DISSERTATION

MACHINE LEARNING AND DEEP LEARNING APPLICATIONS IN NEUROIMAGING FOR BRAIN AGE PREDICTION

Submitted by

Fereydoon Vafaei

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2023

Doctoral Committee:

Advisor: Charles Anderson

Michael Kirby
Nathaniel Blanchard
Agnieszka Burzynska

ABSTRACT

MACHINE LEARNING AND DEEP LEARNING APPLICATIONS IN NEUROIMAGING FOR BRAIN AGE PREDICTION

Machine Learning (ML) and Deep Learning (DL) are now considered as state-of-the-art assistive AI technologies that help neuroscientists, neurologists and medical professionals with early diagnosis of neurodegenerative diseases and cognitive decline as a consequence of unhealthy brain aging. Brain Age Prediction (BAP) is the process of estimating a person's biological age using Neuroimaging data, and the difference between the predicted age and the subject's chronological age, known as Delta, is regarded as a biomarker for healthy versus unhealthy brain aging. Accurate and efficient BAP is an important research topic, and hence ML/DL methods have been developed for this task.

There are different modalities of Neuroimaging such as Magnetic Resonance Imaging (MRI) that have been used for BAP in the past. Diffusion Tensor Imaging (DTI) is an advanced quantitative Neuroimaging technology that gives insight into microstructure of White Matter tracts that connect different parts of the brain to function properly. DTI data is high-dimensional, and age-related microstructural changes in White Matter include non-linear patterns.

In this study, we perform a series of analytical experiments using ML and DL methods to investigate the applicability of DTI data for BAP. We also investigate which Diffusivity Parameters, which are DTI metrics that reflect direction and magnitude of diffusion of water molecules in the brain, are relevant for BAP as a Supervised Learning task.

Moreover, we propose, implement, and analyze a novel methodology that can detect age-related anomalies (high Deltas), and can overcome some of the major and fundamental limitations of the current supervised approach for BAP, such as "Chronological Age Label Inconsistency". Our proposed methodology, which combines Unsupervised Anomaly Detection (UAD) and super-

vised BAP, focuses on addressing a fundamental challenge in BAP which is how to interpret a model's error. Should a researcher interpret a model's error as an indication of unhealthy brain aging or the model's poor performance that should be eliminated? We argue that the underlying cause of this problem is the inconsistency of chronological age labels as the ground truth of the Supervised Learning task, which is the common basis of training ML/DL models. Our Unsupervised Learning methods and findings open a new possibility to detect irregularities and abnormalities in the aging brain using DTI scans, independent of inconsistent chronological age labels. The results of our proposed methodology show that combining label-independent UAD and supervised BAP provides a more reliable and methodical way for error analysis than the current supervised BAP approach when it is used in isolation.

We also provide visualization and explanations on how our ML/DL methods make their decisions for BAP. Explainability and generalization of our ML/DL models are two important aspects of our study.

ACKNOWLEDGEMENTS

I would like to thank my PhD advisor Professor Charles Anderson, whose excellent supervision and fabulous support have made this exciting academic journey very pleasant and joyful. Professor Anderson encouraged me to explore and pursue my interdisciplinary research interests during my PhD program which created a lot of learning opportunities for me, and this dissertation is the result of that exploration. I cannot be more grateful for Professor Anderson's exceptional mentorship and caring supervision.

I would like to thank my graduate committee members, Professor Agnieszka Burzynska, Professor Michael Kirby, and Professor Nathaniel Blanchard.

I would like to thank the researchers and professors at University of Illinois Urbana-Champaign who shared their datasets from the clinical study (identifier NCT01472744), supported by the National Institute on Aging at the National Institutes of Health (R37 AG025667), funding from Abbott Nutrition through the Center for Nutrition, Learning, and Memory at the University of Illinois (PIs Arthur Kramer and Edward McAuley), Lifelong Brain and Cognition Lab at the Beckman Institute, UIUC, Agnieszka Burzynska, Gillian Cooke, Michelle Voss, Jason Fanning, Neha Gothe, Elizabeth Salerno, and a team of research assistants managed by Anya Knecht.

I would like to thank Professor Bruce Draper as I learned a lot from him through his inspiring discussions. I would like to thank Professor Don Rojas for his excellent research seminar in Neuroscience which opened new doors to my research. I would like to thank Professor Andrew Ng for his wonderful courses on Deep Learning and Machine Learning.

I would like to thank Dr. Ehsan Adeli for his research suggestions and comments which helped me a lot in my interdisciplinary research. I would like to thank Dr. David Turner for his insightful comments about Machine Learning and Data Science, and for his helpful career advice.

Finally, I would like to thank my family, my wife, Maryam, and my daughter, Pantea, for their everlasting love and support.

DEDICATION

This dissertation is dedicated to my lovely wife, Maryam, and my intelligent, hard-working, and beautiful daughter, Pantea

TABLE OF CONTENTS

ABSTRACT ii
ACKNOWLEDGEMENTS iv
DEDICATION v
LIST OF TABLES viii
LIST OF FIGURES ix

Chapter 1 Introduction 1
1.1 Brain Age Prediction (BAP): What, Why, How? 1
1.1.1 What Is BAP Using Neuroimaging? 2
1.1.2 Why Does BAP Matter? 2
1.1.3 How Did We Approach BAP Using DTI and ML/DL? 2
1.2 Research Questions of This Dissertation 3
1.2.1 Research Question-1 3
1.2.2 Research Question-2 3
1.2.3 Research Question-3 4
1.2.4 Research Question-4 4
1.2.5 Research Question-5 4
1.3 Challenges 5
1.4 Contributions 6
1.5 Outline of The Following Chapters 7

Chapter 2 Background 8
2.1 Diffusion Tensor Imaging (DTI) 8
2.1.1 Diffusivity Parameters 10
2.2 ML/DL Applications in Brain Age Prediction (BAP) 12
2.3 Limitations of Current Approaches of BAP 15

Chapter 3 Methods 18
3.1 DTI Data 18
3.1.1 Data Acquisition 19
3.1.2 Feature Space and Tensors 19
3.1.3 Target (Dependent) Variable 21
3.1.4 Normalization and Standardization of DTI Data 23
3.1.5 Dimensionality Reduction 24
3.2 Baseline Models 25
3.2.1 Linear and Non-Linear Models 26
3.2.2 Ensemble Methods 28
3.2.3 Model Evaluation Process 32
3.3 Unsupervised Learning Methods 35
3.3.1 Gaussian Mixture Model (GMM) 36
3.3.2 Clustering 38

3.3.3 Unsupervised Anomaly Detection (UAD) . . . 38
3.3.4 Our Proposed Methodology to Combine UAD and BAP . . . 39
3.4 Deep Learning Methods . . . 41
3.4.1 DNN . . . 42
3.4.2 CNN . . . 42
3.4.3 Data Augmentation and Transfer Learning . . . 43

Chapter 4 Results . . . 45
4.1 Benchmark Results of Baseline Regression Models . . . 46
4.1.1 Impacts of Scaling . . . 54
4.1.2 Impacts of Dimensionality Reduction . . . 57
4.2 Feature Selection: Which Diffusivity Parameter is Better? . . . 58
4.3 Age-Group Analysis and Systematic Bias . . . 62
4.4 Unsupervised Learning Results . . . 66
4.4.1 2D Projection of Feature Space . . . 66
4.4.2 Results of Clustering . . . 69
4.4.3 Results of UAD and Our Proposed Methodology . . . 74
4.4.4 Generative Models and Data Generation . . . 77
4.5 Deep Learning Results . . . 77
4.5.1 Results of Data Augmentation and Transfer Learning . . . 83
4.6 Brain Maps and Explainability . . . 84

Chapter 5 Conclusions . . . 91
5.1 Summary of Findings . . . 91
5.2 Limitations of Our Study . . . 96
5.3 Future Work . . . 96

Bibliography . . . 98

Appendix A Hardware and Software Specifications . . . 106
A.1 Hardware Specifications . . . 106
A.2 Software Specifications . . . 106
A.3 Hyperparameter Tuning . . . 107
A.3.1 Random Forest . . . 107
A.3.2 XGBoost . . . 108

Appendix B List of Acronyms . . . 110

LIST OF TABLES

4.1 Benchmark Results of Baseline Models - Dataset-1 48
4.2 Benchmark Results of Baseline Models - Dataset-2 48
4.3 Combination Analysis - Single Parameters 60
4.4 Combination Analysis - Two Parameters 60
4.5 Combination Analysis - Three and Four Parameters 60
4.6 Results of Deep Learning Models 83

A.1 Software Specifications 106

LIST OF FIGURES

2.1 Isotropic vs Anisotropic . . . 10
2.2 Skeletonized FA . . . 11
2.3 Biological Aging vs. Chronological Aging . . . 13
2.4 Supervised BAP Process . . . 14

3.1 Distribution of Diffusivity Parameters Values . . . 20
3.2 Age Histograms . . . 21
3.3 Ratio of Age Groups . . . 22
3.4 Ratio of Genders . . . 22
3.5 Impact of Number of Principal Components on MAE . . . 35
3.6 Unsupervised Anomaly Detection . . . 40
3.7 Feed Forward Neural Network . . . 42

4.1 MAE Variations . . . 50
4.2 Baseline BAP Plots . . . 52
4.3 Distribution of Diffusivity Parameters Values after PCA . . . 56
4.4 Explained Variance Ratio Plots . . . 59
4.5 BAP by AD+RD for Age > 40 . . . 65
4.6 2D Projection of Feature Space . . . 68
4.7 Unsupervised Learning - Clustering-1 . . . 70
4.8 Unsupervised Learning - Clustering-2 . . . 71
4.9 Intra-Cluster Age Distribution . . . 73
4.10 Age Distribution of Generated Samples . . . 78
4.11 Architecture of DNN . . . 80
4.12 Architecture of 3D-CNN VGG-1 . . . 81
4.13 Architecture of 3D-CNN VGG-2 . . . 82
4.14 Brain Map of Feature Importance of Random Forests . . . 84
4.15 Neural Network Brain Maps on AD and RD . . . 87
4.16 Neural Network Brain Map on RD (Zoomed-In Coronal-View) . . . 88
4.17 Reconstruction of FA Scans . . . 89

Chapter 1

Introduction

Brain Age Prediction (BAP) is the process of estimating the biological age of a subject's brain using Machine Learning (ML) and Deep Learning (DL) applied on Neuroimaging data [1, 2]. The difference between the predicted age and the subject's chronological age is defined as "**Delta**", and is considered as a biomarker associated with unhealthy brain aging, which in turn is correlated with neurodegenerative diseases and disorders as well as cognitive decline [3].

Accurate and efficient BAP is of utmost importance for medical and clinical diagnosis and intervention, and hence Machine Learning (ML) and Deep Learning (DL) methods have been developed as an assistive AI technology for BAP because of their success in other neurological applications and modalities of Neuroimaging that require processing and analyzing high-dimensional complex data with subtle and non-linear patterns [4–6]. Different Neuroimaging modalities have been used in the past, but Magnetic Resonance Imaging (MRI) is the most commonly used structural modality for BAP [1].

Diffusion Tensor Imaging (DTI) is an advanced Neuroimaging technology that concentrates on the microstructural characteristics of the White Matter (WM) tracts based on the directions of diffusion of water molecules in the brain [7–10]. DTI has attracted the interest of the research community for a variety of neurological applications [11, 12]. However, to date, few studies have used DTI directly as a single modality for BAP [13, 14].

In this dissertation, we will perform a series of analytical experiments to investigate the relevance of DTI for BAP using ML/DL methods.

1.1 Brain Age Prediction (BAP): What, Why, How?

We include this section with busy and quick reader in mind. By reading this section, the reader should clearly and quickly recognize the answers to the following major questions of this

dissertation: What? Why? How? —What is the problem statement? Why does it matter? How did we approach the problem?

We further address "What" and "Why" questions in Chapter 2. We fully explain "How" in our methodology in Chapter 3.

1.1.1 What Is BAP Using Neuroimaging?

By applying ML/DL methods on Neuroimaging data, we can predict a subject's brain biological age which may differ from their chronological age. Neuroimaging gives researchers a lot of information about the brain, how it functions, and its structural and microstructural characteristics. ML algorithms can model trajectories of healthy brain aging and identify abnormalities of the aging process.

1.1.2 Why Does BAP Matter?

Most importantly, positive Delta has been associated to cognitive decline and different aspects of aging development and its effects on human body and brain specifically. Thus, the result of BAP can give medical practitioners a methodical way to predict the risk of developing neurodegenerative diseases and mortality in older subjects [3].

It is likely that BAP will soon (if not already under way) become a common clinical routine as an extension to the existing procedures and health protocols for checking patients' vital information regularly and routinely to assess the risk of developing neurological disorders. Hence, BAP has grown to an important research topic.

1.1.3 How Did We Approach BAP Using DTI and ML/DL?

We apply different methods and approaches of ML and DL on DTI Diffusivity Parameters, which are metrics that reflect the direction of diffusion of water molecules in the brain. From a high-level perspective, we use two main approaches of Machine Learning: Supervised Learning, defining BAP as a regression task with age as the target variable and using chronological age labels as the ground truth, and Unsupervised Learning: 2D projection of DTI feature space, clustering of

DTI data to recognize potential age differences across samples, and anomaly detection for identifying abnormal scans and irregularities.

1.2 Research Questions of This Dissertation

Now that we have briefly explained the problem statement, its significance and relevance as well as our approach to the problem, we define the following five major research questions as the center of our concentration in this dissertation.

1.2.1 Research Question-1

"Can we use DTI as a single modality of Neuroimaging for Brain Age Prediction (BAP), and not combined with structural MRI or in other multi-modal studies? How competitive are the results of DTI analysis in comparison with other modalities of Neuroimaging for Brain Age Prediction?"

Multi-modal studies are quite complex and computationally expensive. The structure of multi-modal Neuroimaging data is extremely high-dimensional and subtle. Thus, choosing the most relevant Neuroimaging modality that has the highest predictive power for BAP is an important task. Moreover, White Matter may offer some insights to the brain microstructures that other brain properties cannot. Hence, we are going to investigate this question and we directly and completely focus on DTI data analysis for BAP.

1.2.2 Research Question-2

"Is there any significant difference among the four Diffusivity Parameters (FA, AD, MD, RD) of DTI in terms of their contribution to Brain Age Prediction (BAP)? Is a single parameter or any specific combination of parameters a better representation (encoding) of information for prediction? In other words, do diffusivity parameters differ in terms of their predictive power specifically for brain age prediction?"

We will define and thoroughly explain Diffusivity Parameters in Section 2.1.1. Briefly, they are quantitative measures computed directly from DTI data to capture the direction and magnitude of

diffusion of water molecules in the brain, which contain a lot of information about microstructural characteristics of White Matter (WM) properties and their correlations with age [2, 3].

1.2.3 Research Question-3

*"Should a researcher interpret ML/DL model's error in BAP as **Delta**, i.e., an indication of an actual unhealthy brain aging, or the model's poor performance, and how can one distinguish the two possibilities methodically?"*

We will identify and explore major challenges and limitations in the current approaches for BAP. We will explain why, despite recent successes and promising state-of-the-art results, the two challenges "**Chronological Age Label Inconsistency**" and "**Systematic Bias**" remain important for the BAP research community [15, 16], and how our proposed methods help resolving them.

1.2.4 Research Question-4

"How do deep learning models, such as Convolutional Neural Network (CNN), perform on DTI small-sized datasets (298 and 94 subjects) for brain age prediction, and how do the results compare to other non-DL approaches? Can we use deep learning approaches that are used in computer vision such as data augmentation and transfer learning for DTI analysis?"

DTI data is computationally a different data structure compared to structural MRI and other modalities. We will argue that the applicability of DL methods for DTI analysis is limited, especially when they are applied on small datasets.

1.2.5 Research Question-5

*"Can we "**explain**" and "**interpret**" how our models make predictions, and identify the brain regions as biomarkers to which our models are most sensitive for brain age prediction?"*

Explainability has become a very important research topic in different applications of Machine Learning and Deep Learning. We will discuss our approach in detail on how to make ML/DL methods more explainable.

1.3 Challenges

Throughout this study, we have identified major challenges in DTI data analysis specifically with regards to the BAP problem. Some challenges are specific to our datasets while some are more general and prevalent in typical DTI studies as follows.

Challenge 1: Individual and group differences in between-subjects' and within-subjects' brains and across different age groups are subtle and complex. The population of subjects is heterogeneous in terms of demographics, phenotypic differences, unknown health conditions, etc. (inconsistency in individual characteristics).

Challenge 2: Inconsistency of the chronological age labels: As framed in "Research Question-3", after the ML model prediction, it is not clear if the observed error (Delta) is due to the model's error, or the "actual" Delta indicating the gap in unhealthy or delayed brain aging, i.e., deviations from normal aging as specified by Delta.

Challenge 3: The datasets that we use are not only small-sized (298 and 94 subjects), which leaves training data and testing data with few examples, but also high-dimensional as we have to work with huge tensors that have approximately $1.5M$ to $6M$ features (see Section 3.1.2). This would make us deal with "Curse of Dimensionality". [1]

Challenge 4: The preprocessing transformations make the data very sparse, with many values very close to zero. The ratio of non-zero values to the size of tensors that we will work with is less than 1% in some cases. Necessary and standard preprocessing steps of DTI data make this challenge severe.

Challenge 5: The distribution of our datasets is not uniformly distributed with respect to age groups (young/middle-aged/old), and the ratio of young subjects in the datasets whose age is less than 40 is significantly low (less than 30% in both datasets). Thus, the training data is biased with respect to age groups with a heavy bias in favor of $60+$ subjects.

[1] Curse of Dimensionality in Machine Learning refers to different challenges that arise when processing data in high-dimensional spaces. Those challenges typically do not occur in lower dimensional spaces such as 2D or 3D.

Challenge 6: Not only are our datasets skewed with respect to age groups, but also they are skewed within each group as well. For the age group whose age is less than 40, the two datasets that we use are left-skewed, i.e., more samples are in $18-23$ age range (full-range is $18-32$), and for the greater than 60 age group, they are also left-skewed with a range of $60-67$ (full-range is $60-77$).

Challenge 7: Deep Learning models are known to be data hungry, and hence may not be a good fit for our small datasets.

Challenge 8: The results of ML/DL models are difficult to "explain" and to "interpret" as to how the models make decisions and based on which features in the feature space of DTI brain scans.

Challenge 9: Last but certainly not least, DTI analysis is a very expensive task computationally. Consequently, we often experience Out-Of-Memory crashes, dead Jupyter kernel due to GPU memory allocation issues, etc. Resolving and handling those issues is a very tedious and time-consuming task.

1.4 Contributions

Our key contributions in this dissertation are as follows. In addressing our major research questions, we propose and develop novel approaches and methodologies for DTI analysis as it relates to the BAP problem. Specifically, our "Unsupervised Anomaly Detection (UAD)" approach to identify irregular DTI scans is novel to the best of our investigation in the literature.

Moreover, we propose, implement, and analyze a novel methodology which combines our unsupervised UAD method and the supervised BAP method. Our proposed methodology can address the major limitations of the current BAP approaches, and provides a solution for methodical verification and interpretation of BAP errors (Deltas).

We use Diffusivity Parameters as a single modality for BAP for the first time. We provide effective guidelines for preprocessing of Diffusivity Parameters as a result of an extensive set of analytical experiments and analyses. We methodically identify the best combinations of Diffusivity

Parameters for supervised and unsupervised DTI analysis tasks with our novel approach and for the first time. We also provide effective techniques for DTI data augmentation and generation as well as transfer learning of ML/DL models across datasets.

1.5 Outline of The Following Chapters

In Chapter 2, we explain the background of BAP, DTI, and applications of ML/DL, and we review the literature and related work. We also introduce Diffusivity Parameters and the details of their mathematics. In Chapter 3, we thoroughly explain our methods, and the details and mathematics of our ML/DL algorithms with a focus on their application for our DTI analysis. In Chapter 4, we provide the complete details of our experiments with ML/DL methods, analyses on DTI data, results of experiments, discussions of results, and arguments of our approach and novel methodologies. In Chapter 5, we review a summary of our key findings, as well as the limitations of our study, and potential research directions for future work.

In Appendix A, we provide the configurations of our hardware and software tools, as well as the specifications of our Grid Search to fine-tune hyperparameters of some of our models. In Appendix B, we provide the list of acronyms that we frequently use in this dissertation.

Chapter 2

Background

In the previous chapter, we provided an introduction to Brain Age Prediction (BAP) using Machine Learning (ML) and Deep Learning (DL) methods applied on Diffusion Tensor Imaging (DTI) data. In this chapter, we review the background, related work, current BAP approaches, and their limitations. We begin by explaining DTI data and its Diffusivity Parameters as well as how they are calculated mathematically.

2.1 Diffusion Tensor Imaging (DTI)

Diffusion Tensor Imaging (DTI) is an advanced quantitative Neuroimaging modality that concentrates on the characteristics and microstructural changes of the White Matter (WM) tracts [7,8]. White Matter consists of axons, long wires that transmit electric signals between neurons, brain processing cells [9, 11]. Axons are covered by myelin, with a white color, to protect signals and to prevent electric leaks. An analogy in the Computer Science and Engineering context is the network infrastructure (like cables, or fiber optics), that connects the network nodes and hubs, and integrates the whole system. As much as the network infrastructure is critical for the connectivity and functionality of a network, White Matter is important for the organization of the brain and its proper functioning.

DTI generates voxel-wise quantitative measures of the White Matter [7]. DTI gives valuable information about the amount and direction of the diffusion of water molecules in the brain tissues, which are quantified and encoded in "**Tensors**" [8, 10, 11], as described below.

Water diffusion is known to be anisotropic in the White Matter [12], in the sense that it can diffuse in certain directions due to the structures and microstructures of the tissues. Figure 2.1 shows the difference between anisotropic and isotropic shapes. The Einstein diffusion equation, Equation (2.1), specifies how the diffusion coefficient D (in terms of mm^2/s) is calculated [8, 17].

$$D = \frac{< \Delta r^2 >}{2n\Delta t} \tag{2.1}$$

where $\langle \Delta r^2 \rangle$ is the mean squared displacement, n is the number of dimensions, and Δt is the diffusion time. Equation (2.2) shows the molecular water displacement characterized by a Gaussian probability distribution.

$$P(\Delta r, \Delta t) = \frac{1}{\sqrt{(2\pi D \Delta t)^3}} \exp\left(\frac{-\Delta r^2}{4D\Delta t}\right) \tag{2.2}$$

Basser et al. introduced the "**Tensor**" to characterize anisotropic diffusion in DTI [8, 18]. In their work, diffusion is characterized by a multivariate Gaussian distribution, as shown in the following equation.

$$P(\Delta \vec{r}, \Delta t) = \frac{1}{\sqrt{(4\pi \Delta t)^3 |\mathbf{D}|}} \exp\left(\frac{-\Delta \vec{r}^{\mathrm{T}} \mathrm{D}^{-1} \Delta \vec{r}}{4\Delta t}\right) \tag{2.3}$$

where the diffusion tensor, $\mathbf{D}$, is a (3x3) covariance matrix, as shown in Equation (2.4).

$$\mathbf{D} = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{yx} & D_{yy} & D_{yz} \\ D_{zx} & D_{zy} & D_{zz} \end{bmatrix} \tag{2.4}$$

Equation (2.4) explains the covariance of diffusion movements in three dimensions (x, y, z). By Eigendecomposition of the diffusion tensor $\mathbf{D}$, we can get the eigenvalues $(\lambda_1, \lambda_2, \lambda_3)$ and the eigenvectors of the diffusion tensor, i.e., the principal axes of the diffusion, as shown in Figure 2.1. In fact, the eigenvalues are used to calculate the DTI Diffusivity Parameters, as described in the following section, and that provides us with the feature space, and the tensors that we need for our analyses in this study. As we present unsupervised Gaussian Mixture Model (GMM) in the next chapter (see Section 3.3.1), we argue that the Gaussian distribution of the diffusivity makes the GMM assumption that the DTI data is a mixture of Gaussian distributions appropriate and relevant.

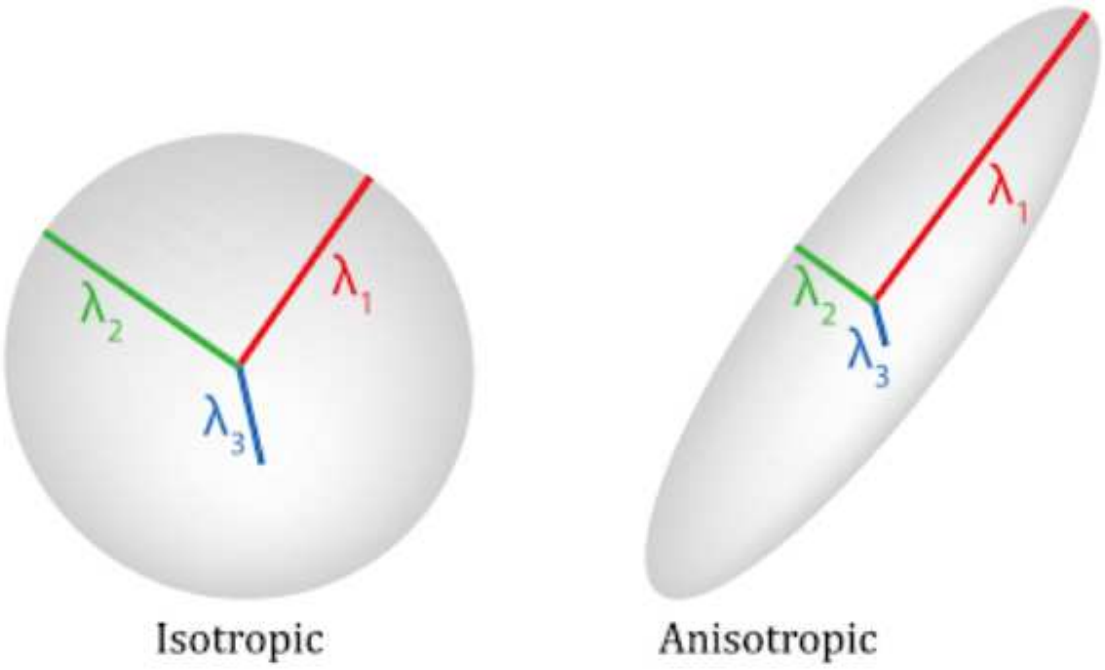


Figure 2.1: Isotropic and Anisotropic Tensors of Diffusivity Parameters. The isotropic tensor on the left has relatively close eigenvalues for each vector whereas the anisotropic tensor on the right has $\lambda_1 > (\lambda_2, \lambda_3)$, and hence an anisotropic diffusion in one particular direction [19].

2.1.1 Diffusivity Parameters

Diffusivity Parameters, Fractional Anisotropy (FA), Axial Diffusivity (AD), Mean Diffusivity (MD), and Radial Diffusivity (RD) are defined in Equation (2.5) [11], and are very important DTI measures which have been shown to be correlated with the brain age and other brain conditions [11, 18, 20–27]. We use them as the input tensors to train our Machine Learning (ML) and Deep Learning (DL) models. We refer to them individually as FA, AD, MD, and RD, or collectively as Diffusivity Parameters throughout this dissertation. They are sometimes known as "Diffusivity Metrics", "Measures", or "Scalars" in the literature. We provide a very brief summary for each of them in the following subsections.

$$
\begin{aligned}
AD &= \lambda_1 \\
RD &= (\lambda_2 + \lambda_3)/2 \\
MD &= (\lambda_1 + \lambda_2 + \lambda_3)/3 \\
FA &= \sqrt{\frac{1}{2}} \frac{\sqrt{(\lambda_1 - \lambda_2)^2 + (\lambda_1 - \lambda_3)^2 + (\lambda_2 - \lambda_3)^2}}{\sqrt{(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)}}
\end{aligned}
\tag{2.5}
$$

where $(\lambda_1, \lambda_2, \lambda_3)$ are the eigenvalues of the diffusion tensor.

Fractional Anisotropy (FA)

FA varies in the range of $[0.0 - 1.0]$, which specifies the level of anisotropy of diffusion (with 0.0 as the maximum isotropy, and 1.0 as the maximum anisotropy, or minimum isotropy), and is associated with microstructural changes in the White Matter. FA has been shown to decrease as adults age [20, 27]. In some DTI studies as well as our study, scans of Diffusivity Parameters are "Skeletonized", as part of the preprocessing steps, by using "The Tract-based Spatial Statistics (TBSS)" pipeline on FSL software, for creating a common White Matter, known as "Skeleton" [23, 28]. A skeletonized standard FA scan (mean) on the standard brain is shown in Figure 2.2.

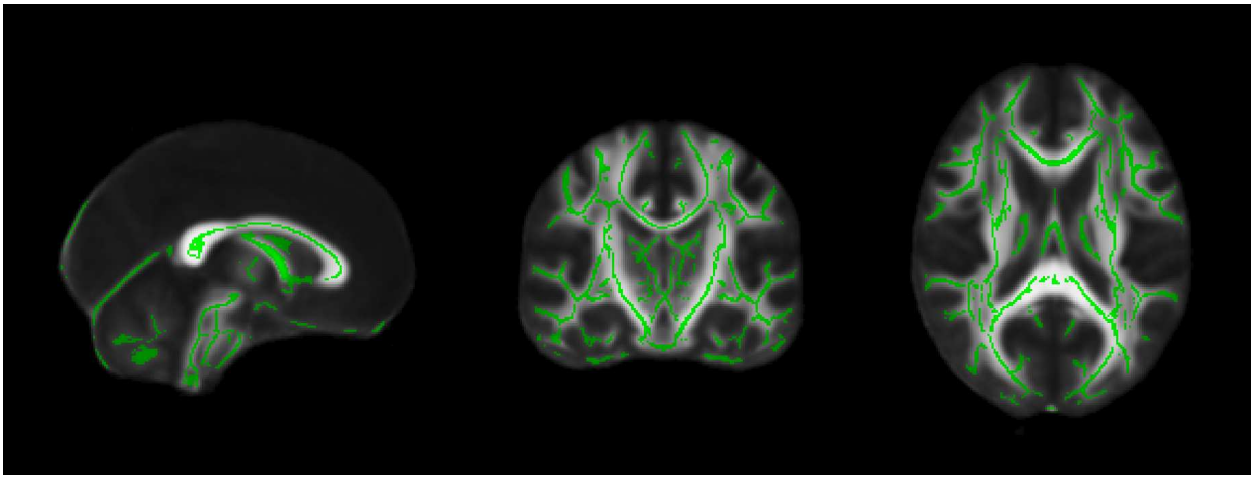


Figure 2.2: Skeletonized FA (1mm) shown in green on the standard brain (1mm) in white in FSL.

Axial Diffusivity (AD)

AD measures the magnitude of diffusion parallel to axonal fibers, as the name suggests, i.e., axial movement. AD is also associated with brain aging as well as injuries, as AD decreases in axonal injury [11, 21]. AD values are close to zero like other Diffusivity Parameters.

Mean Diffusivity (MD)

MD shows the magnitude of the movement of water molecules independent of its axis. MD is shown to be sensitive to damages, injuries, as well as aging [11, 21, 23, 24].

