Machine Learning based biomarkers for neurodegenerative disease classification

PhD

Department of Computer Science

Ali Varzandian

October 2022

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Ali Varzandian

Acknowledgement

The PhD journey was a rewarding journey which did at times become challenging and stressful. I would like to thank my supervisor Prof. Giuseppe Di Fatta for his support. I would also like to thank my family for giving me moral and emotional support throughout this journey.

Abstract

In this thesis a novel classification model framework to predict Alzheimer's Disease is described. In this work a novel brain age feature is proposed, which estimates the biological age of parts of the brain affected by Alzheimer's Disease. This feature can act as a biomarker for medical professionals which together with age, can make an Alzheimer's Disease prediction with high performance. In addition to this feature, a novel interpretable classification framework is proposed for prediction of AD which can achieve high classification performance. Also, a novel interpretability index is also proposed which indicates to the medical professionals why such prediction has been made and which input features had the greatest impact on the final output. The brain age Alzheimer's Disease prediction model is also applied to other type and stages of dementia in a multi-class classification setting as an extension of the work. The results achieved in this thesis in both binary and multi-class classification are comparable to the baseline and relevant previous literature. The binary classification accuracy achieved are 92.84% and 89.74% for female and male subjects respectively.

List of Abbreviations

Abbreviation	Meaning
ABA	apparent brain age
ABA-Clf	the logistic regression model using Age and ABA to classify dementia
ABA-Com	a brain age AD prediction model comprising of ABA-Reg and ABA-Clf
ABA-Reg	the LASSO regression model to estimate ABA
ABOD	angle-based outlier detection
AD	Alzheimer's Disease
ADNI	the Alzheimer's Disease neuroimaging initiative
AIBL	the Australian imaging, biomarker & lifestyle flagship study of ageing
CAD	computer-aided diagnosis
CN	cognitive normal
CNR	contrast-to-noise ratio
ELM	extreme learning machine
EMCI	early mild cognitive impairment
f_i	feature i; referring to single input feature
fMRI	functional magnetic resonance imaging
FTD	Frontotemporal dementia
HC	healthy control
iForest	isolation forest
IXI	information eXtraction from Images
LBD	Lewy body dementia
LMCI	late mild cognitive impairment
LOF	local outlier filter
MAE	mean absolute error
MCI	mild cognitive impairment
MRI	magnetic resonance imaging
NIFD	neuroimaging in Frontotemporal Dementia
OASIS	open access series of imaging studies
PCA	principal component analysis
PET	positron emission tomography
PPMI	Parkinson's progression markers initiative
ROI	region of interest
RVR	relevance vector regression
s_i	proposed feature score for f_i
sMRI	structural magnetic resonance imaging
SPECT	single photon emission computed tomography
SVM	support vector machine
SVR	support vector regression
VBM	voxel-based morphometry
VD	vascular dementia

Abbreviation	Meaning
VFI	voting feature intervals

Table of Contents

Chapter 1

1. Introduction

1.1. Motivation

1.1.1.Introduction to Alzheimer's Disease

Dementia is a neurodegenerative disease with several symptoms such as linguistic dysfunctions, problems in memory, difficulty performing simple every-day tasks and changes to psychiatric and psychological state [1].

Main types of dementia include Alzheimer's disease (AD), Frontotemporal dementia (FTD), Vascular dementia (VD) and Lewy body dementia (LBD).

Definitive diagnosis of dementia is only possible at the brain autopsy after death and this makes the diagnosis in living subjects very challenging [2].

Alzheimer's disease (AD) is a chronic and terminal neurodegenerative disease currently affecting approximately 6% of people aged 65 and older

worldwide. AD is the most common type of dementia and affects the memory of the people affected [1].

At the early stages of AD, patients suffer from a condition referred to as mild cognitive impairment (MCI). This condition can be viewed as a mild case of AD and could get worse and convert to AD or it could get better and disappear [3].

Clinical diagnosis of AD is a time consuming, costly and challenging process involving different types of examinations such as mental and neuropsychological tests, lab tests using blood, urine and cerebrospinal fluid and brain imaging tests such as MRI, CT and PET. This process also involves ruling out other diseases with similar symptoms, in order to avoid misdiagnosis. Although, multiple tests are carried out and different specialists are involved in reviewing the test results, it is reported that in the year 2017/18 only 66% of subjects with AD in the UK were diagnosed with AD and the complexity of the diagnosis process can often be the cause for missed diagnosis [4].

It is important for AD to be diagnosed correctly and in a timely manner [4]. With a large cost involved and limited treatment options, the aim should be to diagnose AD as early as possible in order to slow down the progression of the disease [4].

1.1.2. Benefits of early diagnosis

Early diagnosis of AD has great positive influences on the patients, their carers and the economy in which [4] has studied and reported about these effects.

To name a few advantages of early diagnosis of AD it can be said that when the patients are noticing the very early signs and are unsure of the reason behind them, early diagnosis can stop any doubts the patients may have and give a clear justification and reason behind those early signs.

Not only early diagnosis can provide an answer, but it can also give patients the right to use correct healthcare and medication. This will then enable them to manage the condition and have a longer independent life where they can live on their own while performing everyday tasks unaided. As the result they will preserve a life with good condition and quality for themselves and their family and carers.

At the early stages of the disease life with good condition and quality can be preserved for a few years. Patients who have had the diagnosis at the early stages of the disease and have been advised about their condition, can decide their own futures when they still are capable of making decisions about issues such as the prospective medical care, healthcare and support arrangements, legal matters such as will and financial matters such as properties and investments, and whether the patients want to get the relatives and members of the family involved or make them aware of those decisions.

In healthcare services, the only way to get admission to health and medical care and have access to the medications for AD to improve the life condition and quality is through getting diagnosed. A study was performed on 8995 recently-diagnosed patients with AD from a database in the US. This study divided the patients into two groups of patients with treatment and patients without treatment.

Those patients who received treatment for AD showed a higher survival rate and 20% less chance of being hospitalised and institutionalised [4].

In most cases members of the family or spouses are the ones who care for patients who have been diagnosed with AD and as this disease could last for a long time there could be high pressure on the carers. The care the family members are providing is most of the time unpaid. To estimate the value of unpaid work carried out by 16 million Americans who cared for member of the family with AD the Alzheimer's Association made a report. In this report it is estimated that in the year 2017, unpaid work totalling 18.4 billion hours including different types of support such as emotional, financial and physical, comes to a national value of $232.1 billion [4].

In addition to the benefits of early diagnosis of AD mentioned above, it is worth mentioning that early diagnosis also gives plenty of time to the family members who will be potential carers to adapt to the forthcoming changes as the result of AD in the behaviour and characteristics of the patient and also to adjust to the change in their role from a family member to a supportive carer. Those carers who have resilience and can adjust to their new role have shown to suffer less psychological issues and conditions such as depression and anxiety.

There are great costs associated with the care for AD. These costs could generally be grouped into 3 categories of care; informal care, social care and health care. The first category is informal care which involves caring for the patient with AD by a member of the family. This type of care incurs direct and indirect costs to

the family members. The second category is the social care which includes the care and nursing homes, and respite care and home care. The third category is the health care which includes the hospitalisation and institutionalisation of the patient with AD. Second and third categories have indirect costs both to families and government. The cost from all 3 categories combined associated with dementia in the UK is over £26 billion annually. To give a cost per category: health care, social care and informal care cost £4.3 billion, £10.3bn and £11.6bn respectively. Out of these total costs which are estimated for dementia, 65% are related to AD [4].

Early diagnosis of AD could cost a large amount of money at the beginning but in the long run there will be a reduction in total costs associated with the disease and caring for the patient will be as a result of less need for hospitalisation and care.

Early intervention in diagnosis of AD is the best and most favoured approach to dealing with the disease as it helps the patient maintain independence and function normally for a longer period.

There are no medications to reverse the effect of AD on the brain but early diagnosis can help people related to the patient to adjust to the situation psychologically and mentally and also gives time to the patient to manage all their financial and legal affairs themselves and making decisions for themselves. There are many people who after an early diagnosis carry on living a high-quality life for a number of years while making the use of medications and treatment plans.

1.2. Problem definition and challenges of AD prediction

Although early diagnosis of AD should be the option for every patient, it is not always possible considering the challenges that exist. One of the challenges is that as AD is affecting a large number of people and there are not enough medical professionals to make early diagnoses for all patients. Another challenge is that in order to clearly identify AD symptoms and avoid misdiagnosis specialist knowledge is required but often medical professionals dealing with AD do not have that knowledge therefore there will be a lack of confidence in the correct and timely diagnosis.

Another challenge facing the early diagnosis of AD is that other diseases which are mostly physical are often prioritised before other mental diseases. These challenges were some of the challenges of early diagnosis of AD and these are the reasons why early diagnosis is so hard to achieve. [5] Also, diagnosing AD is a lengthy process involving long waiting times which means the right treatment plans may not always be put in place in time to slow the disease down.

In order to help with the diagnosis of AD so that a novel prediction model is proposed in this thesis. This model can be used by the medical professionals as an indicator to presence of AD while providing an explanation for that indication. This can reduce the overall diagnosis process resulting in an earlier diagnosis.

1.3. Key assumptions and the scope of research

This thesis focuses on prediction of AD using the medical imaging data of the brain belonging to healthy subjects and those with the disease. As AD is a neurodegenerative disease the assumption is that AD will cause faster degeneration and atrophy in the brain. This will cause the age of the brain of subjects with AD to look older than their real age due to this degeneration. This difference between the real age and the age of the brain in the subjects with AD will be used to predict AD in this thesis.

The scope of this research covers analysing brain imaging data for subjects who have voluntarily gone through the scanning process. The data used are downloaded from public repositories which will be explained in the next chapter. The method proposed in this thesis is designed to give an indication to the medical professionals as to whether a subject is suffering from AD while providing an explanation of how such indication is made.

1.4. Aim and Objectives

The ultimate aim and the goal of the research in this thesis is to improve the prediction of AD and have a better understanding of that prediction which is made by machine learning (ML) models. The specific problem that this research is attempting to solve is to improve the detection of AD while maintaining the interpretability of the model. Therefore, following objectives are set out for this research:

- Creation of a descriptive and interpretable linear ML model using brain structural MRI scans for prediction of AD which can achieve comparable predictive performance to the black-box state-of-the-art models.

- Creation of a linear ML model to eliminate the complexity of black-box models.

- Achieving ML model interpretability through combining linear models by defining a linear index while keeping the predictive performance high.

- Application of ML algorithms to create a brain age feature which acts as a biomarker to help in diagnosis of AD.

- Creation of a feature selection model which selects the minimal number of features with maximal performance which are most helpful in predicting AD.

- Application of AD prediction model using brain age feature in a multi-class setting to predict multiple types and stages of dementia while achieving a comparable accuracy to state-of-the-art model.

1.5. Contributions

- A novel feature selection method to identify the most predictive brain regions for prediction of AD.

- A novel brain age feature which is used as a biomarker to AD by showing the difference between real age and brain age in subjects with AD.

- A novel prediction framework for classification of AD.

- A novel interpretability index which explains about the decision making behind AD prediction.

1.6. Introduction to thesis and the proposed method

This thesis focuses mainly on detection of AD as it is the main type of dementia and affects more people than other types and for completeness and as a logical progression of the research, in addition to AD, the proposed framework will be applied to another type of dementia i.e., FTD and three stages of dementia i.e., mild cognitive impairment (MCI), early MCI (EMCI) and late MCI (LMCI).

Considering the impact that AD has on the world population, the importance of correct and timely diagnosis of it and to help with the complex process of AD diagnosis, as discussed previously, the aim of this thesis is to improve the prediction of AD, and to improve the understanding and reasoning behind the that prediction made by machine learning models. This thesis also aims to improve the timely detection of AD by proposing an automatic CAD method with high accuracy.

In [6] a framework is proposed to model the aging of the brain of healthy subjects. This framework aims to estimate the age of subjects which will be referred

to as brain age and for this estimation a regression model is built on selected brain features from all parts of the brain with target variable as real age. This regression model is built using healthy subjects and the aim is to get the estimated (brain) age as close as possible to real age. This way, when the model is applied to subjects with AD, the estimated (brain) age is expected to be higher than real age due to faster aging and atrophy in the brain.

In the proposed method, following the selection of a minimal set of features which are highly affected by AD, a regression model is built on those selected features with target variable as real age. This is to estimate the brain age of those selected features, which will be referred to as Apparent Brain Age (ABA). The aim of ABA-Com method is not to predict the age of the subjects but to estimate age of a subspace of brain features.

For the estimation of ABA, the target variable in regression model should be the age of those selected features when subject is healthy. In order to estimate the age of those features, it is assumed that brain age of a healthy subjects is the same as their real age, so we can use the real age instead. The ABA therefore builds a framework to estimate the age of specific parts of the brain which are affected by AD.

For BrainAGE approach [6] age of subjects is aimed to be estimated and then referred to as brain age (target variable is real age). We however predict the brain age of specific parts of the brain with target variable as brain age (assumed to be as real age).

A journal paper has been published as part of this PhD project on the ABA-Com model which presents the interpretable prediction framework for binary classification of AD [7].

1.7. How this thesis is organised

Chapter 2 explains about the data used in this thesis which includes the description of structural MRI scans and how they are preprocessed to extract numerical measurements from those scans. This chapter also explains about the outlier detection method used to filter out a limited number of images identified as outliers before applying the proposed model.

Chapter 3 provides an overview of the previous work done in the field of automatic CAD of AD. The studies reviewed in this chapter have all used the MRI scans to predict AD. The reason for this selection is to have a similar criterion to what is analysed in this thesis. The prediction of AD using MRI scans are performed using the brain regions directly as features in the machine learning models or indirectly by creating a biomarker and use that as a feature in machine learning models to predict AD. The results reported in this thesis are compared to both types of studies using direct and indirect predictions.

Chapter 4 introduces the proposed novel Apparent Brain Age (ABA) feature. This feature will be used as a biomarker which can be used in a machine learning model as a feature to predict AD. This chapter explains how ABA is different to other brain age features suggested by other previous studies and compares the performance of ABA with previously suggested brain age models. Although ABA

performs worse than other models in predicting age, it has a superior performance to other models when classifying AD. This will be explained in chapter 5.

Chapter 5 introduces the ability of ABA to classify AD. Whilst ABA was introduced in chapter 4 and the regression performance of the feature was analysed, in this chapter a novel classification workflow is proposed where ABA is used as a biomarker or feature in addition to real age (2 features in total) in a classification algorithm in order to predict AD. This chapter shows that the superior performance of ABA in CN vs. AD classification task compared with the baseline method where state-of-the-art SVM algorithm is used.

Chapter 6 provides the details about interpretability and descriptiveness of the proposed workflow. In this chapter a novel feature score is proposed where it will be used to show the contribution of each brain region on the classification of AD. Using this score it is clear that which brain features have been selected and how much they contribute to the classification task.

Chapter 7 presents a multi-class classification of dementia types and stages using the ABA workflow. Up to chapter 7, the ABA model was focused on binary classification of AD vs. CN. A logical extension of that work was to apply the ABA classification workflow to more than two classes therefore a 6-class multi-class classification of CN, EMCI, LMCI, MCI, AD and FTD is proposed in chapter 7.

Chapter 8 provides a conclusion on the results reported in the thesis and outlines the future works identified.

1.8. Summary

This chapter provided an introduction into this thesis including an introduction into the dementia prediction, the challenges involved and the motivation to propose a novel method in this domain. It is also discussed what the research objectives and the original contributions are in this thesis. In the next chapter the data and the preprocessing the data will be discussed.

Chapter 2

2. Data and experiments

2.1. Subjects ROIs data

This chapter explains the data used in this thesis. To perform the analysis and model building process a type of neuroimaging scan of human brain is selected which can be used to detect patterns caused by AD in the brain. The type of scans selected is Magnetic Resonance Imaging (MRI). MRI itself has different types such as functional and structural and as this thesis focuses on detection of patterns from anatomical and morphological changes in the brain, Structural MRI (sMRI) is selected as the data type.

The sMRI scans can be acquired using two weightings: T1 and T2. T1 is used to show fat in tissues (brain structures) and T2 to show fat and water (CSF). As this thesis focuses on analyses of brain structures (white and grey matter) T1 weighting is selected.

Table 2.1 | Distribution of the 3,170 subjects adopted in this study. 1,567 female and 1,603 male.

Gender	Source	Group	Number of subjects	Age Mean	Age Std	Age Min	Age Max
M	ADNI	AD	213	75.82	7.86	55.30	90.40
M	ADNI	CN	317	74.20	6.36	56.20	90.30
M	ADNI	EMCI	179	71.88	7.21	55.00	89.00
M	ADNI	LMCI	97	73.48	7.19	56.00	91.00
M	ADNI	MCI	227	75.35	7.28	54.60	89.80
M	AIBL	AD	32	74.83	8.65	60.40	89.40
M	AIBL	CN	199	74.30	7.85	54.60	89.80
M	AIBL	MCI	52	74.42	6.30	57.80	85.50
M	IXI	CN	90	65.50	7.27	55.09	86.20
M	NIFD	FTD	116	64.42	5.88	55.00	85.00
M	PPMI	CN	81	65.98	7.43	55.00	83.00
F	ADNI	AD	177	74.29	8.07	55.20	91.00
F	ADNI	CN	398	72.10	6.24	55.60	89.90
F	ADNI	EMCI	152	70.26	7.75	56.00	88.00
F	ADNI	LMCI	79	70.57	7.55	55.00	85.00
F	ADNI	MCI	132	73.48	7.64	55.20	86.20
F	AIBL	AD	44	75.34	7.88	56.30	88.40
F	AIBL	CN	272	74.39	7.34	55.20	88.00
F	AIBL	MCI	48	75.62	6.12	60.20	86.60
F	IXI	CN	143	65.10	6.31	55.22	86.32
F	NIFD	FTD	86	65.51	5.63	56.00	79.00
F	PPMI	CN	36	64.86	7.09	55.00	82.00

To acquire sMRI data, there are several public directories of which the following five datasets were selected: Alzheimer's Disease Neuroimaging Initiative (ADNI), Information eXtraction from Images (IXI), The Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL), Neuroimaging in Frontotemporal Dementia (NIFD) which is the nickname for the frontotemporal lobar degeneration neuroimaging initiative (FTLDNI) and Parkinson's Progression

Markers Initiative (PPMI). The distribution of sMRI data acquired from these five datasets are provided in Table 2.1.

The five datasets may contain multiple scans for the same subject taken on the same or different date, or the scan could be taken as different phases such as screening, baseline, month 6, month 12 etc. The scans selected and downloaded for this thesis are the earliest scans taken for given subjects as the aim of this thesis is to improve the AD diagnosis at the earliest stage of the disease. Also, if there are scans taken on the same date (within the same phase), the scan with the highest contrast-to-noise ratio (CNR) is selected.

Following the selection of the subset of data from the five datasets by selecting the earliest scan and the one with the highest CNR in case of multiple scans in the same phase, data containing 3,170 sMRI scans (each scan belonging to a single subject) were downloaded.

The sMRI scan provide a 3D view of the brain showing the morphological structures of different parts including the cortical and subcortical regions.

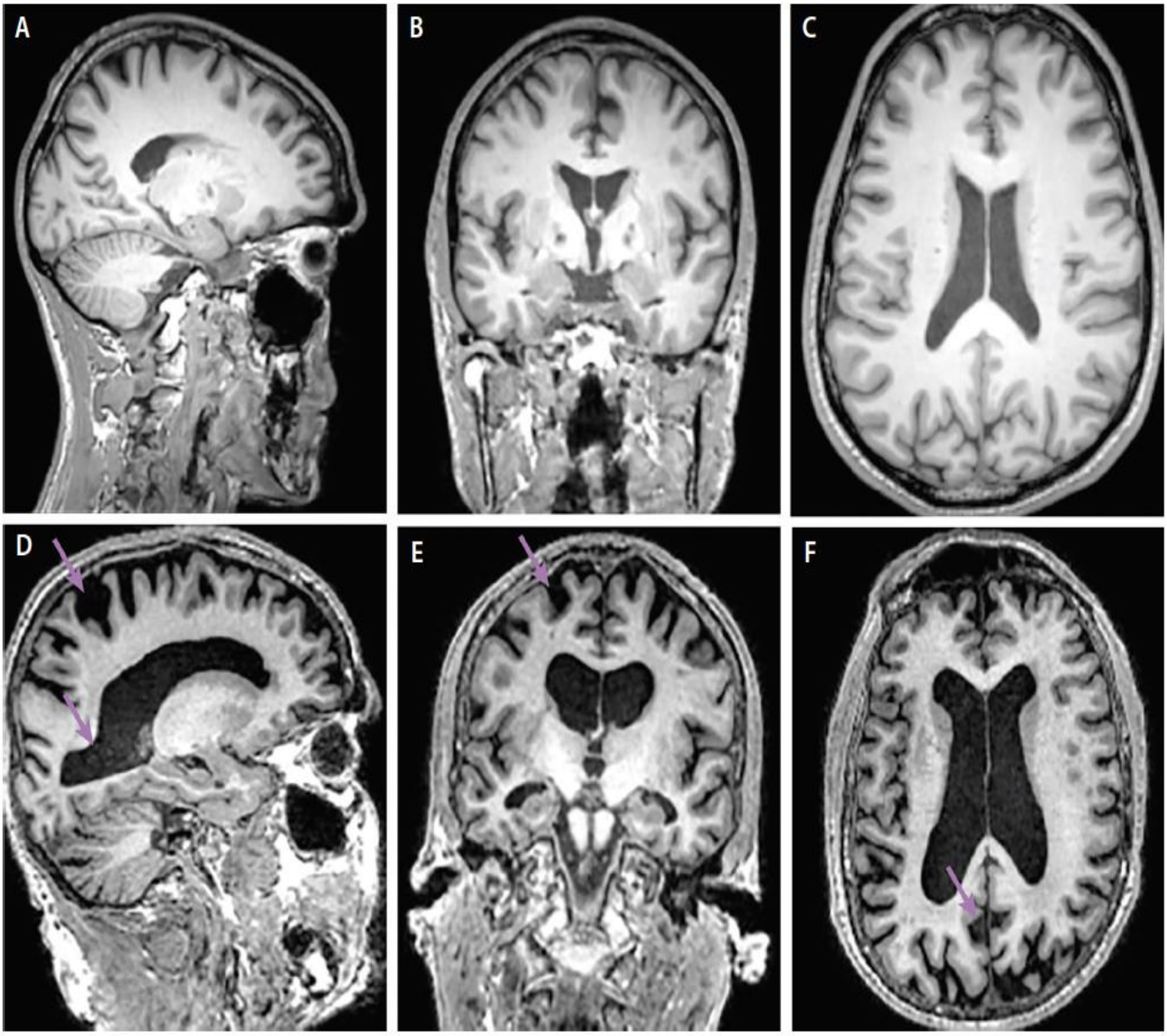

Figure 2.1 | Structural MRI scan of two subjects. The top row represents a CN subject with a typical brain cortical volume and the bottoms row represents an AD subject with atrophy in the hippocampus and gyri. Source: Mary Ellen Koran, MD, PhD https://practicalneurology.com/articles/2019-nov-dec/neuroimaging-and-alzheimers-disease

In Figure 2.1 the hippocampus and gyri region of the brain of two subjects are compared. The three images on the top row belong to a healthy subject with

typical brain cortical volume and the three images on the bottom belong to a subject with AD which show loss of volume and atrophy in the hippocampus and gyri areas. The grey matter thickness on the images on the bottom are significantly thinner than those of images on the top. This shows the atrophy in the brain caused by AD which gives an indication to medical professionals when diagnosing AD.

In this thesis, numerical measurements such as thickness, volume and area of different regions of the brain are extracted and used as features in the analyses. In order to extract those numerical measurements of different regions of the brain, the sMRI data which are 3D scans of the brain are preprocessed using FreeSurfer v.6.0 [8] where processes like image registration, skull stripping, brain segmentation and parcellation and estimation of cortical measurements such as surface area, volume and thickness. Examples of different features could be the thickness of right HATA and the volume of right Hippocampus, each as a numerical value.

