Approaches Towards Advanced Brain Age Prediction Models

Hanzhi Wang

A thesis submitted in partial fulfilment for the degree of Doctor of Philosophy

Cardiff University

School of Computer Science and Informatics

June 2024

Abstract

As the global population ages, it becomes crucial for the early detection and prevention of neurological aspects of ageing, such as cognitive decline. The human brain ageing process is biologically complex and could be affected by various factors. Therefore, the determination of a person's brain biological ageing process holds important clinical implications, which reflect that person's brain health and may indicate the risk of age-associated brain diseases. To quantitatively measure the brain biological ageing process, *brain age*, defined as the biological age of the brain, has been proposed, which has demonstrated huge potential in clinical diagnosis.

Brain age can be estimated by brain age prediction models, which take in brain-ageing-related information, such as brain MR scans, and adopt machine learning models to make predictions. Despite the rapid research advancements in brain age prediction over the past decade, brain age prediction framework has yet to mature before the implementation in clinical practice. In this thesis, we propose three novel approaches, each focused on a distinct perspective, to make the brain age prediction model a more reliable, practical, and accurate tool for clinical diagnosis.

Our contribution is three-fold: firstly, we propose a skewed loss function to correct a commonly-observed regression bias in brain age prediction models. The skewed loss function unifies the model training and bias correction stages, achieving improved accuracy compared with the state-of-the-art practices in the literature. A dynamic training algorithm is further proposed for the skewed loss function. It adopts a heuristic approach to iteratively tune the hyperparameters of the skewed loss function, which has been proven to be robust to different datasets, model architectures and problem domains. The proposed skewed loss function makes the model produce unbiased estimations of brain age, which makes brain age prediction a reliable and trustworthy tool for clinical use. Secondly, we generalise brain age prediction for clinical-grade low-resolution MR images, which makes it a practical and accessible tool for hospital settings. We propose an integrated workflow

by combining brain super-resolution and age prediction models. Clinical MR images are firstly super-resolved, before being fed into a pre-trained age prediction model, which have been proven to achieve negligible differences in predicting age compared with high-resolution images. A non-uniform sampling strategy is also demonstrated to improve the image reconstruction quality especially in the high-frequency regions of the brain. Lastly, we demonstrate the strength of adopting a multi-modal approach for predicting brain age more accurately compared with uni-modal models. Structural T1-weighted MR images and diffusion MRI measures of the brain microstructures are adopted to provide the model with a more complete picture of brain ageing. A tract-wise training approach is also proposed for predicting brain age from diffusion MRI measures, which performs feature selections in the model training process. It prioritises more age-sensitive features and discards less useful ones for brain age prediction, which achieves an improved accuracy on a relatively small dataset.

Contents

List of Figures

List of Tables

Acronyms

MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
DWI	Diffusion Weighted Imaging
DTI	Diffusion Tensor Imaging
fMRI	functional Magnetic Resonance Imaging
FLAIR	Fluid-Attenuated Inversion Recovery
LASSO	Least Absolute Shrinkage and Selection Operator
SVR	Support Vector Regression
RVR	Relevance Vector Regression
GPR	Gaussian Process Regression
XgBoost	Extreme Gradient Boosting
CNN	Convolutional Neural Network
ResNet	Residual Network
DenseNet	Densely connected Network
ADC	Age-Delta Correlation
MAE	Mean Absolute Error

Acknowledgements

Before the start of this thesis, I would like to take a moment to express my gratitude to all who supported me during this long but fruitful PhD journey.

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Dr. Yuhua Li. Beyond his wealth of research expertise, he has also provided invaluable guidance, offering insights into careers and various aspects of life.

I would also like to convey heartfelt thanks to my other supervisors, Dr. Matthias S. Treder, Prof. Derek K. Jones and Prof. David Marshall. Their expertise has been invaluable in guiding me in my research field. Their unwavering encouragement provided much-needed support during challenging moments throughout my studies.

Furthermore, I would like to extend my appreciation to my best friend, Xueji. Her emotional support has always been a cornerstone of my PhD journey. Engaging in conversations with her has been a source of immense joy and comfort, and her encouragement has bolstered my belief in the value of my studies.

I would also like to thank the China Scholarship Council and Cardiff University, for the financial support, necessary resources and academic environment throughout my PhD journey.

Last but certainly not least, I wish to express my thanks to my family. In times of difficulty, I have always found solace and strength in their unconditional love and encouragement. Their constant belief in me has been a guiding light, and I am endlessly grateful for their unwavering support.

Chapter 1

Introduction

As the world's population ages, the early prevention and detection of neurological aspects of ageing is becoming a public priority. Ageing can result in substantial alterations to the brain structures and is closely related to brain function [1]. The neurological ageing process is biologically complex [2] and could be affected by various brain diseases, such as cognitive decline and dementia. Such brain-related diseases can accelerate the biological process of brain ageing, which makes the brain appear older. Therefore, chronological age does not comprehensively represent the complexity and heterogeneity of neurological brain ageing process [3]. To construct the biological age of the brain, **brain age** has been proposed to assess the biological process of brain ageing [3]–[5].

Brain age offers numerous benefits for clinical diagnosis, allowing the assessment of brain health at the individual level [1]. It has been playing an increasingly crucial role in clinical analysis and applications [6], which holds enormous potential to help establish a deeper understanding of brain biological ageing process.

1.1 What is brain age?

Human biological ageing process is complex and heterogeneous, which can not be comprehensively represented using chronological age. To better understand the biological ageing of the brain, **brain age** has been proposed and defined as the biological age of the brain [3]–[5]. It builds on the understanding that human brain undergoes characteristic changes with advancing age and that various factors, such as brain diseases or lifestyle changes, can accelerate or slow down the natural biological ageing process of the brain.

1.2 What is brain age prediction?

Brain age prediction provides a novel framework to measure the brain age. It typically adopts machine learning (ML) in a prediction task where the machine is trained on brain-ageing related information, such as structural magnetic resonance imaging (MRI), of people without psychiatric or neurological diagnoses, and is later applied to new data, resulting in a **predicted brain age**. Brain age prediction aims to build a mapping between brain-ageing related information and brain biological age through machine learning. Figure 1.1 illustrates an overview of the brain age prediction. It takes brain-ageing related information as input and produce brain biological age as prediction based on the input data.

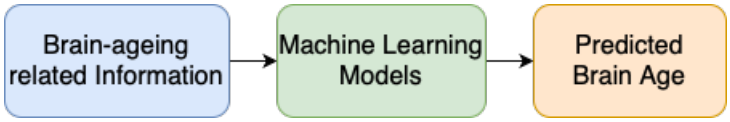


Figure 1.1: An overview of the brain age prediction

Training such machine learning model requires a paired dataset of brain-ageing related data and corresponding brain biological age for each participant. However, there exists no approach to assess the actual brain biological age. Therefore, in the literature, brain age prediction contains two underlying assumptions:

Assumption 1. *People without psychiatric or neurological diagnoses experiences average age-driven changes of the brain trajectories, so that their brain biological age equal to their chronological age.*

Assumption 2. *The brain biological ageing process for people without psychiatric or neurological diagnoses is continuous.*

In that way, brain age prediction model can be formulated as building a machine learning model that takes ageing-related information about the brain, from people without psychiatric or neurological diagnoses as input, and uses chronological ages, as output.

1.3 Why does brain age prediction matter?

The human biological ageing process can be influenced by various internal and external factors, such as genetic differences and brain injury from accidents. Huge variations among people's brain-ageing trajectories may result in the varying ages when people develop age-related brain diseases [4]. Brain age prediction provides an effective approach of measuring brain biological ageing process at the individual level, which make brain age a personalised marker of brain health in the biological ageing process. From Figure 1.1, after the model training phase, the trained model could be applied on new data, resulting in predicted brain age for new participants. Such approach has been widely adopted in research, providing a quantitative approach to investigate whether certain cognitive training or pharmacological treatments can make any changes to the brain biological ageing process [3], [6].

