k subsets is used as the test data and the remaining k-1 subsets are combined to form a training set. The k-fold cross-validation estimate is then computed as the mean of the evaluation metric over the k models.

There is no correct answer to what size k should be chosen, the main point of cross-validation is to ensure that the training and validation splits represent, as much as possible, the variety in the underlying population distribution. For example if the samples are all biased compared to the actual population, cross validation will be of no help.

Leave-one-out cross-validation (LOOCV) is the most extreme form of cross-validation, with K equal to N, the number of data points in the dataset. An attractive property of LOOCV is that it provides an almost unbiased estimate of the test error [13]. However, this improved estimate of model performance comes with a high computational cost and thus may not be suitable for large datasets.

LOOCV is usually the preferred choice when dealing with small datasets, since it allows for the smallest amount of data to be removed from the training data at each

iteration. Considering our dataset is small and an accurate estimate of model performance is critical we chose to use LOOCV.

Chapter 5

Results

5.1 Dataset 1

As expected the model has the ability to classify effectively when the groups differ cognitively, that is, one has cognitive impairment while the other has no cognitive complaints, for example while classifying the Control and MCI groups. This is because the tests being used to make the classification were the tests used to assign the grouping in terms of cognitive impairment.

For the primary classification of the thesis, being the classification of Control and Preclinical, the model fails to classify the groups correctly with an area under the curve value of 0.561 as seen in the below table. This is because by definition the two groups are deemed cognitively normal. The sensitivity is strikingly low at 0.16, indicating that the ability to detect Preclinical subjects is low, and thus there is a high number of people classified as Control that are actually Preclinical. This illustrates that the tests used as features in the classification are not sensitive enough to detect preclinical subjects, which is aligned to many other research papers [9, 4, 27].

	Accuracy	AUC	Sensitivity	Specificity
Preclinical vs Control	0.771	0.561	0.16	0.983
MCI vs Control	0.935	0.923	0.963	0.884

FIGURE 5.1: Dataset 1 performance

The feature importance scores of the tests allow us to understand which tests are most important during the classification. This could provide insights for the clinicians to understand which tests are pertinent and those that are not in order to design new tests for detecting incipient cognitive dysfunction in pre-AD subjects.

Below is a table indicating the importance score of each individual feature of the model, where the higher the score the greater the importance of the feature. From this we see that the bottom 5 features are a cognition test (ideom), comprehension test (compren), gender (Sexo), praxis test (prax_cerad) and a cognition (tdp) indicating they are the least sensible measures for detecting Preclinical. It should be noted that these features also appeared in the bottom for the classification between Control

and MCI which is more reliable given the high accuracy and AUC values. The top 5 features included 3 of the buschke tests, a trail making test (tmtA) and a stroop color test (Stroop_color).

feature	importance
Stroop_color	0.086022
buschke_RDL	0.085814
tmtA	0.063972
buschke_AT	0.059640
buschke_AL	0.059539
bnt	0.058305
edad	0.056946
tam	0.056600
paisajes_tot	0.053669
tmtB	0.048528
Stroop_lect	0.047191
fas_total	0.041240
Clave_num	0.040516
Stroop_I	0.040143
flu_anim	0.034117
años_escol	0.030248
VOSP_num	0.029152
vis_cerad	0.029130
buschke_RDT	0.028984
VOSP_letras	0.015064
tdp	0.014456
prax_cerad	0.013628
Sexo	0.006559
compren	0.000537
ideom	0.000000

FIGURE 5.2: Dataset 1 feature importance

5.2 Dataset 2

In this dataset there are traditional battery tests, similar to what we saw in the previous dataset 1, along with a new test focused on tapping rate and intrasubject variability.

The objective for this analysis is to understand if the new tests (tapping) are more sensitive than that of the traditional battery tests for classifying Preclinical and Control subjects. It should be noted that this dataset does not contain MCI patients that were available in dataset 1.

Firstly we ran the model on the battery tests and found that they presented no predictive power with an AUC of 0.48. This reinforces the conclusion we found in the previous dataset, that battery tests are not sensitive enough to incipient cognitive dysfunction in pre-AD subjects .