Radial Diffusivity (RD)

RD specifies the diffusion particularly on the radial axis, which is in the direction that is perpendicular to the axonal fibers (AD axis). RD increases in WM specifically with demyelination and axonal degeneration, both of which are associated with brain aging [11, 21, 25].

While the Diffusivity Parameters have been shown to be associated with brain aging in various forms, they show different patterns of non-linear changes with respect to aging as well as different brain regions [21]. Based on the mathematics of Diffusivity Parameters and their neurological implications, we argue that taking the combination of them into analysis seems more reasonable and efficient, and is more likely to capture the subtle and non-linear microstructural changes of the aging brain, as each of them is sensitive to certain changes and regions of the WM tracts. This motivates us to perform a combination analysis to identify the best combination(s) of them specifically for BAP, as framed in "Research Question-2" in Section 1.2.2.

2.2 ML/DL Applications in Brain Age Prediction (BAP)

As mentioned earlier, Brain Age Prediction (BAP) has gained the attention of researchers in recent years due to its significance and contribution to the early diagnosis of certain neurodegenerative diseases and cognitive decline as a result of unhealthy brain aging and a gap between biological age vs. chronological age [1–3, 25, 29–33]. Figure 2.3 [3] shows the trajectories of the

brain aging and the associated risk of neurodiseases as age increases, especially with respect to the gap between biological age and chronological age.

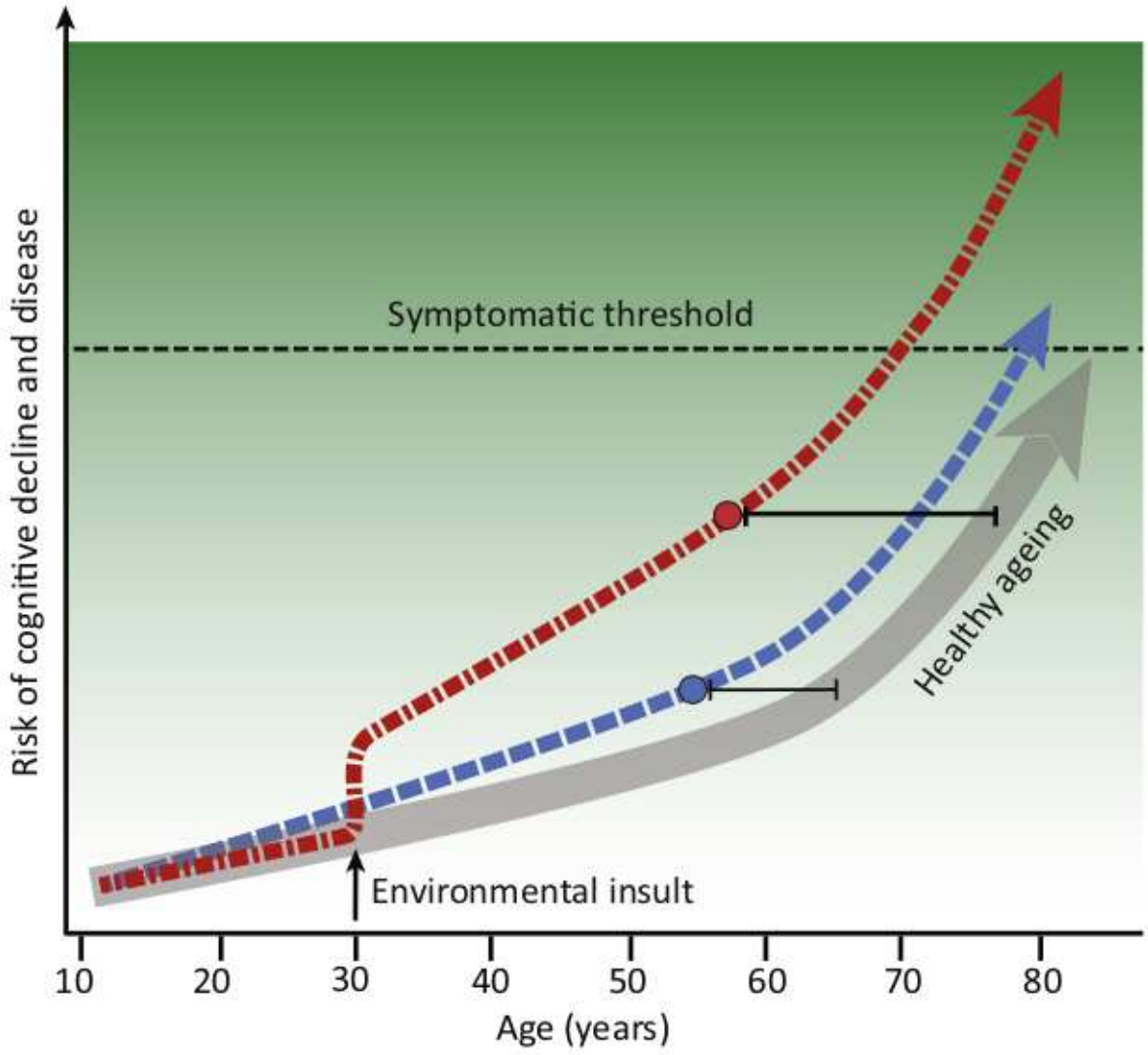


Figure 2.3: Biological Aging vs. Chronological Aging. The risk of cognitive decline and neurodegenerative diseases increases with aging and has been associated with the gap between biological aging vs chronological aging as a biomarker [3].

We have reviewed the literature and read survey papers on the applications of Machine Learning (ML) and Deep Learning (DL) methods for BAP, and we found a diverse set of ML/DL methods which have been applied on Neuroimaging data (mainly structural MRI) for BAP [4–6, 29, 34–37]. However, to the best of our investigation, they have one thing in common, using ML Supervised

Learning methods and training the models with chronological age labels. Figure 2.4 shows the supervised BAP workflow [3]. Models are trained using Neuroimaging data as features along with chronological age labels as ground truth, and then evaluated, validated, and selected to tune their parameters, and finally applied on new samples to estimate their biological age.

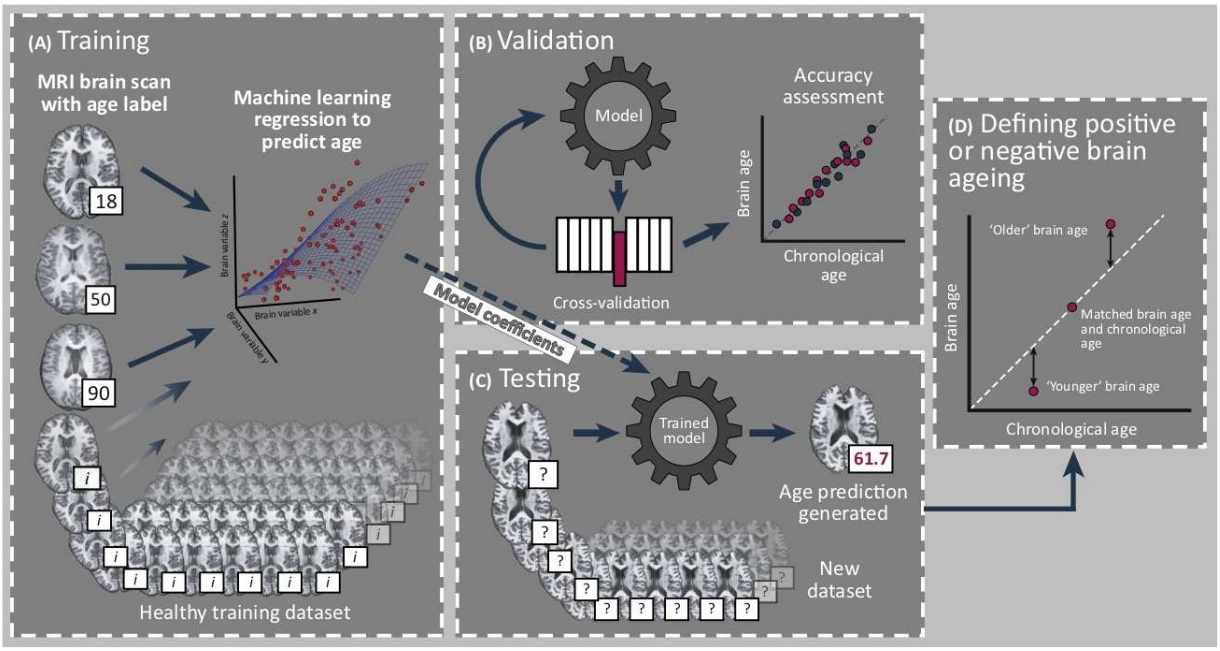


Figure 2.4: Supervised BAP Process. This is the general ML workflow of BAP as a Supervised Learning task where the chronological age labels are presented to the ML model along with Neuroimaging features for training, and then to predict the biological age of new samples [3].

The majority of studies have defined BAP as a regression task on age, and have trained different ML/DL models, including Convolutional Neural Network (CNN) [34, 35, 38], ensemble methods [37], and other ML algorithms. We reviewed a study [39], in which the researchers have used 27 different ML models, including linear and non-linear models, as well as ensemble methods, compared their performance for BAP, and concluded that the models' performance vary on different datasets, and ensemble models do not always outperform the regularized regression models.

Moreover, we reviewed a study in which the researchers have redefined BAP as other forms of supervised learning tasks including soft classification and ordinal regression [34]. They have

also used transfer learning to use pre-trained weights of CNN models applied on publicly available datasets with T1-weighted MRI. In fact, transfer learning has been widely used in multiple MRI studies for BAP and other MRI-based analyses [40–43]. The availability of large pools of MRI datasets as well as the existence of pre-trained DL models (mainly CNN) has made it possible to transfer the models for new similar tasks. However, this prevalence is missing for DTI analyses, although researchers may try to transfer models from MRI domain into their DTI-based analyses.

Despite the general success trend of DL models and ensemble methods for BAP, we reviewed some studies that indicate CNN in particular performs similarly on MRI data compared to other traditional ML algorithms like Support Vector Regression (SVR) or Ridge regression [3, 30]. In the next section, we review a few limitations that we have identified on the current approaches of BAP.

2.3 Limitations of Current Approaches of BAP

To design our methods and experiments and to identify our potential research contributions for the BAP problem, we have identified the major limitations of the current approaches as a Supervised Learning ML task as follows, and we present our methods in the following chapter not only with respect to our five major research questions, but also to address these limitations.

First and foremost, we argue that the current supervised BAP approach is limited and biased, mainly due to the use of inconsistent chronological age labels. "**Chronological Age Label Inconsistency**" has multiple implications. First, despite the inclusion and exclusion criteria that researchers apply to collect data from healthy subjects, there is no methodical way to verify that the subjects are in fact healthy, and there might be unknown health conditions at the time of data collection. This means that a ground truth solely based on chronological age labels is not reliable. To clarify, let us hypothetically assume that our dataset has only two subjects, both 50-year-old, one with healthy brain aging and a biological brain age that matches the chronological age, i.e., 50 or very close, and the other subject with an unknown underlying condition (say diabetes) that affected their biological age, and hence with a wide gap between biological and chronological age.

If we present the Neuroimaging data of these two subjects along with their chronological age labels to train the ML model, the model has no way to learn the differences between them in terms of healthy aging, and would most likely take this variability into future predictions. Again, we do not provide the model with any information that helps it learn the differences. In fact, we do the opposite, meaning, we mislead the model by telling it that the two subjects have the same age! And then later, we expect the ML model to somehow magically learn the differences. Now, for the sake of our argument, let's assume that we keep adding scans from other subjects to our dataset. What would bridge this gap and what would compensate for this chronological age inconsistency to help the model learn the differences effectively? The answer is absolutely nothing! If we add 1000 new samples, or $10,000$ new samples, we would have the same problem that we started with, chronological age label inconsistency, which we argue is the root cause. The issue just gets worse by adding more samples, not at all resolved.

This label inconsistency then spreads to error analysis as we framed in "Research Question-3". How can the researcher distinguish if an observed error is due to the model's poor performance or an actual unhealthy brain aging? The answer is there is currently no methodical way in the literature, to the best of our investigation, that can address this fundamental problem.

Another implication of the issue with the age labels is the imbalanced distribution of different age groups which is very common in Neuroimaging studies. This imbalance has been identified as one of the possible causes of a very common phenomenon in various BAP studies called "**Systematic Bias**", which refers to the fact that the BAP models often overestimate the age of younger subjects (or subjects who are younger than the mean), and underestimate the age of older subjects [15, 16, 44].

Researchers have identified possible causes for the "Systematic Bias" in the BAP context. For instance, it might be due to general statistical characteristics of the supervised regression analysis [44], or it might be due to imbalanced nature of Neuroimaging datasets, i.e., the sample size imbalance distribution across the age groups, or it might be due to the inconsistency of noise distri-

bution across the lifespan [3]. Researchers have also suggested ways to correct the bias [44]. Yet, the fundamental problem of "Chronological Ge Label Inconsistency" remains in the field.

Another limitation in the related work is that the focus of studies is always on lowering the performance metric of the regression model, i.e., Mean Absolute Error (MAE), as much as possible, for obvious reasons, such as getting published, etc. However, this leads to a decrease of the generalization of the existing models, and overfitting of the trained models. In fact, the BAP models' performances in different studies and across different datasets highly vary, as mentioned in our literature review earlier.

Finally, there are far fewer studies with DTI analysis than structural MRI. This makes the comparative and analytical studies more difficult.

Chapter 3

Methods

In the previous chapters, we provided the background, context and limitations of the current approaches for the Brain Age Prediction (BAP) using Diffusion Tensor Imaging (DTI). In this chapter, we present our methods, and we explain how our methodology addresses the limitations of the related work.

First, we explain the DTI data feature space and the target variable (age) as well as the specifications of the two datasets used in our study. We also explain the preprocessing steps that we employ before feeding the data to the ML/DL models.

Next, we explain all the ML/DL models, algorithms, and methods that we use in this dissertation to perform our experiments including our model evaluation and selection process as well as hyper-parameter tuning of the models. The specifications and configurations of the hardware and software used to perform our experiments are provided in Appendix A.

3.1 DTI Data

To the best of our research, we use DTI Diffusivity Parameters specifically for the Brain Age Prediction (BAP) in a novel way and for the first time as described in this chapter. We use two DTI datasets in this study, Dataset-1 ($n = 298$, 173 females, age range $[18-79]$), and Dataset-2 ($n = 94$, 68 females, age range $[20 - 79]$), and we perform our analysis on these two datasets separately and independently. We report the results of our experiments applied on each dataset separately in the following Chapter 4, although for some experiments like transfer learning (see Section 4.5.1), the results of the two datasets are connected, and for a couple of experiments (Table 4.6), results are

reported for Dataset-1 only due to size limitations of Dataset-2. The data acquisition process of the two datasets is as follows.[2]

3.1.1 Data Acquisition

Dataset-1 was collected as a joint research effort by multiple researchers from universities and research institutions across the United States and Europe [25, 26]. Dataset-2 is the current outcome as a part of an ongoing study and data collection process, led by Professor Agnieszka (Aga) Burzynska as the principal investigator who is affiliated with Colorado State University (CSU), which is still in progress at the time of writing this dissertation and not published yet. Both datasets include only cognitively normal adults with no known neurological or psychiatric disorders and no major non-neurological diseases at the time of data collection. Further details of the data acquisition for Dataset-1 are provided in [25]. As mentioned above, the Dataset-2 study is not published yet, but its experimental settings and data acquisition process are similar to Dataset-1.

3.1.2 Feature Space and Tensors

We use DTI Diffusivity Parameters as defined in Section 2.1.1 to train our ML/DL models. Thus, we use FA, AD, MD, and RD data as features. It should be mentioned that as part of the preprocessing of the two datasets, the Diffusivity Parameters were "skeletonized" as specified in Section 2.1.1. The distributions of the values of the four parameters are provided in Figure 3.1.

Each sample (subject) in the dataset has four corresponding 3D DTI skeletonized scans, one for each parameter, FA, AD, MD, and RD. Each 3D scan is a rank-4 tensor with the dimensionality of $(1, 121, 157, 80)$. The first dimension, 1, is added to the 3D scans to specify that it is a single parameter. Since we are interested in studying the combination of Diffusivity Parameters as well, the concatenation of all four of them will be a rank-4 tensor with the dimensionality of $(4, 121, 157, 80)$

[2] Although the author of this dissertation had no role and contribution in the data collection and acquisition process of the two datasets, we provide a summary of the process for the reference and citation purposes. We have acknowledged and credited their work and sharing the data with us in the Acknowledgement page of this dissertation.

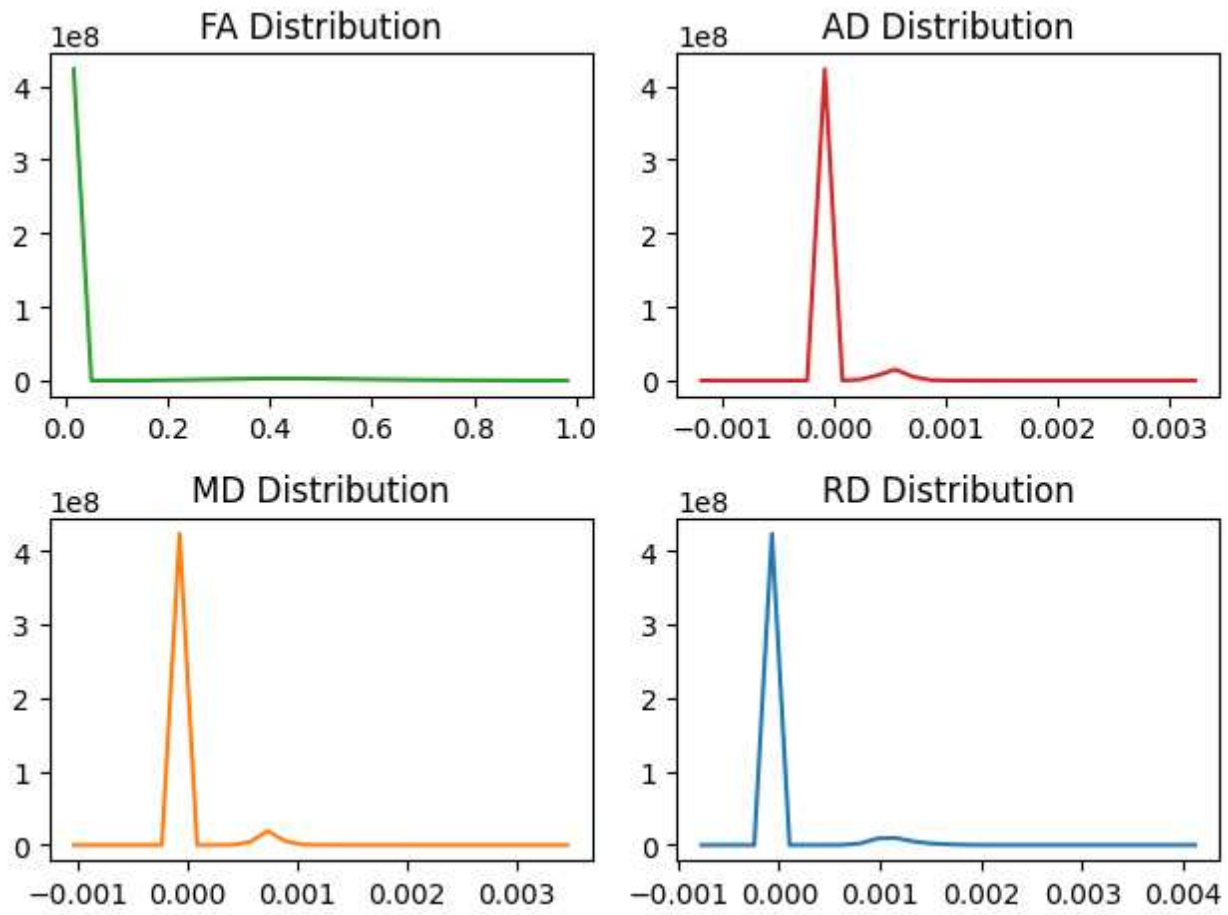


Figure 3.1: Distribution of Diffusivity Parameters (FA, AD, MD, RD) values in Dataset-1. Dataset-2 distributions are almost identical to these plots.

for each sample. If we combine two parameters, say AD and RD, we get a rank-4 tensor with the dimensionality of $(2, 121, 157, 80)$, and so forth. The whole batch of each parameter is a rank-5 tensor with the dimensionality of $(298, 1, 121, 157, 80)$ for Dataset-1, and $(94, 1, 121, 157, 80)$ for Dataset-2, respectively.

3.1.3 Target (Dependent) Variable

Since the common approach for BAP is a Supervised Learning regression method, we use chronological age labels along with the DTI feature space to train the supervised ML/DL models. The histograms of the chronological age labels for both datasets are provided in Figure 3.2. The pie-plots in Figure 3.3 and Figure 3.4 show the ratio of three age groups (young/middle-aged/old), and the two genders (female/male), respectively. As can be seen, the datasets are imbalanced with respect to the age groups. We will do series of experiments with respect to these age groups in Section 4.3.

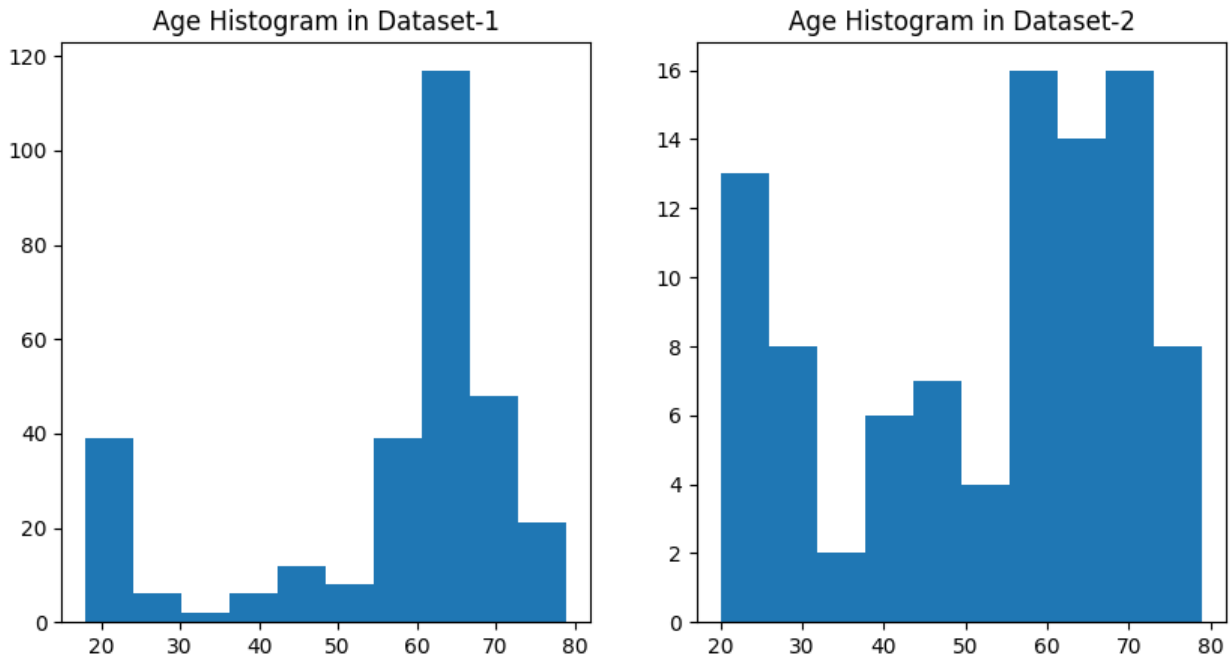


Figure 3.2: Histograms of Age – Target (Dependent) Variable in Dataset-1 and Dataset-2.

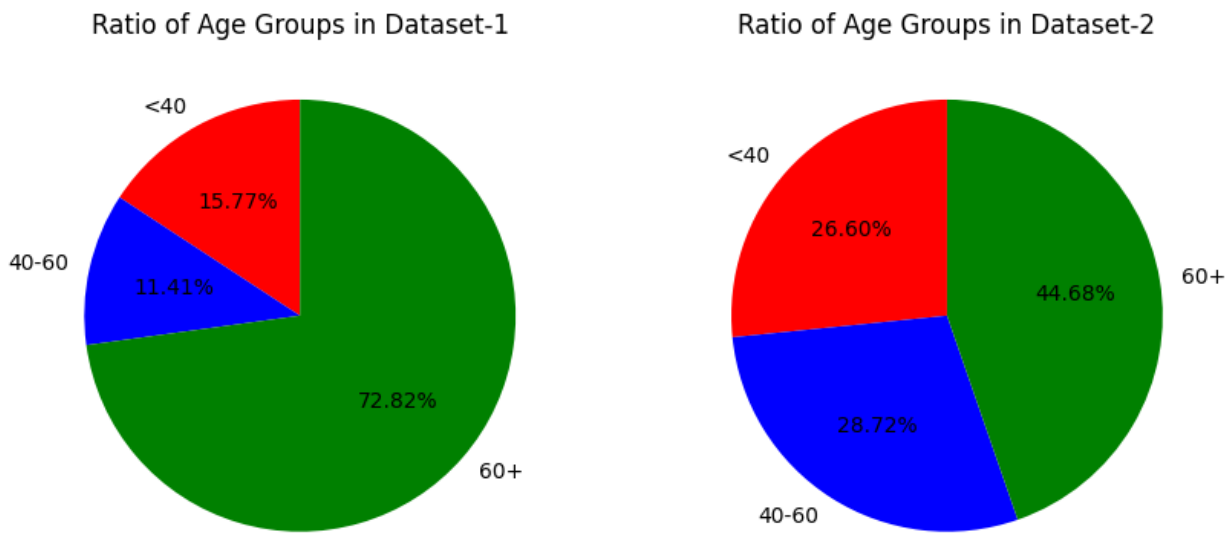


Figure 3.3: Ratio of Age Groups in Dataset-1 and Dataset-2.

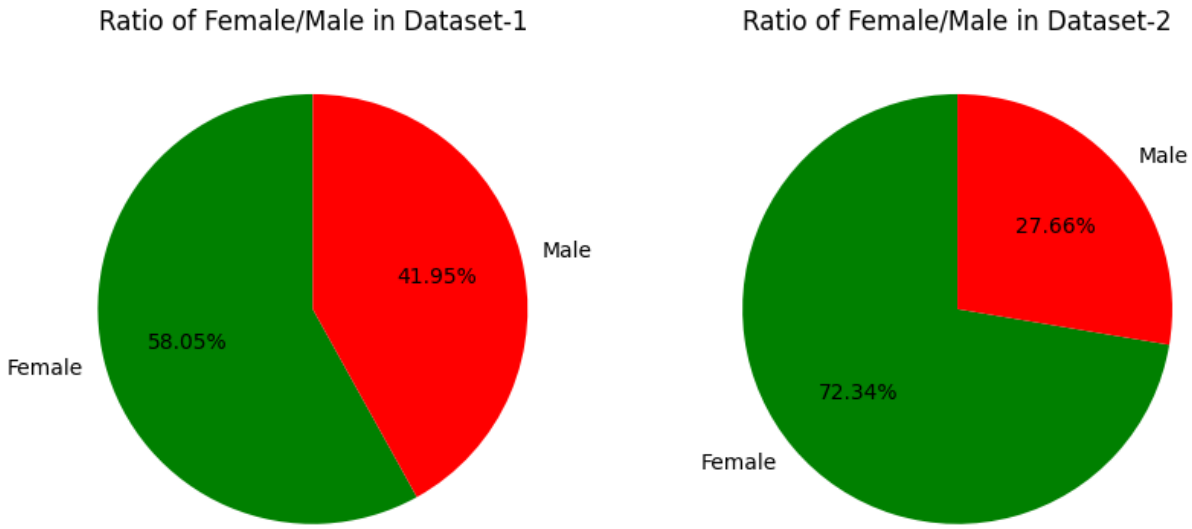


Figure 3.4: Ratio of Genders in Dataset-1 and Dataset-2.

3.1.4 Normalization and Standardization of DTI Data

There are multiple considerations regarding normalization and standardization of DTI data as follows.

First, the DTI scans go through multiple forms of registration, alignment, normalization and scaling during the preprocessing steps, but that doesn't theoretically prevent us from normalizing/standardizing the data further before feeding it to ML/DL models as a common practice in ML workflow and if it can improve the predictions. Given the fact that Diffusivity Parameters are in slightly different scales (see Figure 3.1), normalization/standardization may help with model learning process.

Second, the skeletonized DTI scans go through further preprocessing steps which may nullify the justification for additional normalizing or standardizing, but the same argument as the first one can be made that some forms of scaling may potentially help as long as it is applied properly and consistently to all samples (normalizing/standardizing with respect to training data only).

Third, we argue that it is preferred to preserve the statistical distribution of data in the preprocessing steps, and hence we prefer normalization over standardization, as it scales without distorting the differences in the values, unlike standardization that assumes the dataset is in Gaussian distribution which may or may not be true. As we will report in Section 4.1.1, we see a slight difference between normalization and standardization with an overall advantage for normalization of tensors to the range of $(0, 1)$, and we see improvement by scaling in some of our experiments (see Section 4.1.1).

Fourth, normalization/standardization, if not done properly, meaning if it is done on the whole data, causes information leak as it shares the statistics (mean and standard deviation) of train and validation data with test data which should be totally untouched until inference phase. Thus, we normalize based on training data only, and test data information is not used during normalization/-standardization at all.

Fifth, we perform many of our experiments with and without normalization/standardization to let the ML/DL models decide how the processes impact the results. As mentioned above, we

observe a noticeable difference in the absence and presence of normalizing our tensors for some parameters and experiments (see Section 4.1.1).

Finally, we do not normalize the labels to let the models predict the original unit of the age values (in terms of years) to maintain the consistency of comparison across all of our experiments.

3.1.5 Dimensionality Reduction

The "Curse of Dimensionality" is a very influential factor in working with our DTI tensors, especially given the fact that the dimensionality of feature space as specified in Section 3.1.2, is way higher than the number of samples in both datasets. As a known issue in the literature, high-dimensional neuroimaging data can trigger major issues in a BAP workflow, including high-variance and overfitting of the trained models as well as high computational cost of the training phase [29, 45].

Thus, we use linear and non-linear (kernel) Principal Component Analysis (PCA and kPCA, respectively) with the Gaussian Radial Basis Function (RBF) as the non-linear kernel for dimensionality reduction. PCA is an Unsupervised Learning dimensionality reduction technique. The non-linear kernel choice is based on the result of our hyperparameter tuning by Grid Search, as will be described in Section 3.2.3. The RBF kernel is provided in Equation (3.1).

$$\kappa\left(\mathbf{x_i}, \mathbf{x_j}\right) = \exp\left(-\gamma \left\|\mathbf{x_i} - \mathbf{x_j}\right\|_2^2\right) \tag{3.1}$$

where $\gamma = \frac{1}{2\sigma^2}$ is the kernel hyperparameter to be tuned ($\gamma = 0.02$ turned out to be the best value), σ^2 is the variance, and ($\mathbf{x_i}$, $\mathbf{x_j}$) are two samples in the data to be transformed by the kernel.