Following the preprocessing step, each scan of an individual subject would produce a set of 446 features (or numerical measurements of the brain regions). In the data cleaning stage 45 features were eliminated from the feature set as they were duplicates of other features or had errors. These could be caused by the head movement at the time of scan or the preprocessing of the images using FreeSurfer.

The preprocessed data represents a tabular dataset with 3,170 rows (subjects/scans) and 403 columns (features) which include 401 brain features extracted and produced by FreeSurfer v.6.0 plus age and gender.

ADNI is among the most popular datasets used in automatic CAD for AD and in this study is the primary dataset with greatest number of subjects with AD. The age range of the participants selected to take part in ADNI study is 55-90 therefore any datasets selected and used in addition to ADNI, in order to keep the age consistency among all datasets, subjects with the same age range, 55-90, are selected. As an example, the age range of the participants taken part in IXI study is 19-90 but for consistency only subjects above and including the age of 55 are selected, downloaded and used for this thesis. The reason for selecting all the subjects with the same age range is due to the effect of age on the size of the brain.

2.2. Outlier detection

Among the images downloaded and preprocessed a few could have a significantly different feature distributions compared to the majority of the imaged, which could be the result of a technical error or head movement when taken the MRI scan, or preprocessing of the image using FreeSurfer v.6. These few images are considered to be outliers to rest of the data and should be removed as they will adversely affect the analyses. Therefore, an outlier detection step is performed in order to identify and remove these few outliers before building the proposed model.

Outliers are data points which are unexplainable and different from the rest of the data. Cause of outlier existence can be head movements of the subject or malfunctions of the medical equipment such as MRI machines. One challenge is that on one hand, having outliers in our data can produce a skewed and biased model and it is better if outliers are removed from the data [10] and on the other hand we

cannot afford to lose data by eliminating a great number of instances as outliers. Therefore, there should be a trade-off between the probabilities that some instances are outliers (how different the outliers are from the rest of data) and the number of outliers to remove from the data.

In selection of the outlier detection technique, the number of attributes/dimensions in the data is an important factor as some methods such as Local Outlier Filter (LOF) [11] are only efficient at detection of outliers in a low dimensional dataset. To apply outlier detection (OD) to these data, a high dimensional OD method should be selected. Among high dimensional OD methods, Isolation Forest (iForest) [9] and Angle-based Outlier Detection (ABOD) [12] are regarded as two of the best OD methods [13] where iForest has a much lower computational complexity and therefore is selected as the OD method in this thesis.

iForest is a tree-based outlier detection technique which uses random forests to identify outliers. The intuition behind creation of iForest is handling outliers in high dimensional data with low computational complexity and avoiding profiling normal instances or inliers in order to avoid false alarms (identifying normal instances as outliers). In iForest outliers are detected based on the fact that they are "few and different" therefore iForest isolates outliers rather than profiling inliers.

The process of iForest is explained below as presented by the original paper [9].

To initialise an iForest tree, iTree, one feature q, one split point p in feature q and one sub-sample of ψ instances (referred to as X) are randomly selected. Starting from the root of the iTree, instances X are partitioned into 2 nodes using the split point p; X with q < p are placed into the left child node and instances with q >= p are placed into the right child node. The partitioning of instances into 2 child nodes is performed recursively on all nodes until a termination node is reached. A node will be regarded as a termination node if any of the following criteria occurs: 1- the iTree reaches the height limit l (which is explained in the next paragraph), or 2- all instances in a node have the same values or 3- node contains only 1 instance.

Following the generation of an iTree, every instance x has a path length h(x), which is given by the number of edges an instance x traverses through to reach the termination node. The maximum limit on h(x) is given by height limit l. iForest original paper [9] considers the approximate average tree height to be ceiling($log_2 \psi$) as suggested by [14] and selects this as the height limit l as demonstrated in the equation below:

$$l = \text{ceiling}(log_2 \psi) \quad \text{(E2.1)}$$

where l is the height limit and ψ is the number of instances in the sub-sample.

To create an iForest, multiple iTrees are generated. The number of iTrees in an iForest will be referred to as t. After the creation of iForest, the outlier score is estimated using the following equation:

$$s(x, \psi) = 2^{-\frac{E(h(x))}{c(\psi)}} \quad \text{(E2.2)}$$

where x is an instance, ψ is the number of instances in the sub-sample, h(x) is the path length, E(h(x)) is the average h(x) over a number of iTrees and c(ψ) = 2H(ψ − 1) − (2(ψ − 1)/ψ) where H(i) is the harmonic number and is equal to ln(i) + 0.5772156649 (Euler–Mascheroni constant).

In equation E2.2, s has the following properties:

- when E(h(x)) → 0, s → 1
- when E(h(x)) → ψ - 1, s → 0
- when E(h(x)) → c(ψ), s → 0.5

and therefore, the following conclusions are made:

- an instance x with s very close to 1 is definitely an outlier,
- an instance x with s much smaller than 0.5 is most likely an inlier,
- if all instances have s ≈ 0.5, then there are not any noticeable outliers in the data.

In iForest model, there are three main hyper-parameters, sub-sampling size ψ, height limit l and number of iTrees t.

Following the empirical analysis in [9] the number of 256 instances in the sub-sample ($\psi = 256$) is shown to be sufficient in isolating the outliers where any increase from 256 will not have any improvements on the detection performance and will be less computationally efficient, therefore 256 is selected as the default value for ψ.

Also, as part their preliminary analysis outliers and inliers have proven to have average path lengths h(x) of 4 and 12 respectively when sub-sample of 135 instances were used where ceiling($\log_2 135$) = 8. This has indicated that outliers are isolated well before the iTree height/depth has reached the average tree height of 8. Therefore the default height limit l is selected as ceiling(log2 256) = 8.

It is demonstrated in the same paper that the average path lengths h(x) starts to converge well before the number of iTrees t is reaching 100. Therefore, the default value for t = 100.

In iForest outliers are isolated sooner than inliers being profiled. This means outliers are isolated before the iTree is fully grown and hence the iTree model used is a partial model however, since the aim is isolation of outliers in a time-efficient and computationally-efficient manner, partial model is sufficient to detect outliers.

iForest outlier detection model in this thesis used the default hyper-parameters recommended by the original iForest paper [9] as changing the values do not help the performance of iForest. This could be seen as an advantage, as

iForest could be used as a model with no need for hyper-parameter optimisation. Also, iForest can indicate if all data points are inliers and there are no outliers in our data (when all instances have outlier score s $\approx$ 0.5). And although selecting outliers based on their scores can be challenging, by having a balance between the number of outliers and the degree that they are different from the rest of the data, outliers can be identified and removed.

To apply iForest to our data, sklearn.ensemble.IsolationForest module from scikit-learn Python library [15] is used.

iForest creates a probability or outlier score for each datapoint after being applied to the data. To visualise the result of iForest on the data and outlier score of each data point, the iForest outlier detection technique [9] is applied to all data (on each gender separately), in the unsupervised approach and since the class labels are not used at this stage, this will not cause any overfitting [16].

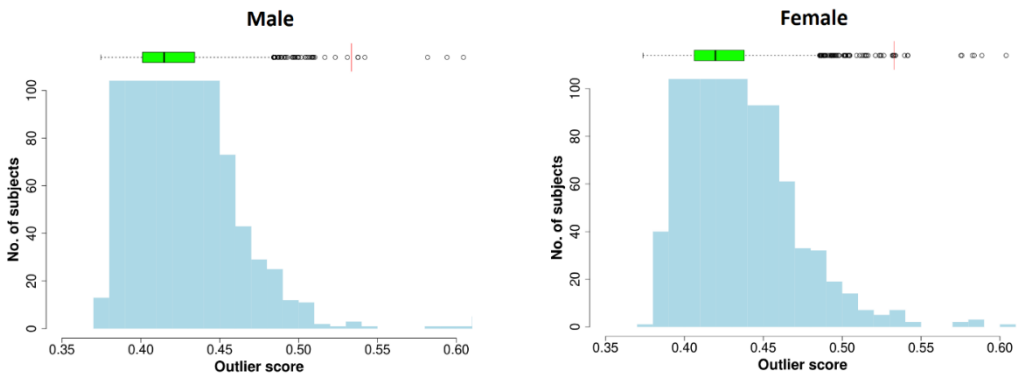

Figure 2.2 | Isolation Forest outlier score distribution for Male group (on the left) and Female group (on the right). The red line represents the cutoff point. Data points on the right of the red line are considered as outliers.

After application of iForest to all data for each gender, the distribution of the generated outlier score for each gender is produced separately and shown in Figure 2.2. On both plots of this figure show positively-skewed distributions presenting a very few data points on the right which can be identified as outliers. To find the cut-off point on x-axis for selecting the outliers, Tukey's method [10] is used:

$$cutoff = \text{Q3} + 3 \,.\, \text{IQR} \qquad \text{(E2.3)}$$

In the box plot, IQR represents inter-quartile range and Q3 represents the third quartile. Any data point with an iForest outlier score greater than the cutoff would be classed as an outlier and deleted.

In the exploratory data analysis using all data, iForest identified 15 male subjects and 14 female subjects to be outliers in their gender groups, as shown in the two plots in Figure 2.2.

The model building in this thesis is performed using a cross-validation. At each fold of the cross-validation a separate iForest model is built and applied. For this purpose, at each fold, both iForest model building and the cutoff point calculation are performed on the training set and applied to both training and test sets.

2.3. Experiments

This section explains about the experiments performed in this thesis. The significant and unique characteristic of the proposed approach is the greedy feature selection of the disease specific features for creation of ABA. However, to show that the effect of feature selection on the performance of ABA, the workflow is also performed without the feature selection as an initial stage. Also, as the proposed approach is data driven and the hypothesis is that more data will result in more accurate performance, in addition to having the initial stage without feature selection and the main proposed method with feature selection the effect of additional data is also investigated by a further stage where additional data is added to the initial stage. Therefore, not only to evaluate the performance of the proposed method but also to compare the effect of adding each component of the additional data and feature selection, the proposed system contains 3 experiments named M1, M2 and M3.

M1 is the initial experiment where the ABA-Com model is built without the proposed feature selection method using single data source (ADNI). The experiment M2 has the same setting as M1 but with additional data sources (ADNI, AIBL, IXI and PPMI) in order to show the effect of data on the ABA-Com performance. M3 is the final experiment which contains the proposed ABA-Com model and the proposed feature selection method in this thesis to classify AD vs. CN in a binary classification setting. M3 uses the same data as M2 but the ABA feature is built on selected features which were selected using the proposed feature

selection method. Therefore, the difference between the results of M3 and M2 shows the effect of the proposed feature selection method on the ABA-Com model performance.

In order to compare the performance of the proposed approach with the performance of a black-box state-of-the-art classification algorithm, two more baseline experiments are performed: B1 and B2. In B1 experiment all features are used to classify AD where only ADNI data is used. The B2 method is also performed with same setup as B1 but with additional data sources (ADNI, AIBL, IXI and PPMI); this is to show the effect of additional data on the result of the SVM classification where no feature selection and regression are used. The change from B1 to B2 can then be compared to the change from M1 to M2 to show that the proposed method is more affected by additional data than when SVM is used.

State-of-the-art could refer to the latest technology or development in a fields but even though the SVM is not a new algorithm, it is still considered a popular state-of-an-art technique which has been extensively used in AD prediction research [17]. Although SVM is considered one of the most powerful machine learning algorithms, the process of how it works is hard to comprehend. These types of algorithms are often referred to as black-box where the user cannot see the inner workings of the algorithms and the output is hard to interpret [18].

In the published journal paper [7] ABA-Com model was evaluated using 1901 subjects acquired from 3 sources of ADNI, AIBL and IXI containing CN and AD subjects, which were at the time of publishing preprocessed using FreeSurfer v.6

and ready for analysis. As the ABA-Com model is a data-driven approach and the performance of the model improves by adding more data, for the purpose of this thesis further CN data were acquired from PPMI and preprocessed using FreeSurfer v.6. Although the PPMI data only contained 81 male and 36 female subjects but the analysis in this thesis confirmed the results achieved in the journal paper. In fact, the performance of ABA-Com model has improved in this thesis compared to the results reported in the paper, due to having more data available. This confirms the fact that ABA-Com model is a data-driven model and by having more data the model performance can improve.

The study proposing the original ABA-Com model [7] was focused on binary classification of AD vs. CN. In this thesis, in addition to the binary classification, multi-classification is also performed where ABA model predicts CN, AD, MCI, EMCI, LMCI and FTD in a 6-class classification setting.

For the purpose of the multi-class classification performance evaluation of ABA-Com model, 12 experiments are performed in chapter 7. These 12 experiments are split into 3 groups of 4 experiments of which refer to different classes being used for that any one experiment.

The 4 experiments are 3-class where CN, MCI and AD are classified, 4-Class where CN, MCI, AD and FTD are classified, 4-Class (E/LMCI as MCI) where EMCI and LMCI are used in the classification as MCI and the 4 classes of CN, MCI, AD and FTD are classified and 6-class is where CN, EMCI, LMCI, MCI, AD and FTD are classified.

Among the 3 groups, Baseline SVM is the first group where all brain features and age are used as input in SVM algorithm to make a classification. SVM in this experiment is used as the state-of-the-art black-box algorithm to provide a baseline classification performance for the proposed method. This experiment is equivalent to B1 in binary classification.

The second group of experiments is ABA_{woFS} where the proposed ABA-Com model classification performance is evaluated while the proposed feature selection method is not used. This experiment is equivalent to M2 experiment in binary classification.

The third group of experiments is ABA_{wFS_x} where the proposed ABA-Com model classification performance is evaluated while using the proposed feature selection method. This experiment is equivalent to M3 experiment in binary classification.

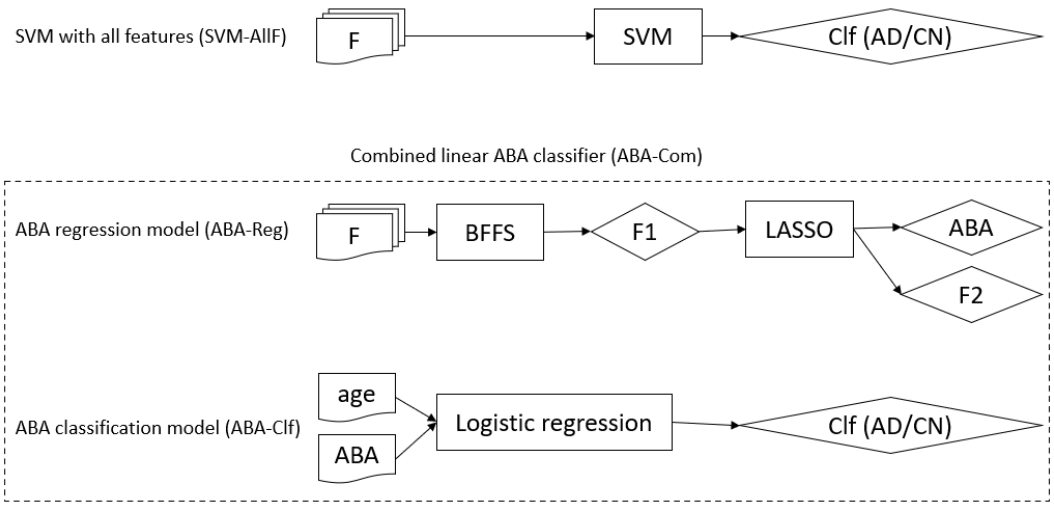

Figure 2.3 | ML models created in this thesis. BFFS: Biased Forward Feature Selection which is a novel method proposed in this thesis, F: all input features, F1: feature subspace selected by BFFS, F2: selected features by LASSO.

The diagram in Figure 2.3 shows a high-view of the models built in this thesis in order to perform the binary classification analyses. SVM-AllF is the baseline model where all brain features and age are used as input features in SVM algorithm to classify AD vs. CN. The proposed dementia prediction model is the Combined linear ABA classifier (ABA-Com) which consists of two ML models; ABA-Reg which is the LASSO regression model to estimate ABA feature and ABA-Clf which is the logistic regression model built on two features of ABA and age to make a binary classification of AD vs. CN.

In a multi-class classification setting in Figure 2.3, SVM-AllF model uses multi-class classification SVM algorithm in order to classify dementia types and stages and for the ABA-Com model, multiple ABA-Reg and ABA-Clf are built, each representing a dementia type/stage.

2.4. Summary

This chapter explained about the data used in this thesis and also the selection, cleaning and preprocessing the data including outlier detection. It is also show in this chapter that what MRI data of the brain are how AD can affect different parts of the brain which helps the prediction of AD. In the next chapter the previous relevant studies are reviewed where MRI data were used to make an AD prediction.

Chapter 3

3. Literature Review

3.1. Computer-Aided Diagnosis of AD using MRI with ML methods

There has been extensive research in the field of machine learning to diagnose AD using automatic computer aided diagnosis (CAD). One approach for CAD is to analyse brain images to find patterns associated with AD in order to help with the diagnosis. Different neuroimaging techniques and modalities which are used for CAD include Positron Emission Tomography (PET), Single Photon Emission Computed Tomography (SPECT) and Magnetic Resonance Imaging (MRI). The MRI scans can include both Functional MRI (fMRI), Structural MRI (sMRI). In this thesis, sMRI is analysed and focused on. sMRI is a non-invasive 3D-imaging of the brain to aid with diagnosis of AD. In this thesis sMRI scans are used to build the machine learning models, therefore this chapter will focus only on the review of the literature on methods using sMRI data, as opposed to fMRI or any other neuroimaging techniques. Hereinafter throughout the thesis, MRI will refer to sMRI.

The MRI scans provide a 3D view of the whole brain but AD does not affect the whole brain equally; it affects some parts much more than the others. In order to analyse MRI scans, most of the times, studies segment the whole brain into smaller parts, in order to identify the parts of the brain which had the highest effect from AD. The segmentation of the brain into smaller parts can be carried out using two main approaches: voxel-wise and region-wise segmentations.

In voxel-wise, the whole brain is uniformly segmented into n equal-sized smaller cubes also referred to as voxels, regardless of which structures of the brain the voxels belong to. The extraction of these voxels from the images are referred to as voxel-based morphometry (VBM) and the type of analysis performed on these voxels is referred to as VBM-based analysis where the voxels are used as features in the ML model.

The region-wise segmentation is when the whole brain is segmented into predefined semantic structures of the brain also referred to as regions. In this analysis, each region may have a different size to other regions and represent a structure such as Hippocampus or Amygdala, and each region has multiple measurements of which are extracted as numerical data at image preprocessing stage. Each of those measurements is referred to as a region of interest (ROI) and the type of analysis performed on those ROIs is referred to as ROI-based analysis where ROIs are used as features in the ML model.

The next two sections, provide a literature review on studies using VBM-based and ROI-based features extraction for classification of AD.

3.2. VBM-based classification of AD

This section contains an overview of some of the relevant previous studies involving VBM-based analysis of MRI data in the context of AD diagnosis. These studies use the raw images, after voxel-wise segmentation and preprocessing, to find patterns associated with AD, as opposed to region-wise numerical measurements of different parts of the brain which will be reviewed in next section.

[6] uses all voxels of the MRI scans as individual features or dimensions in a high dimensional space. SVM and subspace clustering is then applied to find images which are similar to each other and subsequently classify AD from CN. The highest classification accuracy, recall and specificity achieved by this study are all reported to be 95%, using the leave-one-out validation approach. Although the performance metrics of 95% is relatively high for binary classification of AD vs. CN, this study has used a very limited sample, containing 68 subjects with the same proportion of subjects for both classes of CN and AD. This number is relatively small compared to other literature in this field and the size of our dataset used in this thesis. This study has used several datasets from multiple different scanners with different protocols however due to having a small sample size, investigating the effect of different cohorts of data on the results cannot be tested with high confidence.

In [19] voxels in medial temporal lobe are considered, where dense deformation fields and scaled grey-level intensity are used to derive the Jacobian determinants. SVM is then used for the binary classification of AD vs. CN. The

results report the accuracy of 92%, which is also relatively high, however, similarly to [6] the dataset used in this study is of a small size: 150 subjects with the same proportion of AD to CN. Therefore, the high accuracy reported may lack sufficient generalisation. Also, as SVM is a black-box approach, it does not provide any description as to why such accuracy was achieved. Similarly, to [6], this study has used several datasets in order to show the generalisability of the model however as the dataset is a small one, this generalisability test may not be robust. The study also shows that the pre-processing of the MRI data can be adjusted in order to improve the classification outcome.

In [20] three different classification methods are used to separate AD from CN using clustered voxels from the MRI scans. The three classifiers used are Voting Feature Intervals (VFI), Bayes statistics and SVM. The accuracy reported in this study for the binary classification of AD vs. CN is 92%. Similarly to [6] and [19] this accuracy is relatively high but is based on a small dataset of 50 subjects, of which 18 are CN and 32 are AD. In this study by applying density-based spatial clustering to the voxels, a feature is generated that can separate the 2 classes of AD from CN by showing that the brain structure in CN subjects vary greatly from the brain structures in AD subjects.

[21] uses a graph-based method where MRI scans are represented as graphs and subjects are classified to AD vs. CN by SVM using the three factors of gender, level of education and level of cognitive impairment in subjects. The graph-based method adopted in this study represents the shape of the ventricular systems in the brain and also relative to the skull. It is found by this study that AD can greatly

influence the ventricles shape. This work reported 90.9% accuracy for the binary classification of AD vs. CN. This study demonstrated that assessment of the brain could identify small parts of the brain which can be related to an occurrence of life events and the brain functions. Although these small parts of the brain can be identified, there is still a need for a medical professional to review those brain parts to find out what they could be, therefore this can be viewed as a drawback of this study.

[22] is a recent study on binary classification of CN vs. AD where SVM with cubic kernel is used. This study has proposed a novel image preprocessing technique and the dataset obtained from ADNI contains 250 CN and 250 AD. In addition to SVM with cubic kernel, other methods such as Naïve Bayes, Discriminant Analysis, SVM with linear, quadratic and medium Gaussian kernels and K-nearest neighbours have been used but the reported accuracy for SVM with cubic kernel is the highest among all methods, which is 93%. For the validation of the method, 10-fold cross-validation is used. However, as the classifier is a black-box approach, the model lacks descriptiveness and interpretability.

[23] uses VBM analysis to perform binary classification of CN vs. AD using genetic algorithm and SVM. Subjects in this study have been obtained from ADNI which include 162 CN and 160 AD. The proposed method in this paper achieves 93.01% accuracy using SVM. The validation of the method is carried out using 10-fold cross validation.

[24] is among the latest studies suggesting an SVM-based machine learning method to predict AD using MRI scans with VBM analysis in addition to results of neuropsychological tests. The subjects in this study include 353 CN subjects and 296 AD subjects, obtained from ADNI. This study has achieved 93% accuracy for CN vs. AD binary classification using holdout validation and the accuracy is computed using a single test set (20% of all data). Although the study has used a relatively large dataset, as the holdout method is used for validation the performance estimation method is not adequate to provide guarantees of generalisation.

3.3. ROI-Based classification of AD

This section provides an overview of past studies involving the analysis of numerical measurements of brain regions using MRI data. The MRI scans are first preprocessed into numerical data representing multiple measurements such as volume, area and average thickness of different parts or regions inside the brain.

[25] proposes a sparse Bayesian multi-task learning algorithm to help predict AD. In this study MRI scans are pre-processed using FreeSurfer v.4 for image segmentation and for the generation of ROI measurements have been extracted from the scans. This work investigates the connection between the patterns in the brain measurements and the cognitive state of the brain, to see how the physical change in the brain and its structures can affect the cognitive state. This work has reported 73.5% accuracy in classification of AD by prediction of cognitive scores, using 393 subjects, of which 171 are AD and 222 are CN. Multiple biomarkers were

identified in this study that can help determine the cognitive state of the brain and the AD progression in it, which can be an indication that the biomarkers suggested in this thesis, in addition to being able to predict AD, they can also show the progression of AD.