Brain age prediction is also particularly helpful for early prevention and detection of neurological disorders. By comparing the predicted brain age with chronological age, **brain age delta** can be naturally defined by subtracting chronological age from predicted brain age, which can be used to identified whether there is a sign of accelerated or delayed brain ageing. A positive brain age delta represents indicates more prominent brain changes that commonly occur with ageing progression, whereas a negative brain age delta signifies a more youthful brain pattern than expected for that individual's age. Brain age delta has been proven to be correlated with various cognitive disorders, such as mild cognitive decline and dementia [7]–[9].

Therefore, a more complete picture of brain age prediction workflow is demonstrated in Figure 1.2.

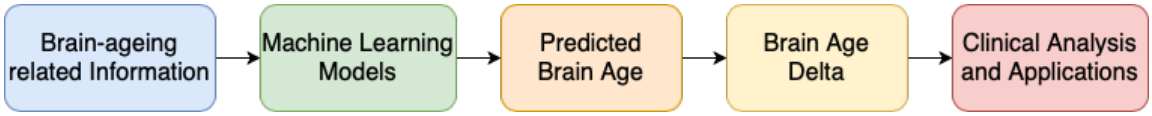


Figure 1.2: An complete overview of brain age prediction

From Figure 1.2, after the model training process, the trained model can be applied on the new participant. Afterwards, the brain age delta of the new participant can be calculated to indicate whether this individual has a higher risk of a certain brain disease. Measuring the brain ageing process at the individual level makes the brain age more crucial in clinical applications for the early diagnosis of neurological diseases, which indicates clear potential for clinical translation.

1.4 Why do brain age and brain age delta matter?

In clinical applications, brain age has been proven to correlate with a range of brain-related disorders and is predictive of mortality risk [10]. It has also been found to be associated with various lifestyle behaviours. Risky behaviours, such as smoking, alcohol consumption [11], high diastolic blood pressure, and obesity [12] have been related to accelerated brain ageing, whereas protective behaviours [13]–[15] show a delay in the brain ageing process. Brain age reflects the biological ageing process of the brain and brain age delta can be adopted to quantitatively measure the process of accelerated or delayed brain biological ageing. Some researches have also identified that brain age delta can be correlated with the severity of cognitive diseases [16]. Patients with Alzheimer's disease experience a high brain age delta (up to +10.0 years), whereas at its earlier stage of mild cognitive decline, patients experience a lower positive brain age delta (up to +6.2 years), which provides clear clinical potential of using brain age prediction as an early-stage screening tool.

Introducing brain age in clinical analysis can help establish a deeper understanding of the relationship between brain biological ageing and neurological diseases. Therefore, measuring brain age can provide a window into general biological ageing as a potential brain-ageing biomarker. It provides a signal indicating the brain health conditions at the individual level, which can help early detection of brain diseases.

1.5 Research landscapes in brain age prediction

Brain age prediction has attracted growing interest over the past decade for its clinical promises in hospital settings. However, despite the rapid development over the past decade, the brain age prediction framework has yet to mature before its implementation in hospital settings and more researches are needed.

From Figure 1.2, the main backbone of the brain age prediction framework consists of three components, **brain-ageing related information (Data)**, **machine learning models (Models)**, and **predicted brain age (Model Predictions)**. In the literature, most studies in brain age prediction focus on at least one aforementioned component to improve brain age prediction framework. In the following subsections, a brief introduction of the current research avenues is illustrated for each aforementioned perspective.

1.5.1 From the data perspective

Brain MRI comes in multiple data modalities, two of the most useful in brain age prediction being structural MRI and diffusion MRI. Structural MRI details the brain's anatomical structures, which is valuable for mapping brain anatomy, detecting structural abnormalities, and tracking changes in brain morphology over time. It also provides detailed images of the brain's gray matter, white matter, and cerebrospinal fluid spaces, allowing researchers to assess the brain volume and cortical thickness. Commonly used MRI sequences in structural MRI include T1-weighted, T2-weighted, and FLAIR (Fluid-Attenuated Inversion Recovery) images, each offering different contrasts to highlight specific tissue types or abnormalities. Diffusion MRI, on the contrary, focuses on the microscopic movement of water molecules, which is influenced by the underlying tissue structure. The most commonly used form of diffusion MRI is Diffusion Tensor Imaging (DTI), which measures the direction and magnitude of water diffusion in the brain. It is critical for mapping white matter tracts, the pathways that connect different regions of the brain. By visualizing these connections, diffusion MRI provides insights into the brain's connectivity and can detect abnormalities in the white matter which may be as-

sociated with various neurological diseases. Different MR image modalities reflect the brain ageing process from different aspects and combining them gives the model a more complete understanding of brain ageing, which typically improves performance. Choosing multiple brain MR image modalities for predicting brain age has been extensively explored in recent years [17]–[19].

Apart from adopting multiple MR image modalities for brain age prediction, adopting large datasets can also improve model performance, especially deep-learning-based models. In the literature, there have been several large-scale brain MR datasets for predicting brain age, such as the UKBioBank [20] and OpenBHB [21] datasets. The UKBioBank dataset includes more than 14,000 participants, with multiple image modalities, such as Structural MRI, diffusion MRI, functional MRI, and Susceptibility-Weighted Imaging (SWI). The OpenBHB dataset gathers more than 5,000 3D T1-weighted brain MRI from Healthy Controls and is highly multi-sites, aggregating more than 60 centers worldwide and over 10 studies. A larger brain MR image dataset essentially increases the diversity of participants, making the model training more stable and increasing the generalisation of models.

1.5.2 From the model perspective

Various models have been proposed over the past decade to predict brain age. Traditional machine learning models and deep learning models have both been adopted in predicting brain age. Generally speaking, traditional machine learning models are more likely to be applied when non-imaging data, such as extracted features from the brain, is adopted for predicting brain age. They are usually fast to train and require significantly fewer computational resources compared with deep learning models. Their predictions are also easier to interpret compared with deep learning models, which is beneficial in clinical environments. Commonly adopted traditional models include ridge regression [18], [22], [23], least absolute shrinkage and selection operator (LASSO) [17], [19], [24], relevance vector regression (RVR) [5], [6], [19], [25], support vector regression (SVR) [18], [19], [24], [26], [27], Gaussian Process Regression (GPR) [1], [18], and extreme gradient boosting

(XgBoost) [19], [28]–[30].

With the rapid advance of deep learning models, various neural network architectures have also been proposed for brain age prediction. Deep learning models are usually adopted when imaging data is used for prediction. Deep learning models demonstrate stronger learning capacities in retrieving information from data compared with traditional machine learning models and do not require any specific data preprocessing techniques. Various deep learning model architectures have been proposed for brain age prediction, such as Convolutional Neural Network (CNN) [1], Residual Network (ResNet) [31], Inception-based network [32], Fully-Convolutional Network [33], Densely-connected network (DenseNet) [34], and Transformer-based network [35], [36]. In the literature, most model architectures for predicting brain age are directly generalised from those in the computer vision field, such as DenseNet [37] and Transformer [38], as the backbones, with additional optimizations for 3D MR images [34], [36]. However, there are also some obstacles in training large deep-learning brain age prediction models. Firstly, MR images are volumetric data, which requires 3-dimensional network layers, increasing the size of models. Secondly, compared with computer vision field, the size of available MR dataset is relatively small, with the largerst MRI dataset having just more than 14,000 images.

1.5.3 From the model prediction perspective

Another research avenue focuses on correcting model predictions after the model training phase to prevent model from being biased or affected by any confounding factors. For example, if the model predictions are biased towards particular groups of participants due to confounding demographic variables, such as gender and age, the validity of brain age as a brain health marker will be compromised.

One of the main prominent biases for brain age prediction models is called “non-zero age-delta correlation (ADC) bias” [1], which has been consistently observed in many studies. It manifests as systematic overestimations of brain age for young individuals and underestimations for elderly participants. Mathematically speaking, the age-delta

correlation (ADC) can be defined as

$$ADC = \text{Corr}(y, f(x) - y) \tag{1.1}$$

where y represents ground truth (chronological age) and $f(x)$ represents model predictions (predicted brain age). And the "non-zero age-delta correlation (ADC) bias" indicates that

$$ADC = \text{Corr}(y, f(x) - y) \neq 0 \tag{1.2}$$

holds true for any brain age prediction models.