Next we conducted the model with the tapping features, tapping rate and tapping variability. Here we found that the AUC improved dramatically with an AUC value of 0.721 as seen in the below table. This indicates that the tapping test features are more sensitive to subtle cognitive issues in Preclinical subjects.

	Accuracy	AUC	Sensitivity	Specificity
Tapping Features	0.772	0.721	0.55	0.892
Battery Features	0.579	0.48	0.15	0.811

FIGURE 5.3: Dataset 2 performance

The only features that show evidence of being sensitive enough to identify cognitive dysfunction in preclinical subjects are the tapping features. However the sensitivity is very low at 0.550 due to the imbalance nature of the dataset and thus we apply SMOTE to make the dataset balanced. The performance metrics improve thanks to SMOTE with an increase of 0.076 in AUC (0.797) and an increase of 0.261 in sensitivity (0.811)

We then computed the feature importance scores and found that intrasubject variability was the most important feature with a score of 0.72 out of 1. This is aligned to other motor dysfunction studies, Verghese et al. (2008) [16] indicated that a group of subjects In the early stages of AD presented greater variability when walking than the control group.

While this result suggests that tapping features are more sensitive than battery tests, we have not been successful at effectively identifying cognitive dysfunction in preclinical subjects with a sensitivity value of 0.811. We would need more new features sensitive enough to detect cognitive dysfunction in preclinical subjects.

Chapter 6

Conclusions

The goal of this thesis was to provide a tool for clinicians to validate if a test is sensitive enough to detect incipient cognitive dysfunction in pre-AD subjects. Furthermore to provide a means of understanding what features are important in the classification. We can conclude that we have achieved these goals by providing a model, in particular Random Forest, that is easily understandable for the clinicians and provides a means to understand the importance of a test thanks to the feature importance score.

We found that imputing the missing values through k-Nearest Neighbor Imputation was the most successful in terms of performance of the model. Additionally we concluded that data augmentation through SMOTE was crucial to correct the imbalanced nature of the datasets and substantially improved the model performance.

We tested the model on two different datasets provided by the clinicians, the first containing neuropsychological tests and the second containing relatively new features based on finger tapping. We found that the neuropsychological tests were not sensitive enough to detect incipient cognitive dysfunction in pre-AD subjects. While the new tapping tests were more successful, none of the tests were sufficient enough to effectively identify incipient cognitive dysfunction in pre-AD subjects, and thus additional testing is needed for this purpose.

Chapter 7

Source code

All the source code used to develop this project is available on GitHub: Link

Appendix A

Model selection performance

Dataset 1

	Accuracy	AUC	Sensitivity	Specificity
Random Forest	0.76	0.56	0.16	0.95
Logistic Regression	0.63	0.56	0.44	0.81
SVM (Linear Kernel)	0.77	0.52	0.04	1.00
SVM (Polynomial Kernel)	0.75	0.54	0.12	0.95
SVM (Radial Basis Function Kernel)	0.76	0.50	0.00	1.00
Naive Bayes	0.38	0.53	0.80	0.25
XGBoost	0.75	0.60	0.32	0.89
Adaboost	0.74	0.57	0.24	0.83
Ensemble Hard Voting	0.70	0.54	0.24	0.84
Stacking	0.76	0.60	0.28	0.91

Dataset 2

	Accuracy	AUC	Sensitivity	Specificity
Random Forest	0.77	0.72	0.56	0.89
Logistic Regression	0.68	0.57	0.20	0.95
SVM (Linear Kernel)	0.56	0.44	0.05	0.84
SVM (Polynomial Kernel)	0.72	0.67	0.50	0.84
SVM (Radial Basis Function Kernel)	0.77	0.72	0.56	0.89
Naive Bayes	0.67	0.64	0.55	0.73
XGBoost	0.65	0.60	0.45	0.76
Adaboost	0.74	0.68	0.50	0.87
Ensemble Hard Voting	0.72	0.65	0.40	0.89
Stacking	0.74	0.69	0.55'	0.84

FIGURE A.1: Model performance during the model selection process

List of Figures

2.1 Alzheimer's Disease stages 4

4.1 Pseudo code for Random Forest algorithm 16

4.2 Graphs for checking overfitting 18

5.1 Dataset 1 performance 23

5.2 Dataset 1 feature importance 24

5.3 Dataset 2 performance 25

A.1 Model performance during the model selection process 31

Bibliography

[1] Gupta A., Ayhan M., and A. Maida. "Natural image bases to represent neuroimaging data". In: *International Conference on Machine Learning* 28 (2013), pp. 987–994.