[26] uses SVM to perform binary classification of CN vs. AD. This study has used FreeSurfer v.4 to preprocess the MRI scans. The number of subjects used include 226 CN and 182 AD from ADNI. The 10-fold cross-validated classification accuracy achieved by SVM is 90.5%.

[27] applies SVM-based recursive feature elimination for feature selection and Extreme Learning Machine (ELM) for binary classification of CN vs. AD. The subjects used were obtained from ADNI including 229 CN and 193 AD subjects. FreeSurfer v.4.5 is used to preprocess the images and extract the measurements of different regions of the brain. The accuracy achieved using 10-fold cross-validation, repeated 10 times, is 92.84% using ELM although standard deviation is not reported in the study. The neural network model is complex and black-box. This approach also achieved a similarly high accuracy than other methods, and lacks descriptiveness and interpretability too.

3.4. Brain Age prediction for classification of AD

The studies presented so far in this chapter were related to classification of AD directly from brain features, VBM-based or ROI-based. An alternative to these types of analyses is to generate biomarkers to help with diagnosis of AD. One type

of biomarker which has helped in classification of AD is “brain age”. Brain age is ML-based estimation of the age of the brain in contrast to the subject age to indicate an accelerated age process in the brain possibly caused by a neurodegenerative disease. Brain age can act as a biomarker to provide an indication of presence of AD. The logic behind brain age is that if the biological age of the brain is higher (older) than the real anagraphical age of the subject then it can be concluded that the brain has been affected by AD, as one of the effects of AD is causing brain atrophy and degeneration in a way that the brain would age at a higher pace than real age.

This section reviews the studies where the brain age of the subject is estimated in order to aid with the classification of AD from CN. Using either VBM-based or ROI-based, previous literature have attempted to use the whole or particular parts of the brain to estimate the age of the brain so it can be compared to the chronological age of the subject. The gap between the brain age and real age is then used to indicate the presence of AD where a bigger positive gap (brain age minus the age) means a higher possibility of presence of AD.

One of the early and main studies on brain age in the context of AD is [6] where a novel feature based on the difference between estimated age of the subjects and their real age is proposed, referred to as brain age gap estimation (BrainAGE) score.

Originally 550 CN subjects from IXI were selected with 3 subjects excluded from the study due to missing age information. Therefore, the analysis was applied

to 497 subjects of age 20-86. The MRI scans were split into 410 training set and 137 test set, using random sampling with stratification on age. In this study VBM analysis was used, where the MRI data were preprocessed using SPM8 package [28] to generate 3700 voxels per image. PCA is then applied to the voxels to reduce their dimensionality to 410 principal components to equal the size of the training set. Following the dimensionality reduction, support vector regression (SVR) and relevance vector regression (RVR) are applied to the voxels data in training set (containing only CN subjects) to generate the suggested model, the BrainAGE score model, where it is also used to generate the BrainAGE score for 2 test samples from ADNI: a sample of 102 AD subjects with age range of 55-88 and a sample of 232 CN subjects with age range 60-90.

The results presented in this study show that BrainAGE score of the 2 samples from ADNI, have a different distribution with mean score for subjects with AD aimed to be plus 10 and for CN subjects aimed to be 0 or under.

The logic behind this score is that a greater score shows a greater pace in the brain degeneration which is an indication of AD and when the score is 10 it means the brain age of the subjects is 10 years older than their real age which is a sign of acceleration in the brain atrophy associated with AD. Subjects with CN on the other hands are expected to have the same brain age as their chronological age as their brain should be aging the same as their real age.

This study uses regression models on a truncated projected space of all voxels (after the dimensionality reduction), from all parts of the brain to predict the brain

age using the real age as the target variable. This means all regions of the brain are used in the model building. Also, to optimise the model, the mean absolute error (MAE) is minimised and the correlation coefficient r is maximised. In other words, the BrainAGE score model is aiming to be as close as possible to the real age by minimising the regression error between BrainAGE score and age, and maximising the correlation between the two, which is in fact the aim of regression models. This study has used a relatively large dataset which makes their method robust. The MAE and r of the test set reported in this study are 4.98 and 0.92 respectively for CN subjects.

There have been several studies on BrainAGE score being used as a biomarker to detect neurodegenerative diseases [29] and one study which provides an accuracy of the BrainAGE score in the binary classification of AD vs. CN is [30], where the BrainAGE score model is built on 561 CN subjects from IXI and applied to the training sample of the CAD Dementia challenge containing 30 subjects. The resulting accuracy of this holdout validation method is 90%, however, as this is a holdout method its generalisation is arguable.

[31] proposes a novel feature estimating the difference between the real age of the CN subjects and their brain age, referred to as brain estimated age difference (Brain-EAD). In this study MRI data for 1128 subjects from four sources of IXI, Open Access Series of Imaging Studies (OASIS), ADNI and PPMI are acquired. The subjects are 839 CN with age range 35-90, 129 AD and 160 PD. This study uses VBM analysis where SPM v12 package is used to generate 3747 voxels for each scan.

A support vector regression (SVR) model is then trained on all voxels of CN subjects to generate Brain-EAD model. This model is then applied to the subjects with AD and PD. Although no accuracy is given in this study, similar to other brain age estimation methods, MAE and r were considered as the metric to estimate the performance of the model, and the aim is to minimise the MAE and maximise r in Brain-EAD model. The MAE and r of the training set reported in this study are 4.38 and 0.92 respectively for CN subjects. The MAE is lower than that of reported in [6] with the same r.

[32] proposes a novel brain age model referred to as DeepBrainNet. In this study 11729 subjects have been acquired from multiple sources, with the aim to create a brain age model with minimal preprocessing using DNN to have an optimum performance in predicting multiple neurodegenerative diseases such as AD, MCI, SCZ (Schizophrenia) and major depression.

This paper is among the very few papers that in addition to MAE and r on brain age model, reported the accuracy. The MAE and r of the training set reported in this study are 3.702 and 0.978 respectively for CN subjects. The accuracy reported for AD vs. CN binary classification is 86%. The possible improvements to this study could be improving the accuracy and descriptiveness, and making the DeepBrainNet model more specific to one neurodegenerative disease.

Multiple studies were reviewed where morphological or numerical features of brain MRI scans were directly used in the machine learning model to classify CN vs. AD. This classification can be viewed as complex and a large number of

features were used in the model. In addition to the complexity of the models generated by these approaches they did not allow for any interpretability as most of the studies used black-box algorithms such as SVM or deep networks. Using a complex machine learning model with many input features and without any description makes the interpretation of the results very difficult if not impossible.

To improve the descriptiveness of the model biomarkers can be used such as brain age where it is used in order to provide a meaningful explanation offering the domain expert the motivation for the classification decision as well as the opportunity to learn useful insights. The concept of brain age could be viewed as a dimensionality reduction approach such as PCA but the difference is that despite principal components brain age has a strong semantic meaning as it is a feature generated using all or particular parts of the brain to represent the effect of AD.

The studies involving generation of brain age as a feature to aid with the prediction of AD are reviewed but the drawback with the methods used in those earlier literature to model brain age is the lack of specificity. Although all methods have been used in the context of AD, but in building the model the characteristics of AD have not been used, therefore the models are generic brain age models, built based on brain of healthy subjects with no information or relation about any specific diseases.

Table 3.1 provides an overview of the results reported in the previous literature. In studies where brain features are directly used in ML model (VBM-based or ROI-based) the consensus shows that most studies using relatively medium

to large-sized datasets, SVM algorithm and cross-validation methods have achieved accuracies of between 92% and 93%. This could be viewed as a baseline performance for such studies. The studies involving the generation and use of brain age are also presented in Table 3.1 but, not all of these studies report accuracy as a measure of performance; instead, they provide MAE and r to show the performance of their brain age regression model. The studies involving brain age used relatively large datasets to generate their models using SVR or DNN and the accuracy for the ones which have reported it is between 86% and 90%.

The models proposed in methods 11-14 in Table 3.1 generate brain age features which can be used by medical professional as biomarkers in detection of AD and could be part of the diagnosis process. But the lack interpretability in the models prevents those using these brain age features from understanding why such features were estimated for a particular subject. This could result in a decreased confidence while using such biomarkers. The work in this thesis therefore proposes a novel brain age estimation model which is also inherently interpretable.

The proposed work in this thesis uses linear regression algorithms to generate a brain age feature which acts as a biomarker and a classification model which can assist medical professionals in predicting AD while providing descriptions of why such prediction has been made.

As a biomarker in a medical setting the interpretability of the brain age estimation model is crucial as it can affect the outcome of the diagnosis.

In the methods 11-14 in Table 3.1 which involve using brain age in the prediction of AD the chronological age of the subject is estimated using the suggested “age estimation framework” to be then compared to subject’s real age. Therefore, the objective is to predict the subject’s age from brain features.

The difference between this method and the proposed method in this thesis is that in the proposed method the objective is to estimate the biological age of the brain with specific reference to a particular neurodegenerative disease, e.g., AD. This means that in the proposed method brain age is actually the estimation of the biological age of particular parts of the brain which are highly affected by a particular neurodegenerative disease, e.g., AD. The classification performance results produced by the proposed model is therefore expected to be better than those models which are based on all parts of the brain which lack specificity to a particular pathology.

3.5. Summary

This chapter provided an overview of the relevant previous literature in the field of brain age and AD prediction using MRI scans. Different methods have been utilised by different studies however the consensus shows that SVR and SVM were the most used methods for the brain age prediction and the AD classification respectively. The next chapter will explain about the proposed brain age model which will be used in the prediction of AD.

Table 3.1 | Summary of results reported in previous studies using VBM-based and ROI-based AD/CN binary classification. Results from previous studies using the "brain age" as biomarker to detection of AD are also reported. Methods have used n-fold (nf) cross-validation (CV), with some repeated (rep) multiple times, leave-one-out cross-validation (LOO) and Hold-out (HO). MAE: mean absolute error; r: Pearson's correlation r; SVM: support vector machine; SVR: support vector regression; ELM: extreme learning machine; DNN: deep neural network. Hyphen is used when the metric was not reported or not applicable.

ID	Ref	Category	Number of Subjects	Algorithm	Validation	Accuracy	MAE	r
1	[33]	VBM	68	SVM	LOO	95.00	-	-
2	[19]	VBM	150	SVM	LOO	92.00	-	-
3	[20]	VBM	50	Bayes	LOO	92.00	-	-
4	[21]	VBM	186	SVM	10f CV	90.90	-	-
5	[22]	VBM	500	SVM	10f CV	93.00	-	-
6	[23]	VBM	322	SVM	10f CV	93.01	-	-
7	[24]	VBM	649	SVM	HO	93.00	-	-
8	[25]	ROI	393	Bayes	5f CV	73.50	-	-
9	[26]	ROI	408	SVM	10f CV	90.50	-	-
10	[27]	ROI	422	ELM	10f CV rep 10	92.84	-	-
11	[6]	Brain Age	550	SVR	HO	-	4.98	0.92
12	[30]	Brain Age	591	SVR	HO	90.00	5.10	0.92
13	[31]	Brain Age	1128	SVR	HO	-	4.38	0.92
14	[32]	Brain Age	11729	DNN	5f CV	86.00	3.70	0.98

Chapter 4

4. Apparent Brain Age

This chapter explains in detail the proposed feature in this thesis, referred to as Apparent Brain Age (ABA) which can be used as a biomarker in identifying AD. The brain age is the estimation of the age of the brain using a machine learning algorithm based on selected brain features in order to show the acceleration of aging of the brain as the result of a neurodegenerative disease. The proposed ABA is the estimation of age of parts of the brain which are highly affected by a neurodegenerative disease i.e., AD.

4.1. Brain Age and AD

The brain age can be used as a biomarker in prediction of Alzheimer's Disease. Several studies have studied brain age to help predict Alzheimer's Disease [6] [31] [32]. In these studies, the biological age of the brain is predicted using MRI scans, where the whole brain and the ROIs from all regions of the brain contribute to the prediction of AD. The deviation between the brain age and the chronological age could be used as an indication to presence of AD. The use of whole brain

morphometry creates a uniformed model for predicting the subject's brain age, regardless of their health, therefore this model lacks specificity to a particular disease.

In previous studies on brain age prediction for prediction of AD [6] [31] [32] a regression model is built on the morphological features of the whole brain. The features are then used in PCA to reduce the dimensionality of the features.

Following the application of PCA, brain age model is built on the those features outputted by PCA of cognitive normal (CN) subjects only. This is to model the morphological brain structure of the healthy subjects and assign different brain ages to those healthy brains. The brain age model is then applied to subjects with AD with the aim of identifying the difference to CN brain and predict the right age for the brain. The brain age for CN subjects is assumed to be the same as the subject's biological age whereas for AD subjects, due to the atrophy in parts of the brain caused by AD, the brain age is expected to be older than the biological age and therefore the brain is expected to be predicted by the brain age model to be older.

Those studies have a general-purpose brain age based on the whole brain morphometry which may not be the most effective way to classify AD from CN. Therefore, by adding the specificity to a particular disease and building the brain age model in the context of a specific disease, the performance on the classification task could be improved.

Those approaches also suffer from lack of descriptiveness by using non-linear models such as PCA and black-box models such as support vector regression (SVR). The descriptiveness of the brain age model could also be improved by using linear and interpretable models. The challenge is to build such linear models in a way that does not affect their accuracy with regards to more complex models.

To add specificity and improve the prediction performance of the brain age models suggested in previous studies [6] [31] [32], in this chapter a novel brain age model is suggested that is specific to a particular neurodegenerative disease (mainly AD in this study), referred to as Apparent Brain Age (ABA). This model uses only those brain ROIs that are highly affected by the particular neurodegenerative disease to estimate ABA, as opposed to all ROIs.

The studies mentioned above, use regression models to predict the chronological age of the subject in order to build the brain age model from the whole brain and the metric used to maximise the performance of the model is the Mean Squared Error (MSE) or the Mean Absolute Error (MAE). In those models, the target of the regression task is to predict the brain age to be very similar to the chronological age of the subject.

On the contrary, in the proposed ABA model the aim of the regression model is not to predict the chronological age, nor the biological age of the entire brain, but to estimate the biological age of the brain regions specifically affected by the target disease. In other words, in this study ABA represents the brain age as predicted by only those parts of the brain highly affected by AD. Therefore, the performance of

the ABA regression model is optimised not by minimising the regression error but by maximising the classification ability of the model in combination with a greedy and aggressive feature selection method.

4.2. Apparent Brain Age Model

To estimate ABA, a regression model is required to be built with target variable being age and independent variables being the brain features. Linear regression is a non-complex form of regression algorithm, but if there are multicollinearity in data and the feature space contain highly correlated features the consistency and stability of the model is reduced due to high variance in the resulted regression model coefficients. To handle this issue a Least Absolute Shrinkage and Selection Operator (LASSO) regression model is used. LASSO is an L_1 penalised regression method which can reduce the coefficient of a feature to zero (and therefore discards the feature from the model) if that feature has no effect on the model, therefore in the case of highly correlated features, it only keeps one of those features and discards the rest.

To estimate ABA a set of features are selected which represent parts of the brain which are morphologically different between CN and AD subjects. These differences are due to morphological changes to those parts due to pathological effects of AD and the neurodegeneration which has caused the atrophy. Those particular parts of the brain are assumed not to have neurodegeneration in CN subjects. Therefore, to model the ABA, only CN subjects are used in order to represent the brain structures with typical/normal measurements which are free

from brain atrophy caused by AD. This way, when ABA model is applied to subjects with AD difference of brain structure pattern will be used to differentiate between ABA for CN and AD. To build the ABA model LASSO is used.

The following is the equation of ABA regression model:

$$ABA = a_0 + \sum_{i=1}^{k} a_i \cdot f_i \quad \text{(E4.1)}$$

where a_0 is the LASSO intercept, k is the number of features, f_i is a single feature value and a_i is the LASSO coefficient for f_i.

As mentioned in chapter 2, there are 401 ROIs extracted from the MRI scans, each representing a regional measurement in the brain. To select the ROIs which are highly indicative of AD, a feature selection method is applied. To estimate ABA, no expert knowledge has been used, therefore the suggested feature selection method should be a machine learning method.

The ABA-Reg model is built on a feature subspace of healthy subjects and to maximise the classification capability of ABA, CN subjects from multiple sources are used in building of the ABA-Reg model. Large dataset of CN subjects, from multiple sources ensures the robustness and generalisability of the model.

To analyse the performance of ABA-Reg model, three different experiments are performed. The proposed ABA-Reg model contains the explicit proposed feature selection method, where the LASSO regression model is built on a small

subspace of features which is selected by the proposed feature selection method. This experiment is referred to as M3 and uses AD and CN data from ADNI, AIBL, IXI and PPMI.

In order to assess the effect of different components of the ABA-Reg model building process on the performance, two more experiments are performed. One experiment referred to as M2 is performed where the proposed feature selection is not used in the workflow to build the ABA-Reg model therefore, the ABA-Reg (LASSO regression) model in M2 is built on 402 features (401 ROIs and age), using AD and CN data from all sources (ADNI, AIBL, IXI and PPMI). M2 therefore highlights the effect of the proposed feature selection method on the model performance and uses the same data as M3.

Another experiment performed is M1 where the effect of the additional data on building of the ABA-Reg model is assessed. In M1 the data used is AD and CN subjects from ADNI only. Also, in this experiment the proposed feature selection method is not used the ABA-Reg model is built on 402 features. Therefore, M1 and M2 share the characteristic of not using the proposed feature selection method but with the difference that M1 uses data only from ADNI and M2 uses data from all available sources. Also, the experiments in M1 and M2 resembles the brain age model building frameworks used in previous literature [6] [31] [32] where features from whole brain are used in the model without any explicit feature selection to make the model specific to one pathology.

These three experiments are designed specifically to highlight not only the performance of the proposed model but also the effect of each component of additional data and proposed feature selection method on the performance. Therefore, M1 is performed first to showcase the ability of the ABA-Reg model using single data source (ADNI) and without the use of proposed feature selection. Then, M2 is performed with same setting as M1 but with additional data and finally M3 is performed with the same data as M2 but with the addition of proposed feature selection.

As explained in this section, ABA gives a disease specific brain age which can be compared to real age of the person. ABA higher than the real age shows that the specific parts of the brain which are affected by the disease i.e., AD are showing an older age compared to the subject's real age. To show the difference between ABA and real age, another feature can be used referred to as Age Deviation Score (ADS) with the following equation:

$$ADS = ABA - age \quad \text{(E4.2)}$$

Positive ADS shows an older ABA which could indicate a higher probability in having AD whereas ADS with negative value or closer to zero indicates a healthy brain. The concept of ADS is also used in the [6] [31] [32] where the greater the value of ADS or the gap between brain age and chronological age, the more likely the subjects suffers from AD.

To select the ROIs highly affected by AD, a wrapper forward selection approach is suggested, referred to as Biased Forward Feature Selection, and

explained in the next section and the pseudocode for this method in provided in Pseudocode 1.

To build the ABA-Reg model two feature selection methods are performed. The first method which is an explicit feature selection method is the proposed BFFS which is applied to features in order to select a small subset of features (F1). The second one is the embedded feature selection performed by LASSO when it is applied to F1. The selected feature set by LASSO is referred to as F2.

4.3. Biased Forward Feature Selection

Selecting the ROIs which are highly affected by AD without any prior expert knowledge is a challenging task. In this section a feature selection method is suggested, referred to as Biased Forward Feature Selection (BFFS), where the features selected are the ones which introduce an inductive bias towards the classification of AD.

In the proposed wrapper method, a forward feature selection approach is used. First, a feature ranking is performed based on the correlation of the features to the target variable, where features with higher absolute correlation will have higher rank. The features are sorted by rank and iteratively added to a target feature set, where in each iteration, an ABA LASSO regression model is built and used to classify AD using logistic regression. The criterion for retaining/removing the feature is the classification accuracy. Following this wrapper approach, a target feature set is identified, providing the feature subspace which yields the highest classification accuracy for predicting AD with the minimal number of features. The

feature subspace will then be used in LASSO regression, which performs an additional and embedded feature selection, to estimate ABA, which will subsequently be used to classify AD. This is explained in detail in the next chapter.

4.4. Apparent Brain Age results and discussion

In this chapter ABA-Reg is suggested which is a LASSO regression model to estimate ABA. One way to estimate the performance of this proposed regression model is to use metrics such as MAE and Pearson's correlation r. These two metrics assess the quality of the regression task as used by previous literature in [6] [31] [32] to assess the brain age model performance.

In these three literatures [6] [31] [32], as discussed before, the aim is to minimise the error and maximise the correlation when brain age is compared with age. As the proposed ABA-Reg model is built on only the very few brain features which reflect the impact of the disease, it only estimates the age of those few brain features. This is why ABA could be very different to age and minimising the MAE and maximising the r will not improve the classification performance, which will be discussed in next chapter.

As the results, it is shown in this section that when BFFS is used in M3, the MAE increases and r decreases as the ABA-Reg will be built on only a small subset of the features, as opposed to the features from the whole brain.

The evaluation of the performance of the ABA-Reg model is carried out through a 10-fold cross-validation, with classification accuracy used as the metric. The 10-fold cross-validation is repeated 10 times and the results are averaged in order to validate the robustness of the model performance estimation.

The F1 resulted from BFFS from each fold of each repeat are collected in order to show which brain features are selected by BFFS in each fold. It is expected to see brain features which are highly affected by AD to be present frequently in F1. The results of this analysis are presented in Table 4.1

Table 4.1 | ROIs selected over 100 BFFS runs through a 10-time repeated 10-fold cross-validation, relating to both genders. The ROIs presented in the table below are those with presence frequency of at least 10% over all 100 BFFS runs, in either Right Hemisphere (RH) or Left Hemisphere (LH). ROIs are ordered in a descending order based on the both RH and LH for both genders combined. The frequencies above 50% are shown in bold.