The motivation to use non-linear kPCA is that there is plenty of evidence by multiple studies that the DTI data has non-linear patterns [9, 11, 21, 24]. If we use linear PCA only, those non-linear patterns might get lost as a result of dimensionality reduction, so we try both liner and non-linear kernels of PCA and choose the best results for our experiments, with the exception of the baseline models (see Section 3.2 and Section 4.1) for which we use linear PCA with 0.99 preserved variance

to consistently compare the baseline models' performance on the Diffusivity Parameters reduced by the same linear kernel.

We tune the number of principal components for PCA and kPCA (see Figure 3.5). While we preserve 0.99 of variance which would give us approximately the number of samples for each dataset as the number of components (see Figure 4.4), we include PCA number-of-components hyperparameter in our grid search (see Section 3.2.3).

Moreover, as a preprocessing step, scans are masked with a binary mask (a 3D array containing 0 and 1) to reduce the dimensionality and filter out the surrounding areas of the White Matter which have zero values and no relevance to our analysis. Masking MRI and DTI scans is a common technique to differentiate the brain from non-relevant tissues [28]. The precision and accuracy of masking is of utmost importance for the reliability of further analyses of neuroimaging data. Without masking, the dimensionality of the 3D DTI scans will be $(182, 218, 182)$ which would create huge tensors that are very difficult to process even for GPU-Accelerated hardware that we use for our experiments.

3.2 Baseline Models

We use baseline regression models to get benchmark results for BAP and as a reference for comparison between our different methods. The selection of the baseline models is based on our thorough investigation in the literature [4–6, 29, 39]. We use both linear and non-linear models to compare their performance on DTI data for the BAP problem, as it is shown that non-linear BAP regression models can recognize non-linear patterns of brain aging such as trend changes that cannot soley identified by linear models [21]. We also use ensemble methods as recent studies have shown ensemble methods to be effective for DTI and MRI analysis [13, 46–49]. We explain our baseline models in the following sections.

3.2.1 Linear and Non-Linear Models

As mentioned above, we use both linear and non-linear models and compare their performance for the supervised BAP. On the one hand, linear models are simpler to implement and train, are often more computationally efficient, have a lower variance, and hence are less prone to overfitting. On the other hand, non-linear models can recognize more subtle non-linear patterns, have lower bias and are less prone to underfitting, but have higher variance and likelihood of overfitting. Therefore, there is a trade-off of using linear versus non-linear models, and we compare their performance.

Ridge

Ridge regression is our simplest regression model for supervised BAP. Ridge is the $l2-$regularized version of the simple linear regression model. The loss function of linear regression is Mean Squared Error (MSE) as provided in Equation (3.2).

$$\text{MSE}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{3.2}$$

where $\boldsymbol{\theta}$ is the vector of the model's trainable parameters, n is the number of samples, y_i is the i-th element in the target vector, and $\hat{y}_i$ is the predicted value of the i-th sample. Thus, the Ridge loss function, $J(\boldsymbol{\theta})$, is MSE plus the $l2$ regularization term, as provided in Equation (3.3).

$$J(\boldsymbol{\theta}) = \text{MSE}(\boldsymbol{\theta}) + \lambda\frac{1}{2}\sum_{i=1}^{n}{\theta_i}^2 \tag{3.3}$$

where λ is the regularization parameter and controls the effect of regularization (the level of parameter constraints). The regularization term is used to reduce the chance of overfitting of the linear regression model.

The loss function is minimized by optimizers, usually Gradient Descent (GD). The gradient vector of the loss function is provided in Equation (3.4).

$$\nabla_{\boldsymbol{\theta}} \mathrm{J}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial}{\partial \theta_0} \mathrm{J}(\boldsymbol{\theta}) \\ \frac{\partial}{\partial \theta_1} \mathrm{J}(\boldsymbol{\theta}) \\ \vdots \\ \frac{\partial}{\partial \theta_n} \mathrm{J}(\boldsymbol{\theta}) \end{pmatrix} \tag{3.4}$$

In each step of the Gradient Descent, the parameters are updated as follows.

$$\boldsymbol{\theta}^{(\text{next step})} = \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathrm{J}(\boldsymbol{\theta}) \tag{3.5}$$

where η is the learning rate (step size), and should be tuned.

Linear and Non-Linear Support Vector Regression (SVR)

Support Vector Regression (SVR) is our next baseline model and has two forms, linear and non-linear. First, let us provide the optimization problem for the linear SVR as follows.

$$\begin{aligned} &\underset{w,\xi,\xi^*}{\text{minimize}} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i + \xi_i^* \\ &\text{subject to:} \\ &y_i - w^T x_i \leq \varepsilon + \xi_i^* \quad i = 1 \dots N \\ &w^T x_i - y_i \leq \varepsilon + \xi_i \quad i = 1 \dots N \\ &\xi_i, \xi_i^* \geq 0 \quad i = 1 \dots N \end{aligned} \tag{3.6}$$

where w is the weight vector, x_i is the i-th sample in the the feature vector, ϵ is the margin width, (ξ_i, ξ_i^*) are the slack variables for the i-th sample to make the SVR "soft margin" by allowing some errors and to regularize the model, and C is the soft margin hyperparameter that controls the regularization. In other words, (ξ, ξ^*) slack variables determine how many data points are tolerated outside the margin (also known as ϵ-tube).

The non-linear SVR uses a non-linear kernel, such as RBF kernel as provided in Equation (3.1), to transform the data samples to a higher-dimensional space by the kernel trick (which actually doesn't need the transformation computational cost).

Linear and Non-Linear Neural Network

The next models are Feed Forward Neural Networks (NN). The linearity or non-linearity of the NN is determined by the activation functions of the neurons. We use linear (no activation function), and Rectified Linear Unit (ReLU) activation function which is non-linear, to make the NN linear and non-linear respectively. ReLU was selected after tuning, in comparison with Sigmoid and Hyperbolic Tangent (tanh) activation functions. The linear NN is optimized by Stochastic Gradient Descent (SGD) which is a modified version of GD such that in each iteration, a random sample is picked, the gradients are computed solely based on that single sample, and the parameters (connection weights) are updated accordingly, instead of using the whole batch for gradient vector as it is used in GD. SGD is typically faster, but may have more oscillations before convergence to the optimal solution. The non-linear NN is optimized by the Adam optimizer using early stopping as regularizer. We include the results of linear NN in the baseline models and the results of non-linear NN in the DL models, denoted as Deep Neural Network (DNN), because it has a deeper architecture with more hidden layers compared to the linear NN.

3.2.2 Ensemble Methods

Ensemble methods employ the "Wisdom of the Crowd" by training multiple models (learners or estimators), and in case of regression, making an average on their predicted values as the final prediction, or letting them correct the prediction of other learners, and/or focusing on fixing other learners' mistakes. The mechanism for aggregating the results of the learners in the ensemble and their arrangement in sequence, in parallel, or stacked, determine the category of the ensemble method, mainly bagging, boosting, and stacking. In general, ensemble methods trade a slightly higher bias for less variance, and hence are less prone to overfitting. They are often known to be

more scalable and generalized. We use two ensemble methods based on our investigation in the literature [46–49], Extreme Gradient Boosting (XGBoost), and Random Forests.

XGBoost

We use the Extreme Gradient Boosting (XGBoost) algorithm proposed by [50], and we use the optimized and parallelized implementation of it [51], which is compatible with our GPU-Accelerated Computing hardware (see Appendix A). Using GPU and training trees in parallel improves our training time significantly. It is also regularized to reduce the chance of overfitting. XGBoost has non-linearity, and hence can detect non-linear patterns in the DTI feature space. It is scalable and can handle our high-dimensional rank-4 and rank-5 tensors.

Our XGBoost model is an ensemble of regression decision trees, adapted from "Classification and Regression Trees (CART)", where the trees in the ensemble try to complement each other [50, 51]. The objective function $\mathrm{obj}(\theta)$ of the ensemble consists of two terms, the loss function $l\left(y_i, \hat{y}_i\right)$ (similar to MSE which calculates the difference between y_i, the age label, and $\hat{y}_i$, the predicted age), and the regularization term, and is written in Equation (3.7).

$$\mathrm{obj}(\theta) = \sum_{i}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \omega\left(f_k\right) \tag{3.7}$$

where $\omega\left(f_k\right)$ is the complexity of the k-th tree in the ensemble with K trees, f_k, which is controlled in the regularization term, and with n samples in the training data. Interestingly, this is the same objective function that is used for our other ensemble method, Random Forest (see the next subsection), with a difference in training strategy though, as described below.

"Additive Training" is used to train the XGBoost ensemble of trees by fixing what a tree f_i (which contains the structure of the tree and the leaf scores) has learned at each step t, and adding a new tree, and this is the core concept of "Boosting" as an ensemble method, as shown in Equation (3.8),

$$\mathrm{obj} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{i=1}^{t} \omega\left(f_i\right) \tag{3.8}$$

where $\hat{y}_i^{(t)}$ is the predicted age at step t as follows.

$$\begin{aligned}
\hat{y}_i^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
\hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\dots \\
\hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned}$$

Now the question is which tree is the best at each step t, and the answer is the one that minimizes the objective function at that step, $\mathrm{obj}^{(t)}$.

$$\begin{aligned}
\mathrm{obj}^{(t)} &= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{i=1}^{t} \omega(f_i) \\
&= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \omega(f_t) + \text{ constant}
\end{aligned}$$

Since our supervised BAP is a regression task, we use MSE as our loss function l, and hence the objective function at step t becomes,

$$\begin{aligned}
\mathrm{obj}^{(t)} &= \sum_{i=1}^{n} \left(y_i - \left(\hat{y}_i^{(t-1)} + f_t(x_i)\right)\right)^2 + \sum_{i=1}^{t} \omega(f_i) \\
&= \sum_{i=1}^{n} \left[2\left(\hat{y}_i^{(t-1)} - y_i\right) f_t(x_i) + f_t(x_i)^2\right] + \omega(f_t) + \text{ constant}
\end{aligned}$$

The generalized form of the objective function for Supervised Learning tasks, which can accept various differentiable convex loss functions, denoted by l, that can measure the difference between the prediction $\hat{y}_i$ and the target y_i (it could be other than MSE, such as log-loss or cross entropy), is provided in Equation (3.9).

$$\mathrm{obj}^{(t)} = \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \omega(f_t) + \text{ constant} \tag{3.9}$$

where the g_i and h_i are defined as,

$$
\begin{aligned}
g_i &= \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \\
h_i &= \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)
\end{aligned}
\tag{3.10}
$$

If we remove all the constants, the final objective function at step t becomes,

$$
\sum_{i=1}^{n}\left[g_i f_t\left(x_i\right)+\frac{1}{2} h_i f_t^2\left(x_i\right)\right]+\omega\left(f_t\right) \tag{3.11}
$$

To define the complexity of the tree $\omega(f)$ in the regularization term, we first define the tree $f(x)$ as,

$$
f_t(x)=w_{q(x)}, w \in R^T, q: R^d \rightarrow\{1,2, \cdots, T\} \tag{3.12}
$$

where w is the vector of leaf scores, q is a function that takes each sample x and assigns it to the corresponding leaf, and T is the total number of tree leaves. The complexity of the tree is then defined in Equation (3.13).

$$
\omega(f)=\gamma T+\frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \tag{3.13}
$$

where γ is the minimum loss reduction required to make a further split on a leaf node of the tree; the larger value of γ, the more constrained the ensemble, or as the author of XGBoost paper refers to it, "complexity cost by introducing additional leaf" [50], and finally λ is the regularization parameter and controls the effect of it. Multiple optimization algorithms to minimize the above loss function are proposed in [50], which we have tried and selected the distributed greedy algorithm that is most compatible and optimized with parallel computing of our GPUs (see Section 3.2.3).

Random Forests

We use Random Forest as another ensemble method for our supervised BAP regression baseline models. Random Forest is a bagging method. It is an ensemble of several decision trees, and the final prediction is the aggregated prediction of the trees in the ensemble by voting (in classification)

or averaging (in regression). Each tree in the random forest uses a random subset of the original feature space to train (could be either subset of samples, or subset of features, or both), and hence, just like XGBoost, random forest trades a higher bias for a lower variance to overcome overfitting by diversifying the learners in the ensemble.

Another important attribute of random forests, which we use to create our brain maps (see Section 4.6), is that they can measure feature importance by computing the average impurity reduction of the trees that use each feature. The formula to calculate the feature importance of random forest is provided in Equation (3.14).

$$fi_i = \frac{\sum_{j:\text{ node } j \text{ splits on feature } i} ni_j}{\sum_{k\in \text{ all nodes}} ni_k} \tag{3.14}$$

where fi_i is the importance of feature i, ni_j is the importance of node j, and is calculated by the following equation.

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{\text{right}(j)} C_{right(j)} \tag{3.15}$$

where w_j is the weighted number of samples reaching node j, C_j is the impurity value of node j as measured by Gini impurity, $left(j)$ is the child node from left split on node j, and right(j) is the child node from right split on node j.

For high-dimensional datasets, the feature importance scores are usually sparse and close to zero; nevertheless, we are able to map them on the standard brain to visualize the areas on the White Matter to which the BAP Random Forest model is most sensitive.

3.2.3 Model Evaluation Process

In this subsection, we explain our model evaluation and selection process as well as the performance metrics of the baseline models. We also present the settings of our hyperparameter Grid Search, including the range of values we try, and the selected best value for each hyperparameter we search on.

Cross Validation and Performance Metrics

We use 5-fold Cross Validation (CV) for model evaluation and selection in accordance with the size of our datasets. For evaluation of the baseline models, we repeat the 5-fold CV 10 times, and get the mean of the CV score across all iterations, to account for the randomness of the algorithms and splits. In 5-fold CV, data is split into 5 partitions stratified by age, from which 4 partitions are used for training in each iteration, and 1 partition is used for testing which changes in each iteration. We apply scaling (when we use it) with respect to training data in each iteration of CV, and we also train PCA and kPCA with respect to training data of each iteration of CV. This is to ensure that there is no information leak from training phase to testing phase in our model evaluation and selection process.

As the performance metric (score) of CV, we use Mean Absolute Error (MAE) as defined in Equation (3.16).

$$\text{MAE}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{3.16}$$

where $\boldsymbol{\theta}$ is the vector of the model's trainable parameters, n is the number of samples, y_i is the i-th element in the target vector, and $\hat{y}_i$ is the predicted value of the i-th sample.

The other common choice for the performance metric of a regression model is Root Mean Square Error (RMSE), as defined in Equation (3.17) (it is just root of MSE). However, we prefer MAE over RMSE for two main reasons, RMSE is more sensitive to outliers which exist in our datasets, and the MAE unit is in years compatible with Delta, and hence researchers from other fields in the BAP literature are more familiar with it.

$$\text{RMSE}(\boldsymbol{\theta}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{3.17}$$

where the terms are the same as described in MSE and MAE equations.

Fine-Tuning Hyperparameters

When there are a few hyperparameters for an ML/DL methods to be tuned, Grid Search is a thorough and systematic way to find the best combination of values for the hyperparameters. However, Grid Search is computationally expensive, as it should try all combinations of the values, and train and evaluate the model in question with each specific combination. Therefore, Grid Search is not recommended when there are so many hyperparameters to be tuned such as NN and XGBoost. In those cases, alternative strategies such as Randomized Search is preferred which narrows the search space by randomly selecting a portion of the search space and focusing on the areas that turn out to be more promising.

We use Grid Search for Ridge and SVR as follows. Ridge regularization parameter λ is tuned in the range of $[0.0, 0.1, 0.01, 0.001, 0.0001, 0.00001]$, along with the fit-intercept (which is whether to fit the intercept for this model). The best values are, $\lambda = 0.00001$, and fit-intercept=True. Linear SVR is tuned in Grid Search for C and ϵ with the ranges $C = [1.0, 10.0, 100.0]$, and $\epsilon = [0.1, 0.5, 1.5]$, and the best values are $C = 10.0$ and $\epsilon = 1.5$.

Also, we tried different settings for linear NN. Since it was linear NN, there was no other choice for the activation function other than being linear, i.e. no activation function. We tried different optimizers and SGD and the Adam optimizer gave similar results although SGD was slightly faster, and hence was selected. For non-linear, we tuned it as DNN along with CNN which resulted to the architecture and settings as shown in Figure 4.11, and Section 4.5.

For XGBoost and Random Forest, since they have so many hyperparameters with a wide range of values, we have included the results in Appendix A.3. Most importantly for XGBoost, we use the GPU-Accelerated version of the greedy algorithm for the optimization of the loss function.

We also tuned PCA hyperparameter, number of principal components, based on the variance ratio in the range of $(0.01 - 0.99)$. Figure 3.5 shows the impact of the number of principal components on MAE of Random Forest trained on AD+RD.

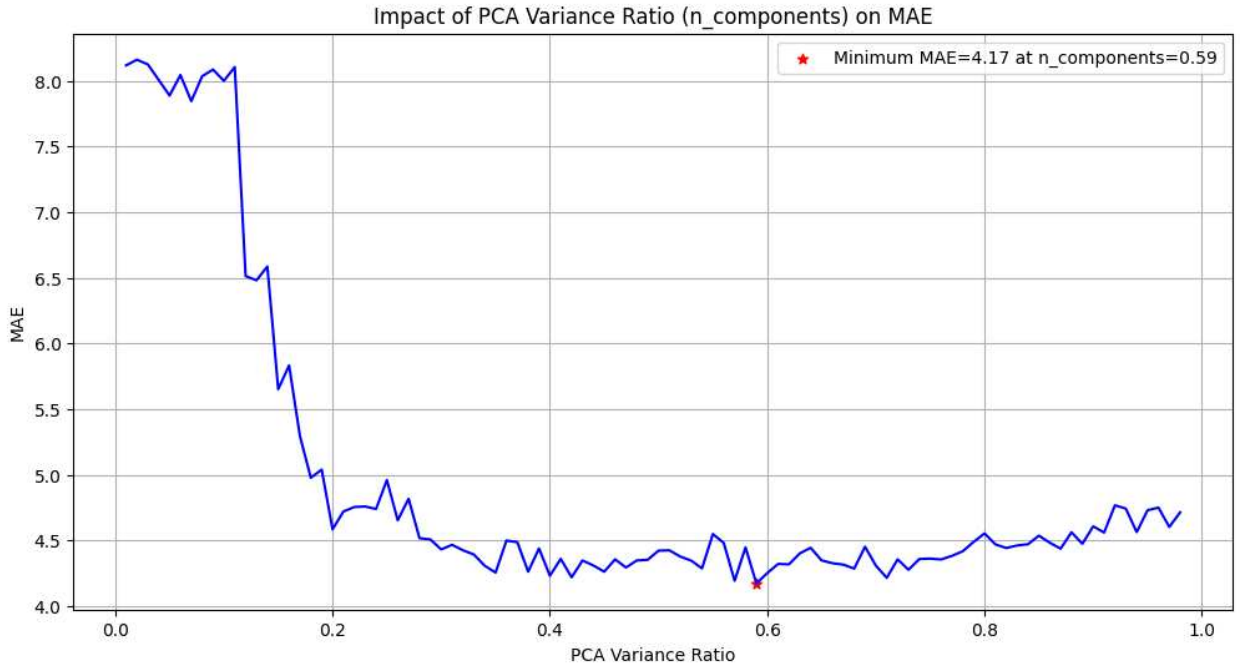


Figure 3.5: Impact of Number of Principal Components on MAE. Linear PCA is used to generate this plot. The scale of increments of the variance on X-axis is 0.01 in the range of $[0.01 - 0.99]$. The MAE scores are for the Random Forest model trained and tested on AD+RD data.

3.3 Unsupervised Learning Methods

The famous Deep Learning scientist and pioneer Yann LeCun said that "if intelligence was a cake, unsupervised learning would be the cake, supervised learning would be the icing on the cake, and reinforcement learning would be the cherry on the cake [52]". The significance, the breadth and the power of the Unsupervised Learning methods mainly come from their independence from the ground truth, labels, as opposed to Supervised Learning methods which require plenty of labeled samples for training. This is critical in the BAP context because as we mentioned in Section 2.3, chronological age labels are inconsistent and do not reflect the biological age for a considerable portion of the population, and that is the main motivation to use BAP models in the first place. However, as we argued in Section 2.3, the current supervised BAP approaches, which solely rely on the inconsistent chronological age labels for training, are limited and biased.

Another advantage of Unsupervised Learning methods is that we can use the whole data to train them, as splitting to train/test subsets is no longer required. We use Unsupervised Learning meth-

ods to perform six ML tasks by four algorithms as follows. We perform dimensionality reduction and 2D projection of the DTI feature space using PCA and kPCA, and we perform clustering, Unsupervised Anomaly Detection (UAD), as well as data generation using Gaussian Mixture Model (GMM) and Bayesian Gaussian Mixture Model (BGMM). We have already explained our dimensionality reduction techniques in Section 3.1.5. Next, we explain the algorithms that we use for clustering and Unsupervised Anomaly Detection (UAD) of the DTI data.

3.3.1 Gaussian Mixture Model (GMM)

We use the probabilistic and generative Gaussian Mixture Model (GMM), with the "**Expectation Maximization**" algorithm to train it, to perform clustering, Unsupervised Anomaly Detection (UAD), and data generation. The GMM assumes that the data is a mixture of a finite set of Gaussian distributions and tries to estimate their parameters, mean and variance. We argue that this assumption is appropriate given the Gaussian distribution of water diffusion as captured by Diffusivity Parameters of DTI, and as described in Equation (2.2). The algorithm is explained below with a summary from [53].

The multivariate Gaussian distribution is defined in Equation (3.18).

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right) \tag{3.18}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are vectors of means and variances, $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$, and D is the number of dimensions $\boldsymbol{x} \in \mathbb{R}^D$. By changing $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, we can shift and change the shape of the Gaussian respectively.

Next, the responsibility r_{nk} for the k-th Gaussian component and the n-th sample is defined in Equation (3.19).

$$r_{nk} = p\left(z_k = 1 \mid x_n\right) = \frac{\pi_k \mathcal{N}\left(x_n \mid \mu_k, \sigma_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(x_n \mid \mu_j, \sigma_j\right)} \tag{3.19}$$

where r_{nk} is also the posterior distribution, because the responsibility r_{nk} corresponds to $p\left(z_k = 1 \mid x_n\right)$, the probability that the sample x_n has been generated by the k-th Gaussian component of the Gaussian mixture.

The Expectation Maximization (EM) algorithm is an iterative method to estimate the maximum likelihood, or Maximum A Posteriori (MAP) of a set of parameters. The algorithm has two main steps, the "Expectation" step, in which the Gaussian parameters are randomly initialized and then "responsibilities" or posterior distributions are estimated, and the "Maximization" step, in which the estimated parameters in the Expectation step are maximized. Each EM iteration increases the log-likelihood function until a certain threshold is met, or the maximum number of iterations is reached.

Since finding the optimal number of Gaussian components is difficult, a variant of the GMM has been proposed, Bayesian GMM [54], which can find the optimal number of components by "**Variational Inference**", an extension of "Expectation Maximization" algorithm, that maximizes a lower limit on the model evidence instead of the likelihood, and adds regularization to the GMM.

The fitness of GMM models is determined by two information criterion scores, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). AIC and BIC are defined in the following equations.

$$\mathrm{AIC} = -2 \log L(\hat{\boldsymbol{\theta}}) + 2k \tag{3.20}$$

where $\boldsymbol{\theta}$ is the vector of the model parameters, $\log L(\hat{\boldsymbol{\theta}})$ is the log-likelihood of the candidate model given the data, when it is evaluated at the maximum likelihood estimate of $\boldsymbol{\theta}$, and k is the number of estimated parameters in the candidate model. The lower the AIC, the better fit the model, and that includes the negative values, meaning a model with $AIC = -996$ is a better fit than a model with $AIC = -995$.

$$\mathrm{BIC} = -2 \log L(\hat{\boldsymbol{\theta}}) + k \log n \tag{3.21}$$

where n is the number of samples, and the other terms are the same as described in AIC.

The fitness implication of BIC is similar to AIC, the lower the better, even with a negative sign.

3.3.2 Clustering

Clustering is grouping similar samples together in an unsupervised manner. Since it is an unsupervised method with no access to the labels, "Similarity" should be defined. We perform clustering using GMM to identify age groups, subjects with "similar" age, without even knowing their actual chronological age. This, if performed successfully, would help us identify outliers, or "Anomalies" as explained in the following subsection. Since we do not have prior knowledge of the optimal number of clusters, we use Bayesian GMM to find it, as the number of cluster. See the results of clustering in Section 4.4.2.

3.3.3 Unsupervised Anomaly Detection (UAD)

Once the GMM is trained on the data (and we use the whole data to train it), we have the log-likelihood of each sample, and we can set a threshold on it to identify the samples with low log-likelihood. This means that we can identify the samples in low-density regions of our clustering model, and regard them as "Anomaly".

The significance of the Unsupervised Anomaly Detection (UAD) is that if it turns out that these detected anomalies have any relationship or commonality with the subjects that yield high Delta in supervised BAP, we can resolve the issue of inconsistent chronological age labels, because we can detect anomalies, i.e., subjects with a wide gap between their chronological age and biological age, in an unsupervised manner with no need to know their chronological age labels. Besides, we can address "Research Question-3", as specified in Section 1.2.3, by a methodical way to verify whether the observed error in the supervised BAP has an actual irregularity or abnormality, as detected by our UAD method. To the best of our investigation, we utilize UAD in a novel way for the first time in the context of BAP.

Figure 3.6 shows 2D projection of FA scans of Dataset-1 on the upper plot, along with the clustering model and the detected anomalies, marked by red star in the bottom plot. The cluster of young age group is clearly recognized by the GMM.

3.3.4 Our Proposed Methodology to Combine UAD and BAP

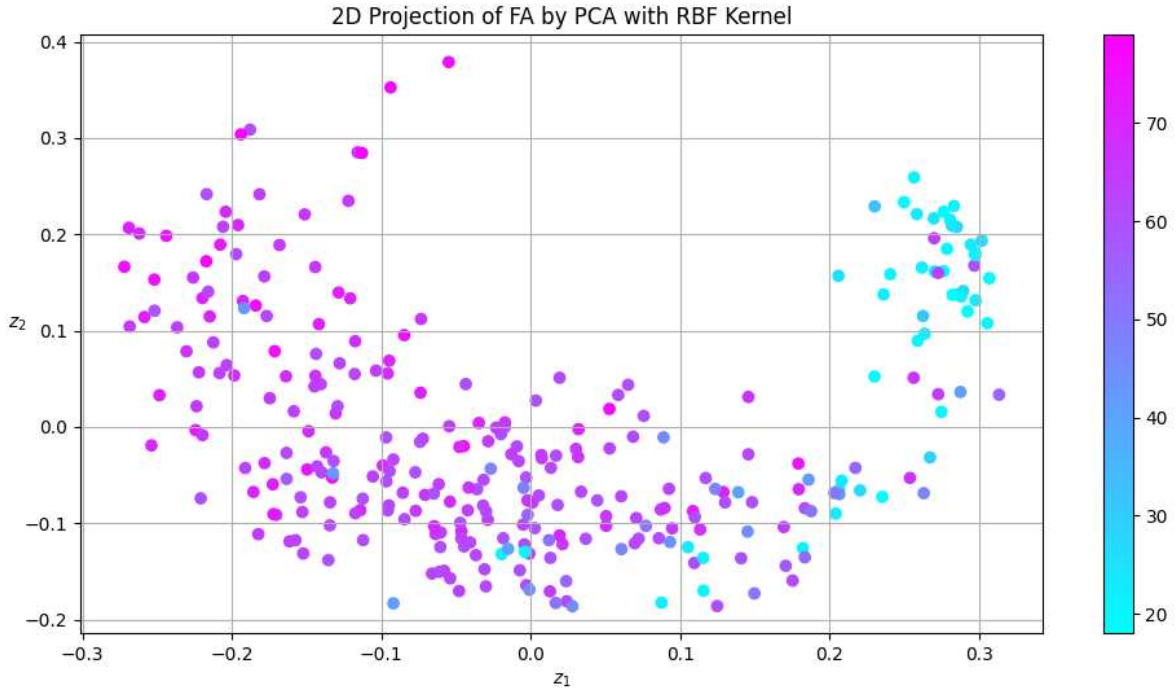
Now that we have explained our supervised and unsupervised methods, we propose our methodology to combine the two approaches in three steps as follows.

Step-1: UAD

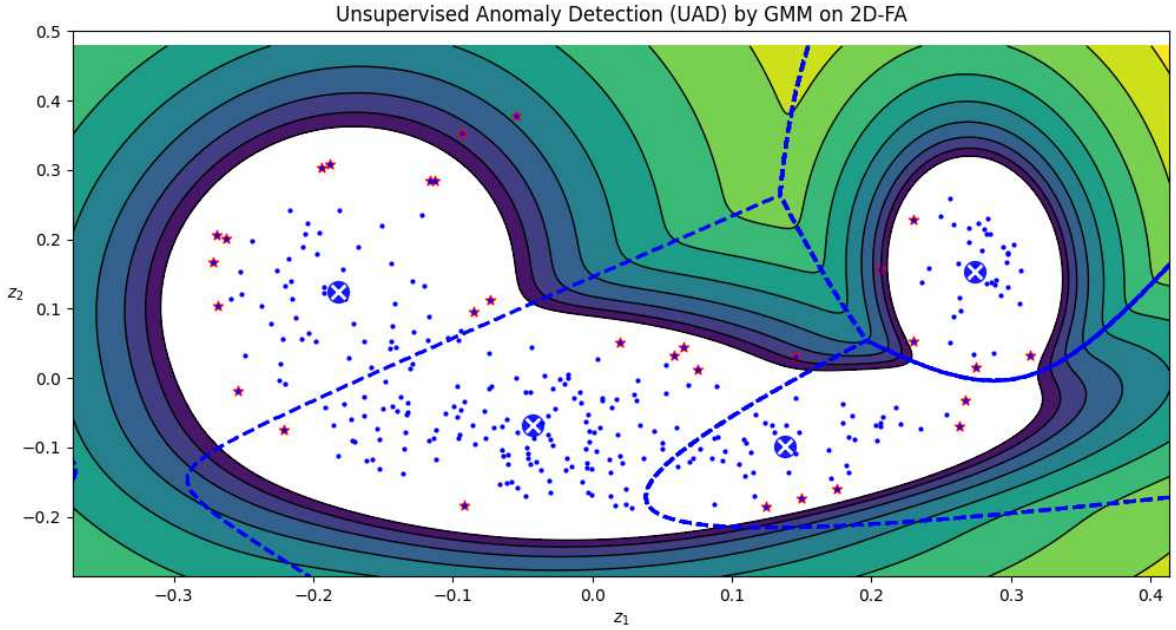
We train our unsupervised GMM model on the whole dataset (without chronological age labels) and identify anomalies by setting a density threshold such that $5\% - 10\%$ of the lowest density (log-likelihood as determined by GMM) samples are detected as anomalies. We call this set "**Anomalies**". The percentile range is an approximate guess on the overall ratio of unhealthy brain aging in the whole population, and can be adjusted based on the researcher's specific information and analyses.