Such "non-zero age-delta correlation (ADC) bias" significantly weakens the clinical use of brain age delta by introducing a spurious relationship between brain age delta and chronological age, which typically requires further corrections to derive unbiased predicted brain age [23], [39], [40]. The potential cause of this bias and correction approaches will be more thoroughly discussed in Section 2.2.

1.6 Aims and objectives of the research

As is discussed in previous sections, brain age prediction can be performed at the individual level, which makes brain age a personalised marker indicating brain health, demonstrating huge potential in clinical diagnosis. To validate brain age as the marker, a reliable, practical and accurate brain age prediction model is the essential prerequisite. Therefore, this research aims to develop novel algorithms and methodologies for brain age prediction problems to make brain age prediction models more reliable, practical and accurate, facilitating its clinical use.

To achieve this aim, the objectives of the thesis are summarised as follows:

- Propose algorithm especially for improving the reliability of model predictions, which prevents potential regression bias in predicting brain age and makes the model produce unbiased estimations of brain age.
- Propose algorithm especially for enhancing the practicality of brain age prediction,

which makes brain age prediction accessible in clinical environments, allowing the use of clinical-grade low-resolution MR images for predicting brain age.

- Propose algorithm especially for boosting the accuracy of predicting brain age, which adopts the brain macro- and micro-structural data for multi-modal learning.

1.7 Thesis contributions and publications

The thesis presents a substantial body of work for brain age prediction methodologies from three perspectives, which aim to achieve more reliable, practical and accurate brain age prediction models. The main contributions of the thesis are summarised as follows:

- The first contribution of this thesis is about correcting a commonly observed regression bias in brain age prediction models to make the models more reliable. A skewed loss function is proposed to correct the regression bias directly in the model training phase, replacing the conventional symmetric loss functions, such as the L1 or L2 loss function. The skewed loss assigns different loss values depending on whether the model overestimates or underestimates a participant's age and is subject to different participants' chronological ages. It is the first integrated bias correction approach that can be adopted for deep learning models. A dynamic training algorithm is further developed to search the optimal hyperparameters for the skewed loss function. The proposed algorithm adopts a heuristic approach to tune the hyperparameters iteratively, which has been proven to be robust to different datasets, data distributions, model architectures, and problem domains.

- The second contribution of this thesis is about brain MRI super-resolution and its application in brain age prediction models to make brain age prediction easily accessible and practical. A direct and effective non-uniform sampling algorithm is proposed for improving the reconstruction quality of high-frequency information details in brain MRI super-resolution. This algorithm is the first work that treats different brain regions separately, depending on their structural differences, to improve the super-resolution quality of brain MR images. A modified brain age

prediction workflow is then developed by integrating the brain MRI super-resolution techniques before predicting brain age, which extends brain age prediction for low-resolution clinical-grade MR images. The super-resolved MR images achieve indistinguishable accuracy compared with original high-resolution MR images, which makes it feasible for clinical-level brain age prediction.

- The third contribution of this thesis is about multi-modal brain age prediction using brain microstructural data and structural images to boost the predictive accuracy of the model. A tract-wise approach is proposed for brain age prediction using diffusion MRI measures and quantitative MRI measures of the white matter tracts. The tract-wise approach prioritises more informative tracts of the brain in the model training process, providing insights into more ageing-sensitive white matter tracts. Unlike some schemes that perform feature selection before model training, the tract-wise approach preserves the most complete information, keeping all white matter tracts in consideration. A performance comparison is carried out between structural images and microstructural measures in predicting brain age. A two-stream framework is then proposed by combining information from structural images and white matter microstructure measures, which provides the model with a more complete picture of the brain, achieving improved accuracy compared with any single data modality.

One journal paper is published based on the first contribution of the thesis, one conference paper is accepted based on the second contribution of the thesis, and another journal paper is under review based on the third contribution of the thesis. The papers are as follows:

- **Wang, H.**, Treder, M., Marshall, D., Jones, D. and Li, Y., 2023. A skewed loss function for correcting predictive bias in brain age prediction. IEEE Transactions on Medical Imaging, 42(6), pp.1577-1589.

- **Wang, H.**, Marshall, D., Jones, D. and Li, Y., 2024. Improving High-Frequency Details in Cerebellum for Brain MRI Super-Resolution (Accepted by 4th IEEE

International Conference on ICT Solutions for eHealth).

- **Wang, H.**, Mcnabbs, C., Messaritaki, E., Laguna, P., Jones, D. and Li, Y., A Tractwise Approach for Brain Age Prediction using Measures of White Matter Microstructure (under review by IEEE Journal of Biomedical and Health Informatics).

1.8 Thesis outline

The remainder of the thesis is organised as follows:

- Chapter 2 presents a literature review. First, the development of the brain age prediction problem and its clinical applications are reviewed. Then, the state-of-the-art methodologies for correcting regression model bias for brain age prediction are thoroughly studied. Afterwards, the recent advances in super-resolution techniques for 2D natural images and 3D brain MR images are discussed. Finally, studies that adopt multi-modal brain MRI data for predicting brain age are summarised.

- Chapter 3 focuses on a commonly-observed regression model bias in brain age prediction problem. A skewed loss function approach is proposed to correct such bias to derive unbiased estimations of brain age. To stabilise model performance, a dynamic algorithm is also developed to search for the optimal hyperparameters for the skewed loss. The chapter concludes with a performance comparison between the proposed skewed loss and the state-of-the-art correction approaches.

- Chapter 4 emphasises the usability of integrating the super-resolution techniques into the brain age prediction, extending the brain age prediction to low-resolution MR images. A non-uniform sampling strategy for brain MRI super-resolution is also demonstrated to make the model focus more on the under-represented regions of the brain, which improves the reconstruction quality for high-frequency details in the brain.

- Chapter 5 highlights the potential of adopting measures of white matter microstructure, other than structural MR images, for brain age prediction. A tract-wise ap-

proach for predicting brain age is proposed by prioritising more informative white matter tracts, which provides insights into the more ageing-sensitive white matter tracts. Brain age prediction using structural images is also adopted as the contrast. A two-stream approach is further proposed by fusing the information from structural images and microstructure measures to produce more accurate estimations of brain age.

- Chapter 6 concludes the thesis and summarises the main contributions of this research. It also suggests potential future research avenues for brain age prediction.

1.9 Summary

This chapter presents an overview of the background of brain age prediction and its research avenues. Brain age prediction aims to estimate the biological ageing of the brain using brain-ageing related information, such as brain MR images, with the help of machine learning models. Despite the rapid research advancements over the past few years, there is still a huge gap for its clinical use in hospital settings. This thesis focuses on three distinct aspects of brain age prediction to make it a more reliable, practical, and accurate tool,facilitating its application in hospital settings. In the next chapter, a thorough literature review on the evolution of brain age prediction and its clinical importance is presented.

Chapter 2

Literature review

In this chapter, we begin by exploring the evolution of brain age prediction over the past decade, focusing on key techniques and their clinical significance. We then review recent advancements aimed at enhancing the accuracy of brain age prediction, examining these improvements from three distinct perspectives, which correspond to the detailed discussions in Chapter 3, Chapter 4, and Chapter 5.

2.1 Brain age prediction

In this section, we review the development of model architectures for brain age prediction and clinical applications of brain age.

2.1.1 Brain age prediction models

In the literature, both traditional machine learning models and deep learning models have been adopted for predicting brain age over the past decade. Traditional machine leaning models and deep learning models have their strengths and weaknesses, and no one significantly outperforms the other in terms of model accuracy. It should also be stressed that although there are some public datasets that have been adopted for brain age prediction research, there is no public benchmark datasets for model evaluations. Also, different brain MR image modalities have been adopted for predicting brain age in different studies, which makes it difficult to fairly compare model performance across different studies. There are also many studies that propose different models using private datasets, which further makes it difficult for public evaluations. Another factor that significantly affects model performance is the age range of the training dataset. Models

trained on a dataset with large age range tend to demonstrate larger mean absolute error (MAE) compared with models trained on a smaller age range dataset.

Generally speaking, traditional machine learning models are more likely to be applied when non-imaging data, such as extracted features from the brain, is adopted for predicting brain age. They are usually fast to train and require significantly fewer computational resources compared with deep learning models. Their predictions are also easier to interpret compared with deep learning models, which is beneficial in clinical environments.