	F&M			**F**			**M**		
ROI	**LH& RH**	**LH**	**RH**	**LH& RH**	**LH**	**RH**	**LH& RH**	**LH**	**RH**
Amygdala	**53%**	**73%**	33%	**66%**	**78%**	**54%**	40%	**68%**	12%
Hippocampal_tail	44%	**57%**	30%	42%	**50%**	34%	45%	**64%**	26%
Subiculum	43%	43%	43%	**86%**	**86%**	**86%**	0%	0%	0%
Whole_hippocampus	41%	**57%**	24%	8%	16%	0%	**73%**	**98%**	48%
Inferiorparietal_thickness	31%	13%	48%	38%	26%	**50%**	23%	0%	46%
Middletemporal_thickness	26%	33%	19%	16%	14%	18%	36%	**52%**	20%
CA1	24%	35%	13%	0%	0%	0%	48%	**70%**	26%
Inferiorparietal_thickness	19%	13%	25%	38%	26%	**50%**	0%	0%	0%
Entorhinal_thickness	18%	35%	0%	0%	0%	0%	35%	**70%**	0%
Inferiortemporal_thickness	17%	6%	27%	22%	0%	44%	11%	12%	10%
Precuneus_thickness	16%	20%	12%	32%	40%	24%	0%	0%	0%
Precentral_volume	13%	18%	8%	14%	12%	16%	12%	24%	0%
Middletemporal_volume	13%	17%	9%	17%	16%	18%	9%	18%	0%
Molecular_layer_HP	12%	14%	10%	0%	0%	0%	24%	28%	20%
Bankssts_thickness	10%	8%	11%	19%	16%	22%	0%	0%	0%

	F&M			F			M		
ROI	**LH& RH**	**LH**	**RH**	**LH& RH**	**LH**	**RH**	**LH& RH**	**LH**	**RH**
Superiortemporal_meancurv	8%	10%	5%	0%	0%	0%	15%	20%	10%
Inferiorparietal_volume	8%	7%	8%	7%	14%	0%	8%	0%	16%
Supramarginal_thickness	7%	14%	0%	0%	0%	0%	14%	28%	0%
Paracentral_volume	7%	6%	8%	14%	12%	16%	0%	0%	0%
Parsopercularis_thicknes_std	7%	7%	6%	0%	0%	0%	13%	14%	12%
Superiorparietal_volume	7%	0%	13%	13%	0%	26%	0%	0%	0%
Lateraloccipital_thickness	5%	5%	5%	0%	0%	0%	10%	10%	10%
Lingual_volume	5%	9%	0%	0%	0%	0%	9%	18%	0%
Right_Pallidum	5%	0%	9%	9%	0%	18%	0%	0%	0%
Superiorfrontal_thicknes_std	4%	8%	0%	0%	0%	0%	8%	16%	0%
Paracentral_thicknes_std	4%	0%	8%	0%	0%	0%	8%	0%	16%
Brain_Stem	4%	0%	8%	8%	0%	16%	0%	0%	0%
Inferiorparietal_volume	4%	7%	0%	7%	14%	0%	0%	0%	0%
Postcentral_thickness	4%	7%	0%	7%	14%	0%	0%	0%	0%
Lateralorbitofrontal_volume	3%	6%	0%	0%	0%	0%	6%	12%	0%
Parsopercularis_volume	3%	0%	6%	0%	0%	0%	6%	0%	12%
Bankssts_thicknes_std	3%	6%	0%	6%	12%	0%	0%	0%	0%
Insula_volume	3%	6%	0%	6%	12%	0%	0%	0%	0%
Precuneus_volume	3%	6%	0%	6%	12%	0%	0%	0%	0%
Lateraloccipital_volume	3%	0%	6%	6%	0%	12%	0%	0%	0%
Temporalpole_volume	3%	5%	0%	0%	0%	0%	5%	10%	0%
Cuneus_volume	3%	5%	0%	0%	0%	0%	5%	10%	0%
Inferiortemporal_volume	3%	5%	0%	0%	0%	0%	5%	10%	0%
Optic_Chiasm	3%	0%	5%	0%	0%	0%	5%	0%	10%
Paracentral_thickness	3%	0%	5%	0%	0%	0%	5%	0%	10%
Posteriorcingulate_thickness	3%	0%	5%	0%	0%	0%	5%	0%	10%
CA3	3%	0%	5%	0%	0%	0%	5%	0%	10%
CA4	3%	0%	5%	0%	0%	0%	5%	0%	10%
Bankssts_volume	3%	5%	0%	5%	10%	0%	0%	0%	0%
Fusiform_thickness	3%	5%	0%	5%	10%	0%	0%	0%	0%

Table 4.2 presents the results of Mean Absolute Error (MAE) and Pearson's correlation r, performed for both genders, and both holdout and cross-validation variations of all methods of M1, M2 and M3. The general trend in results across

both genders and validation methods show that there is an increase in the MAE and decrease in r, from M1/M2 to M3 when BFFS is added. This is because BFFS improves the AD classification results whereas it worsens the regression results and the regression line quality.

Table 4.2 | Overview of regression results (MAE, r) for both genders: three incremental variants of the proposed method using distinct settings in order to show the effect of multiple factors. The information about Mean Absolute Error (MAE) and Pearson's correlation coefficient (r) for LASSO ABA regression model are presented. The reported results show the performance related to different data partitions (training and test sets) based on different validation methods used; 10-fold cross-validation (10f CV) and holdout.

		MAE			r		
Data Partition	Group	M1	M2	M3	M1	M2	M3
(M) 10f CV test	CN	4.05	5.25	5.81	0.60	0.55	0.40
(M) 10f CV test	AD	6.33	5.72	6.78	0.55	0.56	0.25
(F) 10f CV test	CN	3.61	5.03	5.69	0.68	0.54	0.33
F) 10f CV test	AD	6.69	5.90	6.81	0.49	0.51	0.09
M) holdout training	CN	3.65	4.60	5.86	0.70	0.69	0.39
(M) holdout test	CN	4.13	5.19	5.79	0.63	0.57	0.38
(M) holdout test	AD	6.22	5.82	6.25	0.42	0.41	0.22
(F) holdout training	CN	2.95	4.69	5.77	0.82	0.65	0.33
(F) holdout test	CN	3.46	4.69	5.38	0.59	0.54	0.37
(F) holdout test	AD	7.12	4.92	5.65	0.51	0.45	0.13

The effect of the feature selection method in M3 on MAE and r is the opposite of the objectives achieved in previous studies on brain age for prediction of AD [6] [31] [32], this is because to create the brain age model in those studies the aim is to minimise the error (MAE) between chronological age and the predicted brain age and maximise the correlation (r) between the two. Using the proposed BFFS in M3 worsens the quality of the regression task by increasing the error and decreasing the

correlation between chronological age and the ABA but this is in the expense of improving the performance of the classification task. The ABA in M3 therefore is not a good prediction of the biological age of the subjects but an indication to subject's brain age in the context of the neurodegenerative disease i.e., AD.

In method M3, for each fold of the 10-fold cross-validation of each of the 10 repeats, the BFFS method is performed and an ABA model is built. The BFFS method selects a feature subspace F_1 from the full feature space F . The subspace F_1 will then be used as the input features to LASSO model. Due to possible multicollinearity and in order to avoid redundancy among the input feature subspace F_1 , LASSO performs its own regularisation and penalisation system in order to remove those features with great correlation to each other. Among the correlated features, one feature is selected and the rest will be discarded by having coefficients of zero. The feature subspace F_2 are then selected by LASSO as the features with coefficients above zero, which are also used to estimate ABA. The relationship between the three feature spaces can be demonstrated as $F_2 \subseteq F_1 \subseteq F$. After repeating the 10-fold cross-validation 10 times, there will be 100 sets of F_1 and F_2 .

According to minimum description length (MDL) principle although the feature subspace F_2 could be the right choice for the classification task due providing the minimum number of features which are efficient and necessary in the prediction, the feature subspace F_1 can provide a richer level of details and information. While the LASSO generated feature subspace F_2 provides enough details to support the optimal classification decision making using a model built on

a particular training set, the BFFS generated feature subspace F_1 holds information with a greater level of useful details in order to advise a domain expert.

Over the all 100 folds, in the feature subspace F_1 the average number of features for females is 15 with $9 \leq |F_1| \leq 20$, and the average number of features for males is 16 with $10 \leq |F_1| \leq 23$. Similarly, for F_2, over the 100 folds, in the feature subspace F_2 the average number of features for females is 12 with $5 \leq |F_2| \leq 19$, and the average number of features for males is 12 with $4 \leq |F_2| \leq 22$.

The results presented in Table 4.1 demonstrate the frequencies of the presence of ROIs in the selected feature subspace over 10 folds of the cross-validation, over 10 repeats. The features which are present in the selected feature set in most folds are regarded as important for our model. The frequency of presence of a feature in the feature subspace F_1, provides an important insight into the significance and relevance of the feature when it is used in a particular classification task. These features have also been identified by previous studies to be effective in predicting AD; most of the top features are comprised of the hippocampal parts of the brain [33] where the morphological change and decay in those parts are of characteristics of AD. Other features which are selected many times and are also those connected to AD include amygdala [34], entorhinal cortex [36], cortical regions surrounding the superior temporal sulcus (bankssts) [37] and medial temporal lobe [24]. This proves that the feature selection method suggested, BFFS, has correctly identified multiple ROIs which are affected by AD, without any prior or expert knowledge.

The efficiency of the biased forward feature selection method (BFFS) is verified by automatically recognising and determining the ROIs and brain regions which are affected as the early symptoms of AD [27], such as amygdala, hippocampal regions, entorhinal cortex and temporal lobe regions.

Figures 4.1 and 4.2 provide distributions of ABA vs. Age and ADS vs. Age for methods M2 and M3 respectively for female subjects. Method M2 could be viewed as having the same logic as BrainAGE method [6] when no feature selection is used and all ROIs have been used in LASSO model to estimate ABA, whereas method M3 has the addition of the proposed feature selection method BFFS and ABA LASSO model is built on the features which are highly significant and important to AD prediction. In the left plots, ABA vs. Age, the slope on the plot for M2 has a slightly sharper angle that that of M3 where the distribution is more horizontal. This is the effect of the BFFS on the distribution in order to improve the classification capability of ABA. In Method M3, there is a better separation between AD and CN in both plots of Figure 4.2, when compared to the ones in Figure 4.1. This is also due to the presence of the BFFS method, when although the quality of the regression task is worsened, the classification task performance is improved in favour of the AD prediction.

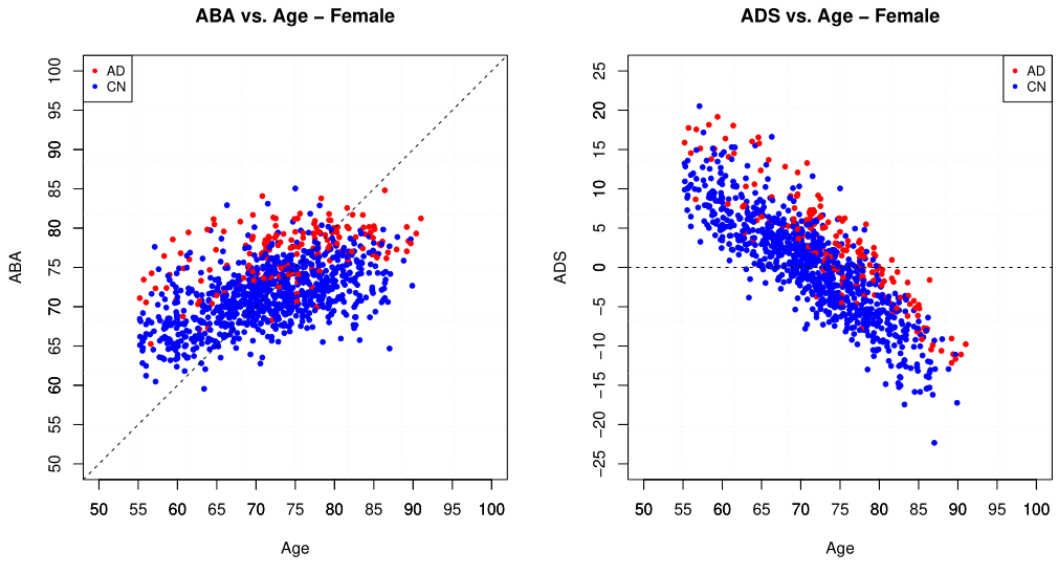

Figure 4.1 | Method M2 – Female subjects: Plot on the left shows ABA vs. Age distribution with dashed line representing y = x and the one on the right shows the ADS vs. Age distribution with dashed line representing y = 0. Both plots used the same data and method. The ABA and ADS used in these plots have been produced using aggregated results on test sets of single run of 10-fold cross-validation. BFFS is not used in M2.

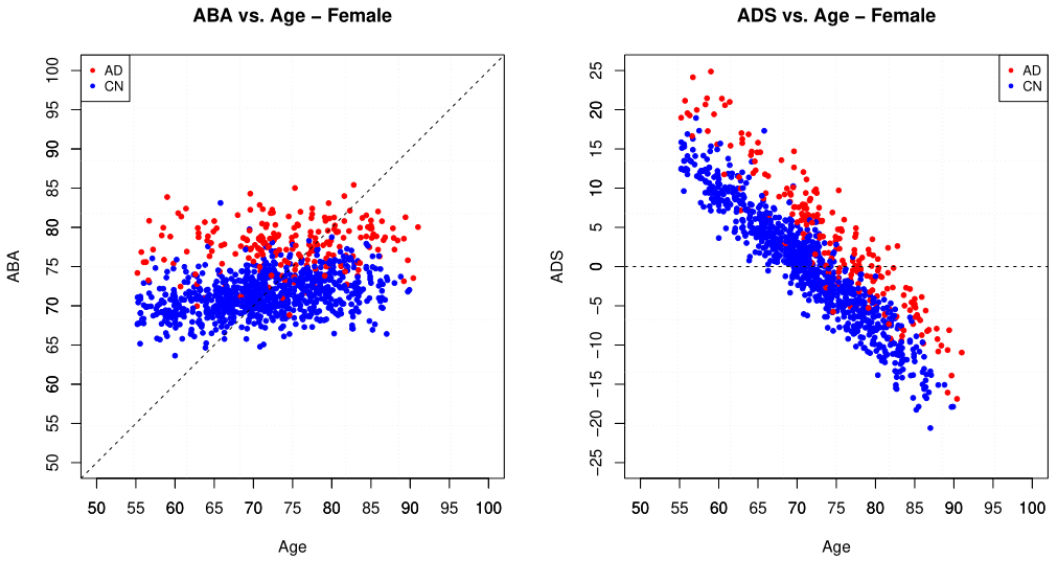

Figure 4.2 | Experiment M3 – Female subjects: Plot on the left shows ABA vs. Age distribution with dashed line representing y = x and the one on the right shows the ADS vs. Age distribution with dashed line representing y = 0. Both plots used the same data and method. The ABA and ADS used in these plots have been produced using aggregated results on test sets of single run of 10-fold cross-validation. BFFS has been used in M3.

In Pseudocode 1 the proposed method of BFFS is outlined. For simplicity of demonstration and easier reading of the pseudocode, validation methods such as holdout and cross-validation are omitted and the models are built and applied to whole dataset available for the method (D) without data partitioning, as a resubstituting method. Also, when $subspace_f$ contains only one feature, linear regression is used instead of LASSO in line 6 to estimate ABA. This is due to the fact that for LASSO algorithms there must be at least two features present in the input features.

Pseudocode 1: Biased Forward Feature Selection (BFFS) approach

Input data : $D = D_{adni,cn} \cup D_{adni,ad} \cup D_{ixi,cn} \cup D_{aibl,cn} \cup D_{aibl,ad} \cup D_{ppmi,cn}$
Output : F_1

1 Initialisation: Create empty $subspace_f$ and set $accuracy_{clf}$ to 0
2 Using D, sort F based on the absolute correlation to class (CN/AD)
3 **for** *(each f with highest absolute correlation)* **do**
4 Add f into $subspace_f$
5 # Regression using LASSO, R cv.glmnet package used for LASSO
6 Build a LASSO model on D_{hc} using $subspace_f$ to estimate ABA (dependent variable: age) $\rightarrow$ $model_{LASSO}(subspace_f)$
7 Apply $model_{LASSO}(subspace_f)$ to all data $D \rightarrow$ ABA generated
8 # Classification using Logistic Regression
9 Build a Logistic Regression model (dependent variable: class) on ABA and age using $D \rightarrow$ $model_{LogReg}(subspace_f$ ABA and age)
10 Test $model_{LogReg}(subspace_f$ ABA and age) on $D \rightarrow TP_{f_i}, TN_{f_i}, FP_{f_i}, FN_{f_i}$
11 # Evaluate performance
12 $accuracy_{LogReg}(subspace_f$ ABA and age) = $\dfrac{TP_{f_i} + TN_{f_i}}{TP_{f_i} + TN_{f_i} + FP_{f_i} + FN_{f_i}}$
13 **if** *($accuracy_{LogReg}(subspace_f$ ABA and age$) > accuracy_{clf}$)* **then**
14 Retain f in $subspace_f$
15 Let $accuracy_{clf}$ = $accuracy_{LogReg}(subspace_f$ ABA and age)
16 **else**
17 Remove f from $subspace_f$
18 Remove f from F
19 Let F_1 = $subspace_f$

4.5. Summary

This chapter explained about how the proposed brain age feature, ABA, was built and how the feature space was selected using the proposed BFFS method. It was shown in this chapter that for building the ABA model the classification ability

was optimised rather than the regression quality of ABA model. In the next chapter, the ABA will be used to predict AD in a binary classification setting.

Chapter 5

5. Classification/prediction of AD

5.1. ABA in classification of AD

In the previous chapter a brain age model, ABA, was proposed, which was built on the features selected to be highly indicative of and affected by AD. The ABA LASSO regression model was not optimised to improve quality of the regression task, but to enhance the subsequent classification task. The metrics such as MAE and r were shown to be worsened as the result of adding the classification-biased feature selection component BFFS: the ABA model is designed not to optimise the age regression accuracy, but rather the AD classification accuracy. In this chapter the effect of the ABA model in the AD classification task will be investigated.

Previous studies have proposed using the brain age as generated feature for the classification of AD [6] [31] [32], however in all the cases brain age was estimated using the whole brain, with the intention to optimise the quality of the regression model and improve its metrics (MAE/MSE and r). Those methods which were used to estimate the brain age, lack specificity to the disease, i.e. AD.

What was proposed in the previous literature on brain age model is that regardless of the specific neurodegenerative disorders to be predicted (Alzheimer's disease, Lewy body dementia, frontotemporal dementia, Huntington's disease, Parkinson's disease, ataxia, etc.), the estimated brain age is the same, as no subjects with the disease are used to train the age model. In other words, in a multinominal classification problem, for neurodegenerative diseases, the same brain age model would be generated and used to estimate the gap with regards to the real age to predict the presence of one of the diseases.

This lack of specificity to the disease makes brain age as biomarker arguable or at least less effective in the prediction task for any neurodegenerative disease. This challenge may not be as apparent when brain age biomarker is used in a binary classification task with a single target disease, e.g., for discriminating patients with AD from healthy subjects (CN). But in a multi-class classification task, which includes multiple positive labels (diseases) and one negative label (healthy) it becomes apparent that using the same brain age model may not be effective at all in distinguishing between types of positive labels. While this chapter focuses mainly on the binary classification, chapter 7 presents the benefits of the proposed ABA model for multinominal classification tasks.

As explained in the previous chapter ABA is the estimation the biological age of the brain with specific reference to a particular neurodegenerative disease, in this case AD. In other words, ABA is the biological age of a subset of regions (features) in the brain (measurements of different parts of the brain) which are selected to be typically affected by AD. Since the proposed approach is meant to be based on machine learning, no expert knowledge was used to select these regions, but rather an automatic feature selection method was specifically designed for this task. This way the proposed approach can be easily applied to other neurodegenerative diseases and, potentially, to other domains directly.

In this chapter, the estimated ABA is used together with the real chronological age to predict and classify AD. The logic behind this approach is that for subjects with AD, the estimated ABA is expected to be higher than the subject age because AD caused faster aging and degeneration in the selected subspace of features: these particular features show a more advanced aging than other features not affected by AD. For healthy subjects the estimated ABA is expected to be similar to the subject age as the selected subspace of features were subject to a normal aging process.

In order to classify AD using ABA and age, a linear logistic regression model is used, where the variable ‘Group’ (diagnosis) is selected as the target/dependent variable and ABA and age are selected as independent variables. The reason for the selection of logistic regression is to maintain linearity of the model to enable the model to be intrinsically interpretable. This linearity will help to create a feature score which shows the direct impact of a single input feature on the classification outcome. The interpretability aspect of the model will be outlined in detail in

Chapter 6. The combination of the ABA linear regression model and the logistic regression model results in an overall model that is also linear with an intrinsic interpretability. This property is exploited in chapter 6, where a feature score definition is derived and used to provide interpretability to the classification model.

As mentioned in chapter 1, three different variants of the main method M (M1, M2 and M3) were tested to show the effect of the different components, such as additional data sources and the feature selection method. Moreover, two methods are tested to provide some baseline performance (B1 and B2). In all experiments discussed in this chapter the aim is the classification of AD; while the three main experiments (M1, M2 and M3) classify AD using the two features ABA and age, the two baseline methods classify AD using all 401 brain features extracted using FreeSurfer v.6 plus age. As the two baseline methods use high-dimensional data (402 features) in the classification task, the SVM algorithm is particularly selected as baseline classifier for its ability to handle a large number of features and for its known excellent performance as shown in [24,22, 27,23 and 26] for this particular classification task.

Experiments performed as part of the preliminary analysis confirmed that selection of linear kernel for the SVM algorithm resulted in the best classification for SVM. This is also confirmed in a similar study on AD classification where Gaussian, cubic, quadratic and linear kernels were tested for SVM and the linear kernel provided the best classification results [38].

In order to evaluate the performance of the classification task in this chapter various metrics are utilised. The goal of a classification task is to predict the label for unseen data and to optimise the classification task, the number of correctly classified data should be maximised. To measure how well each instance in the data is predicted the confusion matrix is used.

The confusion matrix provides an overview of how the data label was predicted in comparison to the original label. As shown in Table 5.1, the confusion matrix is a table with rows and columns representing the actual and predicted labels respectively. These two axes help to identify the number of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) cases in the classification task. In this chapter, for the classification task of CN vs. AD, the AD label is the positive label and the CN is the negative label. TP is when a person with AD is correctly classified as AD. TN is when a healthy (CN) person is correctly classified as CN. FP is when a healthy (CN) person is misclassified as AD. FN is when a person with AD is misclassified as CN.

Table 5.1 | Confusion matrix.

		Predicted label	
		Positive (AD)	Negative (CN)
Actual label	Positive (AD)	TP	FN
	Negative (CN)	FP	TN

Accuracy is a metric which provides an indication of how many instances of the data were correctly classified compared to all predicted data. This is a popular metric to show how well the classifier predicted the correct labels. Equation 5.1 below gives the formula for accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{(E5.1)}$$

While accuracy provides a measure to evaluate classification performance by setting a threshold, area under the curve (AUC) provides an aggregate evaluation measure across all thresholds. AUC is the area under the receiver operating characteristic (ROC) curve, which is visualised by plotting true positive rate (TPR) against false positive rate (FPR) across all thresholds.

Recall, sensitivity or true positive rate, as the name suggests is a metric that provides an overview of the proportion of correctly predicted positive instances over all correctly classified and misclassified positive instances. This metric can show how successful has the classifier been in detecting and identifying the disease cases. Equation 5.2 below gives the formula for recall:

$$Recall = \frac{TP}{TP + FN} \tag{E5.2}$$

Precision or positive predictive value is a metric that shows the proportion of correctly classified positive cases compared to all instances which were predicted as positive (correctly or incorrectly). This metric can show how well the classifier distinguished between positive and negative cases. Equation 5.3 below gives the formula for precision:

$$Precision = \frac{TP}{TP + FP} \tag{E5.3}$$

Specificity or true negative rate is a metric that shows the proportion of correctly classified negative cases compared to all negative cases. This metric can show how well the classifier correctly classified the negative cases. Equation 5.4 below gives the formula for specificity:

$$Specificity = \frac{TN}{TN + FP} \tag{E5.4}$$

The metrics used mainly in previous studies include accuracy, recall and precision. In order to produce results which can be compared to previous literature, those three metrics are reported for the classifiers proposed in main methods M1, M2 and M3 and for the two baseline methods the accuracy is reported.

In the previous chapter it was explained in addition to implicit feature selection in LASSO, the novel proposed ABA-Reg model uses a novel greedy and aggressive feature selection method to select a small number of features which gives the maximum information about the target disease in order to be used in the classification task. Therefore, to build the ABA-Reg two feature selections are performed: proposed BFFS and implicit feature section by LASSO. Although the performance of this method can be achieved by reporting the performance metrics such as accuracy, the individual effect of the feature selection method is not shown when the overall performance is reported. Therefore, in order to evaluate the effect of the proposed feature selection method on the classification performance, ABA-Reg model is built without the BFFS method (using all 402 features).

The proposed method uses data from multiple sources to build the ABA model and perform a classification. It is considered that having data from multiple sources improves the generalisability of the model and ABA is a data driven feature which will perform better in the classification task when more data is used to train the ABA model. But to show that more data improves the classification task, a method is performed where only single source of data i.e., ADNI is used. This method and settings are set up as the initial setting which used ABA in the classification task. As the aim is to evaluate the effect of additional data and feature

selection on the classification performance, these initial settings will be using ADNI data without the proposed feature selection. This setting will be referred to as M1. To view the effect of additional data, other sources are added and this setting will be referred to as M2. And as the final addition, the proposed feature selection is added and this setting will be referred to as M3.

Three setting or experiments have been suggested to not only show the classification performance of ABA feature but also show the effect of data and feature selection on it. Although these three settings show a comparison among different settings using ABA, there should be a setting where classification is performed without ABA using all features in a state-of-the-art classifier. This setting will then be used as a baseline to the proposed method. In this baseline method all (401) brain features and real age (402 feature combined) will be used in SVM to build a model to classify AD vs. CN.

In addition to the baseline settings, it is investigated that how data affects the performance of the classification task, similar to transition from M1 to M2. As the result there will be two baseline settings; B1 where 402 features used in SVM classifier using ADNI data only and B2 where data from other sources are added with the same setup as B1. The transition from B1 to B2 is like that of M1 to M2, where the only difference is data used and the effect of additional data is investigated.