Step-2: BAP

We train our best supervised BAP model using the best combination of Diffusivity Parameters (see Section 4.2) as feature tensor, and evaluate the model by 10 iterations of 5-fold CV, i.e., 50 times of training on 4 partitions (80%) and testing on 1 partition (20%). In each iteration of CV, we identify the subjects whose error is greater than 10 years. Error is defined as either positive or negative difference between the predicted age versus chronological age. Choosing 10 as the error threshold is based on the gap between our age groups, and it can be tuned depending on the distribution of the chronological age labels. These high-error (Delta) subjects are appended to a set of "**Hard Subjects**". At the end of 10 iterations, the set of "Hard Subjects" is pruned by getting the unique IDs of the repeated subjects who appear at least 5 times on the list of high-errors out of 10 iterations.



(a)



(b)

Figure 3.6: (a) 2D Projection of FA scans of Dataset-1. The "Cool" colormap represents chronological age of the data points. Each data point is a reduced FA scan to 2D, as represented by the first two principal components (z_1, z_2). (b) Unsupervised Anomaly Detection (UAD) by GMM. Detected anomalies are marked by red stars. Cluster centroids are marked by white-crosses on blue-circles. The well-separated cluster of younger subjects on the top-right is clearly recognized by the GMM.

Step-3: Intersection between Anomalies and Hard Subjects

As the final step, the intersection between the two sets, "Anomalies" and "Hard Subjects", is identified as the "Confirmed Anomalies by UAD and BAP", or simply, the "**Intersection**" set. The ratio of the cardinality of the intersection set over the maximum of the cardinality of the two sets, "Anomalies" and the "Hard Subjects", determines the "Abnormality Score" of the subjects on the intersection, ranging in $(0.0-1.0)$, as well as the agreement between the two approaches, as shown in Equation (3.22).

$$\text{Abnormality Score} = \frac{|\text{Intersection}|}{\max(|\text{Hard Subjects}|, |\text{Anomalies}|)} \tag{3.22}$$

The higher the "Abnormality Score", the higher the probability of the irregularity and abnormality of brain aging of the subjects who appear on the intersection set. The density threshold of UAD and error threshold of BAP can be tuned to maximize the "Abnormality Score" and agreement of the methods.

Now we have a methodical way of addressing "Research Question-3" because we can verify if an observed error of a BAP model is also detected by UAD as anomaly. Of course, there is still a chance that both unsupervised UAD and supervised BAP methods make errors on specific subjects, but we argue that the probability of erroneous results of our methodology is lower than relying on BAP alone which is highly dependent on the inconsistent chronological age labels. See the results of our proposed methodology in Section 4.4.3.

3.4 Deep Learning Methods

In this section, we explain our Deep Learning (DL) methods. As mentioned in Section 2.2, there are some studies that show CNNs do not outperform on DTI data compared to non-DL algorithms like SVR or even Ridge regression [3, 30]. CNNs are popular in MRI analysis as the vast majority of Computer Vision techniques are applicable in processing MRI scans as well, due to the "image" nature of MRI scans, and the similarity of the tasks. For instance, it is shown that CNNs with pre-trained weights on ImageNet, or pre-trained on an MRI dataset with a similar task,

can be used for Transfer Learning on another set of MRI images for different tasks. However, we perform experiments to investigate whether the same performance can be expected on DTI Diffusivity Parameters, which are quantitative measures with very narrow and sparse distributions close to zero, as shown in Figure 3.1, different from MRI images or Computer Vision studies.

3.4.1 DNN

Our Deep Neural Network (DNN) is a non-linear Feed Forward NN with at least three hidden layers, and non-linear activation functions for the neurons, as shown in Figure 3.7. The exact architecture of our DNN after hyperparameter tuning is provided in Section 4.5.

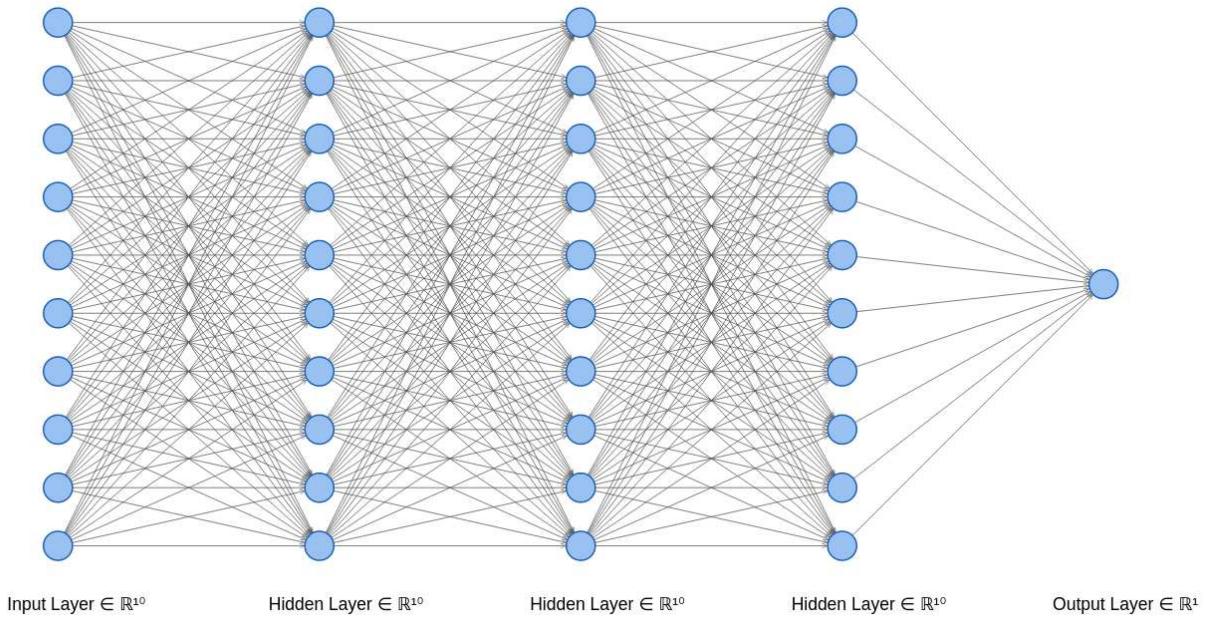


Figure 3.7: Feed Forward Neural Network. This NN architecture has 3 hidden layers, each with 10 neurons (units), and 1 output for regression suitable for BAP task. Figure is drawn with the visualization tool provided by [55].

3.4.2 CNN

We use 3D Convolutional Neural Network (CNN) like the examples that were cited in Section 2.2. The 3D convolution operator is shown in Equation (3.23).

$$S(i,j,k) = (K * I)(i,j,k) = \sum_{h} \sum_{w} \sum_{d} I(i-h, j-w, k-d) K(h,w,d) \tag{3.23}$$

where S is the output of the convolution (feature map), K is the kernel (filter), and I is the input image, each with three dimensions. The architectures of our CNN models are provided in Section 4.5.

We use different regularization techniques for both DNN and CNN to avoid overfitting, including Early Stopping, Dropout, $l1$, and $l2$ regularization, and Batch Normalization. After some pilot experiments to identify the best set of hyperparameters, we use the Adam optimizer with fixed and variable learning rate, and with and without weight decay. All the DL models are implemented by Tensorflow-GPU (see Appendix A).

3.4.3 Data Augmentation and Transfer Learning

When the training datasets are small with few samples like our study, Data Augmentation is a common technique in training DL models, as the models require a lot of training data. Typical methods for data augmentation include generating new instances of the existing data by affine transformations (like shearing, rotating, clipping, etc.) as well as adding random noise to the existing samples to add new ones. Data augmentation techniques have been used for MRI analysis [56], so we investigate if they are applicable to DTI analysis as well.

Transfer Learning is another common approach in MRI analysis, typically by using a pre-trained DL model (usually CNN) for a similar task. Transfer learning can help with the variations in different studies that are caused by scanner technology, experimental settings, etc. However, the key in transfer learning is the similarity of the tasks. For instance, using the pre-trained weights of a Recurrent Neural Network (RNN) model trained on voice data for speech recognition, and transferring it to a vision domain for performing an object detection task would most likely not yield good results. Transfer learning has been used for MRI analysis in multiple studies, and there are publicly available pretrained CNN models for transfer learning in the MRI analysis domain

[40–43], and hence we try them for BAP by using DL pre-trained models with the understanding of the differences between their tasks and our domain of DTI analysis. We also perform a transfer learning experiment by setting Dataset-1 as training data, and training the DL model on it as BAP Task-1, and then using the trained model for transfer learning on Dataset-2, both as testing data with all trainable parameters frozen, and as a pre-trained model with freezing the lower layers only and letting upper layers to be trained.

Chapter 4

Results

In the previous chapters, we explained the background, motivation, related work, our methodology on how we approached the Brain Age Prediction (BAP) using DTI data, and the specifications of our methods in detail. In this chapter, we present and discuss the results of experiments using our methods.

We begin by presenting the results of BAP as a regression problem (supervised learning with age as the target variable) using our baseline models participating in a competition to predict brain age. See Chapter 3 for a full description of our baseline models. The choice of our baseline models is based on a thorough literature review and prior published success of the models as explained in Chapter 2. We use both linear and non-linear models and we fine-tune each model's set of hyperparameters using a grid search and 5-fold Cross Validation (CV). The results of baseline models are provided and discussed in the following Section 4.1. We will use the results of baseline models as benchmarks and an empirical reference to get preliminary insights and to address our major "Research Question-1" as specified in Section 1.2.1.

Next, the winner of the competition of baseline regression models will go through further combination analysis with the Diffusivity Parameters to validate our insights, to draw more solid conclusions, and to address "Research Question-2" as specified in Section 1.2.2 on the potential preferences among diffusivity parameters with regards to their predictive power for BAP.

We then present a series of experiments using "Unsupervised Learning" and "Supervised Learning" methods to identify what we call "Anomalies" and "Hard Subjects" respectively, and we compare the two sets and get their intersection to be reviewed by an expert neuroscientist. Our goal is to propose a methodical way to address "Research Question-3" as specified in Section 1.2.3 to distinguish whether a high Delta is due to a model's poor performance or an actual unhealthy brain aging due to irregularities and abnormalities.

Moreover, we use the benchmark results and the results of combination analysis as the basis for comparison with the results of the following experiments with Deep Learning (DL) models and methods to address our "Research Question-4" as specified in Section 1.2.4. We evaluate the DL methods' performance on DTI data for the brain age prediction problem and with our small-sized datasets.

In performing our analyses as well as designing and implementing our experiments, we strive to comply with two major principles of Machine Learning: **Generalization**, and **Explainability**. To ensure our commitment to these principles, we will base our validation and evaluation methods on Generalization. We will also investigate how our models make their decisions and how their decisions can be explained so that researchers from other interdisciplinary fields (e.g., Neuroscience) can hopefully get some insights from our results. Towards this goal, we will create brain maps that visualize the relative importance of features for some of our models, and we will use them to address our "Research Question-5" as specified in Section 1.2.5 and to find the regions of interest in the brain to which our models are most sensitive.

4.1 Benchmark Results of Baseline Regression Models

As mentioned in Section 1.2.1, our first major research question is whether DTI (and specifically Diffusivity Parameters) is a suitable modality for the BAP regression problem. Moreover, we investigate whether the concatenation of the four Diffusivity Parameters (FA, AD, MD, RD) would impact the BAP results. Thus, we train and evaluate each baseline model once with each diffusivity parameter individually, and once with all four parameters concatenated as a rank-5 tensor: $(298, 1, 121, 157, 80)$ for Dataset-1 ($n = 298$), and $(94, 1, 121, 157, 80)$ for Dataset-2 ($n = 94$), to identify the winner of competition among all runs of experiments.

Recall from Section 3.1.2 that each DTI scan has the dimensionality of $(121, 157, 80)$ after preprocessing, and since we have four Diffusivity Parameters and (298 and 94) subjects, if we concatenate the four parameters altogether for each subject, each sample in training batch of Dataset-1 will be a rank-5 tensor, $(298, 4, 121, 157, 80)$, and $(94, 4, 121, 157, 80)$ in Dataset-2. The tensors

are flattened for our baseline models due to their configurations that accept rank-2 tensors. CNN models on the other hand, will receive rank-5 tensors as input.

The Mean Absolute Error (MAE) of predicted age in years of 5-Fold Cross Validation (CV) of baseline regression models applied on Dataset-1 and Dataset-2 are provided in Table 4.1 and Table 4.2, respectively. The selected hyperparameters of the baseline models and the range of the values of the grid search and randomized search to fine-tune them are specified in Section 3.2.3 and Appendix A.3. MAE was chosen as the performance metric of CV because it is consistent with recent similar studies in the literature [5, 34, 35], it is less susceptible to outliers, and it has a value in years. All models were trained on the data with reduced dimensionality by PCA (linear kernel with 0.99 preserved variance). PCA was performed with respect to training data. More details on PCA specification is provided in Section 4.1.2.

We emphasize once again that the analysis on the two datasets were performed completely independently, and we believe that this independence highlights the reliability and generalization of our results. We also emphasize that in fine-tuning hyperparameters and for model selection, we use three subsets of each dataset: training set, validation set and (unseen) test set. Models are trained on train set, fine-tuned and selected on validation set, and evaluated on unseen test set. Once the best model is selected, cross validation is repeated 10 times to account for randomness of models and splits, and the final score is aggregated over all iterations (mean MAE of 10 full iterations of 5-fold CV).

As can be seen in the results, ensemble methods Random Forest and Extreme Gradient Boosting (XGBoost) outperform other models overall across both datasets, although Ridge particularly performs well with FA, especially for the smaller dataset, Dataset-2 ($n = 94$). Random Forest is the winning model of Dataset-1 ($n = 298$) for all Diffusivity Parameters. Although XGBoost results rank second on Dataset-1, its performance is very close to Random Forest which is a confirmation of better performance of ensemble methods over other baseline models, especially if compared to linear models. For Dataset-2 both ensemble models perform relatively well, but there is no absolute winner, which is another indication of the challenge with smaller datasets.

Table 4.1: Results of Baseline Regression Models on Dataset-1

	FA	AD	MD	RD	ALL
Ridge	7.58	12.42	12.60	12.57	7.49
Linear SVR	12.30	10.65	10.66	10.70	12.04
Non-Linear SVR	7.85	8.85	9.10	9.75	7.99
NN (Lin. MLP)	13.21	12.90	13.17	12.76	12.13
XGBoost	6.74	7.21	7.43	7.14	6.81
Random Forest	**6.58**	**7.01**	**7.13**	**6.96**	**6.43**

Cross Validation MAE of Test Sets

Table 4.2: Results of Baseline Regression Models on Dataset-2

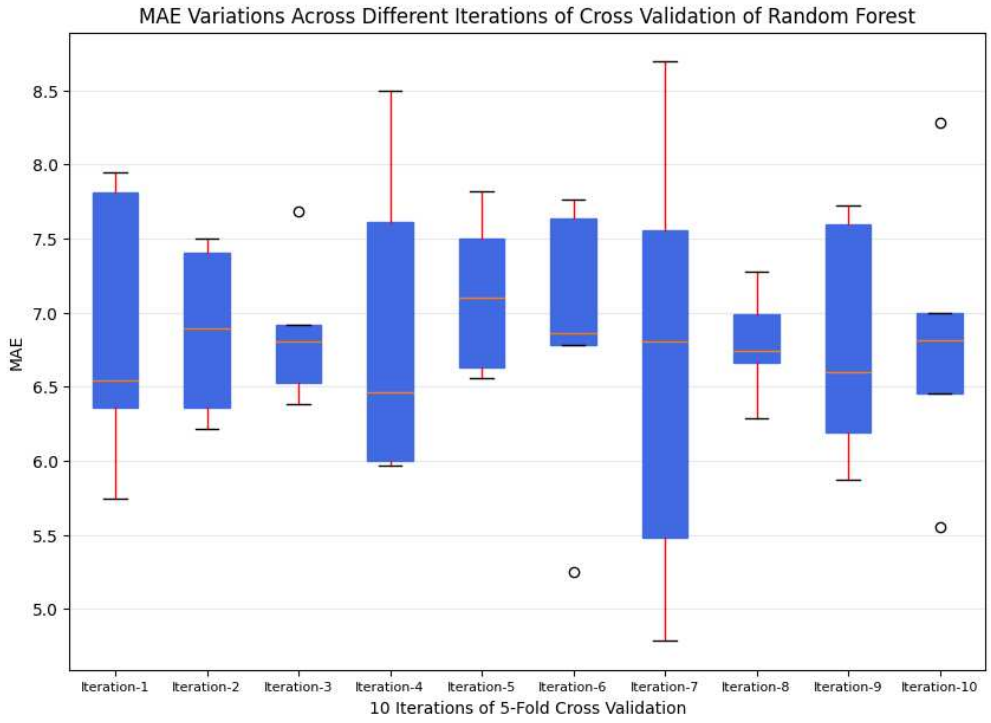
	FA	AD	MD	RD	ALL
Ridge	**8.70**	15.89	15.79	15.92	10.02
Linear SVR	19.12	15.60	15.25	15.31	17.43
Non-Linear SVR	11.55	15.23	13.17	12.43	12.27
NN (Lin. MLP)	15.91	16.88	16.28	17.77	16.88
XGBoost	10.58	8.71	**8.48**	**8.84**	11.21
Random Forest	12.78	**8.25**	9.62	9.96	**9.93**

Cross Validation MAE of Test Sets

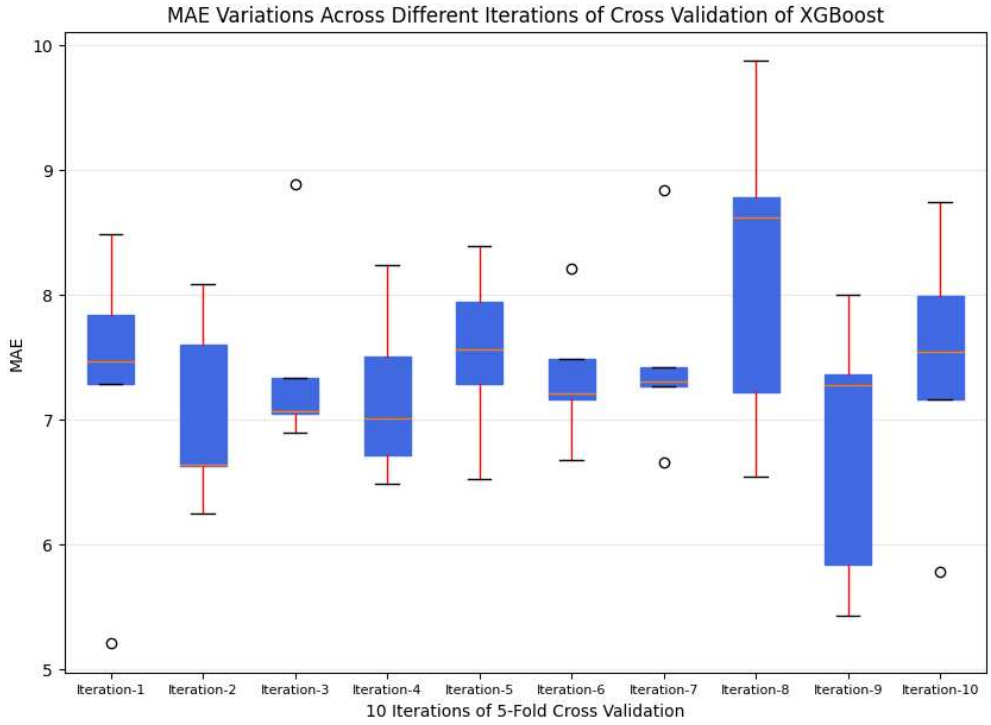
XGBoost has a more consistent and reliable performance across different iterations of repeated CV, as well as across different hyperparameters of preprocessing steps (scaling and dimensionality reduction). Random Forest on the other hand, has a higher variance, but outperforms on Dataset-1 on average. In conclusion of the baseline model selection, while we suggest to researchers that they try different models and evaluate their performance for their particular dataset, we recommend XGBoost as a good starting prototype model to get baseline results more quickly and reliably. We select XGBoost for the remaining experiments in this section (Figure 4.2) and the next Section 4.2 due to its higher consistency, reliability, and scalability as well as less variability due to preprocessing steps. For the following sections in this chapter though, we try both ensemble methods, XGBoost and Random Forests, and report the best performance of them.

Figure 4.1 shows the variations of MAE across different iterations of 5-Fold CV for Random Forest and XGBoost, respectively, when they are trained on AD+RD scans of Dataset-1 (see Section 4.2 for the reason as to why this combination is chosen). As mentioned above, while Random Forest provides lower MAEs, XGBoost performance is slightly more consistent, and has lower variance overall. However, both models perform pretty well overall, and the standard deviation of their MAE across different iterations of CV is typically less than or very close to 1.0 year.

Despite the facts that Dataset-2 is almost $1/3$ of the size of Dataset-1 ($n = 94$ vs $n = 298$), and that there are some differences between the two datasets in terms of scanner technology, acquisition protocol, and age/gender distribution of subjects (see Section 3.1), the selected model's performance on both datasets is comparable with recent BAP results in the literature using ML/DL models applied on other modalities (mainly structural MRI) [5, 34, 35, 39]. As mentioned in Section 2.2, recent studies have compared the performance of popular models like Support Vector Regression (SVR), Ridge Regression, Neural Networks (Regression MLP), and multiple other ML/DL models to provide guidance for model selection for BAP problem [4–6, 29, 39]. Typical MAE results of those studies vary in the range of about $2.5 - 7.7$ years [5, 34, 35, 39] and in some studies more than 11 years [57], and hence our comparable results suggest that DTI can be used for BAP.



(a)



(b)

Figure 4.1: MAE Variations across Different Iterations of 5-Fold Cross Validation. (a) Random Forest. (b) XGBoost. While Random Forest provides lower MAEs, XGBoost performance is slightly more consistent, and has lower variance.

We interpret our result with caution in reference to the "**No Free Lunch (NFL) Theorem**" in Machine Learning [58], which states that the performance of different models depends on the specifications of the problem as well as the characteristics of datasets used for training, and hence there is no single best model for a certain ML task that is guaranteed to always outperform other models. NFL reinforces the fact that a researcher may train different models on their particular problem and data before selecting the best model. Since our study concentrates on two small-sized DTI datasets, it is difficult to reliably generalize our results and insights for model selection. However, we have shown in this study that DTI has predictive power for BAP even with few training samples.

As can be seen in Figure 4.2, our selected model performs well overall; however, it is worth investigating why the model's error is so high for some subjects, particularly younger subjects (chronological age less than 40), while its performance is relatively stable for older subjects for all four parameters and their concatenation in both datasets. The outliers in our selected model predictions might be due to the skewed nature of the ground truth (chronological age values as the target variable) as can be seen in the age histograms of both datasets in Figure 3.2.

In other words, the distribution of age is not uniform with respect to different age groups (young/middle-aged/old), and the ratio of young subjects in the dataset, whose age is less than 40, is significantly low as can be seen in Figure 3.3. Thus, the training data is biased with respect to age groups with a heavy bias in favor of the "over-60" subjects.

We did not find any clear pattern or distinction between the two genders (male vs female) in BAP results. The small size of our datasets and imbalanced distribution of genders restricted our ability to draw reliable conclusions for gender differences.

It is worth mentioning that Root Mean Square Error (RMSE) is usually higher than Mean Absolute Error (MAE) (though not way higher) in our experiments although they are both in years, and that is because of the sensitivity of the Mean Squared Error (MSE) metric (and consequently RMSE) to outliers which do exist in both datasets as the younger subjects (< 40) may be considered as outliers within the training data by models given the age distribution.

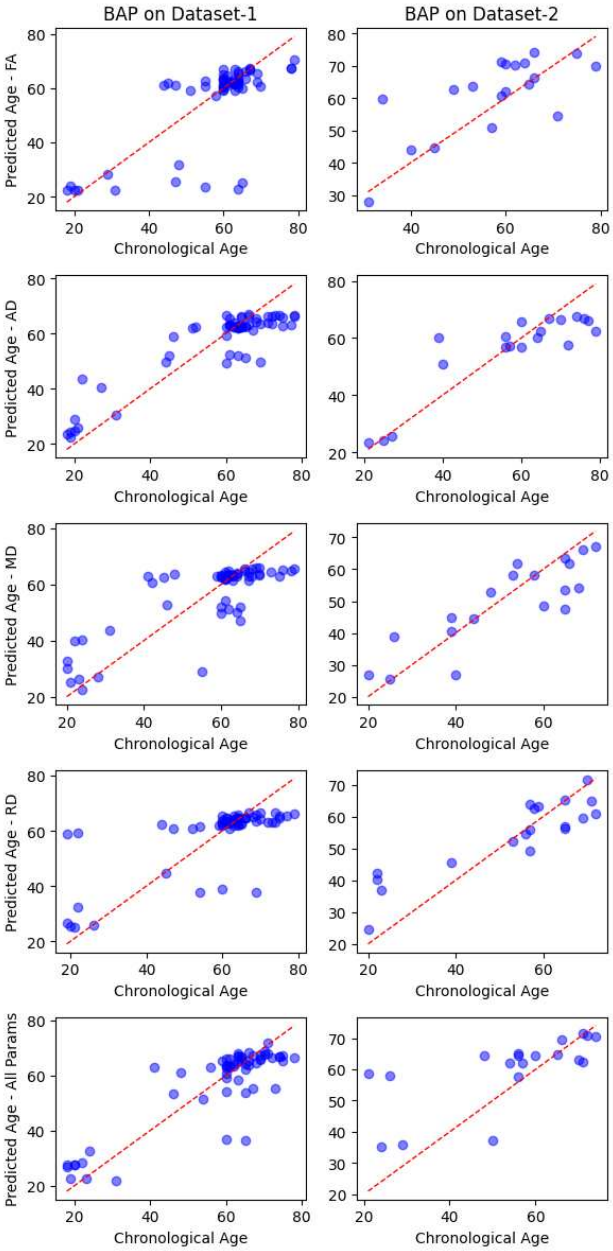


Figure 4.2: Brain Age Prediction Plots of The Selected Baseline Model (XGBoost) on Test Sets of Dataset-1 and Dataset-2. The plots depict Predicted Age vs Chronological Age when the model is fed by Diffusivity Parameters, and their concatenation (FA+AD+MD+RD) at the bottom row. The plots on the left column are for Dataset-1 and on the right column for Dataset-2. While the plots vary for different parameters, they have two things in common, overestimation of age for younger subjects and underestimation for older subjects, a known issue in the literature as mentioned in Section 2.3, which we will address by our proposed methodology in Section 4.4.3.

The error analysis of our results leads to the fundamental challenge in interpreting the results of BAP using ML/DL methods as we framed in our "Research Question-3" which is repeated below:

Should a researcher interpret ML/DL model's error in BAP as Delta, i.e., an indication of an actual unhealthy brain aging, or the model's poor performance, and how can one distinguish the two possibilities methodically?

To address the above question and challenge, we identify these so-called outliers with high MAE and call them "**Hard Subjects**". We perform specific error analysis on "Hard Subjects" and we further investigate whether we can gain any insights from their high MAE by comparing them with the set of "**Anomalies**" obtained by our Unsupervised Learning methods. Due to the novelty and importance of this methodology, we assign a separate section to describe it and discuss its results (see Section 4.4.3). The sensitivity of this challenge is even more for middle-age group $(40 - 60)$ with fewer samples which also have higher MAE more often. They might be relatively more prone to unhealthy aging and the negative impacts and early onset and symptoms of neurological conditions.

Another challenge is that not only is our data skewed with respect to age groups, but also the data is skewed within each age group as well. For age less than 40 it is left-skewed, i.e., more samples are in $18 - 23$ age range than $23 - 40$, and for "over-60", also left-skewed with more density in $60 - 67$ range (full-range of "over-60" is $60 - 79$). This skewness makes the training process for ML/DL models even harder. Thus, interpreting the results and error analysis should be done with caution and we should take into account that the errors might be due to imbalanced datasets with respect to gender and age groups, as well as the small size of datasets.

Since we focus on model selection in this section, we leave the differences in the results of Diffusivity Parameters and their concatenation to the following section which addresses our "Research Question-2". However, it is worth noting that the concatenation of four parameters seems to serve the prediction task at least for Dataset-1. This concatenation has a price which is higher dimensionality and exponentially increased runtime of experiments if dimensionality reduction methods

are not used. We discuss the impacts of dimensionality reduction in detail in the following Section 4.1.2.

To make our results more generalized and less impacted by overfitting, we have reported the average of multiple runs of 5-fold Cross Validation for each model, but it is worth noting that we have obtained MAE scores as low as 3 years in some iterations of our experiments. Moreover, for certain age groups (over 40 and over 60) our average MAE scores are in the range of $4 - 5.5$ years (see Section 4.3).

We argue that the novelty of our study has an important aspect. To the best of our investigation in the literature, we have not found any studies that use DTI measures (Diffusivity Parameters) individually and specifically for BAP. Related studies have used DTI combined with other modalities for BAP [1, 3], and there are also studies that have investigated the changes in White Matter (WM) tracts in aging brain over life span [2, 25]. As stated in our "Research Question-1" in Section 1.2.1, our concentration in this study is to check whether DTI as a single modality can be used for BAP. The significance of DTI data and specifically DTI Diffusivity Parameters is that they reflect microstructural changes of WM tracts in aging brain.

Based on the results of our baseline models, we conclude that the answer to our "Research Question-1" is positive. We will investigate further and address multiple additional research questions in the following sections.

4.1.1 Impacts of Scaling

As mentioned in Section 3.1.4, standardization and normalization of DTI data should take into account multiple considerations as we do in our preprocessing methods. We have investigated the impacts of standardization and normalization (broadly referred to as "scaling" in this document) of DTI measures on the BAP results.

In summary, we found that Diffusivity Parameters (and especially FA) are highly sensitive to scaling. This sensitivity has both positive and negative impacts on the results as follows. When training SVR and NN models after dimensionality reduction with PCA applied on FA scans, scal-

ing significantly improve the results. On the other hand, scaling often hurt the results when AD, MD, and RD are used. Thus, the results of SVR and NN models are obtained "without scaling" AD, MD, and RD values after dimensionality reduction and "with scaling" for FA as model's error would be too high otherwise. Ridge results are hurt with scaling for all four parameters as well as their concatenation, so Ridge results are reported without scaling.

On the other hand, when training XGBoost models after PCA, scaling did not significantly impact the results in any way (sometimes slightly hurts, sometimes slightly improves, which might be due to just randomness). So the results of XGBoost models are obtained without scaling the data after dimensionality reduction. This insensitivity to scaling is another advantage for XGBoost.

We also found that normalization to $[0-1]$ range is better than standardization. This could be due to the fact that normalization does not distort the distribution of Diffusivity Parameters' values and just scales it while preserving the original distribution. Since AD, MD and RD values range from negative to positive values, we expected that scaling to $[-1.0-+1.0]$ might be a better choice than $[0.0-1.0]$ range, but it turned out that $[0.0-1.0]$ is actually better.