Commonly adopted traditional models include ridge regression [18], [22], [23], least absolute shrinkage and selection operator (LASSO) [17], [19], [24], relevance vector regression (RVR) [5], [6], [19], [25], support vector regression (SVR) [18], [19], [24], [26], [27], Gaussian Process Regression (GPR) [1], [18], and extreme gradient boosting (XgBoost) [19], [28]–[30]. Considering the wide range of brain-ageing datasets adopted in these studies, the model performance also varies across different studies and different datasets. Xiong *et al.* [19] adopted four models, LASSO, RVR, SVR, and XgBoost on the diffusion MRI measures data from the UKBioBank dataset [20] and achieved an MAE around 3 years. Cole *et al.* [17] adopted LASSO on multiple MR image modalities and achieved an MAE of around 4 years.

With the rapid advance of deep learning models, various neural network architectures have also been proposed for brain age prediction. Deep learning models are usually adopted when imaging data is used for prediction. Deep learning models show stronger learning capacities in retrieving information from data and do not require any specific data preprocessing techniques. MR images, even after being minimally preprocessed, such as image registration, can achieve satisfactory accuracy using relatively simple deep learning models [1]. Cole *et al.* [1] first demonstrated the feasibility of performing brain age prediction on raw 3D structural images using a plain convolutional network with repeated 3D convolutional layers. Li *et al.* [31] later applied a 3D ResNet architecture on the resting state functional MRI data to predict brain age. Rao *et al.* [32] developed an

inception-based model [41] with global average pooling. Peng *et al.* [33] built a fully convolutional network, named Simple Fully Convolutional Network (SFCN), together with KL-divergence as the loss function, which successfully achieved a new state-of-the-art performance and ranked first place in the PAC2019 competition. Wood *et al.* [34] adopted the 3D DenseNet architecture and applied brain age prediction on the clinical-grade MR images, which generalises the brain age prediction on real-world MR images. Yin *et al.* [42] introduced demographic variables for more accurate brain age prediction and evaluated the interpretability of 3D convolutional networks. Transformer-based models [38] have also been applied in recent years. He *et al.* [36] applied the transformer to extract global information from whole brain images and local information from image patches and introduced a global-local transformer for brain age prediction. Hu *et al.* [35] combined the transformer and the squeeze and excitation block [43] to help the transformer to learn better local details. However, due to the computational cost, transformer-based models are typically trained on 2D slices rather than 3D volumes.

Comparing model performance across different studies is also difficult for deep learning models. Cole *et al.* [1] trained their CNN on the Brain-Age Healthy Control (BAHC) dataset, with 2001 subjects, achieving an MAE around 5 years. Peng *et al.* [33] trained their SFCN using UKBioBank dataset and achieve an MAE between 2 to 3 year. Wood *et al.* [34] trained their DenseNet on private T2-weighted MRI dataset, achieving an MAE below 2 years.

Although various deep neural network architectures have been introduced in brain age prediction, the backbones of most models are generalised from their 2D versions, with some additional optimizations for 3D volumetric data. Therefore, some standard models, such as SFCN [33] and DenseNet [34], have still been widely adopted nowadays, considering their computational speed and learning efficiency.

2.1.2 Clinical applications of brain age

Brain age has been adopted in various clinical assessments, aiming to help the early detection of brain-related disorders and may lead to more timely and targeted interventions.

It has been adopted in various neurological conditions, such as the Alzheimer's disease (AD) [7], [9], [44], [45], mild cognitive impairment (MCI) [7]–[9], [44]–[46], and Traumatic brain injury [4], [47]. Also it has been linked to psychiatric disorders such as schizophrenia [48]–[51], including major depressive disorder (MDD) [51]–[56] and autism spectrum disorder [56]–[58]. Mohajer *et al.* adopted brain age prediction to confirm that gray matter atrophy is associated with accelerated biological brain aging in patients with mild cognitive impairment and Alzheimer's disease [7]. Beheshti *et al.* demonstrated that the brain biological ageing process accelerates along with the severity level increase of Alzheimer's disease using a brain-age score [44]. Participants with mild cognitive impairment experience +2.38 years accelerated brain ageing, whereas participants with Alzheimer's disease experience +5.36 years accelerated brain ageing. Koutsouleri *et al.* investigated how schizophrenia, major depression, personality disorder, and individuals in at-risk mental states for psychosis affect the neuro-anatomical age using brain age prediction [51]. Group analysis demonstrated that patients with schizophrenia experience the highest level of brain ageing acceleration, achieving +5.5 years on average.

Apart from the relationship with brain-disorder diseases, brain age has also been linked to various markers of health. Smoking, alcohol consumption [11], high blood pressure [12], and obesity [12], [59], [60] tend to be associated with a positive brain age delta, whereas regular meditation [13], practising music [14], and higher education level [15] have been linked with delayed brain ageing. Ronan *et al.* demonstrated that overweight and obese individuals are associated with a greater degree of brain volume atrophy, with maximal effects in middle-age corresponding to estimated 10 years brain ageing acceleration [59]. Luders *et al.* applied brain age prediction for meditators' brain and found out at the age of 50, there exists an average 7.5 years delay in brain ageing process compared with control groups [13]. Also, for every additional year over fifty, meditators' brains were

estimated to be an additional 1 month and 22 days younger than their chronological ages.

Adopting brain age in clinical assessments can effectively indicate the risk of potential disease, even if no obvious clinical symptoms are observed [6]. If a high positive brain age delta is observed, further tests could be conducted, which may reveal any pre-symptomatic disorders. Alzheimer's disease has been found to result in a high positive brain age delta [7], [9], [44], [45], whereas its preclinical stage, mild cognitive impairment, has been found to show a smaller brain age delta [7]–[9], [44], [45]. Therefore, adopting brain age can be used to detect brain disorders in the early stage before the neurological conditions become worse.

Despite the wide adoption of brain age prediction in brain ageing analysis, the brain age prediction framework has yet to mature before the implementation in clinical settings. To make brain age prediction applicable in hospital settings, some key problems need to be solved [3]. Firstly, this field lacks consensus regarding the construction and evaluation of developed brain age prediction models. Different datasets are adopted for model building and testing, which make it difficult to compare different models. Secondly, most brain age prediction models are trained and tested on high-resolution data from curated databases that do not represent current practices in clinical imaging. Thirdly, brain-age estimation can be subjected to confounding factors related to input characteristics deriving from differences in scanners, image acquisition protocols, image quality, and preprocessing pipelines [3]. Lastly, some statistical age-bias correction approaches, which can in certain cases artificially inflate the model's prediction accuracy or, under specific circumstances. Therefore, care must be taken when performing different bias corrections.

2.2 Regression model bias

In the previous section, we highlighted the wide range of clinical applications of brain age, demonstrating its strong potential as a biomarker of brain health. In this section, we turn to the first objective of this thesis and present reviews on the regression bias

in brain age prediction models. We first explain the reasons that cause this bias and then demonstrate why such bias matters. Afterwards, we review the existing correction approaches to tackle such bias.

In Section 1.5.3, we introduced the so-called "non-zero age-delta correlation bias". Such bias has been frequently observed in brain age prediction models [1], [10], [22], [26], [33], [34], [39], [40], [61]–[64]. The predicted brain age tends to become older than the actual chronological age for young participants and younger for older participants. In Section 1.5.3, we noted that such bias behaves as a strictly non-zero correlation between chronological age and brain age delta.

In the literature, this correlation is usually measured using Pearson's r correlation [33], [34], [62] or Spearman's rank correlation [1], [39]. The Pearson's r is defined as

$$r = \frac{n\sum y\Delta - \sum y \sum \Delta}{\sqrt{(n\sum y^2 - (\sum y)^2)\,(n\sum \Delta^2 - (\sum \Delta)^2)}} \tag{2.1}$$

where n, y, Δ represents the number of participants, chronological age, and brain age delta respectively. The Spearman's rank correlation is defined as

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{2.2}$$

where d_i is the difference between the ranks of chronological age and brain age delta, and n is the number of participants.