To evaluate the performance of the classification task three performance metrics have been used: accuracy, recall and precision. These three performance

metrics have been selected to compare the three experiments in this thesis (M1, M2 and M3) and also to compare the proposed method to previous relevant literature. In order to compare the proposed method to baseline methods (B1 and B2) accuracy is used.

5.2. Classification results and discussion

Multiple settings (M1, M2 and M3) have been used in order to show the effect of addition of data and the suggested feature selection method (BFFS) on the AD classification workflow. Also, to have a baseline method for the workflow, SVM is used, where all ROIs and age are used in classification of AD, using single and multiple sources of data in order to evaluate the effect of additional data on the baseline method too.

Table 5.2 | Overview of the classification results reported for each gender separately. B1 is the baseline experiment using ADNI data only while B2 is the baseline experiment with more than one data source. The proposed model is evaluated in experiment M3 where multiple settings in experiments M1 and M2 have been used to show the contribution and effect of additional data sources and the proposed feature selection method. Reported evaluation metrics are accuracy (for all 5 experiment) and precision and recall for the three experiments of M1, M2 and M3. The reported values are in the form of average value percentage and standard deviation in brackets. The highest classification value for each gender across the 5 experiment is marked in bold. For each method, 10-time repeated 10-fold cross-validation is used.

Method ID	**B1**	**B2**	**M1**	**M2**	**M3**
Method Category	SVM-AllF	SVM-AllF	ABA-Com	ABA-Com	ABA-Com
Data Sources	ADNI	All sets	ADNI	All sets	All sets
Feature Selection	Not used	Not used	Not used	Not used	biassed FFS
ABA Regression	Not used	Not used	LASSO	LASSO	LASSO
Classification Features	F, age	F, age	ABA, age	ABA, age	ABA, age
Classification	SVM	SVM	LogReg	LogReg	LogReg
(M) Accuracy % (SD)	87.81 (1.01)	86.73 (0.48)	82.99 (0.10)	85.22 (0.10)	89.74 (0.25)
(M) AUC % (SD)	93.83 (0.45)	90.81 (0.66)	90.6 (0.1)	88.69 (0.1)	93.31 (0.24)
(M) AD Specificity % (SD)	91.99 (0.73)	94.23 (0.56)	88.2 (0.17)	93.8 (0.1)	94.7 (0.1)
(M) AD Recall % (SD)	82.33 (1.21)	69.85 (1.09)	75.21 (0.34)	57.67 (0.51)	75.92 (0.71)
(M) AD Precision % (SD)	87.35 (0.96)	81.3 (1.31)	81.04 (0.19)	74.32 (0.27)	83.69 (0.35)
(M) AD F1 % (SD)	84.76 (0.71)	75.13 (0.41)	78.02 (0.17)	64.94 (0.33)	79.61 (0.54)
(F) Accuracy % (SD)	91.26 (0.28)	92.35 (0.33)	85.05 (0.22)	87.31 (0.04)	92.84 (0.17)
(F) AUC % (SD)	95.08 (0.27)	95.23 (0.26)	90.52 (0.16)	87.59 (0.14)	95.63 (0.22)
(F) AD Specificity % (SD)	95.98 (0.47)	96.42 (0.15)	91.95 (0.22)	96.18 (0.04)	96.64 (0.08)
(F) AD Recall % (SD)	81.52 (0.58)	77.24 (1.5)	69.48 (0.31)	44.52 (0.28)	78.19 (0.62)
(F) AD Precision % (SD)	90.0 (1.03)	84.84 (0.41)	79.28 (0.48)	70.69 (0.15)	85.78 (0.35)
(F) AD F1 % (SD)	85.54 (0.52)	80.86 (0.77)	74.06 (0.34)	54.63 (0.19)	81.81 (0.47)

Table 5.2 outlines the classification results of ABA-Clf model in multiple experiments as well as the baseline experiments. Method IDs starting with B are baseline experiments whereas the ones stating with M are main or ABA experiments. All experiments are performed using 10-time repeated 10-fold cross-validation.

In B1 an SVM is used to classify AD using all (401) ROIs and age (402 feature combined), with only ADNI data. This has achieved accuracies of 91.26% (0.28) and 87.81% (1.01) for females and males respectively. In B2 while the workflow is exactly the same as B1, more data from different sources are added in order to assess the effect the data on the baseline method. The results achieved in B2 has had a decrease in accuracy of 1.08% in male subjects and in increase in accuracy of 1.09% for female subjects. This shows that the additional data from multiple sources has had a mixed effect on the baseline, while it has only changed the accuracy by approximately 1%, it caused an improvement to performance of female subjects and worsened the performance for male subjects.

In M1 a LASSO model (ABA-Reg) is built on healthy subjects of ADNI where the estimated ABA and age are used in a logistic regression model (ABA-Clf) to classify AD in ADNI. Although only 2 features are used in the classification task, the accuracies achieved were 85.05% (0.22) and 82.99% (0.10), for females and males respectively. To further our analysis and experiments, more data sources are added in M2. The analysis performed in M2 is identical to M1, with the difference of additional data sources. By adding more data, the accuracies of our model improve by 2.26% and 2.23% for females and males respectively. These

increases are greater than those from B1 to B2. This shows that our model is greatly data driven and more affected by the change in amount of data than the baseline method which uses SVM. In M3 method, the suggested feature selection method, BFFS, is added. This causes substantial increases of 5.53% and 4.52% in the accuracies for females and males respectively compared to M2. These results show that the proposed ABA-Com model, using two features, achieved better or similar results to the state-of-the-art classification methods using all the 402 features.

The proposed method not only simplifies and eliminates the curse of dimensionality posed by high-dimensional data used in the classification task but by using linear models the descriptiveness of the model is maintained throughout the whole classification workflow. The interpretability of the model is explained in detail in the next chapter.

Figure 5.1 also shows the results of the model evaluation, by plotting accuracies of each method on a bar chart for each gender separately. This is to enable easier comparison between accuracies achieved among different methods. The y-axis is showing the accuracies between 75-95% as all reported accuracies are in that range. As can be seen in this figure, the effect of data is greater on the proposed method than the baseline method. It can also be seen that for both genders, the proposed method has achieved comparable accuracies to the baselines.

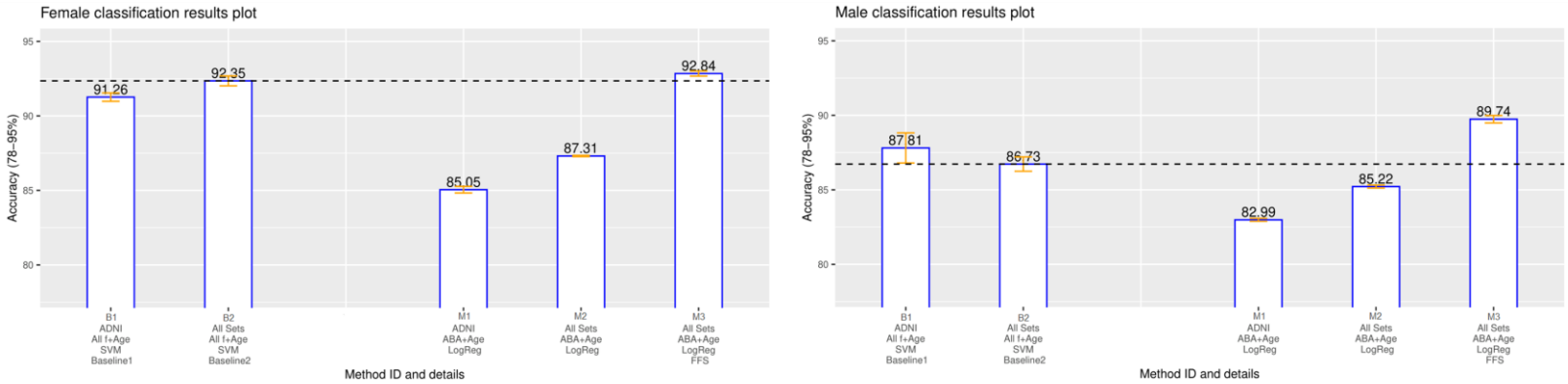

Figure 5.1 | Overview of the classification accuracies reported for each gender separately in a bar chart format. All the accuracy details are the same as the ones provided in tabular format in Table 5.2. This is to give an easier comparison between the accuracies achieved by each of the 5 experiments.

Figures 5.2 and 5.3 show the ABA vs. age and ADS vs. age for female and male subjects respectively, using the aggregated validation sets of 1 run of 10-fold cross-validation from M3 method. In these figures, the plot on the left, ABA vs. age plot, subjects with AD (represented in red dots) are placed mainly on top of the healthy subjects (represented in blue dots), this shows that ABA is correctly estimated to be higher than age for subjects with AD. The green solid lines represent the logistic regression boundaries from each fold of the cross-validation, whereas the solid black line is the boundary of a logistic regression model built on all data. The solid lines show that how logistic regression can classify AD from CN, using only the two features of ABA and age. As these two figures are produced for the method M3, the proposed features selection method is also applied before the classification.

A similar plot to those in Figures 5.2 and 5.3 is presented in Figure 5.4. In this figure, the left plot represents the ABA vs. age and ADS vs. age for female subjects, using the aggregated validation sets of 1 run of 10-fold cross-validation from M2 method. It can be seen that the distribution in left plots in Figures 5.2 and 5.3 are more horizontal than the distribution in left plot of Figure 5.4. This shows the effect of the proposed feature selection method on the classification.

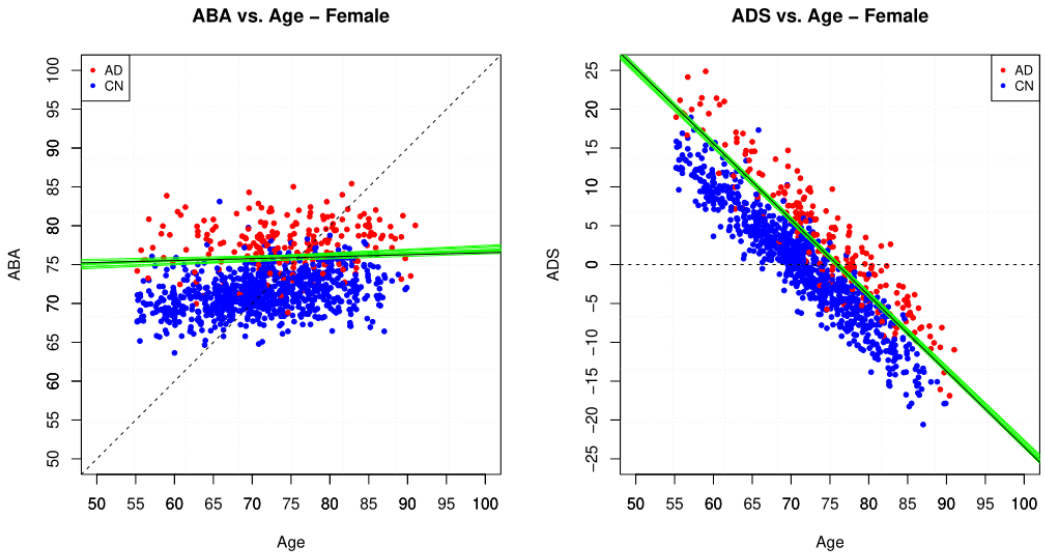

Figure 5.2 | Female M3: The two plots presented in this figure show the ABA vs. age (left) and ADS vs. age (right). Both plots were produced using the aggregated validation sets of one run of 10-fold cross-validation from M3 method using data from female subjects. AD subjects are shown in red dots and CN subjects are shown in green dots. Green solid lines are the logistic regression boundaries from 10 folds of the cross-validation. The Black solid line is the logistic regression boundary using the final model which has been trained on all the data. The black dashed line is y = x and is used to reference.

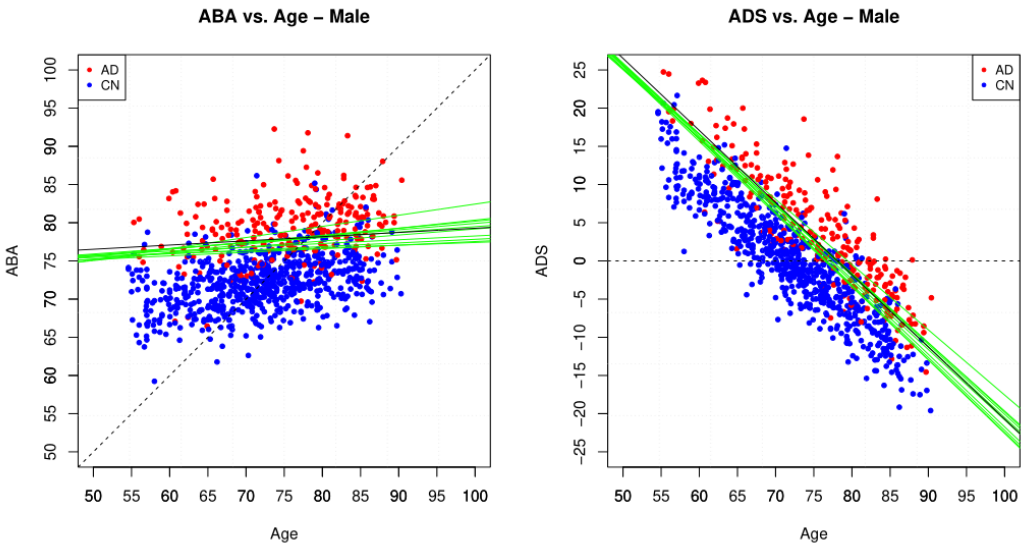

Figure 5.3 | Male M3: The two plots presented in this figure show the ABA vs. age (left) and ADS vs. age (right). Both plots were produced using the aggregated validation sets of one run of 10-fold cross-validation from M3 method using data from male subjects. AD subjects are shown in red dots and CN subjects are shown in green dots. Green solid lines are the logistic regression boundaries from 10 folds of the cross-validation. The Black solid line is the logistic regression boundary using the final model which has been trained on all the data. The black dashed line is y = x and is used to reference.

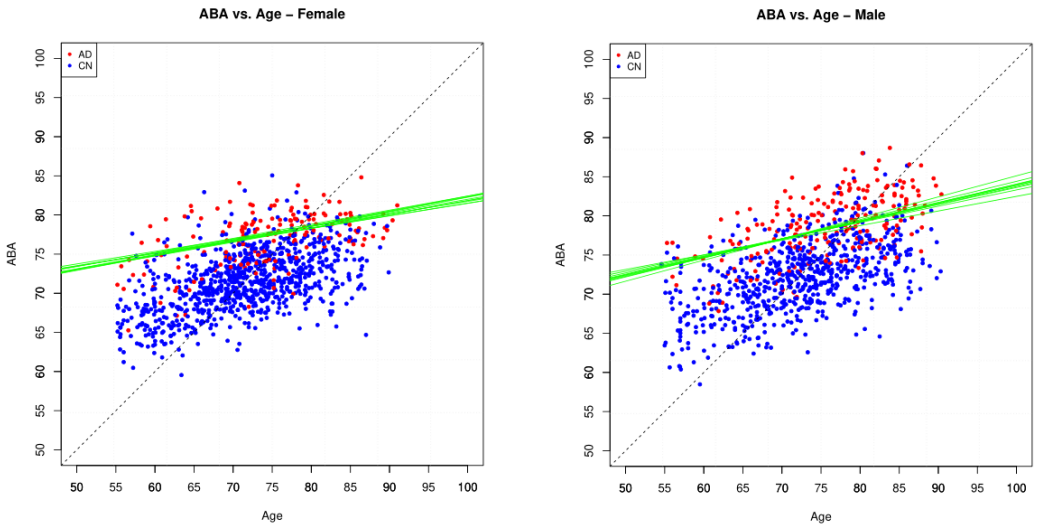

Figure 5.4 | Female and Male M2: The two plots presented in this figure show the ABA vs. age (left) and ADS vs. age (right). Both plots were produced using the aggregated validation sets of one run of 10-fold cross-validation from M3 method using data from male subjects. AD subjects are shown in red dots and CN subjects are shown in green dots. Green solid lines are the logistic regression boundaries from 10 folds of the cross-validation. The black dashed line is y = x and is used to reference.

5.3. Summary

In this chapter the classification ability of the proposed method is presented and the performance of this method is evaluated in a binary (CN vs. AD) classification setting. In order to provide an explanation as to why a prediction decision has been made an interpretability index is proposed in the next chapter.

Chapter 6

6. Interpretability

There has been a focus in the field of AI to make the output of the machine learning models more explainable in a way that it is understandable to the human mind. This has led to the emergence of the concept of Explainable Artificial Intelligence (XAI) where explainability of the model is improved and focused on.

In the field of machine learning, the two terms of explainability and interpretability are used interchangeably and there isn't any clear definition for them. However, some authors have suggested that there are clear distinctions; [39] explains that interpretability refers to intrinsic characteristics of the machine learning model when it is clear and understandable to human mind why and how such output has been produced. However, the concept of explainability applies when the model is black box and not intrinsically interpretable, so there is an attempt to explain the output of the model for the human minds to comprehend.

There are explainability frameworks such as SHAP [40] or LIME [41] which are used to explain what the output of the machine learning model means and

provide an explanation on those outputs generated by black-box and non-linear approaches such as DNN. [39] also advises that when the machine learning models are used in high stake decision makings in criminal justice and healthcare fields, it is better to make the models intrinsically interpretable rather than providing an extrinsic explanation on the output of the models which are intrinsically black-box. This factor has been part of the motivation in making the proposed method intrinsically interpretable as it can be used in a healthcare domain as a biomarker where human lives are at stake.

Although it is challenging to have a model with interpretability and high predictability performance at the same time [42] the proposed method in this thesis has achieved both. The proposed method is intrinsically interpretable for its specific setup and healthcare domain. One improvement to the method would be to make it applicable to other domains. Also, the proposed method cannot be applied to other black-box approaches to provide an explanation, such as what SHAP and LIME do, therefore the method cannot be applied to models such as DNN to explain the output.

6.1. A descriptive feature score for AD classification

The ABA-Clf model has been described and its performance in binary classification of AD vs. CN evaluated in the previous two chapters. In this chapter, the interpretability of the ABA-Clf model is outlined. The LASSO and logistic regression models are both linear models and this helps to have an interpretable

model where input features (ROIs) can be directly linked to the classification outcome.

As part of the machine learning model building for a classification task it is always good if the reasons behind decision made by the ML model were transparent and could be explained and interpreted in a way that the final classification outcome is descriptive. The interpretability and descriptiveness of the model is particularly important in the field of computer aided diagnosis of AD as this involves health of human subjects and the medical professionals need to make sure they understand the reasons behind an outcome from an ML model before being able to use the prediction.

Unfortunately, most of the state-of-the-art algorithms, despite achieving high performance, have black-box approaches and lack interpretability and descriptiveness. This makes the ML model outcome challenging to interpret. In previous studies on classification of AD using brain age [6] [31] [32] although state-of-the-art algorithms have been used, the models proposed lack descriptiveness. In the proposed ABA-Com model, one of the main objectives, in addition to achieving high accuracy in prediction of AD, is to produce an explanation as to why such prediction has been produced. That explanation can then be used by medical professionals to help with diagnosis of AD by identifying the specific regions of the brain that are affected by AD and by how much.

In order to make the AD prediction model interpretable, there needs to be a linear relationship between the input (brain features and age) and the classification

outcome (class label). In the field of analysis of brain MRI scans for detection of patterns in AD patients, there are a large number of regions and therefore features present for the ML analysis. This is referred to as high-dimensional data and therefore there is the curse of dimensionality. To handle the curse of dimensionality and reduce the number of features used in the model different methods can be used.

In previous studies on classification of AD using brain age [6] [31] [32] dimensionality reduction technique, PCA, is used. However, in PCA, by producing and replacing the brain features with principal components the descriptiveness is removed from the model. In the method proposed in this thesis, a novel feature selection approach is used to reduce the dimensions of the feature space to a minimal by retaining the maximal classification performance.

The brain features used in the analysis in this thesis come from the numerical measurements such as thickness, volume and area of different regions of the brain e.g. the thickness of right HATA and the volume of right Hippocampus. Each of the features have direct relationship to a region in the brain and therefore have a meaning to medical professionals. The proposed classification workflow ensures that there is a direct relationship between those semantic measurements inputted to ML model and the outcome. To make that direct relationship transparent and identify the features which are most useful in making the prediction, a novel score is given to each of the features or brain region measurements which are present in classification of AD (selected by both the proposed feature selection method and LASSO). This score can then show how much each feature contributed to the final ML classification outcome, which can be helpful for medical professionals in

decision making on AD diagnosis. As an example, when five features are selected to be present for positive AD classification, one those features are Left-Hippocampus with the highest score, then the professionals know that based on the ML analysis Left-Hippocampus has the highest impact on the AD classification and AD has possibly affected that region the most compared to other regions of the brain.

As part of the proposed ABA-Com building method (M3) the proposed feature selection method (BFFS) is performed followed by the internal feature selection by LASSO. As a result a limited number of features are selected for the ABA-Reg model to be built on. The ABA feature will then be used in addition to age (two features) to classify AD. The features used in building the ABA-Reg model all have different impacts on the ABA feature and therefore have indirect impacts on the classification task. The extent of the impact each feature has on the ABA can demonstrate the impact they have on classification decision.

The main aims and objectives of the proposed workflow of ABA-Com model and AD classification using ABA is to have a linear model and workflow while maintaining high performance and high level of descriptiveness and interpretability. Having a linear model with reduced complexity helps with the level of descriptiveness and interpretability which ultimately gives medical professionals more confidence in using this model as an indication to AD diagnosis.

In order to find the impact of each feature on the building of ABA-Com model, LASSO coefficients are used. Those coefficients with larger values have

greater impact on determining the value of ABA and ultimately the classification decision. In order to use the information provided by LASSO in building a linear regression model for ABA (ABA-Reg), a novel feature score s_i is proposed to incorporate the coefficient given by LASSO for a single feature i.

A novel feature score s_i is proposed, where the score of a feature measures its relative contribution towards the classification outcome. The feature score can be used to rank the input features and determine their importance for a single specific subject.

In this thesis, as explained in previous chapters, ABA is estimated by building a LASSO model (ABA-Reg), which is a penalised linear regression model, on a selected set of features referred to as F1 and resulted from the BFFS. The equation in E4.1 shows the linear relationship between the input feature values and ABA. Using the estimated ABA and chronological age, a logistic regression model (ABA-Clf), which is also a linear model, is built to predict the class label (AD vs CN). The logistic regression model creates a linear decision boundary to classify AD from CN and the inequation of that boundary is given below in E6.1:

$$c_0 + c_1 \cdot age + c_2 \cdot ABA < 0 \qquad \text{(E6.1)}$$

where c_0 is the intercept and the c_1 and c_2 are the coefficients. Considering the equation given in E4.1 and the decision boundary inequation E6.1 given above can be rewritten as:

$$c_0 + c_1 \cdot age + c_2 \cdot \left(a_0 + \sum_{i=1}^{k} a_i \cdot f_i \right) < 0 \qquad \text{(E6.2)}$$

Assuming $c_2 < 0$ and $(c_0 + c_1 \cdot age + c_2 \cdot a_0) < 0$, equation E6.2 can be shown as:

$$\sum_{i=1}^{k} - \frac{c_2 \cdot a_i \cdot f_i}{c_0 + c_1 \cdot age + c_2 \cdot a_0} > 1 \qquad \text{(E6.3)}$$

A feature score s_i associated to the feature f_i is introduced in E6.4 as the contribution of that feature in the summation of inequation E6.3 and can be used to measure the relative contribution made by the feature to the classification outcome. The higher the s_i the more f_i contributes to an AD classification and, vice versa, the lower the s_i the more f_i contributes to a CN classification. In other words, for a given subject a higher value of s_i than the average could mean there is an abnormal atrophy in f_i.

$$s_i = - \frac{c_2 \cdot a_i \cdot f_i}{c_0 + c_1 \cdot age + c_2 \cdot a_0} \qquad \text{(E6.4)}$$

The inequality in equation E6.1, which is used for the classification of AD, can be given in the following form, as the summation of all feature scores:

$$\sum_{i=1}^{k} s_i > 1 \qquad \text{(E6.5)}$$

The summation of all feature scores from those features present in F2 (features selected by LASSO) would determine the classification outcome. The feature scores are designed in a way that the summation greater than 1 would give a classification of AD where the summation of below 1 results in classification of CN. Although 1 is the boundary for the classification using the aggregate of all feature scores, some features may have scores much greater or smaller than 1. For example, when an AD patient is diagnosed as AD (TP) it means the summation of feature scores has been above 1, in this case features scores of some features may be much greater than 1 and some may be close to 1. Those features with score much greater than 1 are the ones which are having the highest impact and result in an AD classification. These features could also be interpreted as being affected by AD more than other features. In fact, the proposed feature score can be a direct indication to the brain regions and features which have been highly affected by AD. This would help medical professionals in understanding the reasons behind such classification which ultimately help with the AD diagnosis.