We investigated why FA is more sensitive to scaling after PCA. Figure 4.3 sheds some light on this observation. As can be seen in the distribution plots, AD, MD, and RD values range and scale are almost preserved while FA values significantly change after PCA (from $[0.0-1.0]$ range to $[-26.88-30.41]$), and this could be the reason as to why FA should be scaled after PCA for some models whereas other parameters do not require scaling even after PCA. Compare these plots with Figure 3.1 which shows the distributions before applying PCA.

The impact of FA scaling or lack thereof is more noticeable when it is concatenated with other parameters when their differently-scaled values are mixed.

In conclusion, we recommend that each parameter should go through suitably selected and tuned preprocessing steps which are also aligned with the used models. We found that applying the same preprocessing steps on Diffusivity Parameters would negatively impact the BAP results.

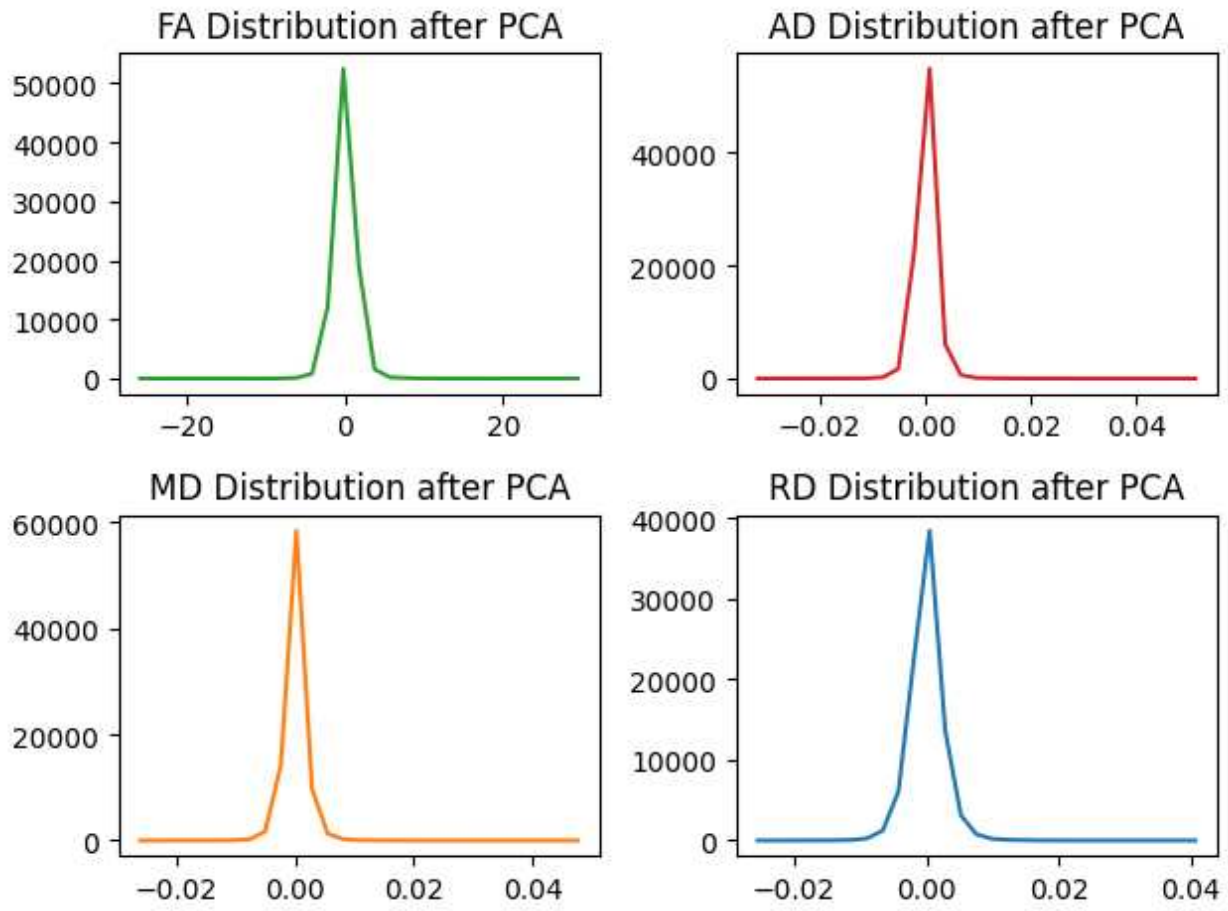


Figure 4.3: Distribution of Diffusivity Parameters (FA, AD, MD, RD) values after PCA in Dataset-1. PCA is performed with a linear kernel while preserving 0.99 of variance. Compare these plots with Figure 3.1 which shows the distributions before applying PCA.

4.1.2 Impacts of Dimensionality Reduction

We use Dimensionality Reduction methods (linear PCA and kPCA with non-linear kernels RBF and Sigmoid) for two purposes, to reduce training time, and to boost BAP model's performance. We found that both goals are served using PCA and kPCA applied on DTI Diffusivity Parameters.

Recall from Section 3.1.2 that each sample of DTI scans has the dimensionality of $(1, 121, 157, 80)$ for a single parameter and $(4, 121, 157, 80)$ for concatenation of four parameters, and hence, when they are flattened to be fed to ML models, the feature space has the dimensionality of $1,519,760$ for a single parameter and $6,079,040$ for concatenation. This means that the model has to adjust about $1.5M$ trainable parameters when working with each of FA, AD, MD, or RD and $6M$ trainable parameters working with their concatenation.

We found that dimensionality reduction reduces BAP models' training time severely and exponentially. While training and cross validation of baseline models with a single diffusivity parameter as training data on High Performance Computing (HPC) processors with high-memory capacity and GPU-Accelerated capability take up to several hours, training them after PCA takes only a few minutes. Moreover, MAE results are usually lower, or at least are very similar to the results without dimensionality reduction. The effect of dimensionality reduction is even more significant when combination of Diffusivity Parameters are used as training data.

We also found that non-linear kernel PCA (kPCA) sometimes works better than linear PCA, and hence kPCA needs hyperparameter tuning as well. We used RBF and Sigmoid kernels, and we tuned their gamma and number of principal components hyperparameters. While we report the results of baseline models with linear PCA by preserving 0.99 variance for consistency as well as minimizing information loss, we use both PCA and kPCA and tune them for the best results of the following sections. Using 0.99 preserved variance in PCA yields 290 features after dimensionality reduction for Dataset-1 and 90 features for Dataset-2 (close to their number of samples).

Figure 4.4 shows how many principal components are required to preserve 0.99 variance ratio. This requirement is important to minimize information loss (reconstruction error). The plots

also show that Diffusivity Parameters are quite similar in terms of required number of principal components to preserve variance ratio. The numbers for both datasets show how significant the dimensionality reduction is, which in turn indicates the data is so sparse that with a minimized information loss, its dimensionality can be reduced from approximately $1.5M$ to 290 for Dataset-1 and 90 for Dataset-2.

Interestingly, we noticed that FA is more sensitive to PCA and kPCA tuning than other parameters just like it is more sensitive to scaling.

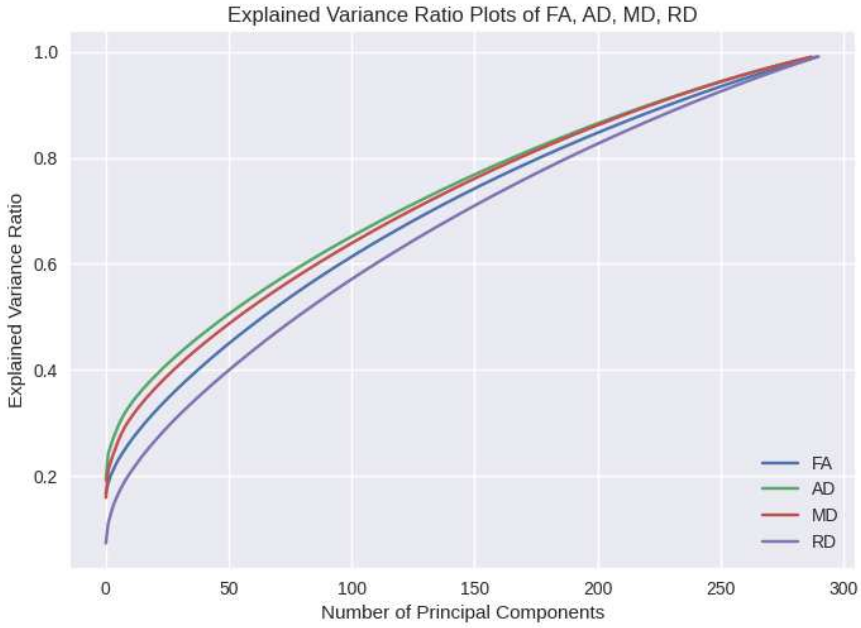
The other advantage of dimensionality reduction is that we can project the DTI feature space in 2D as we do in Section 4.4.1 to gain more insights about the feature space and its implications.

4.2 Feature Selection: Which Diffusivity Parameter is Better?

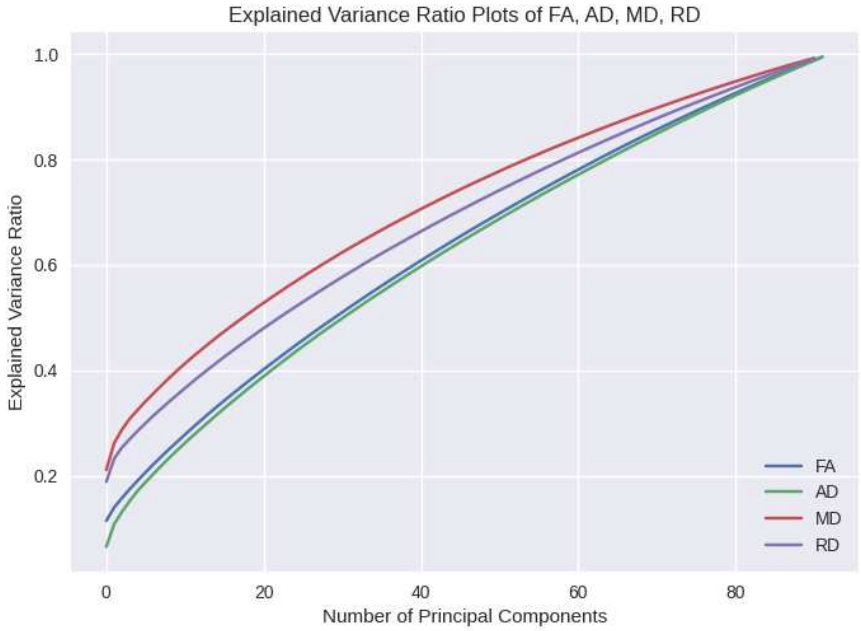
In this section, we design and perform experiments to answer our feature selection "Research Question-2" as stated in Section 1.2.2. The more verbose and complete phrasing of the title of this section is: Which of the four Diffusivity Parameters or which combination(s) of them would be better for BAP? In other words, our research question is not merely which parameter would be better for BAP. The question should also take into account which combination(s) of parameters might be better for BAP, and that is what we will investigate by a combination analysis over all Diffusivity Parameters. Recall from Section 2.2 that Diffusivity Parameters differ in the aging brain over life span [21, 25], and hence they may reflect microstructural changes in White Matter (WM) differently. Therefore, each parameter may have the potential to contribute to BAP problem.

The winning model of the competition which is XGBoost goes through the next round of training and evaluation on all possible combinations of (FA, AD, MD, RD) scans in addition to training and evaluation on individual (FA, AD, MD, RD) scans, as follows:

- $\binom{4}{4}$ one combination concatenating all four parameters (FA+AD+MD+RD),
- $\binom{4}{3}$ four combinations of three selected parameters like (FA+AD+MD, FA+MD+RD, ...),
- $\binom{4}{2}$ six combinations of two selected parameters like (FA+AD, FA+RD, ...),
- $\binom{4}{1}$ four individual diffusivity parameters (FA, AD, MD, RD).



(a)



(b)

Figure 4.4: Explained Variance Ratio Plots of FA, AD, MD, RD. (a) Dataset-1. (b) Dataset-2. The plots show how many principal components are required to preserve a certain variance ratio. As we aim to preserve 0.99 of variance after dimensionality reduction to minimize information loss, Dataset-1 needs roughly 290 principal components, and Dataset-2 requires about 90 principal components, close to their sample size, to minimize information loss. The plots also show that Diffusivity Parameters are quite similar in terms of required number of principal components to preserve variance ratio.

By this thorough investigation and full combination analysis, we are able to address "Research Question-2" as specified in Section 1.2.2 by identifying which diffusivity parameter or which combination(s) of diffusivity parameters would be a better choice for Brain Age Prediction (BAP) as a regression problem in a supervised learning manner. As stated above, the main justification for this combination-based analysis is that each of the diffusivity measures (FA, AD, MD, RD) captures the microstructural characteristics of the White Matter differently (see Section 2.1), and hence each parameter may contribute separately or in combination and correlation with features obtained from other parameters for brain age prediction. Research has shown that combinations of (FA, AD, MD, RD) may improve the results of DTI analysis because each diffusion measure shows different changes in aging brain microstructure [21,25]. Thus, a thorough analysis is required as we do here to make conclusions about "Research Question-2" as specified in Section 1.2.2.

The full results of the combination analysis to address "Research Question-2" are provided in Table 4.3, Table 4.4 and Table 4.5.

Table 4.3: BAP MAE Results of Combination Analysis - Single Parameters

	FA	AD	MD	RD
Dataset-1	6.74	7.21	7.43	7.14
Dataset-2	10.58	8.71	8.48	8.84

Table 4.4: BAP MAE Results of Combination Analysis - Two Parameters

	FA+AD	FA+MD	FA+RD	AD+MD	**AD+RD**	MD+RD
Dataset-1	6.76	6.74	7.18	7.53	**6.59**	7.26
Dataset-2	10.47	11.39	14.43	9.27	**7.59**	9.63

Table 4.5: BAP MAE Results of Combination Analysis - Three and Four Parameters

	FA+AD+MD	FA+AD+RD	FA+MD+RD	AD+MD+RD	FA+AD+MD+RD
Dataset-1	7.00	6.97	7.23	7.31	6.81
Dataset-2	10.53	10.32	11.32	10.76	11.21

Concatenated AD+RD outperforms all other combinations of Diffusivity Parameters for both datasets. Interestingly, the difference in MAE of the two datasets for AD+RD is only one year. Getting close and consistent BAP results for both datasets reinforces our conclusion. In addition to combination analysis, we observed that the feature selection of AD+RD data is scalable and consistent. When we pool both datasets together, AD+RD is also the winning combination with very close results to each dataset individually.

Please note that the results of the combination analysis are provided in three tables due to the page-width-limitation, but the three tables should be considered as one table, and that is why there are bold numbers in Table 4.4 only, because the bold numbers of Table 4.4 apply to all three tables, meaning AD+RD is the best combination across all combinations in the three tables.

AD and RD are the best single parameters among the four Diffusivity Parameters considering the results of both datasets, so it makes sense that the combination of them is the winner as their total predictive power is the superposition of each parameter's contribution in BAP. RD increases in WM with Demyelination and Axonal degeneration which may happen in the aging brain [11], so that may explain its success with BAP. FA alone also seems a good choice, and MD ranks fourth as it is mostly redundant with AD and RD, aligned with the findings of other recent studies [25].

Another interesting observation of the results is that the combination of AD+RD used for training models outperforms AD and RD when they are used individually. While the results in Table 4.3 and Table 4.4 are aggregated results of repeated iterations of 5-fold cross validation, it is worth noting that the best MAE result across all experiments and iterations of CV is 3.88 years for Dataset-2 ($n = 94$) which is obtained with AD alone, reduced by kPCA (kernel=RBF, number of components=10, gamma=0.05) followed by RD which is the second best choice when it is used individually. This shows the significance of AD in Brain Age Prediction along with RD.

Performing BAP with 2D feature space of AD+RD (PCA using only two principal components of reduced AD+RD) gives MAE results that are very close (and in some iterations lower) to PCA with 0.99 preserved variance, meaning with AD+RD a dimensionality of 2 is sufficient (and in some iterations better) for BAP using our selected models. Figure 3.5 may explain the reason for

these results, as the plot shows that by preserving only 0.2 of variance, MAE results are comparable and remain fairly stable as the variance ratio increases to 0.99. The plot is generated by training and testing the BAP model on AD+RD features as the winning combination.

On the other hand, FA, when it is used alone and not in combination with other parameters, seems to be a good choice; however, we found that the results of FA often varies and as mentioned in earlier sections, it is highly sensitive to preprocessing steps.

The impact of feature selection and a good choice on FA, AD, MD, RD is higher on Dataset-2 than Dataset-1. We argue that with fewer samples, it becomes more important to pick the best combination of Diffusivity Parameters. This might be due to the reduced ratio of samples over dimensions for smaller datasets. For high dimensional datasets with few samples like the datasets we use in our study, our results and analysis provide good insights for feature selection and making the best choices on Diffusivity Parameters.

Another implication of our feature selection is dimensionality reduction. Concatenating four parameters is computationally expensive and an effective feature selection strategy can avoid working with huge tensors in training, evaluation, and inference of BAP models.

Finally, our finding on AD+RD's good performance indicates that both radial and axial diffusivity are important and should be taken into account in BAP using DTI. This means that the combination of all three eigenvalues $(\lambda_1, \lambda_2, \lambda_3)$ corresponding to the eigenvectors of the diffusivity tensor, may contribute to BAP (see Equation (2.4), Equation (2.5), and Figure 2.1). On the other hand, FA's individual good performance (with caution on its sensitivity to preprocessing steps) indicates that if proper preprocessing steps are taken, patterns of changes in WM tracts as measured by Fractional Anisotropy may also contribute to prediction.

4.3 Age-Group Analysis and Systematic Bias

In this section, we focus on the results with respect to the differences among three age groups as follows.

- Group-1: Subjects with chronological age less than 40 years old.

- Group-2: Subjects with chronological age equal to or greater than 40 and less than 60.

- Group-3: Subjects with chronological age equal to or greater than 60.

We want to further investigate the differences in our models' performance in three age groups as can be seen in Figure 4.2, and to address the "**Systematic Bias**" as mentioned in Section 2.3. This investigation can help us better understand our models' behavior and may potentially improve their performance by identifying the root cause of their poor performance in certain age groups (particularly overestimation of younger subjects and underestimation of older subjects).

There are two main differences in models' performance with respect to age groups. The first difference is that the models generally perform better for older subjects (60+) whereas perform particularly poorly on younger subjects (<40). The second difference in the results is that the models generally overestimate the age for younger subjects (<40) while underestimate the age for older subjects (60+). As mentioned in Section 2.3, this is known in the literature as "Systematic Bias" [15, 16]. As mentioned earlier, these differences might be due to skewness of the datasets which are biased towards older subjects (60+) as can be seen in the age histograms and pie charts (See Figure 3.2 and Figure 3.3). However, we perform the following series of experiments to get to the bottom of these differences.

The first experiment we do is filtering out Group-1 subjects (<40) and train the model on Group-2 and Group-3 using the combination AD+RD and selected baseline model to see if this change makes any difference in the model's performance. The BAP plot of this experiment is provided in Figure 4.5. The MAE drops by an average of 2 years (4.5 years for Dataset-1 and 5.5 for Dataset-2) as a result of filtering out Group-1 which is not surprising because the highest errors in the complete datasets were for Group-1. However, both plots clearly show overestimation of age for younger (in this case middle-aged) subjects and underestimation for older subjects, the "Systematic Bias" as mentioned in Section 2.3 [15, 16]. Moreover, as can be seen in the plots, there are a couple of data points in both datasets whose chronological age is in Group-2, while their predicted age is in Group-3 (MAE > 20). One justification for this observation is that middle-aged subjects in Group-

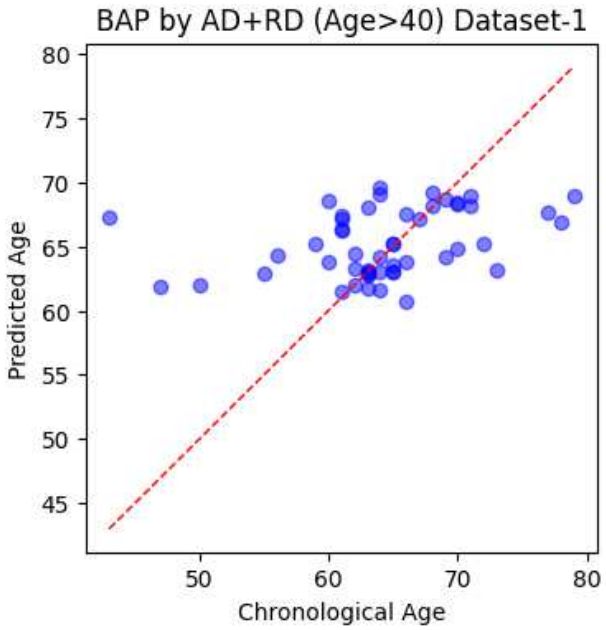
2 are as underrepresented in both datasets as younger subjects in Group-1. Another possibility is that their DTI scans might have shown a higher age than what their actual chronological age is.

Those two possibilities refer us back to the challenge we identified in Section 4.1, and as stated in our "Research Question-3", that implies further analysis is required to distinguish whether high MAE is due to the model's poor performance as a result of imbalanced dataset and/or "Systematic Bias", or a correctly detected anomaly (unhealthy aging brain) by the model as recognized through Neuroimaging scans.

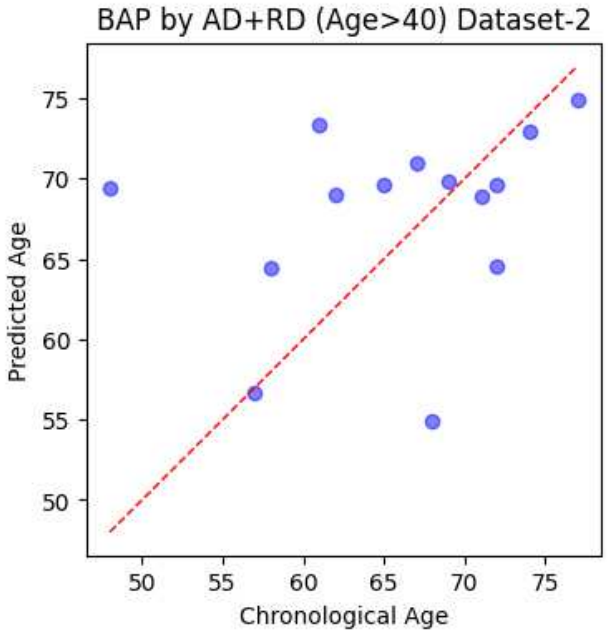
We perform another experiment, this time using Group-1 only. The MAE for Group-1 is lower than the pool of Group-2 and Group-3 (an average of 4.0 between the two datasets) despite the fact that there are only a few samples for training the model. This is probably due to the narrower age range. However, as the sample size of this experiment is too small, we do not make any conclusions about Group-1 results specifically to comply with our Generalization principle.

Our next experiment is to check whether Diffusivity Parameters have the predictive power to recognize age groups and age differences in DTI scans by multi-class classification of the three age groups mentioned above, Group-1, Group-2, and Group-3. We trained and evaluated an XGBoost classifier on RD scans and we got a balanced accuracy of 74%, with a weighted average precision of 0.80 and recall score of 0.83. Binary classification of two age groups $(50+, 50-)$ yielded a balanced accuracy of 92%, AUC score 0.96, precision and recall 0.93, and 0.92, respectively. These results confirmed our hypothesis that Diffusivity Parameters indeed have the predictive power to recognize the differences of age groups.

Next, we want to perform further analysis to gain more insights using "**Unsupervised Learning**" methods for projection of feature space, clustering of data samples, and anomaly detection to check whether we can detect suspected anomalies more methodically to address "Research Question-3".



(a)



(b)

Figure 4.5: Predicted Age vs Chronological Age Plots of BAP using AD+RD for Age > 40. (a) Dataset-1 test set ($n = 51$). (b) Dataset-2 test set ($n = 14$). The highest error, (Delta > 20), in both datasets is for two subject whose age is in range $[40 - 50]$. The MAE dropped by an average of 2 years (~ 4.5 years for Dataset-1 and ~ 5.5 for Dataset-2) as a result of filtering out younger subjects (<40). Still, both plots clearly show overestimation of age for younger (in this case middle-aged) subjects and underestimation for older subjects, a known issue in the literature as mentioned in Section 2.3, which we will address by our proposed methodology in Section 4.4.3.

4.4 Unsupervised Learning Results

Our main Machine Learning task in this study, Brain Age Prediction (BAP) is a supervised learning regression task. However, as mentioned in Section 3.3, Unsupervised Learning methods can offer many advantages in different Machine Learning tasks including our study with small datasets. Unsupervised Learning methods can work well with small datasets, and do not need labeled data, unlike Supervised Learning methods like regression and classification that require large amount of labeled training data. We can also use the whole data instead of splitting it to train/test subsets in Unsupervised Learning.

As we will provide the results and discussions in the following subsections, it turns out that Unsupervised Learning is indeed relevant and provides valuable insights for our Brain Age Prediction (BAP) study. Interestingly, the results of our Unsupervised Learning converge and agree with the results of Supervised Learning methods.

First, we begin with 2D projection of DTI data to check whether we can visually recognize any patterns in the feature space.

4.4.1 2D Projection of Feature Space

As mentioned in Section 4.1.2, our experiments benefited from dimensionality reduction in two ways: reducing training time exponentially, and boosting BAP results. In this section, we will use PCA and kPCA again, this time for 2D projection of DTI data to investigate the existence of data patterns in an unsupervised learning manner, and to observe the distribution of feature space in 2D. Another motivation for this investigation was mentioned in the previous section as identification of age groups.

Figure 4.6 shows the 2D projections of feature space for FA, AD, MD, and RD. We use PCA with the first two principal components (z_1, z_2) and we train and project by two kernels: linear and RBF ($\gamma = 0.02$). (z_1, z_2) are scaled to $(-1, 1)$ to make the plots more aligned and comparable. Each data point is colored with respect to its chronological age according to the provided colormap.

The figure is for Dataset-1 only, to avoid repetition of plots as Dataset-2 plots are similar although with fewer data points.

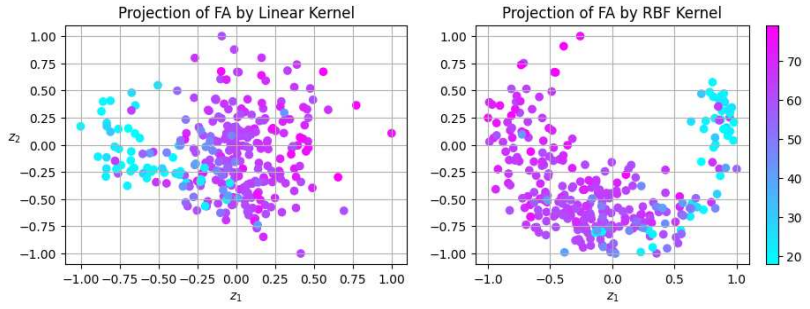
There are three interesting observations in these plots. First, data points seem to be in separate clusters with respect to the age groups we presented in the previous Section 4.3, especially on FA, AD and RD projections. Second, the plots show that there are outliers, data points with different color (age) that are far from any density and/or groups, or data points that visibly have a different color (age) from their neighbors. Third, by this dimensionality reduction technique and 2D projection, we can encode the (reduced) brain (to be precise part of the brain, and to be more precise White Matter) and represent it by just two numbers (z_1, z_2), which are the first two principal components of the PCA.

The following two numbers are the first two principal components of a reduced-dimensionality sample (2D projected) FA scan in Dataset-1:

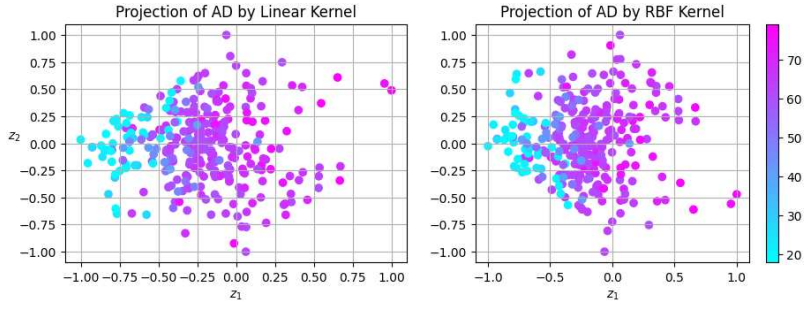
$$[0.1795, -0.0646]$$

As we will show later in this chapter, if we do an inverse transform of PCA on these two numbers and map it on the brain, we can verify that the information loss as a consequence of dimensionality reduction is minimized on the White Matter skeleton and would not negatively affect BAP, and as far as Machine Learning and Deep Learning BAP models are concerned, the information loss is negligible.

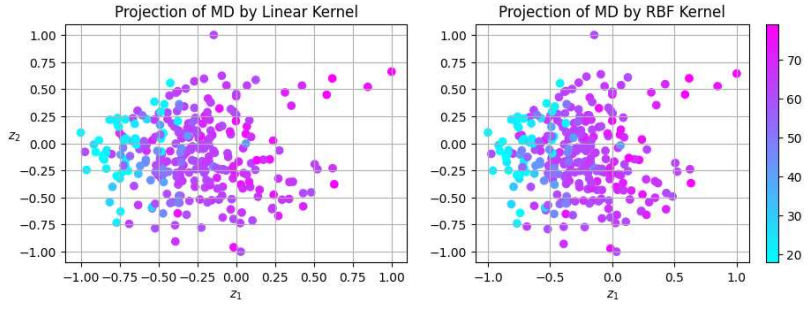
The non-linear kernel seems to work better on FA (as the age groups seem to be well separated) while there is no noticeable difference between the linear kernel and RBF for other Diffusivity Parameters. Recall that each data point on the plots of Figure 4.6 represents a subject's DTI scan reduced to 2D, a sample in the dataset, and the 2D projector (PCA) had no access to or information about the labels, chronological age of the subjects. The 2D projection of DTI feature space suggests that there might be clusters of age groups, as they can be seen in the plots based on the colormap. Thus, we will train a clustering algorithm in the following subsection to investigate this possibility.