Pearson's r measures statistical dependence between two sets of data in terms of the linear correlation of the variables. Spearman's rank correlation measures the linear correlation between the rank values of two variables and hence quantifies monotonic relationships. The range of both correlation coefficients lies between -1 (negative correlation) and 1 (positive correlation). In this way, the bias manifests as a negative age-delta correlation (ADC) value.

2.2.1 Why does the bias arise?

There have been different explanations of the cause of this bias in the literature. Liang et al. [22] found that a negative ADC value consistently arises in a range of ageing datasets regardless of the regression models being used. Le et al. [26] argued that bias is inevitable for regression models due to the Regression to the Mean (RTM) phenomenon and hence not limited to ageing datasets. Smith et al. [62] observed that a penalized regression model and a non-Gaussian distribution of the participants' chronological age can also cause the model to make predictions towards the mean age. In Section 3.2.1, we will provide another potential cause of such bias based on mathematical evaluations.

In real-world experiments, a relatively large ADC may be observed more frequently when adopting relatively simple models and smaller datasets [22], [33], [39], [62]. There are also some recent studies claiming that their models achieve a negligible bias using more advanced model architectures [34], [42] and larger datasets. It may imply that such age-related bias can be alleviated using more sophisticated model architectures.

2.2.2 Why does the bias matter?

Another question we would like to answer about the bias is why does it matter for subsequent analyses.

The purpose of the brain age delta is to indicate overall brain health in the ageing process. When brain age delta is adopted as the independent variable in the subsequent analysis on a covariate A, one should always check whether A is associated with chronological age or not. If covariate A is not related to chronological age, then it is safe to evaluate the bivariate relationship between brain age delta and variable A. However, if A is correlated with chronological age, a spurious relationship will naturally arise between the brain age delta and variable A. The apparent relationship between brain age delta and variable A might be more driven by the chronological age and not the brain age delta [62]. In that case, the chronological age confounds the relationship between brain age delta and variable A [65]. Therefore, the inclusion of chronological age dependence of brain age

delta reduces its effectiveness as a biomarker of brain health.

2.2.3 Bias correction approaches

Although the observed bias may be alleviated by adopting more sophisticated models and larger training datasets [34], the bias still needs to be eliminated, especially in situations with low computing resources. To address such bias in brain age delta, a two-stage correction approach has been proposed and widely adopted in the literature [10], [22], [39], [40], [62], [63], [66]. It introduces a second-stage correction to correct model predictions from the first stage, i.e., the brain age prediction model and the resultant corrected brain age delta will then have no linear relationship with chronological age [39]. The pipeline of the two-stage correction approach is demonstrated in Figure 2.1.

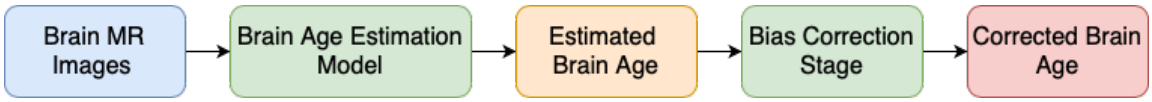


Figure 2.1: The pipeline of the two-stage correction approach. The "Corrected Brain Age" block represents the final estimation of brain age and the corrected brain age should have no linear relationship with chronological age.

The detailed two-stage correction approach implementations can be summarised below:

1. Stage 1 - Brain Age Prediction: Train a regression model f to predict chronological age (Y) given brain MR images (X). The uncorrected brain age delta is defined as

$$\delta = f(X) - Y \tag{2.3}$$

2. Stage 2 - Bias Correction: Remove the dependence of uncorrected brain age delta δ on the chronological age (Y).

Two different approaches have been proposed in the bias correction stage:

- Approach 1 [22], [62], [63], [66]:

    1. Fit a linear regression between uncorrected brain age delta δ and chronological

age Y

$$\delta = \beta_1 \times Y + \beta_0 \tag{2.4}$$

2. The corrected predicted brain age is defined as

$$f(X)_{corrected} = f(X) - (\beta_1 \times Y + \beta_0) \tag{2.5}$$

3. The corrected brain age delta is then defined as

$$\delta_{corrected} = f(X)_{corrected} - Y \tag{2.6}$$

- Approach 2 [10], [39]:

1. Fit a linear regression between predicted brain age $f(X)$ and chronological age Y

$$f(X) = \beta_1 \times Y + \beta_0 \tag{2.7}$$

2. The corrected predicted brain age is defined as

$$f(X)_{corrected} = (f(X) - \beta_0)/\beta_1 \tag{2.8}$$

3. The corrected brain age delta is then defined as

$$\delta_{corrected} = f(X)_{corrected} - Y \tag{2.9}$$

Regarding the first approach, there have been several studies [23], [64], [67] arguing that this approach essentially corrects the target label (Y) and the model prediction is still biased. Moreover, in a machine learning framework, the target value (Y) from the test set should remain unknown when the correction approach is applied. Otherwise, one can always make corrections towards the target value (Y) to achieve a significantly lower error, which may be regarded as a data leakage problem [64]. The first approach uses the chronological age (target value) to make corrections in (2.5), which violates the principles of this predictive framework.

To apply the second approach in the brain age prediction framework, the linear regression (2.7) is performed on the validation set and the resultant parameters (β_0, β_1) are assumed to be generalised to the test set. In that way, the predictions from the test set can be

corrected using (2.8).

The two-stage correction approach has become prevalent due to its robustness and user-friendly nature. In Chapter 3, we will conduct a more detailed analysis of the two-stage correction approach and we refer to the second approach [10], [39] as the default two-stage correction approach in the rest of this thesis.

Despite the wide adoption of the two-stage correction approach, there also exist other correction approaches to tackle the bias. Zhang et al. [40] argued that the two-stage correction approach performs a group-level correction and ignores the bias for each age group. They proposed age-level bias correction which can be integrated as an additional stage of the two-stage correction approach. It performs a second correction for each age group, which can be summarised as

$$\delta_i^{ac} = (\delta_i - u_a)/\sigma_a \tag{2.10}$$

where δ_i^{ac} denotes the age-level corrected delta over a sample of age a, u_a and σ_a denote the mean and standard deviation of the corrected delta from (2.6) and (2.9) over a sample of age a.

In that way, the bias in each age group can be eliminated since

$$\mathbb{E}_a[\delta_i^{ac}] = \mathbb{E}_a[(\delta_i - u_a)/\sigma_a] = (\mathbb{E}_a[\delta_i] - u_a)/\sigma_a = 0 \tag{2.11}$$

However, the age-level bias correction [40] also needs to use the target value (Y) in the correction, as is shown in (2.10).

Another argument about the two-stage correction approach is that it does not take ADC directly into account. Instead of correcting the predictions of a biased model, Treder *et al.* [23] proposed to fuse both stages by integrating a correlation constraint into the model training stage. This results in a regression model that is unbiased to start with and hence does not require post-hoc correction. To this end, the overall constrained optimization problem can be defined as

$$\begin{aligned} \min \quad & \widetilde{\mathcal{L}}(y, \hat{y}) \\ \text{s.t.} \quad & |\text{Corr}(y, \delta)| \leq \rho \end{aligned} \tag{2.12}$$

However, the authors use an analytical solution for ridge regression and kernel ridge regression which does not extend to deep learning models.

2.2.4 Regression model bias summary

To sum up, the observed bias in brain age prediction models prevents the direct use of the brain age delta as a biomarker and such bias needs to be addressed. The bias arises due to the irreducible level of noise in the dataset, resulting in the model's incapability to model target labels accurately. Using more advanced models may alleviate such bias whereas it requires more computational resources. Since chronological age is usually a crucial factor in many brain-related clinical studies, the inclusion of this bias may introduce a spurious relationship with other variables of interest if other variables are also associated with age. The apparent relationship between the brain age delta and variables of interest might be more driven by chronological age and not the brain age delta [62]. To address this bias, different approaches have been proposed. A two-stage correction approach has been widely adopted, considering its robustness and user-friendly nature. By adding a correction stage after the model training process using linear regressions, the resultant brain age delta will have no linear relationship with chronological age.