6.2. Interpretability results and discussion

For the interpretability analysis, the ABA-Com model was tested on the female subjects using a holdout validation method, where the 80% of the data are used for training and 20% for testing. Following the application of the outlier detection method, 7 subjects were identified as outliers and removed. Therefore, 1063 subjects were used in this analysis, where 850 are in the training set and 213 are in the test set. The accuracy achieved on the test set using this holdout validation method is 89.67% (TP = 27, TN = 164, FP = 5, FN = 17).

The reason for using the holdout validation instead of cross-validation is to have a single set of selected features for the interpretability visualisation. Using cross-validation could result in different feature sets produced in each fold.

Figure 6.1 provides the ABA vs. age plot for the subjects in the test set for the analysis performed in this chapter. In this plot the solid black line is the logistic regression boundary separating the two classes of CN and AD, with the points below the line classified as CN and points above the line classified as AD. The colour of the point in the plot are based on the original class labels (CN as blue and AD as red). As the plot shows, 22 subjects are misclassified: 5 CN subjects are misclassified as AD and placed above the solid black line and 17 AD subjects misclassified as CN and placed below the solid black line.

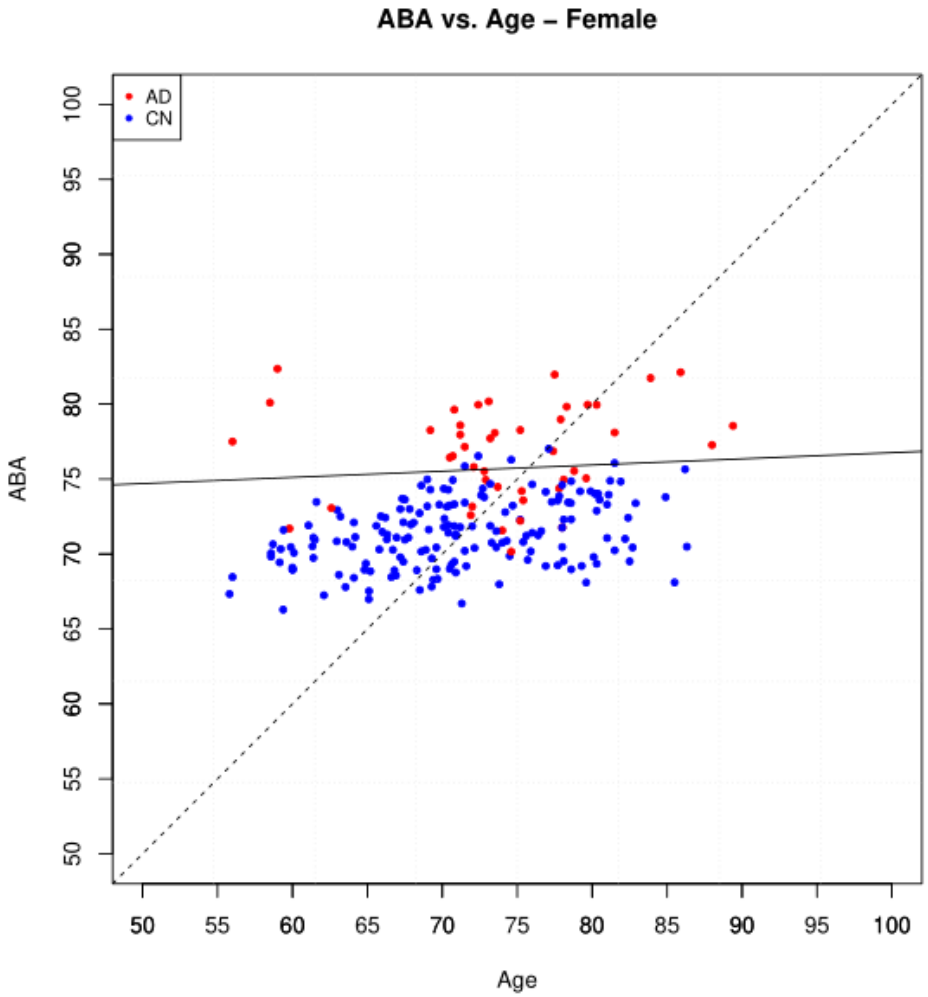

Figure 6.1 | Distribution of age vs. ABA produced from ABA-Reg model in experiment M3 using hold-out method (data used for this plot are from the test set). Solid black line represents the ABA-Clf model boundary line. The diagonal dashed line represents y = x. Female subjects data used.

To showcase how the interpretability analysis works and show the level of descriptiveness in the feature scores, box plots are used. In these plots the distribution of score s_i for each f_i present in F2 is plotted separately for each of the classes of AD and CN using the subjects in the training set. This is to show where the distribution of each class for the training set data are positioned on the plot. The s_i for each f_i present in F2 for a single selected specific subject from the test set is then plotted as dots on the box plots in order to be compared to the distribution of box plots or data from the training set. Distributions are plotted in red and blue to represent AD and CN respectively.

To perform the plotting and showcase the interpretability and distribution of s_i , multiple subjects are selected, one true positive and one true negative, in order to show how their feature scores differ. To view the feature scores for the two selected cases of TP and TN, Figure 6.2 and Figure 6.3 are provided respectively.

The features showing in Figure 6.2 and Figure 6.3 are the features selected by the LASSO algorithm when building the ABA model, also identified as F2. The distribution of feature scores for the features present in F2 are provided in these two figures in the form of box plots with 2 colours of blue and red representing CN and AD subjects respectively in the training set. Both figures show that the feature scores for CN are mainly below 1 while the feature scores for AD are mainly above 1, as stated in E.6.5. This is in line with the logic behind the feature score, where the higher the feature score, the more likely that that feature belongs to an AD subject.

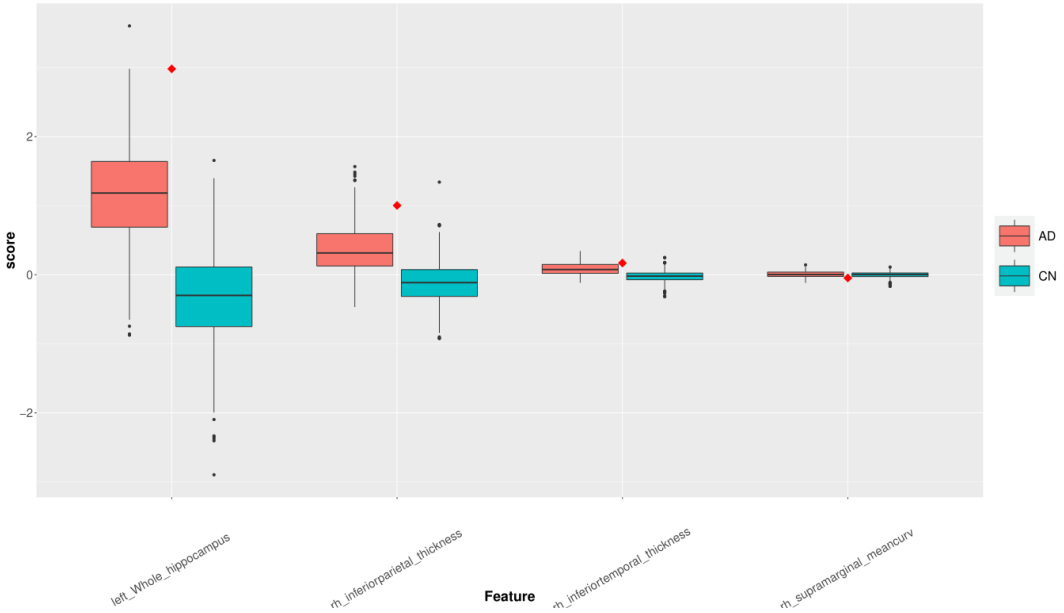

Figure 6.2 | Distribution of the feature scores for a single case of True Positive (TP), selected from test set of hold-out method are plotted using red dots. The box plots represent feature scores of all subjects in the training set of hold-out method. Female subjects data used.

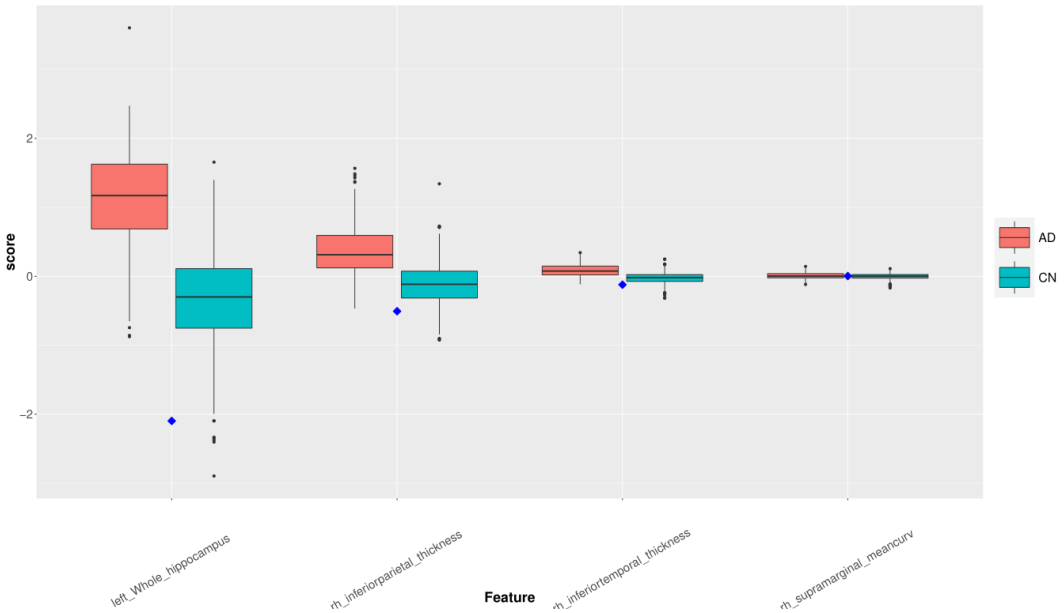

Figure 6.3 | Distribution of the feature scores for a single case of True Negative (TN), selected from test set of hold-out method are plotted using blue dots. The box plots represent feature scores of all subjects in the training set of hold-out method. Female subjects data used.

Figure 6.2 shows a single red point for each feature, representing the feature scores for a single selected TP case. For this subject, the feature scores, from left to right, are 2.98, 1.00, 0.17, -0.05, and the sum of the scores is 4.10. As this sum is greater than 1, the subject is classified correctly as AD. Individual feature score distributions and the place of the red dots also show how badly a feature is affected by AD. In this figure, Left Whole Hippocampus has the feature score of 2.98, which is much higher than 1 and even the distribution of feature scores for this feature for subjects with AD. This shows that this feature (region of the brain) has been affected greatly by AD.

In contrast to Figure 6.2 and the distributions of feature scores for a TP case, the feature scores of a TN case are also presented as blue points in Figure 6.3. The values blue points represent for the features, from left to right are, -2.10, -0.51, -0.122, 0.00 with the sum of -2.72. As can be seen in this figure, all the individual feature scores for this subject (blue points) are below 1 and as the sum of scores is also below 1, the subject has correctly been classified as CN.

In order to provide more examples of the of distributions of feature scores eight female subjects and two male subjects are selected. The eight female subjects provide two TP, two TN, two FP and two FN cases which are presented each in Figures 6.4 to 6.11. To also demonstrate how the feature scores perform in male subject group, two cases of TP and TN are presented in Figure 6.12 and 6.13 respectively. These are provided in Appendix C.

6.3. Summary

In conclusion, this chapter has shown that the proposed framework can provide explanations as to why a prediction has been made. These explanations are though the interpretability index which shows the presence of each brain feature and the extent of impact of each feature on the outcome. This will then help the medical professional understand why a person has been predicted to have AD and which parts of the brain contributed most to that prediction. This is a significant improvement compared to the black-box methods where the decision making behind the prediction cannot be easily explained.

Up to this chapter the focus of the proposed method has been on the binary classification. In the next chapter, the proposed method will be applied in a multi-class classification setting.

Chapter 7

7. Multi-Class classification

In this thesis the focus has been on binary classification of AD vs. CN. This type of classification is what has mainly been studied in the previous literature which can ultimately assist the medical professionals in their decision making and clinical diagnosis when attempting to decide whether a person has AD or is a healthy person. This type of binary classification to detect AD has been most researched compared to other types of dementia because AD affects more people than other types of dementia and, consequently more data are also available.

In the previous three chapters, ABA-Com, the full interpretable workflow of binary classification of AD vs. CN using the proposed Apparent Brain Age (ABA) feature is defined. In those three chapters the ABA-Com was applied to AD and CN data and the aim was to choose a minimal number of features while maintaining the maximal performance and interpretability when predicting AD from CN. The logical progression from this analysis is to apply the same ABA-Com workflow in a multi-class setting where in addition to AD, other types and stages of dementia

are also classified. This can be particularly helpful when there is a risk of misdiagnosis among dementia types due to some shared and similar symptoms. Misdiagnosis could result in the prescription of wrong drugs and medical treatment plan which does not help with the treatment of the actual disease. Therefore, a multi-class setting for the workflow is proposed.

In this chapter the ABA-Com model workflow, which involves LASSO regression (ABA-Reg model) and logistic regression classification (ABA-Clf model), is applied to the data with healthy subjects and subjects with any of the following selected five types and stages of dementia. The dementia types and stages selected are Alzheimer's disease (AD), Frontotemporal dementia (FTD), early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI) and mild cognitive impairment (MCI).

It is worth mentioning that MCI is not considered a type of dementia clinically as it is a group of symptoms that could convert to AD or other types of dementia overtime or they could get better and disappear without conversion into a dementia type. Therefore, MCI could be seen as an initial stage and a biomarker to development of a dementia type. The MCI itself has two stages of late MCI (LMCI) and early MCI (EMCI). In this chapter, the data from the two types and three stages of dementia are utilised so that ABA-Com model can help to explore the relation between different types and stages of dementia.

In the ABA-Com model workflow for the multi-class classification of dementia types and stages, the proposed ABA feature will be different and specific

to each dementia type/stage and the features selected by the proposed feature selection method BFFS are indicative of that particular dementia type/stage. This characteristic and specificity to the disease is what makes ABA feature different to the brain age features suggested in [6] [31] [32] where the features are not created with specificity to a particular disease and therefore may not have a good performance in a multi-class setting.

In this chapter all available data were used which consists of 3,170 subjects of which 1,603 are male and 1,567 are female subjects. These data have been obtained from ADNI, AIBL, IXI, NIFD and PPMI.

7.1. Detection of dementia types and stages using ABA

As explained in Chapter 4, the proposed feature selection method BFFS aggressively selects minimal number of features which are highly specific to a particular disease. In other words, ABA-Reg is built on parts of the brain which had the highest effect from the disease. This makes ABA feature specific to a particular disease. This means that ABA feature is the biological age of those parts of the brain. This specificity to the disease helps ABA achieve a better AD classification performance compared to generic whole-brain approaches proposed in previous literature [6] [31] [32].

Up to this chapter binary classification of AD vs. CN was aimed and therefore a single ABA-Reg model was needed to be built for AD patients using the brain features (ROIs) which are highly affected by AD. In this chapter five dementia

types/stages will be classified in a 6-class classification setting of AD vs. EMCI vs. LMCI vs. MCI vs. FTD vs. CN. This is to show the classification performance of ABA in a multi-class setting.

The ABA-Com workflow in this chapter is different to the ABA-Com workflow for binary classification. In this multi-class classification workflow, due to specificity of the ABA-Reg model to the disease, a separate ABA-Reg model is built for each disease and as there are five diseases or types/stages of dementia, five ABA-Reg models are built.

Let ABA_x be an ABA feature which was created using two groups of subjects combined: healthy subjects and subjects with disease x. The ABA-Reg model is always built on healthy subjects only but the proposed feature selection method BFFS selects the features using the two groups of subjects. In other words, to estimate ABA_x, BFFS is applied to healthy subjects and subjects with disease x combined in order to select the brain features (ROIs) which are highly indicative of disease x. Those selected features are then used to build ABA-Reg for that specific disease x.

After creation of ABA_x for each disease x, the two features of ABA_x and age are used to build a logistic regression binary classification model (ABA-Clf model) using the two groups of healthy subjects and subjects with disease x. This ABA-Clf model is referred to as $ABA - Clf_x$. The generated logistic regression model $ABA - Clf_x$ is then applied to all subjects to predict the probability of each subject having disease x.

After applying all five $ABA - Clf_x$ models to all subjects, each subject will have a set of five probabilities resulted from the logistic regression models referred to as $Prob_x$ which provides the probability of a subject having disease x. The disease x with the highest $Prob_x$ will then be selected as the predicted class. If probabilities of all diseases are below 0.5, the subject is classified as healthy and CN is predicted.

The five probabilities $Prob_x$ resulted from the logistic regression models are the result of a binary classification of healthy vs disease x. Therefore, when the probability of a subject having disease x is $Prob_x$, the probability of that person being healthy is worked out as $1 - Prob_x$. As an example, when there is a probability of a person having MCI is 0.65, which is determined by binary logistic regression model, the probability of the same person being healthy is 0.35. So, when the probabilities of all diseases are below 0.5 it means that there is less than 50% chance that the subject will have any of the diseases, which as the result means that there is more than 50% chance of the subject being healthy and therefore the subjects is classified as CN.

To show the effect of additional diseases on the classification model, 4 experiments are proposed. Initially, MCI is added to the two classes of AD and CN to create the 3-class model. This combination of classes is what is mainly used in previous literature on multi-class classification of dementia. In this experiment CN vs. AD vs. MCI classification is performed.

To add an additional dementia type, FTD is added and the 4-class model is proposed. In this experiment CN vs. AD vs. MCI vs. FTD classification is performed. An additional 4-class model is also proposed when all EMCI, LMCI and MCI are used as MCI in the model. This is to see the effect of E/LMCI on the model. And in the final experiment the 2 labels of EMCI and LMCI are also added to create the 6-class classification of CN vs. AD vs. MCI vs. FTD vs. EMCI vs. LMCI.

The four multi-class classification experiments proposed are performed using the proposed BFFS and they are reported in Table 7.1 as ABA_{wFS_x} . To compare these results to scenarios where BFFS is not present, the 4 experiments were performed without BFFS and are labelled as ABA_{woFS} in Table 7.1. Also, to have a baseline for all 4 experiments where all brain features are used in SVM to make classification experiments are performed with the label Baseline SVM Table 7.1. The baseline is to compare the multi-class classification ability of the ABA-Com to SVM.

7.2. Results and discussion

In Table 7.1 for each of the three categories of experiments the accuracy reduces as additional classes are added. In ABA_{wFS_x} category, for the 3-class classification using CN, AD and MCI (W1), the reported accuracy is 69.85% and 77.3% for males and females respectively. This is an increase of 5.88% and 4.41% for males and females respectively compared to that of ABA_{woFS} (WO1). This shows that BFFS has had a significant positive impact on the classification

performance of ABA-Com in a 3-class setting. To compare these results to the baseline, there is an increase of 4.72% and 4.53% in accuracy in from B1 to W1 for males and females respectively. This shows that ABA-Com has performed better in a 3-class setting in ABA_{wFS_x} compared to the baseline. This means two features of ABA and age, in addition to providing interpretability, were able to perform better in the classification task than all brain features and age combined using SVM algorithm.

In Table 7.1 it can also be seen that for each of the four multi-class classification scenarios, ABA_{wFS_x} has performed better than ABA_{woFS} and baseline. The four scenarios in ABA_{woFS} category on the other hand performed poorly on average compared to the baseline. This shows the considerable positive impact of BFFS on the classification performance.

In Table 7.2, the W1 results reported in Table 7.1 were used to be compared to the previous literature. This is because previous literature in the field of multi-class classification of AD mainly used the three classes of AD, MCI and CN. This 3-class classification is also very challenging as AD and MCI have very similar symptoms and affect the same parts of the brain as MCI is viewed as the initial stage of AD. This 3-class setting is therefore selected to show the performance of ABA-Com in this challenging scenario and also be comparable to previous literature.

Table 7.1 | Summary of multi-class classification results for both gender groups using three experiments of Baseline SVM, ABA_{woFS} and ABA_{wFS_x}, each of which has been implemented on four different class groups: four Baseline SVM experiments (B1, B2, B3, B4), four ABA_{woFS} experiments (WO1, WO2, WO3, WO4) and ,the proposed approach, four ABA_{wFS_x} experiments (W1, W2, W3, W4). These results have been achieved using 10-fold cross-validation, repeated 10 times. For the target classification task the estimated accuracy is reported (average and standard deviation over 10 repeats) which is also equal to, precision, sensitivity and F1 due to using of micro averaging method.

		Multi-Class Classification Method											
		Baseline SVM				ABA_{woFS}				ABA_{wFS_x}			
		B1	B2	B3	B4	WO1	WO2	WO3	WO4	W1	W2	W3	W4
Gender	Overall performance metric % (SD)	3-Class	4-Class	4-Class (E/LMCI as MCI)	6-Class	3-Class	4-Class	4-Class (E/LMCI as MCI)	6-Class	3-Class	4-Class	4-Class (E/LMCI as MCI)	6-Class
Male	Accuracy	65.13 (0.18)	63.06 (0.65)	54.63 (0.4)	51.52 (1.08)	63.97 (0.56)	59.59 (0.33)	54.49 (0.52)	49.54 (0.59)	69.85 (0.88)	64.94 (1.35)	55.01 (1.3)	52.36 (0.52)
Female	Accuracy	72.77 (0.88)	70.35 (0.71)	60.26 (0.51)	58.83 (0.75)	72.89 (0.24)	67.85 (0.18)	57.87 (0.19)	57.94 (0.29)	77.3 (0.35)	73.39 (0.22)	63.66 (0.67)	61.94 (0.6)

Table 7.2 | Comparison of 3-Class experiments in this study with relevant previous studies. The algorithms used for classification in the previous four studies include extreme learning machine (RELM), regularised extreme learning machine (RELM), nonlinear graph fusion (NGF), random forest (RF), and stacked autoencoder (SAE). All methods have used n-fold (nf) cross-validation (CV), with some repeated (rep) multiple times.

ID	Study	Dataset	Subject	CN	MCI	AD	Validation	Classifier	Acc	Sens	Spec	F1	Prec	Prec CN	Prec MCI	Prec AD
1	[43]	ADNI	746	200	441	105	5f CV rep 100	ELM	61.5 (1.2)			59.0 (1.4)				
2	[44]	ADNI	147	35	75	37	4f CV rep 100	NGF, RF	56.3							
3	[45]	ADNI	210	70	70	70	LOO CV	RELM	61.58	54	62.25					
4	[46]	ADNI	800	200	400	200	10f CV	SAE	46.30 (4.24)	66.14 (10.57)	77.78 (4.48)			52.40 (8.43)	41.25 (7.16)	46.89 (4.40)
5	Baseline SVM	ADNI AIBL IXI NIFD PPMI	1250	849	180	221	10f CV rep 10	SVM	72.77 (0.88)	72.77 (0.88)	86.39 (0.44)	72.77 (0.88)	72.77 (0.88)	89.26 (0.32)	27.86 (1.94)	55.91 (2.28)
6	Proposed ABA-Com ABA_{wFS_x}	ADNI AIBL IXI NIFD PPMI	1250	849	180	221	10f CV rep 10	Logit	77.3 (0.35)	77.3 (0.35)	88.65 (0.18)	77.3 (0.35)	77.3 (0.35)	86.48 (0.61)	32.17 (2.82)	58.53 (2.04)

Four previous studies have been selected and presented in Table 7.2 to compare to the reported results in this chapter. These past four studies in the multi-class classification of AD, MCI and CN [43] [44] [45] [46] have been selected based on their relevance to the work in this chapter. The data set used by all four is ADNI and they are relatively recent studies.