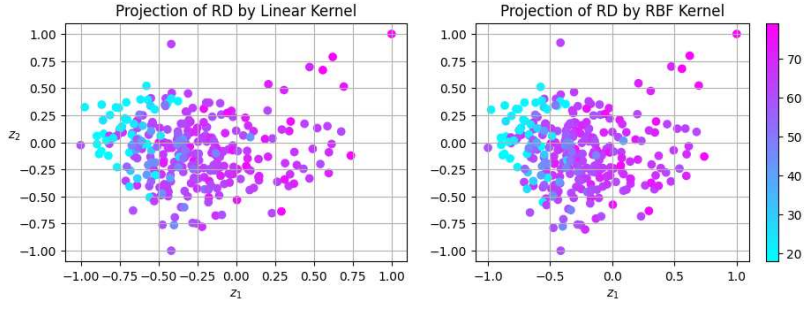
(a)



(b)



(c)



(d)

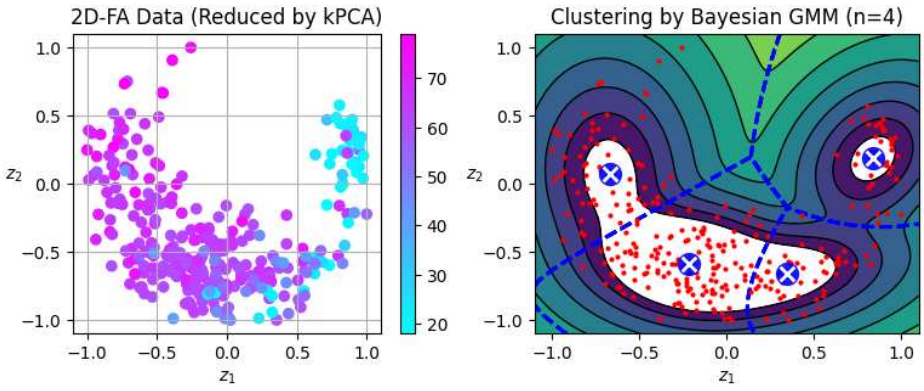
Figure 4.6: 2D Projection of Diffusivity Parameters' feature space using PCA: (a) FA (b) AD (c) MD (d) RD. Colormap represents chronological age. Each data point represents a sample and (z_1, z_2) are principal components after dimensionality reduction. The two principal components (z_1, z_2) are scaled to $(-1, 1)$. The plot on the left shows projection using linear kernel and the plot on right is projection by RBF kernel $(\gamma = 0.02)$.

4.4.2 Results of Clustering

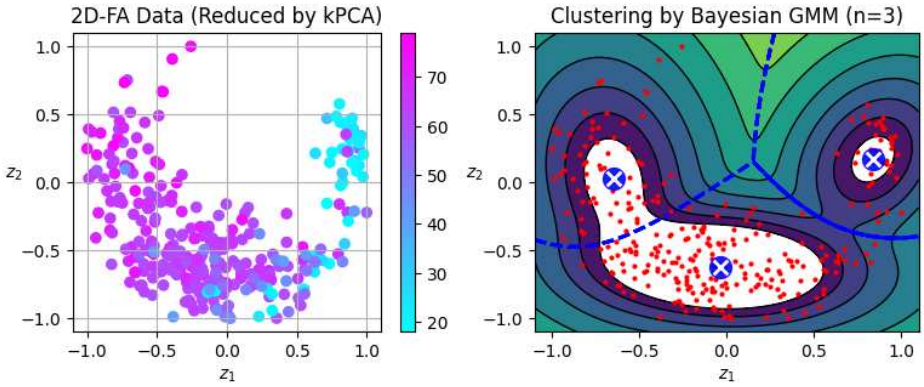
Our goal now is to find whether the clusters of similar data points, potentially the age groups, which can be seen in the 2D projection (see Figure 4.6) can be recognized by unsupervised clustering algorithms. We tried different algorithms, including K-Means, DBScan, and t-Distributed Stochastic Neighbor Embedding (t-SNE) to no success for different reasons. While K-Means and DBScan did not converge to any solution, t-SNE algorithm crashed on memory and CPU/GPU errors even on High Performance Computing GPU-Accelerated hardware that we use for our experiments. Following up on those experiments, we applied two generative models in a completely unsupervised manner: Gaussian Mixture Model (GMM) and Bayesian Gaussian Mixture Model (BGMM) which turned out to be successful applications. We argue that the success of GMM and BGMM is probably due to the Gaussian distributions of water diffusion in the White Matter (see Section 2.1), which make GMM and BGMM a good match.

The results of clustering are provided in Figure 4.7 and Figure 4.8. There are multiple observations that are worth discussing. First and foremost, clusters seem to match with age groups as younger data points (represented by blue points on the lower edge of the colormap) fall into clusters that are further from older subjects (shown by magenta on the other side of the colormap), and middle-aged group almost fall in-between those two groups.

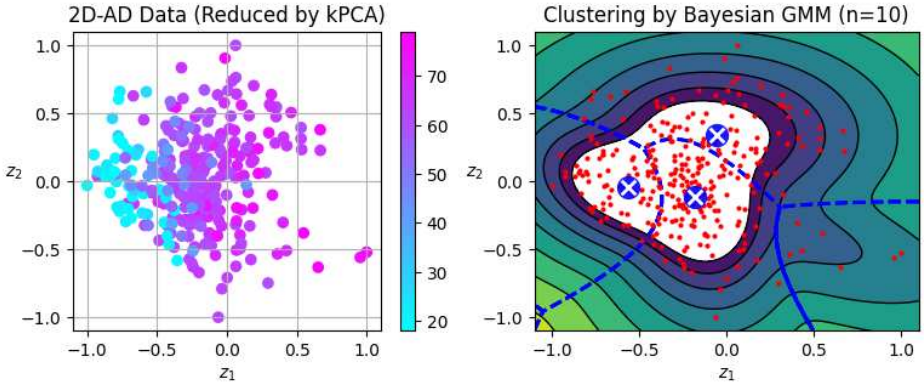
While the two algorithms GMM and BGMM converge to very similar results, it should be mentioned that the optimal number of clusters was determined by BGMM. When we set the desired number of clusters to any number higher than three or four clusters (like 10), BGMM determined that the optimal number is three (or four depending on the Diffusivity Parameters' choice) and assigned a zero score to all other clusters. Both algorithms are very robust in their results and convergence across multiple runs and iterations, and they always detect similar clusters that can be interpreted and recognized by a human user. The algorithms are specially powerful in recognizing and well-separating the young group in all clustering models using different Diffusivity Parameters. In other words, parameter choice does not negatively impact identification of age groups.



(a)

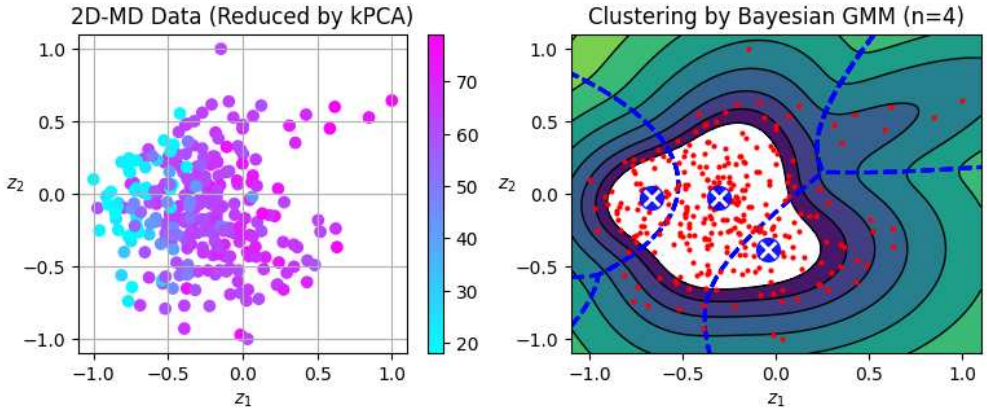


(b)

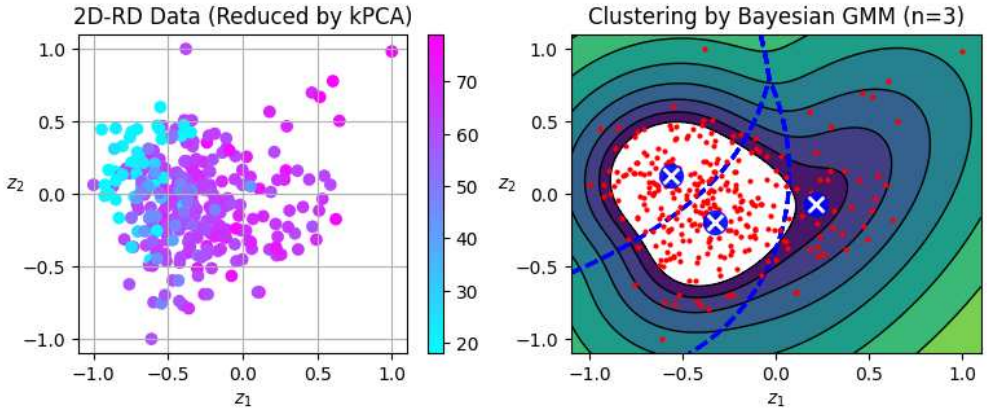


(c)

Figure 4.7: Clustering by Bayesian GMM: (a) FA with 4 clusters (b) FA with 3 clusters (c) AD with 4 clusters. Left plots are 2D projections of FA and AD. Each data point represents a sample and (z_1, z_2) are principal components after dimensionality reduction. Colormap represents chronological age. Right plots are clusters as recognized by Bayesian GMM algorithm. Cluster centroids are marked by white crosses on small blue circles. Number of clusters was determined by Bayesian GMM in a completely Unsupervised Learning manner. For FA, the algorithm determined $n = 4$ is a good number but we also plot with 3 clusters to see the effect. For AD, we set $n = 10$ but the model determined that 4 is an optimal number of clusters and assigned the scores of other clusters to zero.



(a)



(b)

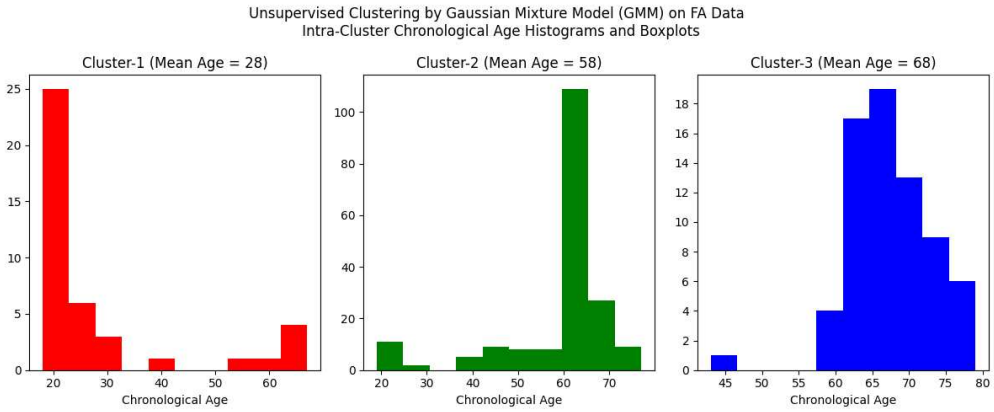
Figure 4.8: Clustering by Bayesian GMM: (a) MD with 4 clusters (b) RD with 3 clusters. Left plots are 2D projections of MD and RD. Each data point represents a sample and (z_1, z_2) are principal components after dimensionality reduction. Colormap represents chronological age. Right plots are clusters as recognized by Bayesian GMM algorithm. Cluster centroids are marked by white crosses on small blue circles. Number of clusters was determined by Bayesian GMM in a completely Unsupervised Learning manner.

Before we make any further conclusions or assumptions about the identified clusters, we need to do an intra-cluster analysis to check if the members of a cluster are actually similar in age and hence can represent age groups. The results of our intra-cluster analysis is provided in Figure 4.9.

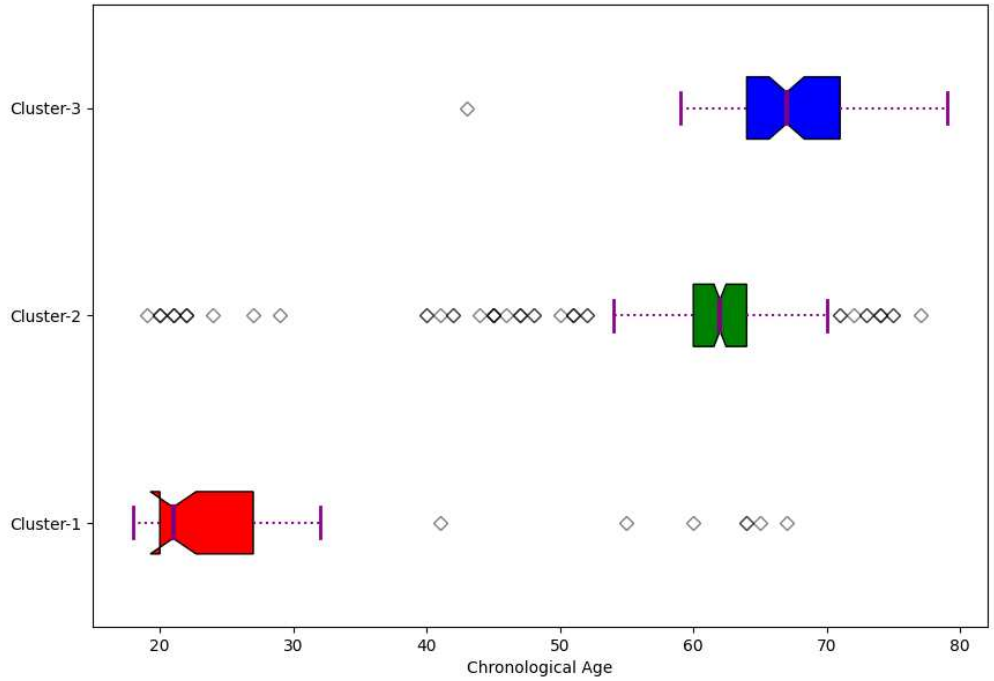
Intra-cluster chronological age distributions clearly show that the age groups are well-separated by clustering, and more importantly, the optimal number of these groups was determined by the algorithm itself. Clusters are generated by Gaussian Mixture Model (GMM) applied on FA scans in an unsupervised learning manner with no access to ground truth. The three clusters are fairly aligned with the age groups outlined in Section 4.3. The cluster of younger subjects, with a mean of 28 years old of chronological age, are further separated from other two clusters with the mean of 58 years and 68 years. In fact, all clustering models recognize this cluster better than the others. This sensitivity in recognition of young cluster helps us later in data augmentation, as our original datasets are not uniformly distributed, and we can use generated data to compensate for the imbalanced distribution of age groups.

As mentioned in Section 3.3.1, the optimality of GMM models is determined by Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) scores. The lower the AIC and BIC scores the better even with negative values. Clustering with FA and RD has the lowest AIC and BIC. Moreover, fitting GMM with 2D-FA and RD has the lowest AIC and BIC scores, meaning the dimensionality reduction actually helps clustering models. The results of GMM and Bayesian GMM are again pretty similar.

Now that we have shown the applicability and relevance of clustering algorithms which were trained in a completely Unsupervised Learning manner without access to the ground truth (chronological age of the subjects), we will provide further applications of these models by using them for "Unsupervised Anomaly Detection" and generating new samples for "Data Augmentation", as they are generative models that can generate new data.



(a)



(b)

Figure 4.9: Intra-Cluster Chronological Age Distribution: (a) Intra-Cluster Age Histograms. (b) Intra-Cluster Age Boxplots. Clusters are generated by Gaussian Mixture Model (GMM) applied on FA scans in an unsupervised learning manner with no access to ground truth. The three clusters are fairly aligned with the age groups outlined in Section 4.3.

4.4.3 Results of UAD and Our Proposed Methodology

Our next objective is using our clustering models to identify outliers, also known as anomalies. As mentioned in Section 3.3.3, this process of detecting abnormal instances in an Unsupervised Learning manner is called "Unsupervised Anomaly Detection (UAD)". Although a different form of UAD has been used in the brain MRI literature in the past [59], it is not used yet with regard to BAP to the best of our investigation, and we employ it in a novel way for the first time in the context of BAP.

UAD is an Unsupervised Learning method because labels (chronological age) of subjects are not used in training the models. UAD is particularly suitable for small datasets unlike the Brain Age Prediction (BAP) regression task that requires large amount of labeled training data. As proposed in Section 3.3.3, our methodology to combine UAD and BAP is performed in three steps as follows.

Step-1: UAD

After training GMM, we compute the log-likelihood of each sample, and we refer to it as "density" estimate. Any instance located in a low-density region is considered an "anomaly".

We need to define what density threshold we want to use for anomaly detection. Setting the density threshold determines what ratio of anomalies will be detected. For example, if we know that typically 5% of subjects have abnormal characteristics in their brain aging, we would set the density threshold to a value that yields roughly 5% of the population of subjects, meaning 5% of the instances located in areas below that density threshold. We may get too many false positives (i.e., healthy subjects that are flagged as abnormal), so we may decide to lower the threshold. On the other hand, if we get too many false negatives (i.e., unhealthy subjects that our model does not flag as anomaly), we can simply increase the density threshold. This is typically referred to as "Precision-Recall" trade-off in Machine Learning. Since we are not certain what percentage of population may have unhealthy brain aging, we set the threshold to give us roughly 10% of our dataset size as anomalies. Applying this rationale on Dataset-1 ($n = 298$) would give us 30 anomalies.

Step-2: BAP

Next, we do a Supervised Learning BAP regression task with 10 full iterations of 5-fold Cross Validation (CV) (50 times of training on train set and testing on unseen test set), and in each iteration, we identify the subjects that give us high error (absolute Delta greater than 10), and hence either positive or negative Delta (overestimation or underestimation of age) will be counted as "Hard Subject" for the regression task. 64 unique subjects that appear in at least 5 iterations of 5-fold CV are listed.

Step-3: Intersection between Anomalies and Hard Subjects

We then get the intersection of this set of "**Hard Subjects**" that we got using Supervised Learning BAP method, and the set of "**Anomalies**" that we got using our Unsupervised Anomaly Detection. Interestingly, a third of the "Anomalies" set, end up being on the list of "Hard Subjects" too with very high Delta. Of course, we could tune the density threshold of our UAD, as well as the Delta threshold of BAP, and get a higher intersection ratio or "Abnormality Score" as defined in Equation (3.22).

Further investigation by a field-expert neuroscientist professor at Colorado State University (CSU), Dr. Agnieszka (Aga) Burzynska, revealed that five out of nine (56%) of the detected anomalies' scans showed large ventricles in their DTI scans to different degrees (large ventricles indicate brain atrophy). There might be other factors involved, such as health conditions and lifestyle including but not limited to, physical activity, history of chemotherapy, low education, undiagnosed (prodromal) cognitive decline, long history of treated hypertension, and high BMI. Thus, we need further investigation on the detected anomalies, and that is listed in Section 5.3 for our future work.

The significance of the results of our proposed methodology is that it shows our Unsupervised Learning method, UAD, and our Supervised Learning BAP models converge and agree on the subjects with a gap between their chronological age and biological age, as measured by their high Delta and confirmed by the field expert assessment. Moreover, we have addressed "Research Question-3" by providing a methodical way for the error analysis of our BAP models.

Another interesting observation is that our UAD method also detects subjects whose brain age is much younger than their chronological age. These are the subjects whose age is highly underestimated by the BAP model. In one case, the chronological age of a subject identified as "anomaly" is 61, while our model predicted their age as 45.5 years, and our expert's assessment was that their DTI scan looked really good. Although we are interested in identifying unhealthy brain aging, as far as our unsupervised method is concerned, any outlier and severe deviation from normality is considered as anomaly, and the model is capable of detecting it. We should mention that the prevalence of underestimated BAP is way lower than overestimation, especially for younger and middle-ages subjects whose age is often overestimated.

Our UAD analysis provides an effective method for detecting abnormal DTI scans with no dependence on their chronological age labels. We argue that this is very important given the "Label Inconsistency" of Brain Age Prediction (BAP) task. Chronological age labels are inconsistent because they fail to represent the actual biological age of the brain for a significant portion of population, and that is the main motivation to use the BAP in th first place.

However, our argument is that the current supervised methodology for BAP which relies on chronological age labels is limited, insufficient, and biased. If a researcher wants to train the models for BAP regression task to identify high Deltas as an indication/biomarker of unhealthy brain aging or irregularities, they should ideally provide the model in supervised training with both "Chronological Age" and "Biological Age" (estimated by brain experts through Neuroimaging or other modalities and data analysis) so that the Machine Learning model can learn the differences effectively.

This is a fundamental dilemma in the sense that to train the BAP regression models, we use chronological age labels that are inconsistent themselves. At the minimum, we think that researchers should provide the BAP regression model with an additional feature like a binary flag (Healthy vs Unhealthy) so that the model can learn the differences better. Since in the common practice and current methods that supplementary information is not provided to ML models, we argue that our proposed label-independent "Unsupervised Anomaly Detection" approach can supple-

ment BAP by combining the results of the two methods as well as cross-examination and methodic verification as described in our methodology.

4.4.4 Generative Models and Data Generation

As GMM and BGMM are generative models, we can use them to generate new samples. This is especially important as our datasets are small, and we can use generated data for augmentation of training data to improve our Supervised Learning models. Figure 4.10 shows the age distribution of 1000 generated samples by GMM. Data generation and GMM training on RD data were performed completely unsupervised with no access to labels (ground truth). The labels of new generated samples were assigned using a trained Random Forest model which was trained on Dataset-1 ($n = 298$) RD scans. Then, the generated samples and their assigned labels were used for Data Augmentation of Dataset-1 ($1000 + 298$) and Transfer Learning on Dataset-2 ($n = 94$). We will discuss Transfer Learning further in the following Section 4.5.1.

To further clarify, we can perform data generation in two ways. We can use data generation with a generative model that has been trained on the data with the original dimensionality (no dimensionality reduction), or we can generate new samples with a generative model that has been trained on the data with the reduced dimensionality (e.g., 2D). In the latter case, we can use inverse transform of the PCA or kPCA model that was used to reduce the dimensionality of the training data of the generative model, and transform the new samples to the original dimensions of the dataset. Of course, if we use the latter way, there will be some information loss proportionate to the preserved variance ratio of the PCA/kPCA.

4.5 Deep Learning Results

Next, we train and evaluate Deep Learning (DL) models (CNN and DNN) to check whether they can achieve our benchmark results, and to address our "Research Question-4" as specified in Section 1.2.4. As mentioned in Section 2.2, there have been some studies that show CNNs performs similarly on DTI data compared to traditional ML algorithms like SVR or Ridge regres-

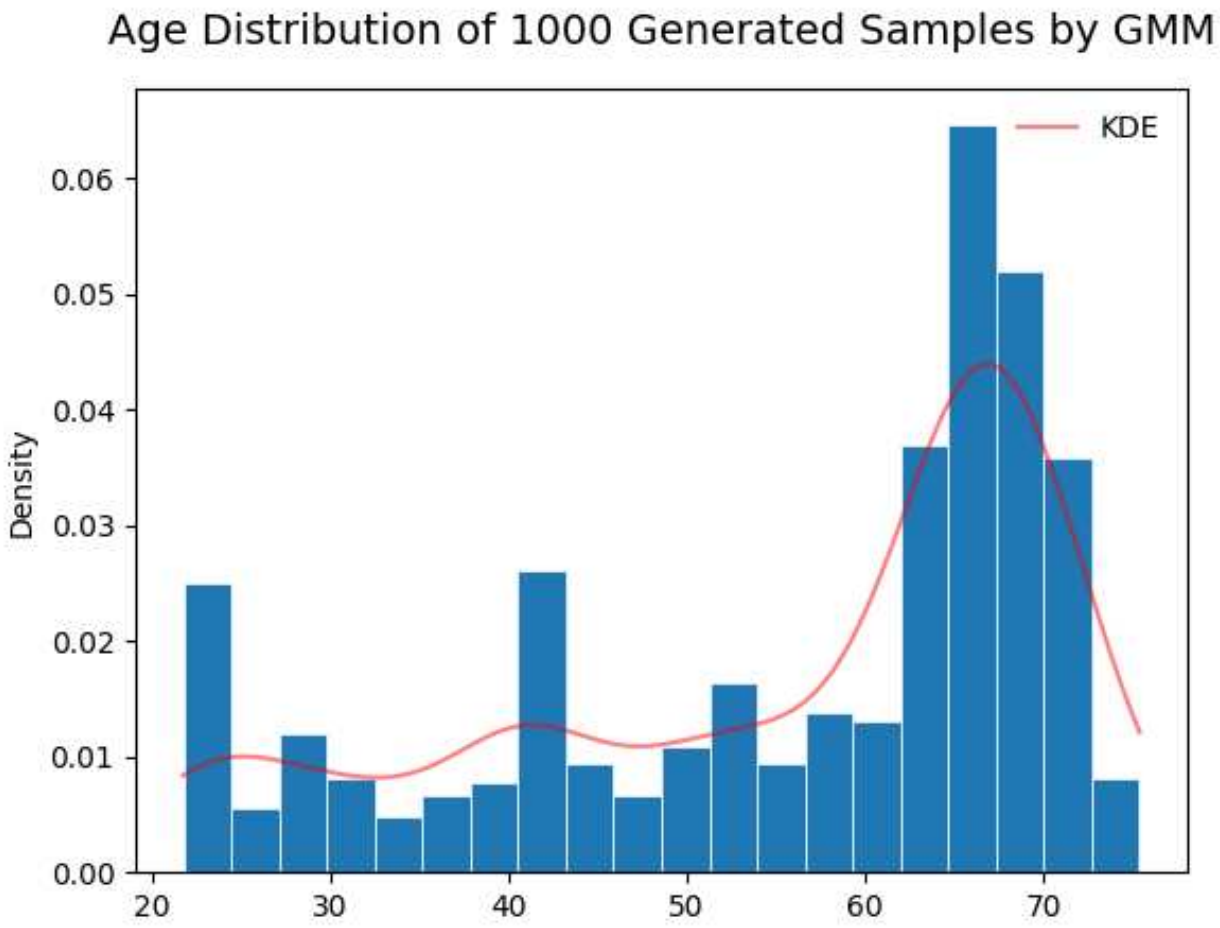


Figure 4.10: Age Distribution of 1000 Generated Samples by Gaussian Mixture Model. Data generation and GMM training on RD data were performed completely unsupervised with no access to ground truth. The labels of new generated samples were assigned using a trained Random Forest model which was trained on Dataset-1 ($n = 298$) RD scans. These generated samples were used for Data Augmentation and Transfer Learning on Dataset-2 ($n = 94$).

sion [3, 30]. However, since we use a rather different modality (skeletonized Diffusivity Parameters), we do not make any prior assumptions about the performance of our DL models and we perform a thorough investigation on the effectiveness of DL models applied on DTI data for BAP.

The architecture of Deep Neural Network (DNN) that we train and evaluate is provided in Figure 4.11. This symmetric architecture with four hidden layers resembles a stacked autoencoder as we can change the number of outputs to the dimensionality of the feature space, and then train the model to replicate itself (same features and targets). Since we have a bottleneck (10 neurons) in the middle, we can use it for non-linear dimensionality reduction (listed as future work in Section 5.3). We also train this DNN for BAP.

We fine-tuned the DNN and CNN hyperparameters and tried various settings on learning rate policies, optimizer, and initialization strategies. The best optimizer was Adam with a constant learning rate of 0.001. We used early stopping and dropout for regularization to avoid overfitting as other regularizes such as $l1$ and $l2$ as well as batch normalization did not work well.

We tried different architectures of 3D CNN for our Brain Age Prediction (BAP) task to no success, including the architecture explained in [34] which is a deep architecture with restriction on predictions. We noticed that simpler and shallower architectures work better with our data, and the best architecture for our task turned out to be VGG based on our experiments. However, we had to modify classic VGG architecture to deal with our high-dimensional data, so we added an inception module in the beginning with (1x1x1) filters for dimensionality reduction. Two VGG architectures that we use are provided in Figure 4.12 and Figure 4.13, respectively. The difference between the two is the number of VGG blocks.

The results of DNN and CNN models for Dataset-1 are provided in Table 4.6. The training of DNN and CNN was problematic and did not converge well due to few training samples, so we report accordingly (we use Dataset-2 for transfer learning, though). We used the same model evaluation and selection protocol that we applied on our baseline models. While the results of Deep Learning models are not promising, they come close to the baseline models, especially with RD and AD.

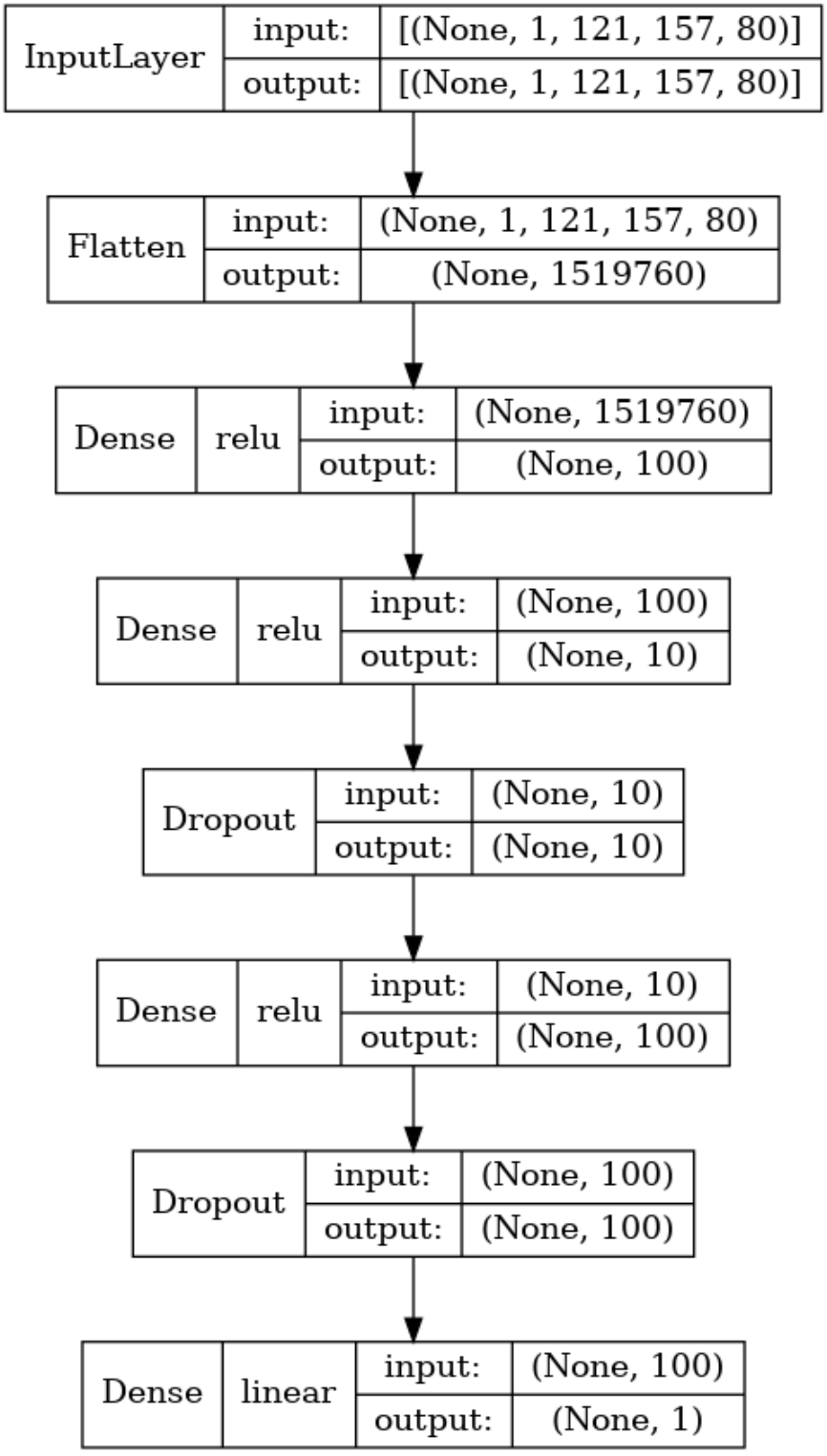


Figure 4.11: Architecture of the DNN (MLP Regressor) with four hidden layers for BAP task. The symmetric architecture of our DNN resembles the architecture of a stacked autoencoder that, with slight modifications, can be used for non-linear dimensionality reduction (latent space representation).