2.3 MRI super-resolution

In the previous section, we focused on methodologies for bias correction in brain age prediction models. In this section, we switch to a different aspect, brain MRI super-resolution, which could be incorporated into the brain age prediction to make it feasible to predict brain age on low-resolution MR images. We first explain the background of super-resolution and how does super-resolution model work Then, we review the development of deep-learning-based super-resolution over the past decade and demonstrate its recent advances in MR imaging applications.

2.3.1 Super resolution background

MRI super-resolution aims to increase the image resolution for low-resolution MR images, by reconstructing more fine-grained details, matching the details in the high-resolution images. There are many different super-resolution strategies, such as single image super-resolution (SISR) and multiple image super-resolution (MISR). In this section, we focus on single image super-resolution (SISR) for brain MR images.

Training SISR models requires paired low resolution (LR) and high resolution (HR) images. The LR images are typically generated from the corresponding HR images. The resolution downgrading process from the HR image X to the LR image Y could be simplified as:

$$Y = f(X) \tag{2.13}$$

where f is a certain type of function that causes a loss of image resolution. The SISR problem is to find an inverse mapping function g which could recover the HR image $\hat{X}$ from the LR image Y:

$$\hat{X} = g(Y) = f^{-1}(Y) + r \tag{2.14}$$

where r is the reconstruction residual.

In the literature, deep learning models have demonstrated superior performance in SISR problem. The training process of SISR are as follows. Firstly, LR images are fed into the training model, producing the model output image, which has the same shape with the corresponding high-resolution image. Secondly, the differences between model output image and target HR image are evaluated using some loss functions, such as L1 loss. Lastly, model parameters are updated based on the loss values from the second step.

2.3.2 Deep-learning-based super-resolution

Super-resolution (SR) using deep learning models has experienced significant progress over the past decade. Dong *et al.* first proposed SRCNN in 2014, establishing the basic structure for SR models [68]. Kim *et al.* later proposed VDSR that learns the residual information between the low resolution (LR) and high resolution (HR) images [69], greatly reducing the learning difficulty. Motivated by the success of the ResNet [70], several ResNet-based SR models have been proposed [71], [72], achieving superior performance with the help of the residual connection structure. In addition, the dense connection structure [37] proposed by Huang *et al.* has also been applied to SR problems. Tong *et al.* applied several dense blocks and added dense connections between different blocks to build the SRDenseNet [73]. Zhang *et al.* later took the advantages of both residual and dense connection to propose a residual dense network to further improve the reconstruction quality [74]. With the help of the Transformer models [38], Lu *et al.* proposed ESRT, which consists of CNN and transformer backbones, achieving competitive results with low computational costs [75].

Generative adversarial network (GAN) [76] has also been applied to SR problems. Ledig *et al.* proposed SRGAN, which generates more realistic images compared with a CNN-based model [72]. Bell-Kligler *et al.* introduced an unsupervised model named KernelGAN [77], using real LR images for model training. In addition, various model design strategies, such as the learning-based upsampling [78], [79], recursive learning [80], [81] have also been proposed for SR.

However, most of these works are designed for natural 2D images and a direct conversion into their 3D versions generally fails to achieve a good trade-off between the model complexity and accuracy.

2.3.3 Super-resolution on MR images

Brain age prediction models are typically trained using research-level high-resolution MR images so that the model can learn more detailed structures from the brain. However, it

becomes an obstacle to applying brain age prediction for clinical diagnosis where high-resolution scanners may not be available. To mitigate the image resolution differences between clinical and research-level images, MRI super-resolution becomes a straightforward solution. Incorporating super-resolution models into the brain age prediction framework makes it feasible to predict brain age in clinical environments where high-resolution MR scanners are not accessible.

MRI super-resolution does not require extra scan time or high-cost scanners to generate a high spatial resolution image and can be used to enhance the image quality from the low-field scanners. The overall pipeline of brain age prediction after the inclusion of MRI super-resolution is demonstrated in Figure 2.2

Figure 2.2: The brain age prediction pipeline with super-resolution included. The LR images are first fed into super-resolution models, generating corresponding super-resolved images, which will be used for brain age prediction.

Unlike natural images, structural MR images contain four dimensions, greatly increasing the model complexity as well as learning difficulty. Therefore, models using 2D image slices or 3D image volumes have both been explored in recent years.

2D models are generally more stable and faster to train, whereas they normally fail to learn the information from the third dimension. Zhu *et al.* proposed a 2D multi-scale GAN with a lesion-focused approach to achieve a more stable training process and better perceptual quality for SR images [82]. You *et al.* presented a 2D GAN model incorporating cycle consistency and residual learning [83]. Du *et al.* built a 2D CNN with residual connections for SR reconstruction of single anisotropic MR images [84]. Lyu *et al.* established an ensemble learning framework using 2D GANs to integrate complementary SR results from each GAN model [85]. Zhao *et al.* introduced a channel splitting block to incorporate different information from different receptive fields to increase the model learning capacity [86].

On the contrary, 3D models can utilize the additional information from the volumetric images and outperform the 2D models at the cost of model complexity. Pham *et al.* first demonstrated that for brain MRI, 3D models outperform 2D models by a large margin [87]. Chen *et al.* applied the dense connection structure extensively and proposed the 3D mDCSRN model [88], which contains relatively few parameters. A GAN-based model [88] has also been built by Chen *et al.*, producing more perceptually better results. Li *et al.* proposed a lightweight network, called VolumeNet, using separable 2D cross-channel convolutions [89], which has fewer parameters but still achieves superior performance. Zhang *et al.* adopted the squeeze and excitation architecture [43] and attention mechanism to make the model learn from the more informative regions of the brain [90]. With the help of the Transformer models [38], Li *et al.* proposed a multi-scale contextual matching method that can learn information from different scales [91]. Zhou *et al.* applied a domain correction network that can super-resolve unpaired MR images [92].

There are also lots of works focusing on different aspects of MRI super-resolution, such as models handling multi-scale LR images [93], [94], image quality transfer [95]–[97], and models trained using real-world LR images [97], [98], and sub-cellular super-resolution techniques [99]–[102].

To sum up, existing works on MRI super-resolution have achieved superior performance. However, even the most advanced models, such as attention-based [90] or transformer-based model [91], fail to handle the whole brain volume and the cerebellum at the same time. The attention mechanism applied in such models [90], [91] offers an indirect way of learning from more informative areas, whereas the performance is poor on the cerebellum.

2.4 Multi-modal MRI for brain age prediction

In the aforementioned two sections, we focused primarily on MR imaging-based models. However, there are also lots of works focusing on predicting brain age using MR non-imaging data.

Brain MRI data comes in many data modalities, such as structural T1-weighted imaging,

structural T2-weighted imaging, functional MR imaging, and diffusion-weighted imaging. Different imaging modalities reflect the brain ageing process from different perspectives and combining various imaging modalities provides the model with a more complete picture of brain ageing, which generally improves the model performance. Among all MR imaging modalities, T1-weighted MRI and diffusion-weighted MRI tend to demonstrate significantly better performance in brain age prediction in the literature [17]–[19], [30], [33], [34], [103]. T1-weighted MRI reveals the anatomical details of the brain structure, whereas diffusion MRI measures the integrity of the brain microstructural tissue compartments and connectivity between different brain regions.

Different data modalities are preferred by different types of models. Deep learning models prefer structural MR imaging data, whereas diffusion-weighted imaging data is first processed to extract relevant features and is typically trained using traditional statistical models.

In this section, we first summarise works focusing on MR imaging data and MR non-imaging data, using structural T1-weighted imaging and diffusion-weighted imaging as examples respectively. Then we demonstrate the effectiveness of adopting multi-modal models in brain age prediction compared with uni-modal models.

2.4.1 Predicting brain age using structural MR imaging data

With the development of deep learning techniques over the past decade, more effective and efficient model architectures have been proposed [37], [38], [70], which attracts growing interest in adopting MR imaging data for accurate brain age prediction [1], [33], [34], [104].