[2] Wyner Abraham J et al. "Explaining the success of adaboost and random forests as interpolating classifiers". In: (2015).

[3] Casamitjana Adrià et al. "MRI-Based Screening of Preclinical Alzheimer's Disease for Prevention Clinical Trials." In: *Journal of Alzheimer's disease* 64.4 (2018), pp. 1099–1112. DOI: http://dx.doi.org/10.3233/JAD-180299.

[4] Tort-Merino Adrià et al. "Early Detection of Learning Difficulties when Confronted with Novel Information in Preclinical Alzheimer's Disease Stage 1." In: *Journal of Alzheimer's disease* 58.3 (2017), pp. 855–870. DOI: http://dx.doi.org/10.3233/JAD-161173.

[5] Serrano-Pozo Alberto et al. "Neuropathological alterations in alzheimer disease". In: *Med* 1.1 (2011).

[6] Buchman Aron S and Bennett David A. "Loss of motor function in preclinical Alzheimer's disease". In: *Expert review of neurotherapeutics* 11.5 (2011), pp. 665–76. DOI: http://dx.doi.org/10.1586/ern.11.57.

[7] Alzheimer's Association. "Alzheimer's disease facts and figures." In: *Alzheimer's Dement* 16.3 (2015), pp. 391–460.

[8] J. Chen, Ji Zhu, and Jieping Ye. "An Attention-Based Hybrid Network for Automatic Detection of Alzheimer's Disease from Narrative Speech." In: *INTERSPEECH* (2019).

[9] Rentz D. M. et al. "Promising developments in neuropsychological approaches for the detection of preclinical Alzheimer's disease: a selective review. Alzheimer's Research Therapy". In: 58 (2013). DOI: http://dx.doi.org/10.1186/alzrt222.

[10] Murti D. M. P. et al. "K-Nearest Neighbor (K-NN) based Missing Data Imputation." In: *5th International Conference on Science in Information Technology* 11.1 (2019), pp. 83–88. DOI: http://dx.doi.org/10.1109/ICSITech46713.2019.8987530.

[11] Siedlecki-Wullich Dolores et al. "Altered microRNAs related to synaptic function as potential plasma biomarkers for Alzheimer's disease." In: *Alzheimer's research and therapy* 11.1 (2019). DOI: http://dx.doi.org/10.1186/s13195-019-0501-4.

[12] Kovács G. "An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets, Appl. Soft Comput". In: 83.2 (2019), p. 109.

[13] Cawley Gavin and Talbot Nicola. "Efficient leave-one-out cross-validation of Kernel Fisher discriminant classifiers." In: *Pattern Recognition* 36 (2003), pp. 2585–2592. DOI: http://dx.doi.org/10.1016/S0031-3203(03)00136-5.

[14] Hardy J A and Higgins G A. "Alzheimer's disease: the amyloid cascade hypothesis." In: *Science* 256.5054 (1992), pp. 184–5. DOI: http://dx.doi.org/10.1126/science.1566067.

[15] Cummings J. and Fox N. "Defining disease modifying therapy for Alzheimer's disease," The journal of prevention of Alzheimer's disease." In: *The journal of prevention of Alzheimer's disease* 4.2 (2017), p. 109.

[16] Verghese Joe et al. "Gait dysfunction in mild cognitive impairment syndromes." In: *Journal of the American Geriatrics Society* 56.7 (2017), pp. 1244–51. DOI: http://dx.doi.org/10.1111/j.1532-5415.2008.01758.

[17] Johnson K.A et al. "Update on appropriate use criteria for amyloid PET imaging: Dementia experts, mild cognitive impairment, and education." In: *Alzheimer's Dementia* 9 (2013), pp. 106–109. DOI: http://dx.doi.org/10.1016/j.jalz.2013.06.001.