The past four papers have provided different performance measurement metrics such as accuracy, sensitivity, specificity and F1-score; some provided these as overall figures for all class labels and some provided these metrics per class in addition to overall figure as shown in Table 7.2. To be able to compare the results in this chapter to previous literature, the reported metrics used in this chapter include both overall accuracy, sensitivity, precision and F1-score, and specificity per class.

To evaluate the performance of multi-class ABA-Com, the four metrics of accuracy, sensitivity, precision and F1-score have been selected. While accuracy is calculated at dataset level, other three metrics are calculated at class level. To get the overall dataset-level results for the three metrics of sensitivity, precision and F1-score, an averaging method can be used to average the values for all classes. The type of averaging methods can be micro and macro averaging. While micro gives different weight to each class based on its size, macro treats all classes the same way. The type of averaging that is similar to calculation of dataset-level overall accuracy is the micro averaging where different classes with different sizes have different effects and importance on the overall average value. Macro averaging is calculated for each class label individually and the class imbalance is

ignored. Selecting micro averaging method is consistent to calculation of accuracy. Therefore, micro averaging is used in this chapter to get the overall dataset-level values for the three metrics of sensitivity, precision and F1-score.

It should be noted that as explained in [47] when micro averaging used, the four overall values of precision, recall (sensitivity), F1-score and accuracy are the same, therefore in this study the value of accuracy, represents all four values.

As shown in Table 7.2, three studies used neural networks [43] [45] [46] and one study used random forest [44] to perform 3-class classification of AD, MCI and CN, while all four studies used cross-validation in their model performance evaluation workflow. Although all four papers used robust methods and algorithms the highest accuracy reported is 61.58% which is 15.72% lower than the accuracy achieved by ABA-Com, 77.3%, while only the two features of age and ABA were used in the classification task. This shows that not only ABA-Com is less complex than neural networks but it can achieve superior performance to them.

As the analysis in past studies may not be identical to what has been performed in this chapter, a baseline SVM method is provided to show the 3-class classification performance of a state-of-the-art algorithm using all brain features and all available data. Table 7.2 shows that SVM has achieved an accuracy of 72.77%, which is 4.53% below the accuracy achieved by ABA-Com. This confirms that the proposed ABA-Com not only had superior performance when compared with relevant previous literature but it has better performance compared to the baseline method.

7.3. Summary

In this chapter the evaluation of the ABA-Com model had been performed when it was applied to multiple dementia types and stages. The results of the 3-class classification have shown that the ABA-Com can achieve better results than the relevant previous studies. Also, the ABA-Com had performed better than the baseline SVM.

Chapter 8

8.Conclusion

One of the main objectives of this research was to create an ML classification model to predict AD using the MRI data while achieving comparable performance to black-box state-of-the-art models. This objective has been achieved as shown in the results section of Chapter 5. The proposed ABA-Com model framework, which is one of the novel contributions of this thesis, used two linear regression models to make a classification of AD vs. CN and achieved better predictive performance than baseline SVM for both female and male subjects.

The next objective was to eliminate the complexity in the black-box models by proposing a linear ML model. The proposed ABA-Com consists of two linear models; ABA-Reg model proposed in Chapter 4 which uses a linear regression model (LASSO) and the ABA-Clf model proposed in Chapter 5 which uses linear logistic regression model to make the final classification. Therefore, by using only linear models in the proposed model the complexity of the model is eliminated and the objective is achieved.

Using linear models throughout the framework of the proposed model has enabled the model to achieve the intrinsic model interpretability by defining a novel feature score s_i outlined in Chapter 6 which is a linear index and links the original feature values to the final model outcome by showing how much each feature affected the prediction outcome. Defining and using of score s_i does not affect the classification performance of the proposed ABA-Com model and it keeps the high performance. The objective of model interpretability has therefore also been achieved.

The ABA feature proposed in Chapter 4, another novel contribution of this thesis, could be used as a biomarker by medical professionals as it is specific to a pathology and gives the biological age of a part of a brain or a small subspace of features in the brain which are highly affected by that pathology. The ABA can therefore give representation of that partial specific brain age to show how different the biological age of those parts is compared to chronological age. The difference, also referred to as ADS is then an indication as to whether those parts of the brain have aged faster and there has been an atrophy in those parts which can indicate the presence of the pathology i.e., AD. Also, in addition to ABA, the medical professionals can use the proposed feature scores to identify which specific brain feature contributed to the value of ABA the most. As the result ABA can be used by medical professionals as a biomarker and indicator of AD.

To train the ABA-Reg model a small subspace of features were used. These features were selected to represent the subspace which have had the highest impact from the pathology. There is therefore a bias introduced in the training of the ABA-

Reg model where the model is not built on all brain features but a very small and specific feature which are indicative of one pathology i.e., AD. This bias deteriorates the regression performance of the model by increasing the MAE and decreasing the Pearson's r. However, this bias is intentionally introduced in the training process not for the purpose of improving the regression task performance, but to maximise the final classification performance of ABA-Clf. Therefore, the bias in the training process has been introduced and it has improved the accuracy of the classification task.

In order to select the features which have had the highest effect form AD, a feature selection model is proposed, referred to as Biased Forward Feature Selection (BFFS), as presented in Chapter 4. BFFS is one of the novel contributions of this thesis. This is an aggressive feature selection method which selects the minimal number of brain features to achieve the maximal classification performance in classifying AD. The ability and the effect of BFFS on the classification task has been assessed in Chapter 5 and it is shown that using of BFFS has a significant improvement on the ABA-Com model classification performance. The objective of creation of a feature selection method has also been achieved.

The logical continuation of the ABA-Com would be to apply the model in a muti-class setting to predict multiple types and stages of dementia. Therefore, in Chapter 7 the ABA-Com model has been applied in a 6-class classification setting with five different dementia types and stages as well as being applied in 3-class and 4-class settings. It is shown in the results of this chapter that the proposed method has achieved better accuracy than the previous literature and the baseline SVM, and

the ABA-Com is successful in being applied in a multi-class classification setting. The objective of multi-class classification while achieving a superior performance has also been achieved.

One of the limitations in this thesis could be related to the preprocessing the MRI scans using FreeSurfer, which is very time-consuming using ordinary computers. Also, FreeSufer segments the brain based on a predefined atlas template of the brain, which can introduce artifacts

Another limitation is the availability of the data. Although the data used in this thesis were downloaded from public repositories around the world the size of data available is very limited. A potential solution would be if public organisations such as the National Health Service (NHS) provided MRI data in anonymised format to be used in research.

Another limitation would be the use of only linear models in the proposed framework in order to achieve the interpretability. It would be good if non-linear models could be used as part of the framework while attempting to retain interpretability.

8.1. Future work

The proposed ABA-Com framework considers morphological brain features from MRI data to detect AD. This prediction is purely based on the biological and morphological characteristics of the brain features and there is no expert knowledge, demographical information and other health biomarkers to reinforce

and improve this framework. Therefore, a potential future work is to combine ABA-Com framework with other biomarkers and information which could improve the prediction.

Also, in Chapter 7 the ABA-Com has been applied as a multi-pathology prediction approach. The accuracy achieved in this chapter was superior to the baseline and relevant other work however, it was significantly lower than the accuracy achieved in binary classification of AD. This could be due to the fact that dementia types and stages used in the multi-class classification task have very similar symptoms especially MCI stages and AD, and they affect similar parts of the brain. As this thesis only focuses on morphological features, distinguishing between similar diseases is a great challenge. Also, the MRI scans used in this thesis are from the initial scans taken from the subject. A possible continuation of this work is not only to incorporate information from other biomarkers and cognitive tests to the model but to perform longitudinal analysis on multiple scans of the same subject at different stages of the disease as the progression of the disease along with expert knowledge can narrow down the symptoms to fewer diseases.

Although some dementia types and stages have been analysed in this thesis the proposed framework could be applied to more dementia and other neurodegenerative disease types. The reason for not applying the proposed model to more diseases is that the data for other diseases are very scarce. A possible progression from this work is therefore to apply the proposed model to other dementia types such as Huntington's and Parkinson’s diseases and also other diseases such as schizophrenia and multiple sclerosis.

In this work the data used were Structural MRI scans where morphological measurements of the brain were then extracted from those scans. A possible addition to this work would be using other types of data such as Functional MRI, CT scans, PET scans, in addition to Structural MRI.

Although the results were compared to previous literature, the findings of this thesis such as the classification results and the selected brain features to predict AD were not presented to the medical professionals and experts. Therefore, a potential future work could involve reviewing the results with a medical experts.

Appendix A

Multi-class classification confusion matrices

In addition to the results provided in Table 7.1, confusion matrices for the four experiments of W1 to W4 which are part of the proposed approach, are provided in this appendix.

Table A.1F | W1 - Female

		Predicted label		
		CN	**MCI**	**AD**
Actual label	**CN**	796 (3)	21 (4)	26 (6)
	MCI	81 (4)	29 (3)	70 (4)
	AD	44 (4)	40 (6)	135 (5)

Table A.1M | W1 - Male

		Predicted label		
		CN	**MCI**	**AD**
Actual label	**CN**	612 (3)	40 (3)	30 (4)
	MCI	104 (3)	106 (5)	68 (6)
	AD	46 (1)	76 (8)	123 (7)

Table A.2F | W2 - Female

		Predicted label			
		CN	**MCI**	**AD**	**FTD**
Actual label	**CN**	786 (4)	23 (6)	26 (2)	8 (2)
	MCI	76 (2)	30 (2)	70 (2)	3 (2)
	AD	42 (2)	35 (6)	137 (5)	5 (1)
	FTD	44 (2)	5 (2)	14 (2)	21 (3)

Table A.2M | W2 - Male

		Predicted label			
		CN	**MCI**	**AD**	**FTD**
Actual label	**CN**	601 (4)	43 (6)	27 (6)	12 (2)
	MCI	104 (5)	103 (7)	66 (7)	6 (1)
	AD	44 (4)	72 (10)	117 (9)	13 (2)
	FTD	45 (3)	5 (2)	27 (5)	37 (6)

Table A.3F | W3 - Female

		Predicted label			
		CN	**MCI**	**AD**	**FTD**
Actual label	**CN**	751 (2)	63 (3)	21 (4)	9 (3)
	MCI	230 (5)	87 (9)	87 (7)	6 (2)
	AD	32 (2)	48 (5)	133 (6)	6 (2)
	FTD	37 (1)	14 (0)	13 (3)	19 (3)

Table A.3M | W3 - Male

		Predicted label			
		CN	**MCI**	**AD**	**FTD**
Actual label	**CN**	518 (6)	136 (7)	20 (3)	8 (2)
	MCI	226 (9)	215 (13)	107 (8)	6 (2)
	AD	24 (2)	95 (6)	118 (5)	8 (2)
	FTD	34 (2)	28 (5)	24 (3)	28 (7)

Table A.4F | W4 - Female

		Predicted label					
		CN	**MCI**	**EMCI**	**LMCI**	**AD**	**FTD**
Actual label	**CN**	782 (4)	21 (5)	3 (2)	3 (2)	25 (5)	9 (2)
	MCI	77 (1)	32 (6)	1 (2)	4 (2)	61 (6)	4 (2)
	EMCI	133 (2)	6 (2)	0 (1)	2 (1)	10 (2)	2 (1)
	LMCI	52 (4)	4 (2)	2 (2)	2 (2)	17 (1)	2 (1)
	AD	40 (4)	34 (4)	2 (0)	11 (3)	125 (2)	6 (1)
	FTD	40 (1)	5 (3)	1 (1)	5 (2)	12 (5)	22 (3)

Table A.4M | W4 - Male

		Predicted label					
		CN	**MCI**	**EMCI**	**LMCI**	**AD**	**FTD**
Actual label	**CN**	581 (5)	39 (3)	19 (6)	7 (3)	24 (3)	11 (2)
	MCI	98 (2)	98 (10)	7 (3)	9 (4)	62 (9)	4 (2)
	EMCI	126 (6)	12 (1)	17 (3)	6 (2)	15 (2)	3 (2)
	LMCI	49 (2)	9 (1)	10 (2)	3 (1)	23 (3)	3 (2)
	AD	35 (3)	65 (8)	10 (3)	20 (6)	109 (9)	7 (3)
	FTD	41 (3)	3 (1)	13 (5)	6 (2)	23 (2)	28 (5)

Appendix B

Selected hyperparameters

In this thesis hyperparameter tuning for was not performed for either SVM and LASSO algorithms. This was due to limitation of time during the research for this thesis and also a fair comparison between the baseline experiment using SVM and the proposed framework using LASSO. The value for C in SVM and lambda in LASSO were selected as 1.

Appendix C

Interpretability index - further examples

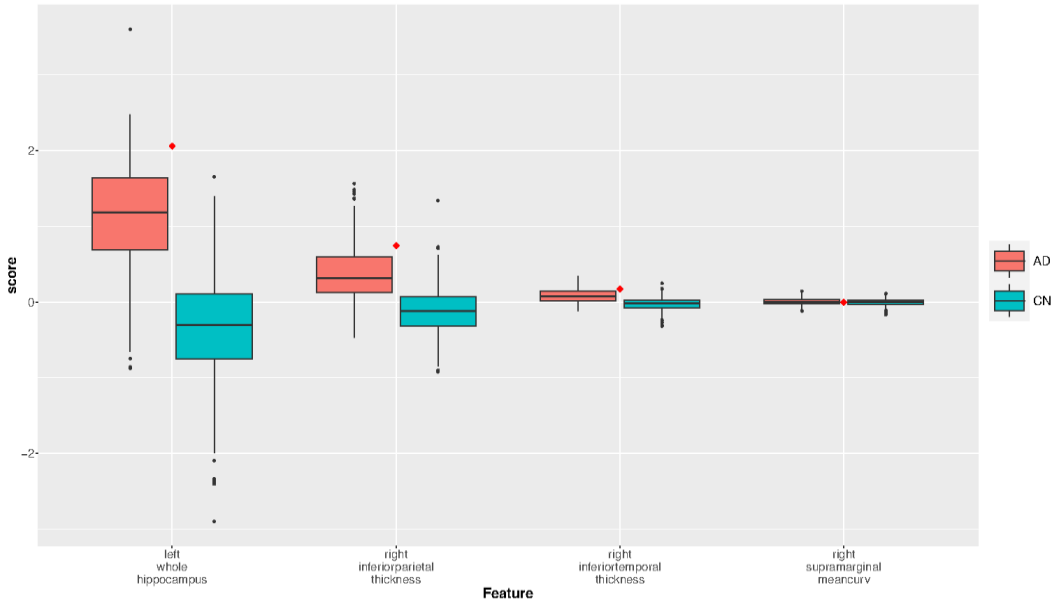

Figure 6.4 | Distribution of the feature scores for a further case of True Positive (TP), selected from test set of hold-out method are plotted using red dots. The box plots represent feature scores of all subjects in the training set of hold-out method. Female subjects data used.

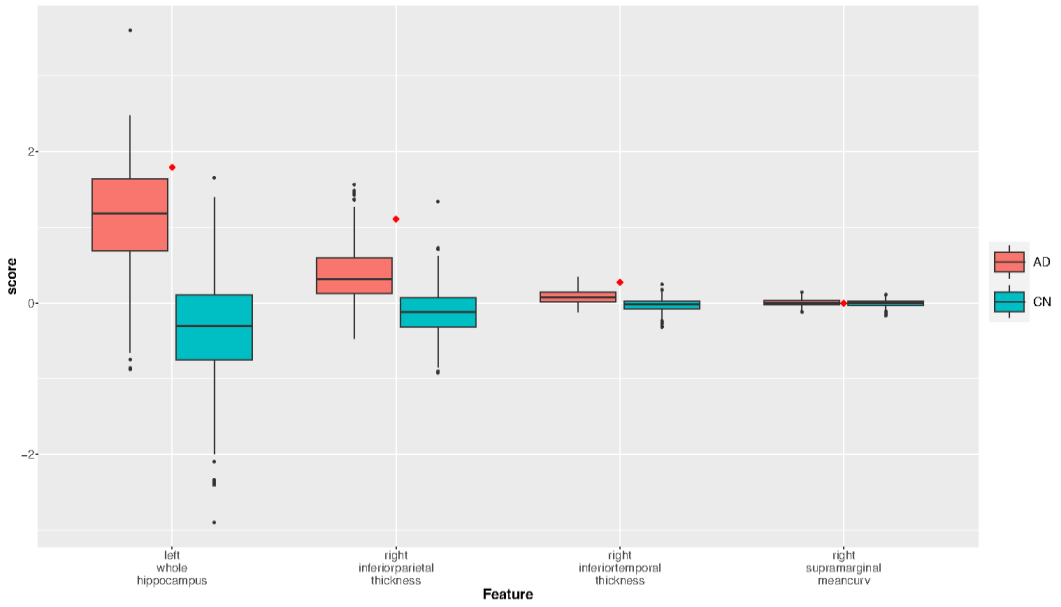

Figure 6.5 | Distribution of the feature scores for a further case of True Positive (TP), selected from test set of hold-out method are plotted using red dots. The box plots represent feature scores of all subjects in the training set of hold-out method. Female subjects data used.

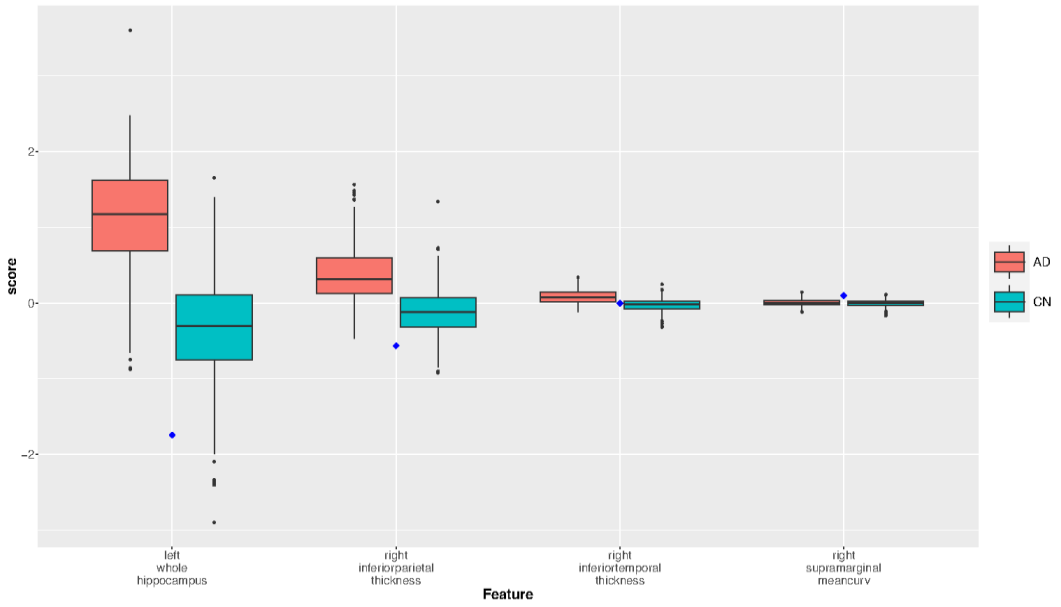

Figure 6.6 | Distribution of the feature scores for a further case of True Negative (TN), selected from test set of hold-out method are plotted using blue dots. The box plots represent feature scores of all subjects in the training set of hold-out method. Female subjects data used.

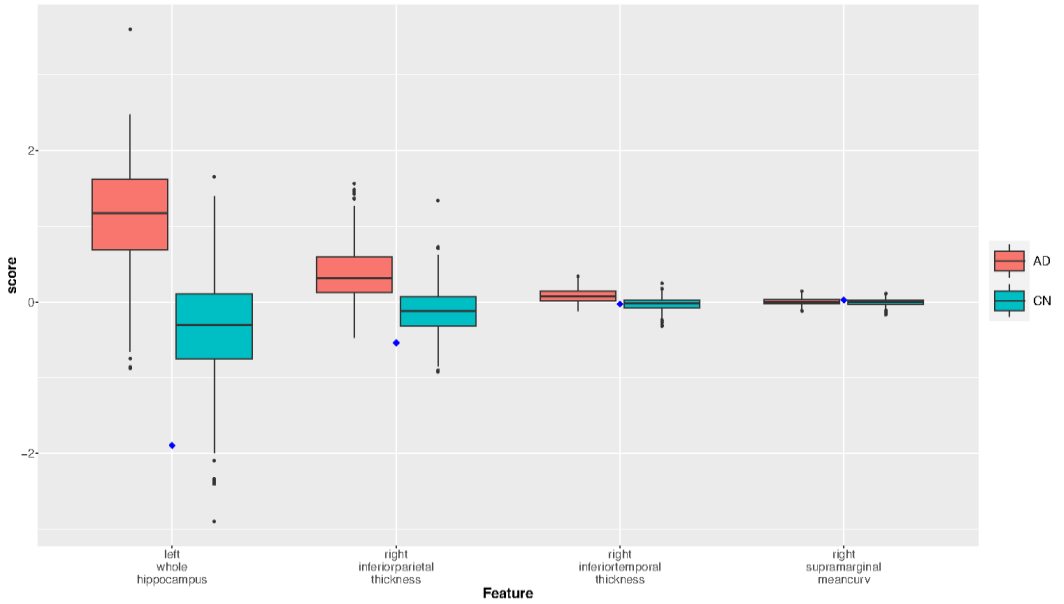

Figure 6.7 | Distribution of the feature scores for a further case of True Negative (TN), selected from test set of hold-out method are plotted using blue dots. The box plots represent feature scores of all subjects in the training set of hold-out method. Female subjects data used.

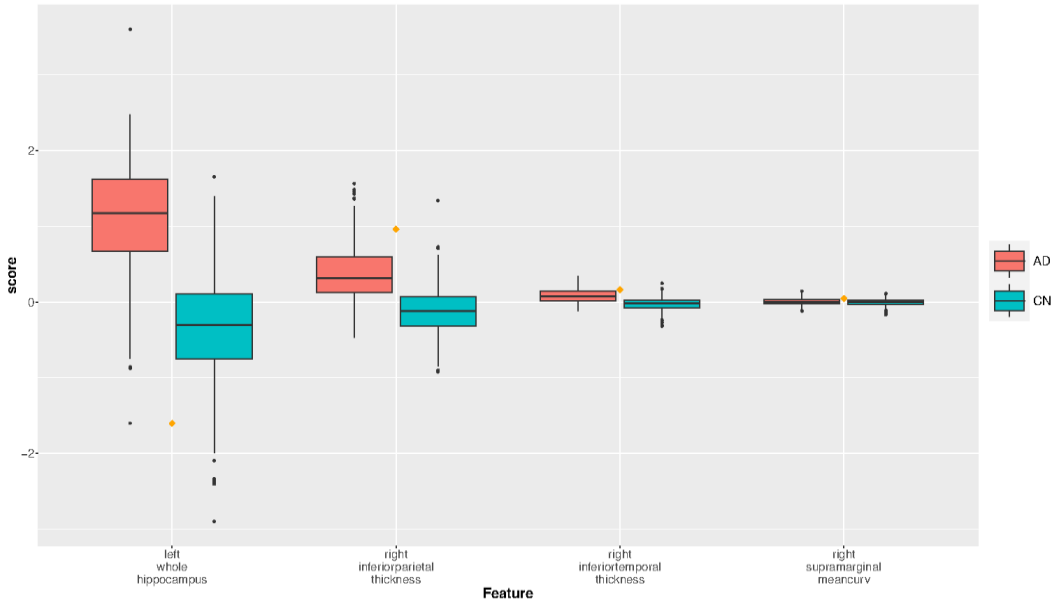

Figure 6.8 | Distribution of the feature scores for a further case of False Positive (FP), selected from test set of hold-out method are plotted using orange dots. The box plots represent feature scores of all subjects in the training set of hold-out method. Female subjects data used.