Adopting imaging data for brain age prediction brings numerous benefits. First, the raw imaging data contains the most complete information about the brain. Further in-depth processing of the imaging data may extract more useful features, whereas it also faces the risk of losing information about the brain. Second, using imaging data naturally introduces deep neural networks as the prediction model. Neural networks

have demonstrated their excellent capacities of learning information from imaging data in the computer vision field, such as object detection [105], image segmentation [106], and image super-resolution [107]. Therefore, adopting neural networks allows the models to learn more high-level abstract brain features, which may discover more valuable insights into brain ageing. Due to the strong learning capacities of neural networks, feeding imaging data into the neural networks to predict brain age has been widely adopted in the literature [1], [17]–[19], [33], [34]. Third, imaging data requires fewer preprocessing steps before it can be used for brain age prediction. Take the structural T1-weighted MR image as an example. The preprocessing of it generally contains skull-stripping, image registration, bias field correction, intensity normalization, and black background removal, which can be effectively performed using open-source libraries, such as FMRIB Software Library (FSL) [108] and Advanced Normalization Tools (ANTs) [109]. On the contrary, non-imaging data requires more steps to extract relevant features.

In Section 2.1, we reviewed the development of brain age prediction models and highlighted several efficient model architectures, such as DenseNet [34], SFCN [33], and transformer-based model [36]. It should be noted that different imaging modalities can always share the same model architecture to make predictions as long as they can be represented as 2D-slice or 3D-volume data.

2.4.2 Predicting brain age using MR non-imaging data

The brain-related non-imaging features mainly come from more in-depth analysis of the brain imaging data, such as diffusion MRI measures of the brain microstructure [30], [110]. Despite the advantages of utilizing MR imaging data for brain age prediction, using non-imaging features also holds its strengths. First, the complexity of the non-imaging feature dataset is significantly smaller compared with 3D volumetric images. The feature dataset usually comes in tabular format, which allows us to use simpler models, such as XGBoost [111] and support vector regression [112], for faster training with a lower requirement of computing resources. Second, it allows us to apply many traditional ML techniques with more theoretical foundations. Commonly adopted techniques, such as

Principal Component Analysis (PCA) for dimensionality reduction [113], SMOTE [114] for data up-sampling, and SHAP [115] for explaining model predictions, can be applied in the model pipeline to derive better performance.

To demonstrate some studies using non-imaging data, we select diffusion measures of brain microstructure as an example. Diffusion MRI has become a powerful imaging modality and provides a unique view into the physical properties of the connections in the white matter tracts. There has been a lot of research on how diffusion MRI measures change as age increases, on some particular white matter tracts [116]. Wang *et al.* [117] adopted the FA map from diffusion MR images and trained a 3D convolutional network to predict brain age. Chamberland *et al.* [113] applied the principal component analysis (PCA) on the diffusion MRI measures data of the white matter tracts and found a relatively strong correlation between the first principal component and age. Cole *et al.* applied the least absolute shrinkage and selection operator (LASSO) to perform feature selection, deriving a sparse linear model [17]. Chen *et al.* [118] applied a cascade network without any feature selection and claimed to have a better performance compared with classic machine learning models. Beck *et al.* [30] investigated the trends of different diffusion MRI measures with age and adopted the XGBoost model to predict age. However, their accuracy is relatively poor, due to the lack of a proper feature engineering process. Richie-Halford *et al.* [110] developed the sparse group lasso method to select more informative features from the tracts by enforcing the global and bundle-level sparsity. Xiong *et al.* further evaluated several machine learning models using predefined dMRI measures but ignored proper feature selections [19]. Cai *et al.* proposed a two-stream framework by combining DenseNet and Transformer for brain age prediction [119]. Instead of adopting conventional diffusion MRI measures, they extracted diffusion MRI features using the DenseNet automatically.

2.4.3 Multi-modal models for brain age prediction

Given various brain-related data modalities in the literature, it is straightforward to consider adopting multi-modal models for brain age prediction. Since different data modal-

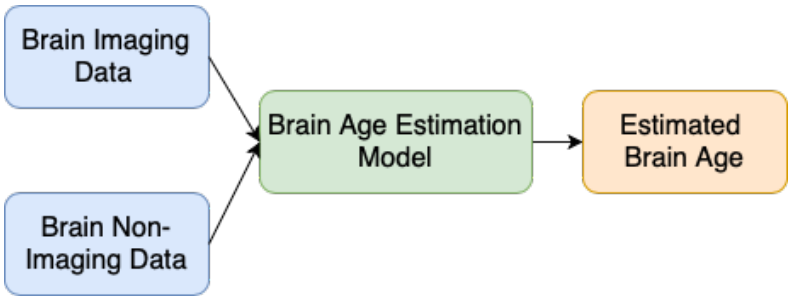


Figure 2.3: The brain age prediction pipeline with multi-modal inputs.

ities reflect brain ageing from different perspectives, combining different modalities of data provides a more complete picture of brain ageing for brain age prediction models. It has been validated in many studies that adopting multi-modal models experience a significant improvement in predictive accuracy compared with uni-modal models [17]–[19], [33]. The overall pipeline of brain age prediction with multi-modal inputs is demonstrated in Figure 2.3. However, it should also be noted that most studies claiming the adoption of multi-modal models do not represent the real multi-modal models. In brain age prediction literature, a more commonly adopted approach is to train different models using different data modalities and combine the model predictions using model ensemble or model averaging. Model architectures, that have multiple inputs, each dedicated for a specific data modality, have been rarely explored.

2.5 Summary

In this chapter, we provide a thorough literature review in four perspectives. We first summarise the evolution of brain age prediction models and demonstrate clinical applications of brain age prediction. Then, we review the potential reasons and correction approaches in the literature for the commonly observed "non-zero age-delta correlation bias". Afterwards, we give a brief introduction on brain MRI super-resolution by demonstrating the evolution of super-resolution techniques for 2D natural images and 3D MR images. Lastly, we summarise different machine learning models for different MRI modalities, such as structural MRI and diffusion MRI, and introduces some recent advancements by adopting multi-modal approach for brain age prediction.

In the next chapter, we turn to the first objective of this thesis to correct the "non-zero age-delta correlation bias" to make brain age prediction more reliable and trustworthy.

Chapter 3

Reliable Brain Age Prediction - A Novel Skewed Loss Function for Debiasing Regression Models

In this chapter, we focus on the first objective of this thesis to debias brain age prediction models to make brain age a reliable marker of brain health.

We propose a novel bias correction algorithm for regression models by introducing a skewed loss function to replace the ordinary loss functions, such as L1 loss, during the model training phase. The regression model then behaves differently depending on whether it makes overestimations or underestimations. This approach works with any type of MR image and no specific preprocessing is required, as long as the image is sensitive to age-related alterations in the brain. The proposed approach has been validated using three classic deep learning models on publicly available neuroimaging ageing datasets. It shows flexibility across different model architectures and different choices of hyperparameters. The corrected brain age delta from the proposed approach then has no linear relationship with chronological age and achieves higher predictive accuracy than a commonly used two-stage approach.

3.1 Introduction

The fundamental reason why brain age delta becomes prevalent in brain-related research is that it shows significant correlations with various measures of ageing, such as the decline in cognitive function, weaker grip strength, and walking speed [6], [10]. To act as a marker of brain health, the brain age delta should remain independent from any potential compounding factors in the brain-ageing analysis. In various brain-ageing research and clinical studies, chronological age is usually the most obvious factor of many age-related

brain diseases, such as cognitive decline and dementia [120]. Therefore, to adopt the brain age delta in the early detection and prevention of brain-associated disease, its association with chronological age needs to be addressed.

A nonzero age delta correlation (ADC) significantly weakens the validity of the brain age delta as a biomarker. It may raise problems for subsequent experiments. For example, when investigating if the brain age delta differs across groups with different degrees of cognitive impairment, the differences in the brain age delta between groups may simply be due to the group differences in chronological age distributions [67]. In that way, the apparent relationship between the brain age delta and variables of interest might be more driven by age and not the brain age delta [62].

In Section 2.2, we have also reviewed different bias correction approaches to mitigate the dependence of the brain age delta on age. Most bias correction approaches can be summarised as a two-stage approach as they involve first training a brain age prediction model and applying the bias correction afterwards on the model predictions. However, such an explicit correction approach is a post-hoc correction of the model predictions (using a biased model) which can lead to sub-optimal results.