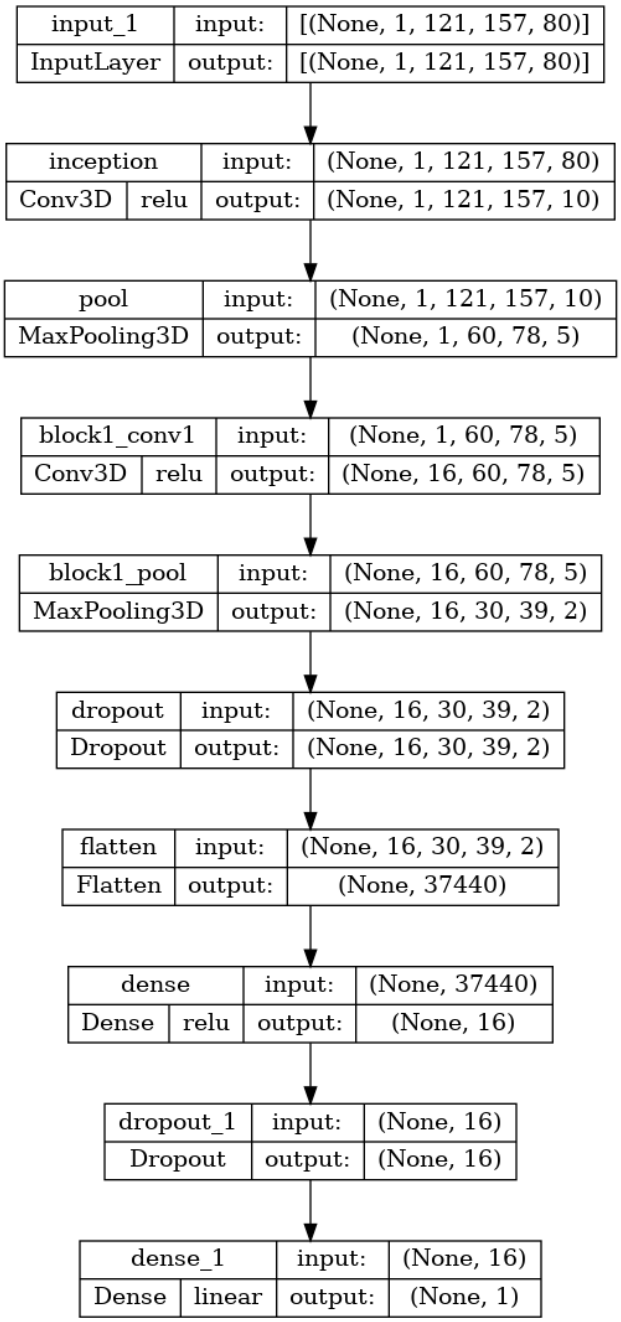


Figure 4.12: Architecture of 3D-CNN VGG-1 for BAP task, with one inception module in the beginning followed by MaxPooling3D for dimensionality reduction, and one VGG convolutional block with two Conv3D layers followed by a MaxPooling3D layer, followed by a Fully Connected (FC) layer before the final output of regression.

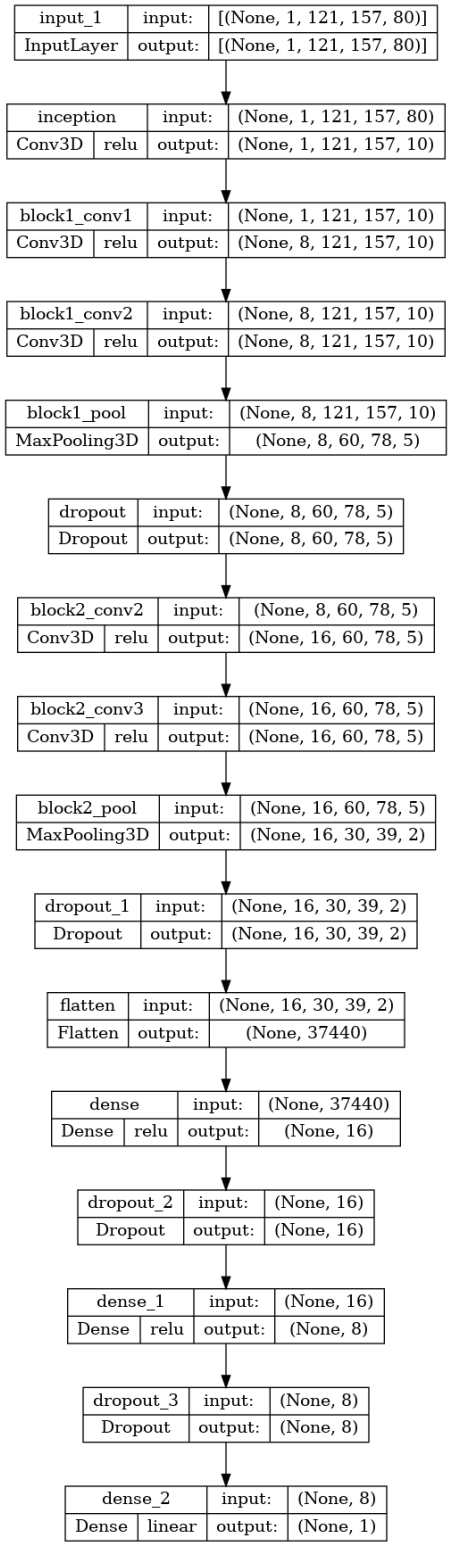


Figure 4.13: Architecture of 3D-CNN VGG-2 for BAP task, with one inception module in the beginning for dimensionality reduction, and two VGG convolutional blocks, each of which with two Conv3D layers followed by MaxPooling3D layer, followed by two Fully Connected (FC) layers before the final output of regression.

We argue that the under-performance of the DL methods has two main reasons: first, insufficient training data due to small size of our datasets, and second, the feature space and the sparsity of voxel-wise value distributions of Diffusivity Parameters (DP) which is not a good fit especially for CNN. As shown in Figure 3.1, the data modality that we work with is very different from structural MRI images that are the most relevant and successful applications of CNNs. Generalization of this argument needs further verification with larger DTI datasets.

Table 4.6: Results of Deep Learning Models on Dataset-1

	FA	AD	MD	RD	ALL
DNN	14.73	8.82	9.91	**8.52**	8.91
CNN VGG-1	13.50	9.65	12.23	8.96	11.32
CNN VGG-2	17.95	12.85	14.10	11.43	12.27
Cross Validation MAE of Test Sets					

4.5.1 Results of Data Augmentation and Transfer Learning

Since using DTI as a single modality for BAP is a novel approach, there are not many similar datasets or pretrained models available for transfer learning to the best of our investigation and research. Also, commonly used data augmentation techniques that are applicable on images for computer vision tasks (such as affine transformations) or using pretrained weights of the models trained on ImageNet did not yield good results due to lack of similarity of the tasks and data structures. Thus, we use the generated data by GMM (see Section 4.4.4) for data augmentation of training set as Deep Learning models are data hungry and our datasets are small. To investigate whether transfer learning is applicable in this problem we do an experiment once with Dataset-1 ($n = 298$), and once with augmented Dataset-1 ($n = 1000 + 298$) as training, and Dataset-2 ($n = 94$) as test set. We train our DNN and CNN (VGG-1) on Dataset-1 and evaluate it on Dataset-2 with similar experimental settings for evaluation as before (repeated 5-fold CV) . Without data augmentation, MAE is approximately 16 years whereas with data augmentation MAE drops by

5 years to approximately 11 years. This means that we freeze all layers of the trained model on Dataset-1 when we apply it on Dataset-2. RD is the best parameter for transfer learning from Dataset-1 to Dataset-2. This is promising as we often need to apply our pretrained models on new data. Dataset-2 in particular is a more recent research with an ongoing data acquisition process, so this method of transfer learning is applicable and relevant.

4.6 Brain Maps and Explainability

We started this chapter with an emphasis on our commitment to two major principles in Machine Learning, **Generalization** and **Explainability**. To explain how our BAP models make their decisions and to address our last research question, "Research Question-5" as specified in Section 1.2.5, we create brain maps that show the relevant areas on the standard brain to which the models are sensitive. These techniques are used in the literature for visualization of the results as well as explainability and interpretability of the models [16, 28, 60, 61].

The first brain map is Figure 4.14 which shows the feature importance of random forest regression model for BAP applied on the standard brain using FSL software.

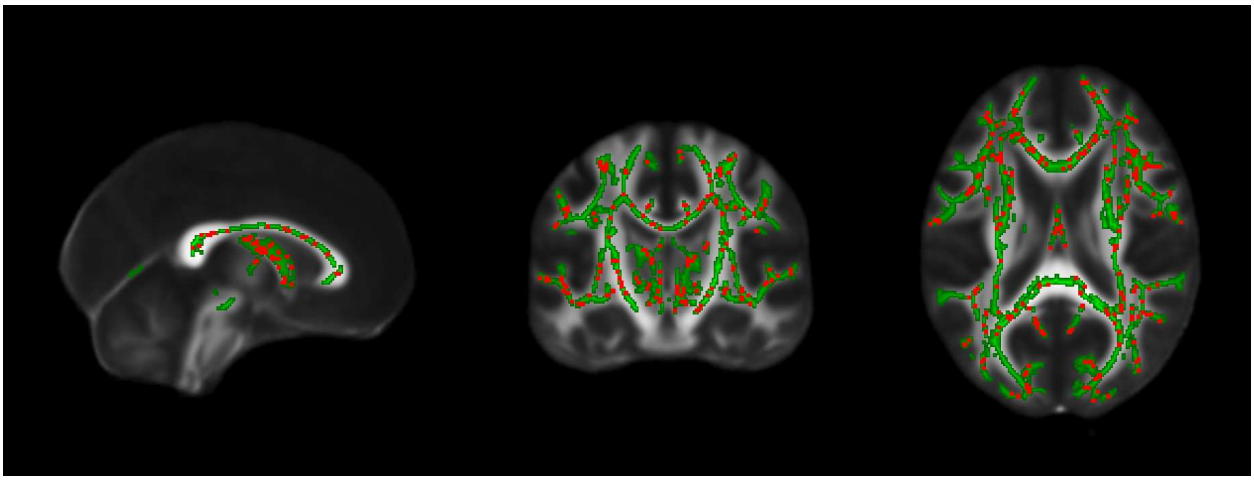


Figure 4.14: Brain Map of Feature Importance of Random Forests Trained on FA scans with no dimensionality reduction. The green skeleton is the mean FA skeleton (the mean of all FA scans of Dataset-1) overlayed on the standard brain (1mm). The red dots are voxel-wise feature importance as obtained by the reduction of impurity in training Random Forest model for BAP. The FA skeleton and the feature importance map have been slightly thickened by the linear interpolation of FSL to better visualize the distribution of feature importance across the skeleton.

The process of creating the feature importance map is as follows.

We train a random forest regressor on FA scans with no dimensionality reduction (training takes hours on AWS GPU-accelerated instances). This is because our goal is to visualize all regions on the brain which may potentially contribute to Brain Age Prediction. Feature importance scores are calculated as a measure of the relative importance of each feature by estimating the average amount of impurity reduction for the tree nodes that utilize that feature across all trees in the random forest. In other words, the importance of a feature is computed as the normalized amount of reduction of the impurity provided by that feature (see Equation (3.14)). The feature importance scores of a trained random forest typically end up being a sparse matrix with so many identical values and/or zeros, which makes it difficult to visualize on the maps. Hence, it is of utmost importance to fine-tune the model and train the model efficiently; otherwise, the ratio of feature importance scores with non-zero values would be too small to show. The other techniques include permutation importance which is computationally more expensive and make take days to be computed for high-dimensional data. Permutation importance is performed by randomly shuffling each feature and evaluating the drop in model's performance. In our case, we were able to calculate the matrix of feature importance scores by a thorough grid search on hyperparameters and then training the random forest model on FA data.

Once the feature importance scores are computed, we do an inverse process of masking on the array by padding to convert it back to its original dimensionality because, as mentioned in Section 3.1.5, the array of scans are masked as a preprocessing step, and if we do not transform them back to their original dimensions, they would not be aligned with the standard brain. Next, we convert the array to a Nifti [3] image and overlay it on the standard brain (1mm) as well as the mean FA skeleton (the mean of all FA scans in Dataset-1) in FSL software, and then tuning the projection parameters to highlight the areas on the brain that model uses to make its decisions. As can be seen in Figure 4.14, the whole skeleton of White Matter is recognized by the model as important, with a higher importance for "corpus callosum". This is aligned with the literature in the

[3]Nifti is an abbreviation of Neuroimaging Informatics Technology Initiative. https://nifti.nimh.nih.gov/

sense that "corpus callosum" microstructural changes (like thickness) over lifespan is associated with brain aging [62].

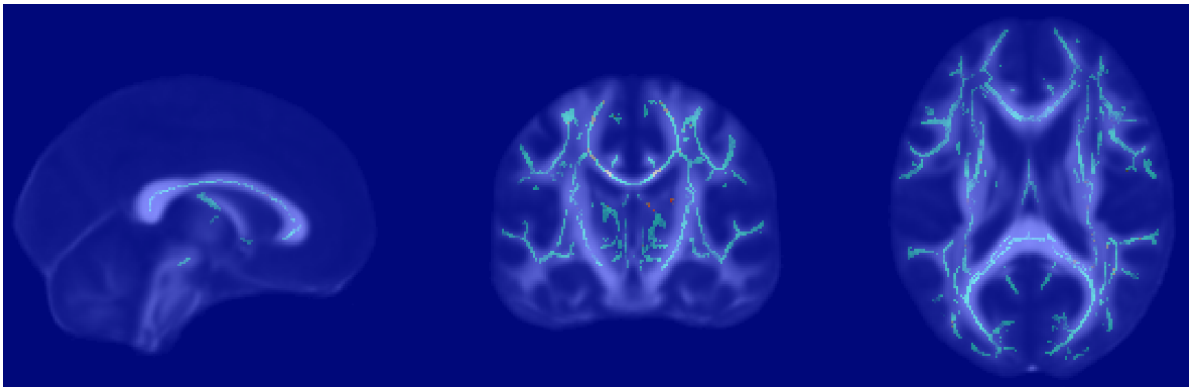
The next set of maps show the magnitude of the mean weights of the neurons of the first layer of our DNN model. Figure 4.15a and Figure 4.15b show two maps of the DNN's mean weights of the neurons of the first layer on the standard brain, when trained on RD scans and AD scans, respectively, and Figure 4.15c shows the feature map of the 10 neurons of the first layer of DNN when it is trained on AD. For better visualization, we inversed the order of the first layer and the second layer of the network architecture depicted in Figure 4.11 with no significant drop in its performance. The method to generate the map is similar to the one we used for random forest feature importance, unmasking (by padding) and reshaping the weights array to the scans' original dimensionality, converting it to Nifti image, and overlaying it on the standard brain.

As can be seen on the both maps of AD and RD, the whole White Matter tract is important for Brain Age Prediction. The corpus callosum is specially important as it is highlighted in both maps. The main difference between DNN and random forest maps is that DNN's map is less sparse, more continuous, and more extended across the White Matter (WM) (skeleton).

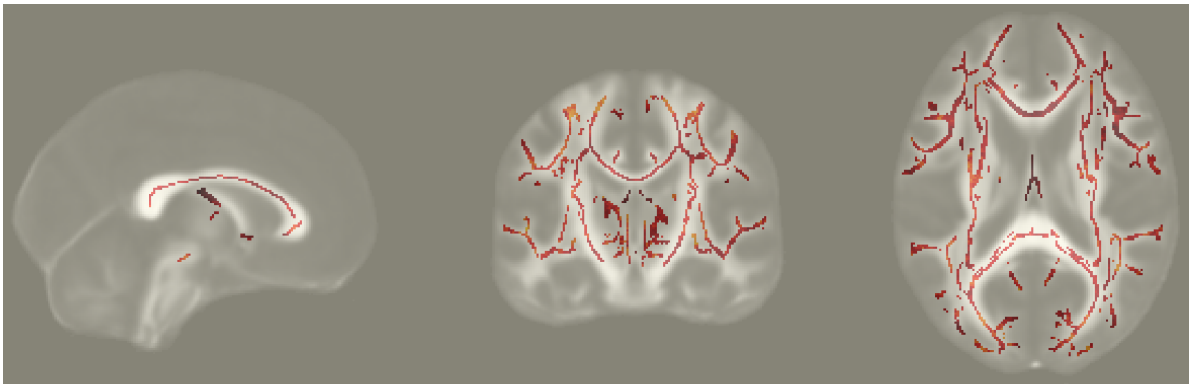
We emphasize that for these particular experiments of getting the feature importance scores of the Random Forest model and the DNN weights, we train the Random Forest and DNN on the data with the original dimensionality, i.e., we do not perform dimensionality reduction before training Random Forest and DNN for the purpose of this experiment.

The last map is provided in Figure 4.17 to visualize the impact of dimensionality reduction by the PCA, and minimization of reconstruction loss. Figure 4.17a and Figure 4.17b show the FA-mean and the reconstructed FA-mean, respectively. The reconstructed FA-mean is created after inverse transform of PCA of all scans and taking the mean of them. Figure 4.17c shows the reconstructed FA-mean and the original FA-mean in Python ecosystem where we perform all our experiments.

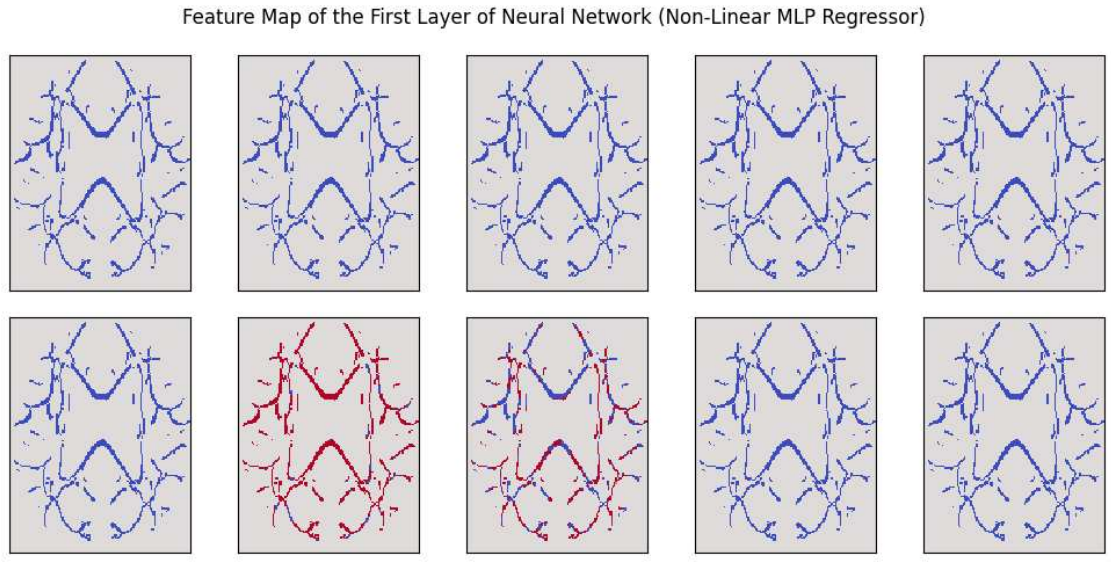
Our goal in the last map is to visualize the impact of PCA and the information loss as a consequence of dimensionality reduction and whether they might be important for Brain Age Prediction



(a)



(b)



(c)

Figure 4.15: Brain Maps of Deep Neural Network (DNN) Weights. (a) DNN's weights of the first layer when it is trained with RD data. The red and gray dots have higher magnitude of weights (see Figure 4.16 for better visualization). (b) DNN's weights of the first layer when it is trained with AD data. The dark-red areas like below corpus callosum have higher magnitude of weights. (c) DNN feature map of the 10 neurons of the first layer when it is trained with AD. The red areas have higher magnitude of weights.

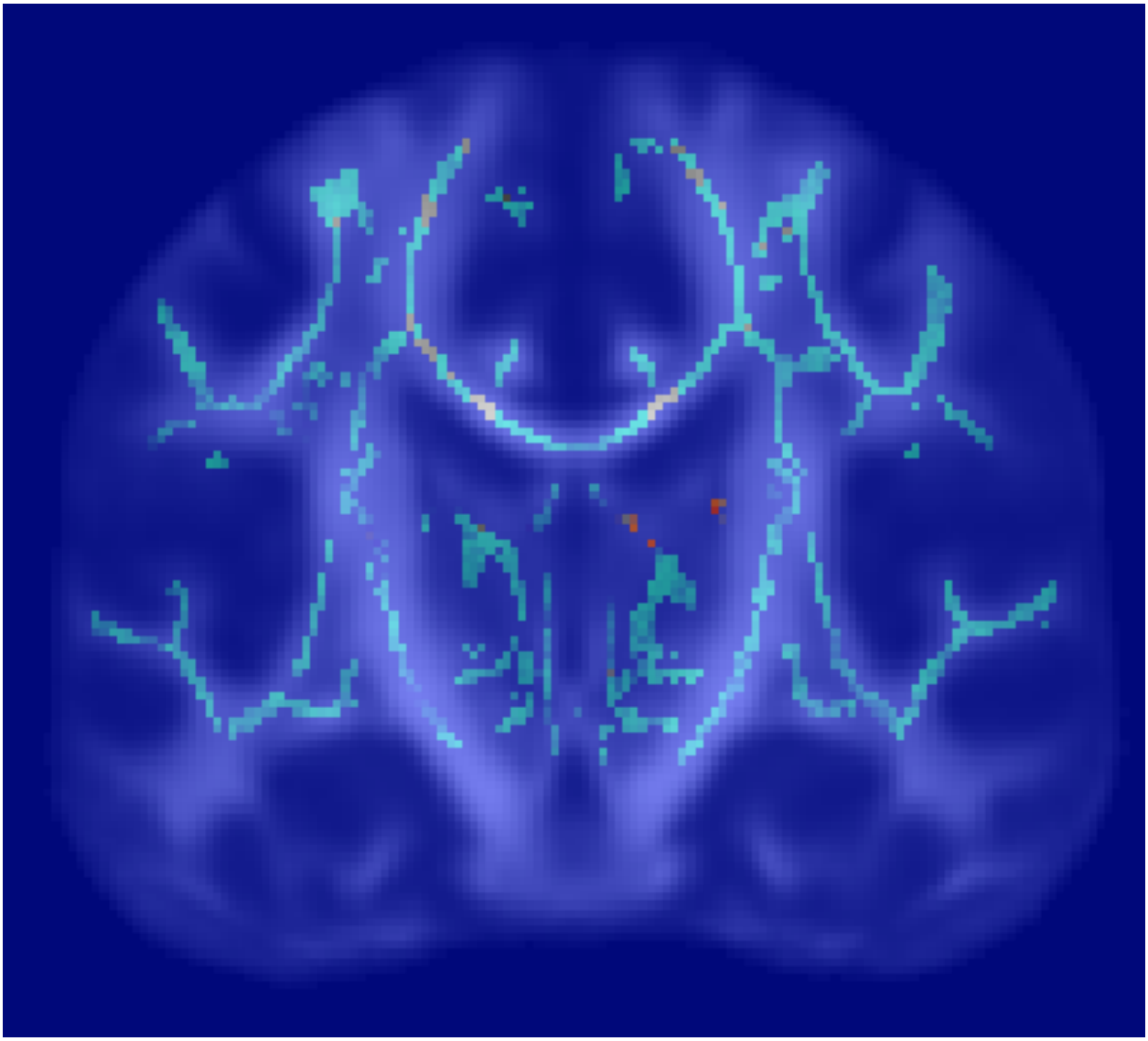
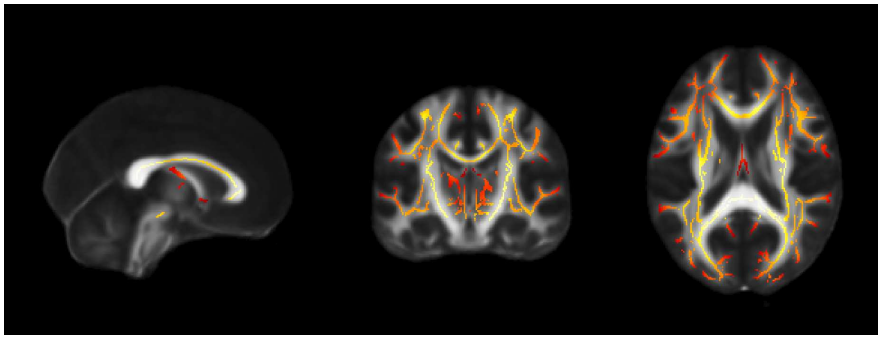
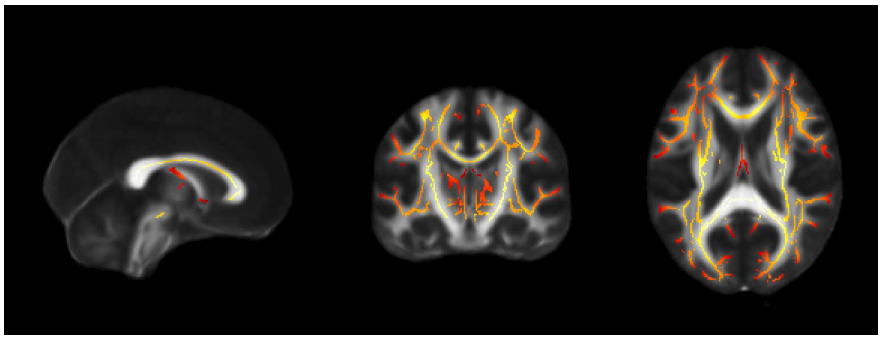


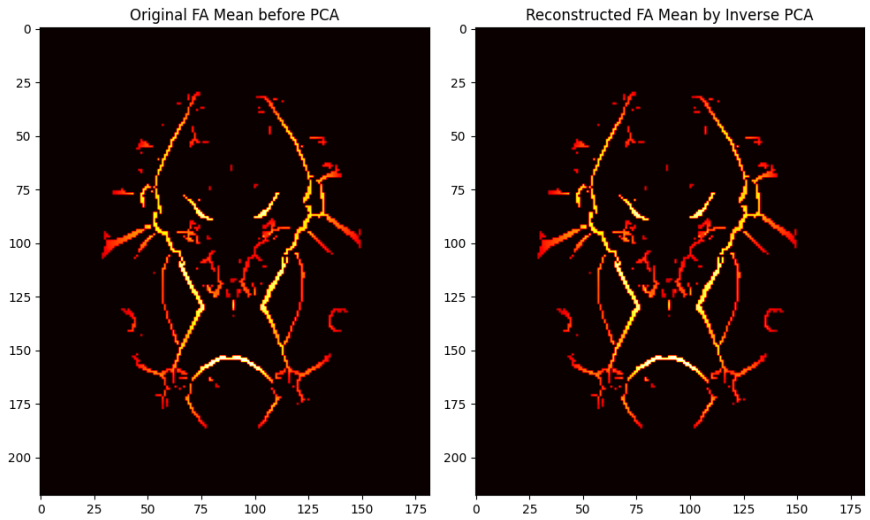
Figure 4.16: The Zoomed-In Coronal-View of Figure 4.15a. This is to better visualize the DNN weights with higher magnitudes, appearing as red and gray dots on the map.



(a)



(b)



(c)

Figure 4.17: Reconstruction of FA Scans and the Impact of Dimensionality Reduction by PCA (linear kernel). (a) Full three-view of the original FA-mean skeleton (mean of all FA scans in Dataset-1) before PCA shown with the "hot-iso" red-to-yellow colormap. (b) Reconstructed FA-mean skeleton after inverse transform of PCA for all scans and taking the mean of them. (c) Axial view of the FA-mean and the reconstructed FA-mean in Python ecosystem. FA scans of Dataset-1 have been projected to 290 principal components, followed by inverse transform of PCA for reconstruction. The map shows that the unsupervised reconstruction loss has been minimized, as there is no noticeable difference between the FA-mean skeleton and the reconstructed FA-mean skeleton. The loss is not noticeable even when the dimensionality is reduced to only two principal components.

as determined by our supervised models Random Forest and DNN. As can be seen, the reconstructed FA-mean skeleton has exactly preserved all the White Matter (WM) tracts with the same distribution.

Interestingly, if we reduce the dimensionality of Diffusivity Parameters to 2D, like the 2D projection we did in this chapter to represent FA scans by just two principal components (z_1, z_2), and then train GMM model on them to generate new samples out of the 2D distribution of data, followed by an inverse transform of PCA on the new 2D samples, we can get brain maps that look the same as the maps shown for the reconstruction after PCA with 0.99 variance. This is aligned with our plot in Figure 3.5, in which we can observe that the impact of number of principal components after 0.2, i.e., close to 2D projection, is relatively stable.

We would like to end this chapter by highlighting what we mentioned earlier in discussions of Unsupervised Anomaly Detection (UAD) and its alignment with the supervised Brain Age Prediction (BAP) results:

The significance of these results is that our Unsupervised Learning methods and our Supervised Learning methods consistently converge, agree and confirm each other.

Chapter 5

Conclusions

In the previous chapters, we provided the details of our experiments, our results, our findings, analytical discussions, implications of our results, the details of our methodology, as well as the background and related work on the applications of Machine Learning (ML) and Deep Learning (DL) in Neuroimaging, specifically Diffusion Tensor Imaging (DTI), for Brain Age Prediction (BAP) to identify unhealthy brain aging early and efficiently for medical interventions and gaining insights about the biomarkers that help the researches toward these goals. We framed five major research questions as specified in Section 1.2, and we addressed the questions in detail and with concrete evidence of our answers. In this chapter, we will review our key findings and conclusions as well as the limitations of our study, and we will identify potential directions for future research by ourselves and other researchers who have followed us in this dissertation thus far.

5.1 Summary of Findings

We list our key findings with regards to our five major research questions as outlined in Section 1.2:

Research Question-1:

Through our research, we provided evidence that Diffusion Tensor Imaging (DTI) and its measures (scalars), Diffusivity Parameters, are a suitable quantitative modality of neuroimaging that can be used for Brain Age Prediction (BAP) with promising results. This neuroimaging data modality is predictable and explainable compared to other modalities that are more qualitative and need extensive subjective descriptions and annotations like structural Magnetic Resonance Imaging (MRI). As an evidence of its quantitative predictability, we refer to our finding on the cause for FA's sensitivity to preprocessing steps, which is due to its value distribution and the mathematical impacts of preprocessing transformation on its values. These impacts can be methodically and mathematically investigated and analyzed, as we did and made conclusions in this study.