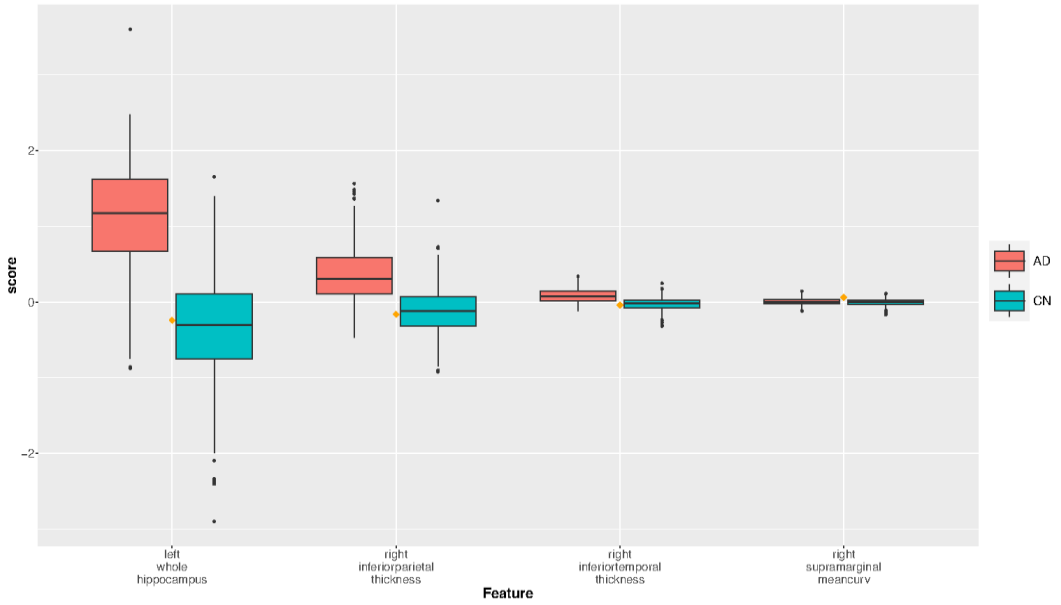

Figure 6.9 | Distribution of the feature scores for a further case of False Positive (FP), selected from test set of hold-out method are plotted using orange dots. The box plots represent feature scores of all subjects in the training set of hold-out method. Female subjects data used.

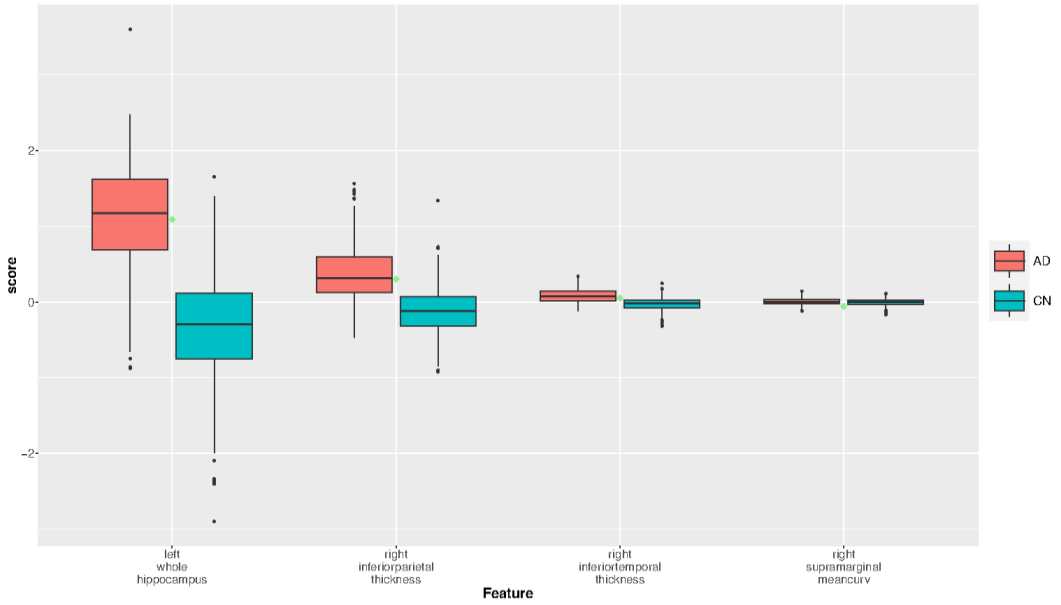

Figure 6.10 | Distribution of the feature scores for a further case of False Negative (FN), selected from test set of hold-out method are plotted using green dots. The box plots represent feature scores of all subjects in the training set of hold-out method. Female subjects data used.

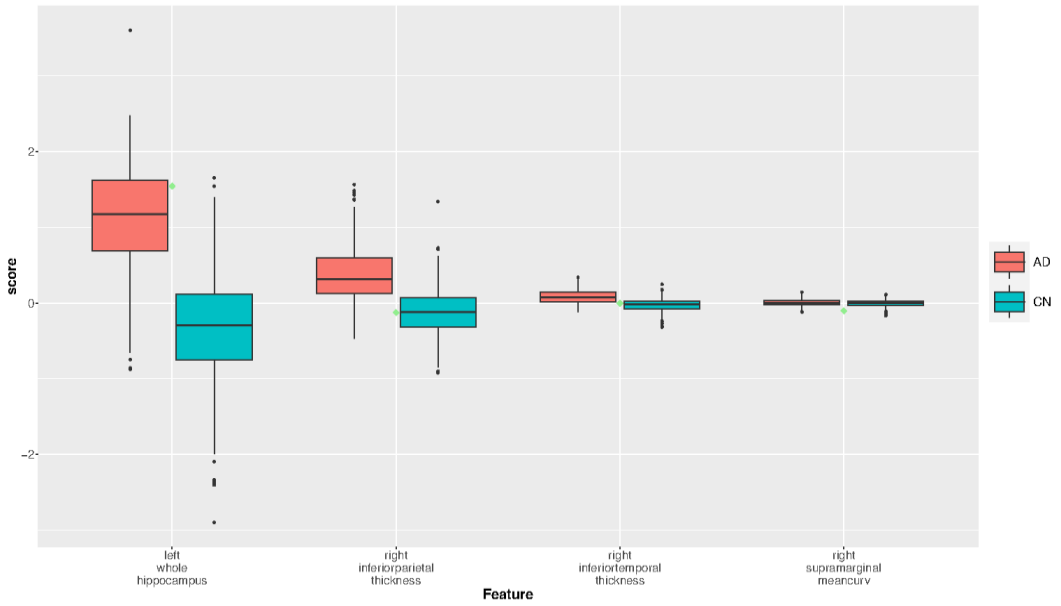

Figure 6.11 | Distribution of the feature scores for a further case of False Negative (FN), selected from test set of hold-out method are plotted using green dots. The box plots represent feature scores of all subjects in the training set of hold-out method. Female subjects data used.

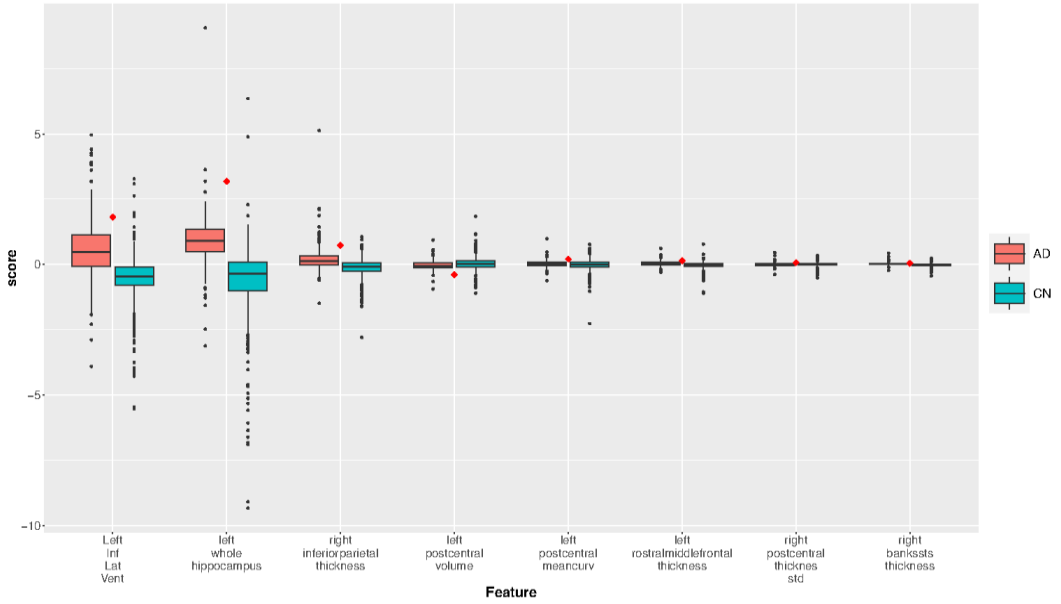

Figure 6.12 | Distribution of the feature scores for a further case of True Positive (TP), selected from test set of hold-out method are plotted using red dots. The box plots represent feature scores of all subjects in the training set of hold-out method. Female subjects data used.

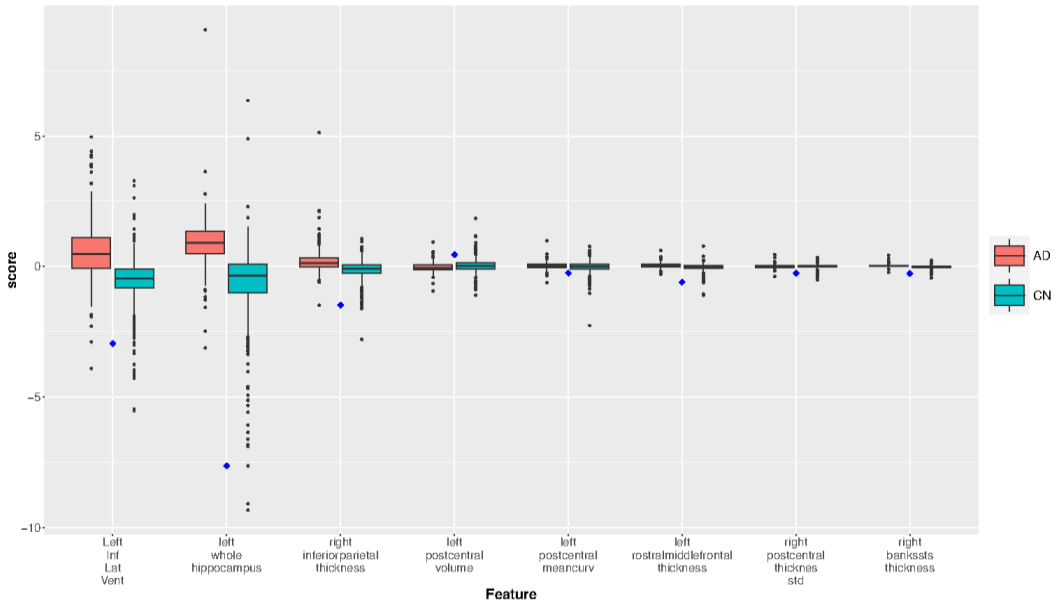

Figure 6.13 | Distribution of the feature scores for a further case of True Negative (TN), selected from test set of hold-out method are plotted using red dots. The box plots represent feature scores of all subjects in the training set of hold-out method. Female subjects data used.

Bibliography

1 Burns, A., & Iliffe, S. (2009). Alzheimer's disease. In BMJ (Vol. 338, Issue feb05 1, pp. b158–b158). BMJ. https://doi.org/10.1136/bmj.b158

2 Blennow, K., de Leon, M. J., & Zetterberg, H. (2006). Alzheimer's disease. In The Lancet (Vol. 368, Issue 9533, pp. 387–403). Elsevier BV. https://doi.org/10.1016/s0140-6736(06)69113-7

3 Duong, S., Patel, T., & Chang, F. (2017). Dementia. In Canadian Pharmacists Journal / Revue des Pharmaciens du Canada (Vol. 150, Issue 2, pp. 118–129). SAGE Publications. https://doi.org/10.1177/1715163517690745

4 Rasmussen, J., & Langerman, H. (2019). Alzheimer's Disease – Why We Need Early Diagnosis; In Degenerative Neurological and Neuromuscular Disease: Vol. Volume 9 (pp. 123–130). Informa UK Limited. https://doi.org/10.2147/dnnd.s228939

5 Dubois, B., Padovani, A., Scheltens, P., Rossi, A., & Dell'Agnello, G. (2015). Timely Diagnosis for Alzheimer's Disease: A Literature Review on Benefits and Challenges. In A. Saykin (Ed.), Journal of Alzheimer's Disease (Vol. 49, Issue 3, pp. 617–631). IOS Press. https://doi.org/10.3233/jad-150692

6 Franke, K., Ziegler, G., Klöppel, S., & Gaser, C. (2010). Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. In NeuroImage (Vol. 50, Issue 3, pp. 883–892). Elsevier BV. https://doi.org/10.1016/j.neuroimage.2010.01.005

7 Varzandian, A., Razo, M. A. S., Sanders, M. R., Atmakuru, A., & Di Fatta, G. (2021). Classification-Biased Apparent Brain Age for the Prediction of Alzheimer's Disease. In Frontiers in Neuroscience (Vol. 15). Frontiers Media SA. https://doi.org/10.3389/fnins.2021.673120

8 Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., & Dale, A. M. (2002). Whole Brain Segmentation. In Neuron (Vol. 33, Issue 3, pp. 341–355). Elsevier BV. https://doi.org/10.1016/s0896-6273(02)00569-x

9 Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. In 2008 Eighth IEEE International Conference on Data Mining. 2008 Eighth IEEE International Conference on Data Mining (ICDM). IEEE. https://doi.org/10.1109/icdm.2008.17

10 Salgado, C. M., Azevedo, C., Proença, H., & Vieira, S. M. (2016). Noise Versus Outliers. In Secondary Analysis of Electronic Health Records (pp. 163–183). Springer International Publishing. https://doi.org/10.1007/978-3-319-43742-2_14

11 Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data - SIGMOD '00. the 2000 ACM SIGMOD international conference. ACM Press. https://doi.org/10.1145/342009.335388

12 Kriegel, H.-P., S hubert, M., & Zimek, A. (2008). Angle-based outlier detection in high-dimensional data. In Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08. the 14th ACM SIGKDD international conference. ACM Press. https://doi.org/10.1145/1401890.1401946
13 Domingues, R., Filippone, M., Michiardi, P., & Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms: Experiments and analyses. In Pattern Recognition (Vol. 74, pp. 406–421). Elsevier BV. https://doi.org/10.1016/j.patcog.2017.09.037
14 Knuth, D. E. (1997). The art of computer programming (Reading, Mass: Addison-Wesley), 3rd ed.
15 Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
16 Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer series in statistics (Springer)
17 Mirzaei, G., & Adeli, H. (2022). Machine learning techniques for diagnosis of alzheimer disease, mild cognitive disorder, and other types of dementia. In Biomedical Signal Processing and Control (Vol. 72, p. 103293). Elsevier BV. https://doi.org/10.1016/j.bspc.2021.103293
18 Dinov, I. D. (2018). Black Box Machine-Learning Methods: Neural Networks and Support Vector Machines. In Data Science and Predictive Analytics (pp. 383–422). Springer International Publishing. https://doi.org/10.1007/978-3-319-72347-1_11
19 Duchesne, S., Caroli, A., Geroldi, C., Barillot, C., Frisoni, G. B., & Collins, D. L. (2008). MRI-Based Automated Computer Classification of Probable AD Versus Normal Controls. In IEEE Transactions on Medical Imaging (Vol. 27, Issue 4, pp. 509–520). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/tmi.2007.908685
20 Plant, C., Teipel, S. J., Oswald, A., Böhm, C., Meindl, T., Mourao-Miranda, J., Bokde, A. W., Hampel, H., & Ewers, M. (2010). Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease. In NeuroImage (Vol. 50, Issue 1, pp. 162–174). Elsevier BV. https://doi.org/10.1016/j.neuroimage.2009.11.046
21 Long, S. S., & Holder, L. B. (2012). Graph based MRI brain scan classification and correlation discovery. In 2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). 2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). IEEE. https://doi.org/10.1109/cibcb.2012.6217249
22 Herzog, N. J., & Magoulas, G. D. (2021). Brain Asymmetry Detection and Machine Learning Classification for Diagnosis of Early Dementia. In Sensors (Vol. 21, Issue 3, p. 778). MDPI AG. https://doi.org/10.3390/s21030778
23 Beheshti, I., Demirel, H., & Matsuda, H. (2017). Classification of Alzheimer's disease and prediction of mild cognitive impairment-to-Alzheimer's conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm. In Computers in Biology and Medicine (Vol. 83, pp. 109–119). Elsevier BV. https://doi.org/10.1016/j.compbiomed.2017.02.011
24 Meng, X., Wu, Y., Liu, W., Wang, Y., Xu, Z., & Jiao, Z. (2022). Research on Voxel-Based Features Detection and Analysis of Alzheimer's Disease Using Random Survey Support Vector Machine. In Frontiers in Neuroinformatics (Vol. 16). Frontiers Media SA. https://doi.org/10.3389/fninf.2022.856295
25 Jing Wan, Zhilin Zhang, Jingwen Yan, Taiyong Li, Rao, B. D., Shiaofen Fang, Sungeun Kim, Risacher, S. L., Saykin, A. J., & Li Shen. (2012). Sparse Bayesian

multi-task learning for predicting cognitive outcomes from neuroimaging measures in Alzheimer's disease. In 2012 IEEE Conference on Computer Vision and Pattern Recognition. 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. https://doi.org/10.1109/cvpr.2012.6247769

26 Nho K, et al. (2010). Automatic prediction of conversion from mild cognitive impairment to probable Alzheimer's disease using structural magnetic resonance imaging. *AMIA Annual Symposium Proceedings*, 2010, 542–546. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041374/

27 Kim, J., & Lee, B. (2017). Automated discrimination of dementia spectrum disorders using extreme learning machine and structural T1 MRI features. In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE. https://doi.org/10.1109/embc.2017.8037241

28 SPM 8, 2009. Wellcome Trust Centre for Neuroimaging. Institute of Neurology, UCL, London, UK. http://www.fil.ion.ucl.ac.uk/spm/

29 Franke, K., & Gaser, C. (2019). Ten Years of BrainAGE as a Neuroimaging Biomarker of Brain Aging: What Insights Have We Gained? In Frontiers in Neurology (Vol. 10). Frontiers Media SA. https://doi.org/10.3389/fneur.2019.00789

30 Franke, K., & Gaser, C. (2014). "Dementia classification based on brain age estimation," in Proc MICCAI Workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data (Boston, MA), 48–54.

31 Beheshti, I., Mishra, S., Sone, D., Khanna, P., & Matsuda, H. (2020). T1-weighted MRI-driven Brain Age Estimation in Alzheimer's Disease and Parkinson's Disease. In Aging and disease (Vol. 11, Issue 3, p. 618). Aging and Disease. https://doi.org/10.14336/ad.2019.0617

32 Bashyam, V. M., Erus, G., Doshi, J., Habes, M., Nasrallah, I. M., Truelove-Hill, M., Srinivasan, D., Mamourian, L., Pomponio, R., Fan, Y., Launer, L. J., Masters, C. L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S. C., Fripp, J., Koutsouleris, N., Satterthwaite, T. D., Davatzikos, C. (2020). MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. In Brain (Vol. 143, Issue 7, pp. 2312–2324). Oxford University Press (OUP). https://doi.org/10.1093/brain/awaa160

33 Kloppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., Fox, N. C., Jack, C. R., Ashburner, J., & Frackowiak, R. S. J. (2008). Automatic classification of MR scans in Alzheimer's disease. In Brain (Vol. 131, Issue 3, pp. 681–689). Oxford University Press (OUP). https://doi.org/10.1093/brain/awm319

34 Schröder, J., & Pantel, J. (2016). Neuroimaging of hippocampal atrophy in early recognition of Alzheimer´s disease – a critical appraisal after two decades of research. In Psychiatry Research: Neuroimaging (Vol. 247, pp. 71–78). Elsevier BV. https://doi.org/10.1016/j.pscychresns.2015.08.014

35 Poulin, S. P., Dautoff, R., Morris, J. C., Barrett, L. F., & Dickerson, B. C. (2011). Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. In Psychiatry Research: Neuroimaging (Vol. 194, Issue 1, pp. 7–13). Elsevier BV. https://doi.org/10.1016/j.pscychresns.2011.06.014

36 Velayudhan, L., Proitsi, P., Westman, E., Muehlboeck, J.-S., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Soininen, H., Spenger, C., Hodges, A., Powell, J., Lovestone, S., & Simmons, A. (2013). Entorhinal Cortex Thickness Predicts Cognitive Decline in Alzheimer's Disease. In Journal of Alzheimer's Disease (Vol. 33, Issue 3, pp. 755–766). IOS Press. https://doi.org/10.3233/jad-2012-121408

37 Wang, L., Goldstein, F. C., Veledar, E., Levey, A. I., Lah, J. J., Meltzer, C. C., Holder, C. A., & Mao, H. (2009). Alterations in Cortical Thickness and White Matter Integrity in Mild Cognitive Impairment Measured by Whole-Brain Cortical Thickness Mapping and Diffusion Tensor Imaging. In American Journal of Neuroradiology (Vol. 30, Issue 5, pp. 893–899). American Society of Neuroradiology (ASNR). https://doi.org/10.3174/ajnr.a1484
38 Altaf, T., Anwar, S. M., Gul, N., Majeed, M. N., & Majid, M. (2018). Multi-class Alzheimer's disease classification using image and clinical features. In Biomedical Signal Processing and Control (Vol. 43, pp. 64–74). Elsevier BV. https://doi.org/10.1016/j.bspc.2018.02.019
39 Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. In Nature Machine Intelligence (Vol. 1, Issue 5, pp. 206–215). Springer Science and Business Media LLC. https://doi.org/10.1038/s42256-019-0048-x
40 Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1705.07874
41 Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why Should I Trust You?': Explaining the Predictions of Any Classifier (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1602.04938
42 Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE. https://doi.org/10.1109/dsaa.2018.00018
43 Lin, W., Gao, Q., Du, M., Chen, W., & Tong, T. (2021). Multiclass diagnosis of stages of Alzheimer's disease using linear discriminant analysis scoring for multimodal data. In Computers in Biology and Medicine (Vol. 134, p. 104478). Elsevier BV. https://doi.org/10.1016/j.compbiomed.2021.104478
44 Tong, T., Gray, K., Gao, Q., Chen, L., & Rueckert, D. (2017). Multi-modal classification of Alzheimer's disease using nonlinear graph fusion. In Pattern Recognition (Vol. 63, pp. 171–181). Elsevier BV. https://doi.org/10.1016/j.patcog.2016.10.009
45 Lama, R. K., Gwak, J., Park, J.-S., & Lee, S.-W. (2017). Diagnosis of Alzheimer's Disease Based on Structural MRI Images Using a Regularized Extreme Learning Machine and PCA Features. In Journal of Healthcare Engineering (Vol. 2017, pp. 1–11). Hindawi Limited. https://doi.org/10.1155/2017/5485080
46 Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., Feng, D., Fulham, M. J., & ADNI. (2015). Multimodal Neuroimaging Feature Learning for Multiclass Diagnosis of Alzheimer's Disease. In IEEE Transactions on Biomedical Engineering (Vol. 62, Issue 4, pp. 1132–1140). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/tbme.2014.2372011
47 Grandini, M., Bagli, E. & Visani, G. (2020), 'Metrics for multi-class classification: an overview'. https://arxiv.org/abs/2008.05756