[18] Tin Kam Ho. "A data complexity analysis of comparative advantages of decision forest constructors." In: *Pattern Analysis Applications* 5.2 (2002), pp. 102–112.

[19] Breiman L. "Random Forests." In: *Machine Learning* 45 (2001), pp. 5–32. DOI: http://dx.doi.org/10.1023/A:1010933404324.

[20] Shaw Leslie M et al. "Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects." In: *Annals of neurology* 64.4 (2009), pp. 403–13. DOI: http://dx.doi.org/10.1002/ana.21610.

[21] Pazhoohesh M. et al. "A Comparison of Methods for Missing Data Treatment in Building Sensor Data." In: *IEEE 7th International Conference on Smart Energy Grid Engineering* (2019), pp. 255–259. DOI: http://dx.doi.org/10.1109/SEGE.2019.8859963.

[22] Stone M. "Cross-validatory choice and assessment of statistical predictions." In: *J. R. Stat. Soc.* 36.1 (1974), pp. 111–147.

[23] Raju Manu et al. In: *IOP Conf. Ser.: Mater. Sci. Eng. 1084 012017* (2021).

[24] Mollica Maria A et al. "Early detection of subtle motor dysfunction in cognitively normal subjects with amyloid-Beta positivity." In: *a journal devoted to the study of the nervous system and behavior* 121 (2015), pp. 117–124. DOI: http://dx.doi.org/10.1016/j.cortex.2019.07.021.

[25] Mollica Maria A et al. "Early detection of subtle motor dysfunction in cognitively normal subjects with amyloid- positivity."Cortex." In: *a journal devoted to the study of the nervous system and behavior* 121 (2019), pp. 117–124. DOI: http://dx.doi.org/10.1016/j.cortex.2019.07.021.

[26] Albert Marilyn S et al. "The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease." In: *the journal of the Alzheimer's Association* 7.3 (2011), pp. 270–9. DOI: http://dx.doi.org/10.1016/j.jalz.2011.03.008.

[27] Albers Mark W et al. "At the interface of sensory and motor dysfunctions and Alzheimer's disease." In: *Journal of Alzheimer's disease* 11.1 (2015), pp. 70–98. DOI: http://dx.doi.org/10.1016/j.jalz.2014.04.514.

[28] Kuhn Max and Kjell. Johnson. "Applied predictive modeling. New York: Springer." In: (2013).

[29] Japkowicz N. *Assessment Metrics for Imbalanced Learning.* Wiley, 2013.

[30] Chawla N.V. et al. "SMOTE: synthetic minority over-sampling technique." In: *Journal of Artificial Intelligence Research* 16 (2002), pp. 321–357.

[31] Del Campo Natalia et al. "Relationship of regional brain -amyloid to gait speed". In: *Neurology* 86.1 (2016), pp. 36–43. DOI: http://dx.doi.org/10.1212/WNL.0000000000002235.

[32] Reitan R. "he halstead-Reitan neuropsychological test battery: Theory and clinical interpretation. Tucson: Neuropsychology Press." In: *Neuropsychology Press.* ().

[33] Sperling Reisa A et al. "Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease." In: *the journal of the Alzheimer's Association* 7.3 (2011), pp. 280–92. DOI: http://dx.doi.org/10.1016/j.jalz.2011.03.003.

[34] Brookmeyer Ron et al. "Forecasting the global burden of alzheimer's disease." In: *Alzheimers.Dement* 3.3 (2007), pp. 186–191.

[35] Wilcockson Thomas D W et al. "Abnormalities of saccadic eye movements in dementia due to Alzheimer's disease and mild cognitive impairment." In: *Aging* 11.15 (2017), pp. 5389–5398. DOI: http://dx.doi.org/10.18632/aging.102118.

[36] Hadoux X et al. "Non-invasive in vivo hyperspectral imaging of the retina for potential biomarker use in Alzheimer's disease." In: *Nat Commun* 10.4227 (2019). DOI: http://dx.doi.org/10.1038/s41467-019-12242-1.

[37] Xu Zhaozhao et al. "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data." In: *Journal of biomedical informatics* 107.103465 (2020). DOI: http://dx.doi.org/10.1016/j.jbi.2020.103465.