We have shown that Brain Age Prediction (BAP) using DTI provides promising results that are comparable with state-of-the-art results of other modalities in the recent literature. The implication of this finding has two aspects: DTI can be used individually for BAP, in the absence of other modalities, or it can be combined in multi-modal studies as our results show that DTI has predictive power for BAP, even when it is used to train Machine Learning models in an Unsupervised Learning manner.

Our results indicate that non-linear models and non-linear kernels generally outperform linear models on both DTI datasets in our study, only with a couple of exceptions – good performance of Ridge for FA specifically, and linear PCA for dimensionality reduction, although non-linear kPCA also performs very well. We argue that this is an indication of the existence of non-linear patterns that would be missed otherwise if linear models are solely used. Non-linear patterns in White Matter data suggests non-linear and subtle microstructure changes in the brain as a result of normal aging as well as other factors. We suggest combining the linear and non-linear models as an ensemble to make the most out of different models' applicability.

We found that ensemble methods outperform all linear and non-linear models including Deep Learning models. Our results with applying ensemble methods on DTI data show that they are scalable, reliable, and less sensitive to parameters and experimental settings. That said, researchers should consider testing and applying other models that might be more suitable for their particular problem with respect to "No Free Lunch Theorem" in Machine Learning.

Research Question-2:

We found that FA, AD, and RD are more useful in different Machine Learning tasks, both supervised and unsupervised, although FA is more sensitive to preprocessing steps. We also found that MD is less relevant in Brain Age Prediction, and that it is mostly redundant and provides less information than other Diffusivity Parameters.

We have shown that AD+RD is the best combination for BAP regression task (Supervised Learning) by our combination analysis. On the other hand, FA is a good parameter for Supervised Learning when it is used individually and not in combination with other parameters with the same

preprocessing transformations. We also found that FA is the best choice in Unsupervised Learning tasks that we performed: dimensionality reduction, clustering and anomaly detection using DTI data. We do not recommend discarding and ignoring FA for Unsupervised Learning tasks.

We found that FA needs scaling after Principal Component Analysis (PCA) while AD, MD, and RD generally do not need to be standardized/normalized as scaling sometimes hurts the results of the models that are trained on AD and RD data, while not scaling rarely hurts their results.

We found that in concatenating DTI scalars, they should be preprocessed separately and differently. For example, if a researcher wants to combine the parameters, a good strategy could be applying PCA and scaling FA, while applying PCA and not scaling AD, MD, and RD.

Research Question-3:

To answer this question methodically, we proposed, implemented and analyzed a methodology to combine and compare the results of our Unsupervised Anomaly Detection (UAD) method and our supervised Brain Age Prediction (BAP) models. We argue that our label-independent UAD is very helpful because "Chronological Age" labels are inconsistent. We believe that the "**Inconsistency of Labels**" in BAP as a regression problem in a Supervised Learning manner, as well as the known "**Systematic Bias**" in the recent work of BAP in the literature [15, 16], raise a fundamental or "philosophical" dilemma of the current BAP approach. We argue that if a researcher wants to train the models for BAP as a regression task to identify high Deltas, and as an indication/biomarker of unhealthy brain aging or irregularities, they should ideally provide the model in supervised training with both "Chronological Age" as well as "Biological Age" (estimated by brain experts through Neuroimaging) so that the model can efficiently learn the differences between healthy and unhealthy brain aging. Otherwise, a supervised regression model has no way to learn the differences between healthy and unhealthy brain aging because it is trained with "Chronological Age" labels only. By providing inconsistent "Chronological Age" labels to the supervised Machine Learning models for BAP, not only we do not help them to learn better, but also may mislead them, and that highlights the relevance of using our label-independent UAD approach that does not require age labels. At least, researchers and practitioners should provide

the model with an additional feature like a binary flag (Healthy vs Unhealthy) so that the model can learn well with supplementary information on the differences between healthy and unhealthy subjects. Since those supplementary information for removing the inconsistency of "Chronological Age" is not provided in usual practice, we argue that our proposed methodology to supplement the supervised BAP by combining its results with our label-independent "UAD" can resolve the issues and the inconsistency of the chronological age labels, and can methodically address our "Research Question-3".

Now a critic may raise a very legitimate question as follows.

*"Isn't the whole point of your work to find a way to estimate **Biological Age** because we don't have good ways of estimating it?"*

Our response to the above question is positive. However, our argument is that the current Supervised Learning approach for Brain Age Prediction (BAP) is limited, insufficient, and problematic. If we use only the supervised BAP to estimate "Biological Age" and the inconsistent "Chronological Age" labels to train the models, we do not have a systematic and verifiable method for error analysis to definitively determine whether the observed errors are due to the model's poor performance or an actual gap between the "Biological Age", as predicted by the model, and the "Chronological Age". Hence, our proposed methodology which combines the results of label-independent Unsupervised Anomaly Detection (UAD) and supervised Brain Age Prediction (BAP) provides a methodical way for error analysis, and bridges the gap caused by "Chronological Age" label inconsistency.

We found that Supervised Learning and Unsupervised Learning results are very similar in our study. On the one hand, Unsupervised Learning methods, Gaussian Mixture Model (GMM) and Principal Component Analysis (PCA) provide insightful results, and on the other hand, ensemble methods agree with those results on the relevance of White Matter tracts and Diffusivity Parameters for brain age estimation.

We found that Unsupervised Learning clustering finds age groups in 2D projected DTI data with intra-cluster members that are similar in chronological age while the models have been trained

with no access to ground truth (chronological age). This is a further confirmation that DTI is in fact a relevant modality for Brain Age Prediction. We also found that Gaussian Mixture Model (GMM) and Bayesian Gaussian Mixture Model (BGMM) algorithms are more robust in recognition and separation of the underrepresented group, which in our study is the population of younger subjects. Our finding can help researchers working with imbalanced datasets to compensate data distribution by generative models that can generate new data as we did in this study for data augmentation and transfer learning.

We found that the Gaussian Mixture Model (GMM) is successful in unsupervised DTI analysis (clustering and UAD), and we argue that this success is caused by a match between the GMM assumption that the data is a mixture of a finite set of Gaussian distributions, and the Gaussian distribution of water diffusion as captured by Diffusivity Parameters of DTI, and as described in Equation (2.2).

Our proposed "Unsupervised Anomaly Detection (UAD)" effectively finds "anomaly" subjects who end up with high-delta in BAP (both positive and negative delta). This can lead to further investigation on the detected anomalies to find possible irregularities and abnormalities in their scans as a supplementary method for the existing supervised BAP methods.

Research Question-4:

DTI is a high-dimensional quantitative data modality with scalar parameters that are in a very limited range, mostly very close to zero, so they are not images like structural MRI. Therefore, Deep Network and specifically CNNs and more complex and sophisticated architectures of Deep Learning did not work well for our study, or at least did not outperform other methods. Our results about DL methods can't be generalized though due to limitations of our dataset size. However, we found that ensemble methods perform well even on our small datasets.

Research Question-5:

We have used brain maps as an effective visualization technique that can help explain how our models make decisions. Applying dimensionality reduction techniques on DTI high-dimensional

data can significantly boost training and reduce training time. Our brain maps show that reconstruction loss is minimal and the maps are aligned with supervised BAP feature importance.

Our brain maps also reveal the regions of interest in the White Matter tracts that are most relevant for Machine Learning and Deep Learning models to make their prediction on the brain age. Those areas highlighted on our maps, like corpus callosum, are aligned with the literature and are known correlations with brain aging. Further investigation is required to analyze our maps and identify other regions that might be relevant for BAP as well.

5.2 Limitations of Our Study

While our findings are interesting and novel, generalization of our models, results, and conclusions should be done with caution given that we have trained on small-sized datasets, and a special kind of neuroimaging modality (skeletonized Diffusivity Parameters of DTI scans) was used.

Our methods seem to be scalable (especially Unsupervised Learning methods). However, we have not tested our methods on larger DTI datasets.

Some models are still hard to "explain" and to "interpret" in the Machine Learning context, especially ensemble methods, although our brain maps can shed some light on their performance. This is a recognized and fundamental challenge in Machine Learning about ensemble methods. They are effective, but hard to explain.

Gender analysis was not performed in this study due to small-sized datasets and imbalanced distribution of genders. Gender is considered a confounding factor in this regression problem.

5.3 Future Work

Finally, we share a list of research directions and topics that we think are potential next steps of this study to follow, and we hope that our study and our findings encourage other researchers to pursue them if they are interested:

We plan to apply our methods on public and large datasets. While our methodology shows promising results, we still need to verify our findings with further studies.

Multi-modal studies have gained popularity in recent years, given the fact that different modalities of Neuroimaging may capture different aspects of how the brain works, matures, and develops over time. Combining DTI with other modalities has a good potential for further research especially given our results that show DTI has the predictive power for age prediction.

We would like to further investigate the applicability of Deep Learning on DTI data, especially Graph Neural Networks (GNN), and Graph Convolutional Neural Networks (GCN). Our hypothesis which needs investigation is that graphs might be a better representation of DTI data given its sparseness. Representing the data as values in a graph might be a better approach than a full 2D image or 3D volume with plenty of zeros. On the graph, two values that are neighbors along the skeleton would be linked, and hence we think that we can alleviate or ideally resolve the sparsity issue of the DTI data by graph representation.

Variational Autoencoders (VAE) and Generative Adversarial Networks (GAN) are generative models that have been shown to be applicable for neuroimaging data reconstruction and generation of new samples that resemble actual brain scans. We have used Gaussian Mixture Model (GMM) and Bayesian Gaussian Mixture Model (BGMM) generative models to generate new data, and we are interested to further investigate applicability of VAE and GAN on DTI data.

Further analysis on anomalies detected by our Unsupervised Anomaly Detection (UAD) is required to check the alignment of their health conditions and lifestyle with our findings. Statistical and medical cross-examination of our findings and clinical diagnosis can lead us to design and develop more efficient algorithms that can be used as an AI/ML technology to diagnose the onset of neurodegenerative conditions such as Alzheimer Disease (AD) and cognitive decline.

As the final note to end this dissertation, Artificial Intelligence, and specifically Machine Learning and Deep Learning, nowadays provide a new generation of assistive technologies for medical, clinical, and neurological applications. We look forward to collaborations and contributions in this exciting interdisciplinary field of research.

Bibliography

[1] James H. Cole. Multimodality neuroimaging brain-age in uk biobank: relationship to biomedical, lifestyle, and cognitive factors. *Neurobiology of Aging*, 92:34–42, 8 2020.

[2] C. Lebel, M. Gee, R. Camicioli, M. Wieler, W. Martin, and C. Beaulieu. Diffusion tensor imaging of white matter tract evolution over the lifespan. *NeuroImage*, 60:340–352, 3 2012.

[3] James H. Cole and Katja Franke. Predicting age using neuroimaging: Innovative brain ageing biomarkers. *Trends in Neurosciences*, 40:681–690, 12 2017.

[4] Shammi More, Georgios Antonopoulos, Felix Hoffstaedter, Julian Caspers, Simon B. Eickhoff, Kaustubh R. Patil, and . Brain-age prediction: a systematic comparison of machine learning workflows. *bioRxiv*, 2022.

[5] Lea Baecker, Rafael Garcia-Dias, Sandra Vieira, Cristina Scarpazza, and Andrea Mechelli. Machine learning for brain age prediction: Introduction to methods and clinical applications. *EBioMedicine*, 72, 10 2021.

[6] M. Tanveer, M. A. Ganaie, Iman Beheshti, Tripti Goel, Nehal Ahmad, Kuan-Ting Lai, Kaizhu Huang, Yu-Dong Zhang, Javier Del Ser, and Chin-Teng Lin. Deep learning for brain age estimation: A systematic review. 12 2022.

[7] Susumu Mori. *Introduction to Diffusion Tensor Imaging*. Elsevier Science, United Kingdom, 2007.

[8] Andrew L. Alexander, Jee Eun Lee, Mariana Lazar, and Aaron S. Field. Diffusion tensor imaging of the brain. *Neurotherapeutics*, 4(3):316–329, 2007.

[9] Andrew L. Alexander, Samuel A. Hurley, Alexey A. Samsonov, Nagesh Adluru, Ameer Pasha Hosseinbor, Pouria Mossahebi, Do P.M. Tromp, Elizabeth Zakszewski, and Aaron S. Field. Characterization of cerebral white matter properties using quantitative magnetic resonance imaging stains. *Brain Connectivity*, 1(6):423–446, 2011.

[10] José M. Soares, Paulo Marques, Victor Alves, and Nuno Sousa. A hitchhiker's guide to diffusion tensor imaging. *Frontiers in Neuroscience*, 2013.

[11] Heidi M. Feldman, Jason D. Yeatman, Eliana S. Lee, Laura H.F. Barde, and Shayna Gaman-Bean. Diffusion tensor imaging: A review for pediatric researchers and clinicians. *Journal of Developmental and Behavioral Pediatrics*, 31:346–356, 5 2010.

[12] L. M. Alba-Ferrara and G. A. de Erausquin. What does anisotropy measure? insights from increased and decreased anisotropy in selective fiber tracts in schizophrenia. *Frontiers in Integrative Neuroscience*, 2 2013.

[13] Dani Beck, Ann-Marie G. de Lange, Ivan I. Maximov, Geneviève Richard, Ole A. Andreassen, Jan E. Nordvik, and Lars T. Westlye. White matter microstructure across the adult lifespan: A mixed longitudinal and cross-sectional study using advanced diffusion models and brain-age prediction. *NeuroImage*, 224:117441, 2021.

[14] Lan Lin, Cong Jin, Zhenrong Fu, Baiwen Zhang, Guangyu Bin, and Shuicai Wu. Predicting healthy older adult's brain age based on structural connectivity networks using artificial neural networks. *Computer Methods and Programs in Biomedicine*, 125:8–17, 2016.

[15] Hualou Liang, Fengqing Zhang, and Xin Niu. Investigating systematic bias in brain age estimation with application to post-traumatic stress disorders. *Human Brain Mapping*, 40:3143–3152, 8 2019.

[16] Pedro L. Ballester, Laura Tomaz da Silva, Matheus Marcon, Nathalia Bianchini Esper, Benicio N. Frey, Augusto Buchweitz, and Felipe Meneguzzi. Predicting brain age at slice level: Convolutional neural networks and consequences for interpretability. *Frontiers in Psychiatry*, 12, 2021.

[17] Albert Einstein. Investigations on the theory of the Brownian movement. Dover Publications, New York, 1956.

[18] Peter J. Basser. Inferring microstructural features and the physiological state of tissues from diffusion-weighted images. *NMR in Biomedicine*, 8(7):333–344, 1995.

[19] Do Tromp. Dti scalars - website: http://www.diffusion-imaging.com, 2023.

[20] David J. Madden, Ilana J. Bennett, Agnieszka Burzynska, Guy G. Potter, Nan kuei Chen, and Allen W. Song. Diffusion tensor imaging of cerebral white matter integrity in cognitive aging. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1822(3):386–400, 2012.

[21] Anna Behler, Jan Kassubek, and Hans-Peter Müller. Age-related alterations in dti metrics in the human brain—consequences for age correction. *Frontiers in Aging Neuroscience*, 13, 2021.

[22] Shengwei Zhang and Konstantinos Arfanakis. White matter segmentation based on a skeletonized atlas: Effects on diffusion tensor imaging studies of regions of interest. *Journal of Magnetic Resonance Imaging*, 40(5):1189–1198, 2014.

[23] Audrey Low, Elijah Mak, James D. Stefaniak, Maura Malpetti, Nicolas Nicastro, George Savulich, Leonidas Chouliaras, Hugh S. Markus, James B. Rowe, and John T. O'Brien. Peak width of skeletonized mean diffusivity as a marker of diffuse cerebrovascular damage. *Frontiers in Neuroscience*, 14, 2020.

[24] Gregory Beaudet, Ami Tsuchida, Laurent Petit, Christophe Tzourio, Svenja Caspers, Jan Schreiber, Zdenka Pausova, Yash Patel, Tomas Paus, Reinhold Schmidt, Lukas Pirpamer, Perminder S Sachdev, Henry Brodaty, Nicole A. Kochan, Julian N. Trollor, Wei Wen, Nicola J Armstrong, Ian Deary, Mark Bastin, Joanna Wardlaw, Susana Muñoz Maniega, A. Veronica Witte, Arno Villringer, Marco Duering, Stephanie Debette, and Bernard Mazoyer. Age-related changes of peak width skeletonized mean diffusivity (psmd) across the adult lifespan: A multi-cohort study. *Frontiers in psychiatry*, 11, May 2020.

[25] Agnieszka Z. Burzynska, Yuqin Jiao, Anya M. Knecht, Jason Fanning, Elizabeth A. Awick, Tammy Chen, Neha Gothe, Michelle W. Voss, Edward McAuley, and Arthur F. Kramer. White matter integrity declined over 6-months, but dance intervention improved integrity of the fornix of older adults. *Frontiers in Aging Neuroscience*, 9, 3 2017.

[26] Agnieszka Z. Burzynska, Karolina Finc, Brittany K. Taylor, Anya M. Knecht, and Arthur F. Kramer. The dancing brain: Structural and functional signatures of expert dance training. *Frontiers in Human Neuroscience*, 11, 2017.

[27] Christian Beaulieu. The basis of anisotropic water diffusion in the nervous system – a technical review. *NMR in Biomedicine*, 15(7-8):435–455, 2002.

[28] Fsl documentation by oxford university - website: https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/tbss/userguide, 2023.

[29] Daichi Sone and Iman Beheshti. Neuroimaging-based brain age estimation: A promising personalized biomarker in neuropsychiatry. *Journal of Personalized Medicine*, 12(11), 2022.

[30] Xin Niu, Fengqing Zhang, John Kounios, and Hualou Liang. Improved prediction of brain age using multimodal neuroimaging data. *Human Brain Mapping*, 41:1626–1643, 4 2020.

[31] Geneviève Richard, Knut Kolskår, Anne Marthe Sanders, Tobias Kaufmann, Anders Petersen, Nhat Trung Doan, Jennifer Monereo Sánchez, Dag Alnæs, Kristine M. Ulrichsen, Erlend S. Dørum, Ole A. Andreassen, Jan Egil Nordvik, and Lars T. Westlye. Assessing distinct patterns of cognitive aging using tissue-specific brain age prediction based on diffusion tensor imaging and brain morphometry. *PeerJ*, 2018, 2018.

[32] Siren Tønnesen, Tobias Kaufmann, Ann-Marie de Lange, Genevieve Richard, Nhat Trung Doan, Dag Alnæs, Dennis van der Meer, Jaroslav Rokicki, Torgeir Moberget, Ivan I. Maximov, Ingrid Agartz, Sofie R. Aminoff, Dani Beck, Deanna Barch, Justyna Beresniewicz, Simon Cervenka, Helena Fatouros Bergman, Alexander R. Craven, Lena Flyckt, Tiril P. Gurholt, Unn K. Haukvik, Kenneth Hugdahl, Erik Johnsen, Erik G. Jönsson, , Knut K.

Kolskår, Kristiina Kompus, Rune Andreas Kroken, Trine V. Lagerberg, Else-Marie Løberg, Jan Egil Nordvik, Anne-Marthe Sanders, Kristine Ulrichsen, Ole A. Andreassen, and Lars T. Westlye. Brain age prediction reveals aberrant brain white matter in schizophrenia and bipolar disorder: A multi-sample diffusion tensor imaging study. *bioRxiv*, 2020.

[33] Jun Ding Zhu, Shih Jen Tsai, Ching Po Lin, Yi Ju Lee, and Albert C. Yang. Predicting aging trajectories of decline in brain volume, cortical thickness and fractional anisotropy in schizophrenia. *Schizophrenia*, 9, 12 2023.

[34] Esten H. Leonardsen, Han Peng, Tobias Kaufmann, Ingrid Agartz, Ole A. Andreassen, Elisabeth Gulowsen Celius, Thomas Espeseth, Hanne F. Harbo, Einar A. Høgestøl, Ann Marie de Lange, Andre F. Marquand, Didac Vidal-Piñeiro, James M. Roe, Geir Selbæk, øystein Sørensen, Stephen M. Smith, Lars T. Westlye, Thomas Wolfers, and Yunpeng Wang. Deep neural networks learn general and clinically relevant representations of the ageing brain. *NeuroImage*, 256, 8 2022.

[35] Nicola K. Dinsdale, Emma Bluemke, Stephen M. Smith, Zobair Arya, Diego Vidaurre, Mark Jenkinson, and Ana I.L. Namburete. Learning patterns of the ageing brain in mri using deep convolutional networks. *NeuroImage*, 224:117401, 2021.

[36] Thorgeirsson TE Ellingsen LM Walters GB Gudbjartsson DF Stefansson H Stefansson K Ulfarsson MO Jonsson BA, Bjornsdottir G. Brain age prediction using deep learning uncovers associated sequence variants. *Nature Communications*, 10, 12 2019.

[37] Chen Yuan Kuo, Tsung Ming Tai, Pei Lin Lee, Chiu Wang Tseng, Chieh Yu Chen, Liang Kung Chen, Cheng Kuang Lee, Kun Hsien Chou, Simon See, and Ching Po Lin. Improving individual brain age prediction using an ensemble deep learning framework. *Frontiers in Psychiatry*, 12, 3 2021.

[38] Huiting Jiang, Na Lu, Kewei Chen, Li Yao, Ke Li, Jiacai Zhang, and Xiaojuan Guo. Predicting brain age of healthy adults based on structural mri parcellation using convolutional neural networks. *Frontiers in Neurology*, 10, 2020.

[39] Juhyuk Han, Seo Yeong Kim, Junhyeok Lee, and Won Hee Lee. Brain age prediction: A comparison between machine learning models using brain morphometric data. *Sensors*, 22(20), 2022.

[40] Juan Miguel Valverde, Vandad Imani, Ali Abdollahzadeh, Riccardo De Feo, Mithilesh Prakash, Robert Ciszek, and Jussi Tohka. Transfer learning in magnetic resonance brain imaging: A systematic review. *Journal of Imaging*, 7, 4 2021.

[41] Zaniar Ardalan and Vignesh Subbian. Transfer learning approaches for neuroimaging analysis: A scoping review. *Frontiers in Artificial Intelligence*, 5, 2022.

[42] Xueyan Mei, Zelong Liu, Philip M. Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E. Link, Thomas Yang, Ying Wang, Hayit Greenspan, Timothy Deyer, Zahi A. Fayad, and Yang Yang. Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5):e210315, 2022.

[43] Karim Aderghal, Karim Afdel, Jenny Benois-Pineau, and Gwénaëlle Catheline. Improving alzheimer's stage categorization with convolutional neural network using transfer learning and different magnetic resonance imaging modalities. *Heliyon*, 6(12):e05652, 2020.

[44] Ann-Marie G. de Lange, Melis Anatürk, Jaroslav Rokicki, Laura K. M. Han, Katja Franke, Dag Alnæs, Klaus P. Ebmeier, Bogdan Draganski, Tobias Kaufmann, Lars T. Westlye, Tim Hahn, and James H. Cole. Mind the gap: Performance metric evaluation in brain-age prediction. *Human Brain Mapping*, 43(10):3113–3129, 2022.

[45] Katja Franke, Gabriel Ziegler, Stefan Klöppel, and Christian Gaser. Estimating the age of healthy subjects from t1-weighted mri scans using kernel methods: Exploring the influence of various parameters. *NeuroImage*, 50(3):883–892, 2010.

[46] Aaquib Syed, Richard Adam, Thomas Ren, Jinyu Lu, Takouhie Maldjian, and Tim Q. Duong. Machine learning with textural analysis of longitudinal multiparametric mri and molecular subtypes accurately predicts pathologic complete response in patients with invasive breast cancer. *PLOS ONE*, 18(1):1–14, 01 2023.

[47] Jonathan Stubblefield, Alan Kronberger, Jason Causey, Jake Qualls, Jennifer Fowler, Kaiman Zeng, Karl Walker, Xiuzhen Huang, and . Study the combination of brain mri imaging and other datatypes to improve alzheimer's disease diagnosis. *medRxiv*, 2022.

[48] Wan Tang, Han Zhou, Tianhong Quan, Xiaoyan Chen, Huanian Zhang, Yan Lin, and Renhua Wu. Xgboost prediction model based on 3.0t diffusion kurtosis imaging improves the diagnostic accuracy of mri birads 4 masses. *Frontiers in Oncology*, 12, 2022.

[49] Sangjin Ahn, Si Eun Lee, and Mi hyun Kim. Random-forest model for drug–target interaction prediction via kullbeck–leibler divergence. *Journal of Cheminformatics*, 14, 12 2022.

[50] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016.

[51] Xgboost library documentation - website: https://xgboost.readthedocs.io/en/latest/index.html, 2023.

[52] Aurelien. Geron. *Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow, Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, Inc, 2022.

[53] Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong. *Mathematics for Machine Learning*. Cambridge University Press, 2020.

[54] Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.

[55] Alexander LeNail. Nn-svg: Publication-ready neural network architecture schematics. *Journal of Open Source Software*, 4(33):747, 2019.

[56] Muhammad Farhan Safdar, Shayma Alkobaisi, and Fatima Tuz Zahra. A comparative analysis of data augmentation approaches for magnetic resonance imaging (mri) scan images of brain tumor. *Acta Informatica Medica*, 28:29 – 36, 2020.

[57] Won Hee Lee, Mathilde Antoniades, Hugo G Schnack, Rene S. Kahn, and Sophia Frangou. Brain age prediction in schizophrenia: Does the choice of machine learning algorithm matter? *Psychiatry Research: Neuroimaging*, 310:111270, 2021.

[58] David H. Wolpert. The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation*, 8(7):1341–1390, 10 1996.

[59] Christoph Baur, Stefan Denner, Benedikt Wiestler, Nassir Navab, and Shadi Albarqouni. Autoencoders for unsupervised anomaly segmentation in brain mr images: A comparative study. *Medical Image Analysis*, 69:101952, 2021.

[60] Daniel T Huff, Amy J Weisman, and Robert Jeraj. Interpretation and visualization techniques for deep learning models in medical imaging. *Physics in Medicine & Biology*, 66(4):04TR01, feb 2021.

[61] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, Los Alamitos, CA, USA, jun 2016.

[62] V.M. Danielsen, D. Vidal-Piñeiro, A.M. Mowinckel, D. Sederevicius, A.M. Fjell, K.B. Walhovd, and R. Westerhausen. Lifespan trajectories of relative corpus callosum thickness: Regional differences and cognitive relevance. *Cortex*, 130:127–141, 2020.

Appendix A

Hardware and Software Specifications

A.1 Hardware Specifications

- Amazon Web Services (AWS) - Elastic Compute Cloud (EC2).
- Accelerated Computing G5 Instances - g5.12xlarge, 4 GPUs, 96 GB RAM, 48 vCPUs.

A.2 Software Specifications

- The Operating System : Ubuntu 20.04 on Deep Learning AMI (DLAMI) Instance with GPU.
- Python Kernel: Jupyter Notebook via secured "ssh" to connect to Jupyter Servers which were run on the AWS EC2 instances.
- Distributed Computing: ***Distributed training with TensorFlow*** by "MirroredStrategy" to distribute training across multiple GPUs. Please see Tensorflow-GPU documentation for further information.
- Further specifications of the softwares that were used for this dissertation is provided in Table A.1.

Table A.1: Software Specifications - As of March 2023

Software/OS	Version
Ubuntu	20.04
Python	3.10.7
Tensorflow-GPU	2.11
XGBoost-GPU	1.7.3
Scikit-Learn	1.2.0
Numpy	1.23.5
Matplotlib	3.6.2
Nibabel	5.0.0
FSL	6.0.6.2

A.3 Hyperparameter Tuning

The following is the result of our Grid search and Randomized Search for fine-tuning of hyper-parameters of Random Forest and XGBoost BAP regression models, respectively.

A.3.1 Random Forest

```
# Grid Search parameters to fine-tune Random Forest
param_grid = {
        'bootstrap': [True, False],
        'max_depth': [5, 10, 20, 40, 100],
        'max_features': [2, 3, 4, 5],
        'min_samples_leaf': [3, 4, 5, 6, 7],
        'min_samples_split': [8, 10, 12, 14, 16],
        'n_estimators': [100, 200, 400, 800, 1000]
}
# Create the Random Forest Regression Model
rf = RandomForestRegressor()
grid_search = GridSearchCV(estimator = rf, \
param_grid = param_grid, \
cv = 5, scoring = 'neg_mean_absolute_error', \
n_jobs = -1, verbose = 2)

# The best values for hyperparameters of Random Forest
{'bootstrap': True,
'max_depth': 10,
'max_features': 5,
'min_samples_leaf': 3,
'min_samples_split': 12,
'n_estimators': 100}
```

A.3.2 XGBoost

```
# Randomized Search to fine-tune XGBoost
# We use "gpu_hist" to make XGBoost parallelized on GPU
# Due to the high number of XGBoost hyperparameters,
# RandomizedSearch is more time-efficient than Grid Search

# Randomized Search to fine-tune XGBoost
param_dist = {
        'eta': [0.1, 0.2, 0.3, 0.4, 0.5],
        'gamma': [0, 1, 2, 4, 8, 10, 20],
        'max_depth': [2, 3, 4, 5, 10],
        'reg_alpha': [0, 1, 2, 5, 10],
        'reg_lambda': [0.1, 0.5, 1.0, 5.0, 10.0],
        'subsample': np.arange(0.5, 1.0, 0.1),
        'colsample_bytree': np.arange(0.4, 1.0, 0.1),
        'colsample_bylevel': np.arange(0.4, 1.0, 0.1),
        'n_estimators': [100, 200, 400, 800, 1000]
}
# Create the XGBoost Regression Model
xgb = XGBRegressor()
random_search = RandomizedSearchCV(estimator = xgb, \
param_distributions = param_dist, \
cv = 5, scoring = 'neg_mean_absolute_error', \
n_jobs = -1, verbose = 2)

# The best hyperparameter values for XGBoost for BAP
# This set is less likely to be generalized well!
{'subsample': 0.8,
```

```
'reg_lambda': 0.1,
'reg_alpha': 1,
'n_estimators': 100,
'max_depth': 2,
'gamma': 20,
'eta': 0.1,
'colsample_bytree': 0.6,
'colsample_bylevel': 0.5}
```

Appendix B

List of Acronyms

Machine Learning (ML)

Deep Learning (DL)

Brain Age Prediction (BAP)

Convolutional Neural Network (CNN)

Deep Neural Network (DNN)

Gaussian Mixture Model (GMM)

Bayesian Gaussian Mixture Model (BGMM)

Unsupervised Anomaly Detection (UAD)

Principal Component Analysis (PCA)

Extreme Gradient Boosting (XGBoost)

Cross Validation (CV)

Mean Absolute Error (MAE)

Diffusion Tensor Imaging (DTI)

Fractional Anisotropy (FA)

Axial Diffusivity (AD)

Mean Diffusivity (MD)

Radial Diffusivity (RD)

White Matter (WM)

Neuroimaging Informatics Technology Initiative (NIFTI)