In this chapter, we propose a novel approach to correct this bias. Compared with existing correction approaches, our approach only consists of a single training stage without the need to apply a bias correction stage. Model predictions at the end of the training process are unbiased. The overall workflow of the two-stage approach and our proposed approach is illustrated in Figure 3.1. Also, we would like to stress that our approach only modifies the loss function and can be combined with any existing differentiable models and any model architectures (e.g. ResNet [70], VGG-19 [121], and DenseNet [37]). The detailed implementations are available in the GitHub repository[1].

To sum up, we propose a novel algorithm that solves the regression bias at the model-building stage for any deep learning models. The proposed approach can be extended to any type of model that could be trained with a symmetric loss function (e.g. linear

[1] https://github.com/hanzhiwangchn/MRI_Age_Prediction

regression, Support Vector Regression).

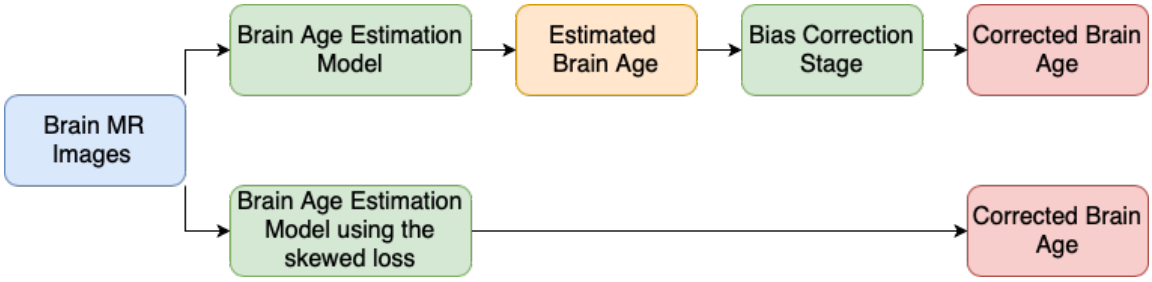


Figure 3.1: A detailed workflow of brain age prediction in addressing regression bias in brain age delta. The upper route represents the widely adopted two-stage approach and the lower one represents our proposed approach. Each "Corrected Brain Age" block represents the final estimation of brain age for each approach and the corrected brain age should have no linear relationship with chronological age. Two estimates of corrected brain age are then compared with chronological age respectively to see which one achieves a lower mean absolute error.

Our main contributions in this chapter are:

1. We propose a novel approach that corrects regression model bias directly in the model training phase. Compared with the commonly used two-stage approach, our approach does not require an explicit bias correction stage. The proposed approach acts as an alternative correction algorithm to the two-stage approach, whereas it achieves significantly better accuracy. To the best of our knowledge, this is the first such approach for deep learning models.

2. We develop a dynamic training strategy to search for the optimal hyperparameters for our skewed loss, which has been proven to be robust to different datasets, data distributions, model architectures, and problem domains.

This chapter is organised as follows. Section 3.2 provides a mathematical evaluation of the cause of the regression bias and introduces the skewed loss function to solve it. Section 3.3 proposes a robust training strategy for the skewed loss to make the model performances consistent. Section 3.4 explains the experiment settings, including datasets, models, and relevant training parameters. Section 3.5 demonstrates experiment results

and model comparisons between the two-stage approach and our approach. Section 3.6 generalises the skewed loss to other problem domains using the apparent age prediction problem. Section 3.7 discusses the advantages of the skewed loss and propose future works. Section 3.8 summarises the overall approach to conclude this chapter.

3.2 Methodology

3.2.1 A mathematical evaluation of the regression bias

From our point of view, this bias comes from the irreducible level of noise in the dataset. A mathematical evaluation is presented as follows:

Proposition 1. *Given a dataset* (x_i, y_i)*, where* $i = 1, 2, ..., N$*. Assuming that we would like to predict the value of* y *given* x *and the hidden relationship between* x *and* y *is*

$$y = h(x) + \epsilon \tag{3.1}$$

where $h(x)$ *is the hidden target function,* ϵ *is the noise term, and* ϵ *follows an unknown distribution* $\Pr(\epsilon|\mu, \sigma)$ *with* μ *represents mean and* σ *represents standard deviation.*

It can be proved that the covariance between ground truth y *and model error* e *is strictly non-zero.*

Proof. Since the model and noise are independent, we have

$$\mathrm{Var}(y) = \mathrm{Var}(h(x)) + \mathrm{Var}(\epsilon) \tag{3.2}$$

If we are trying to model $h(x)$ using an estimator $f(x)$, where $f(x)$ is a generic regression model with any degrees of accuracy, so that

$$y = f(x) + e \tag{3.3}$$

where e is the model residuals.

Then the covariance between target label y and model residuals e is

$$
\begin{aligned}
\mathrm{Cov}(y, e) &= \mathrm{Cov}(y, y - f(x)) \\
&= \mathrm{Var}(y) - \mathrm{Cov}(y, f(x)) \qquad (3.4) \\
&= \mathrm{Var}(h(x) + \epsilon) - \mathrm{Cov}(f(x) + e, f(x)) \\
&= \mathrm{Var}(h(x)) + \mathrm{Var}(\epsilon) - \mathrm{Var}(f(x)) - \mathrm{Cov}(e, f(x)) \\
&= \{\mathrm{Var}(h(x)) - \mathrm{Var}(f(x))\} + \sigma^2 - \mathrm{Cov}(e, f(x)) \qquad (3.5)
\end{aligned}
$$

From (3.4), it can be inferred that

$$
\mathrm{Cov}(y, e) = 0 \iff f(x) = y + \mathrm{const} \iff \mathrm{Var}(y) = \mathrm{Var}(f(x)) \qquad (3.6)
$$

From (3.6), the correlation between target value y and model residuals e stays zero if and only if the target values y and model predictions $f(x)$ have the same variance.

However, in the regression framework, $f(x)$ is designed to estimate the hidden relationship $h(x)$ rather than y, and y is also corrupted with the noise term ϵ. Therefore, the correlation between the target label and model residual is strictly always nonzero.

Take linear regression as an example and we are using Ordinary Least Square (OLS) to solve it, (3.5) can be further simplified as

$$
\begin{aligned}
\mathrm{Cov}(y, e) &= \{\mathrm{Var}(h(x)) - \mathrm{Var}(f(x))\} + \sigma^2 - \mathrm{Cov}(e, f(x)) \\
&= \{\mathrm{Var}(h(x)) - \mathrm{Var}(f(x))\} + \sigma^2 \\
&\approx \sigma^2 \qquad (3.7)
\end{aligned}
$$

Since for OLS, the model residuals and model predictions are orthogonal. However, such orthogonality does not hold for other models, such as neural networks.

□

Even though the correlation value always stays non-zero, it should also be noted that a relatively small ADC value does not prevent brain age delta as an effective biomarker. In the literature, a general rule of thumb is that the age delta correlation should fall between

-0.1 and + 0.1.

3.2.2 Skewed loss function

In this section, we introduce our approach for bias correction using the skewed loss functions. We start by observing that regression models are trained by minimizing a loss function. Commonly used loss functions include L1 or L2 loss, which can be represented as:

$$\begin{aligned} \mathcal{L}_{\text{MAE}}(y, \hat{y}) =& |y - \hat{y}| \\ \mathcal{L}_{\text{MSE}}(y, \hat{y}) =& (y - \hat{y})^2 \end{aligned} \tag{3.8}$$

where y and $\hat{y}$ represent the target and predicted value.

Regression loss functions are typically symmetric, i.e., overpredictions are penalized as much as underpredictions. Therefore, one possible approach to counteract bias is to skew the loss functions and penalize overpredictions more than underpredictions for low values of y and vice versa. We denote this approach as a *skewed loss function* to reflect this characteristic. To this end, let us define a step function $s : \mathbb{R} \to \mathbb{R}$ as

$$s(x) = \lambda_0 \, \mathbb{1}_{\mathbb{R}<0}(x) + \lambda_1 \, \mathbb{1}_{\mathbb{R}\geq 0}(x) \tag{3.9}$$

where $\mathbb{1}_U(x)$ is the indicator function (1 if $x \in U$ and 0 otherwise) and λ_0 and λ_1 are the heights of the steps. This implies that $s(x) = \lambda_0$ for $x < 0$ and $s(x) = \lambda_1$ for $x \geq 0$.

Multiplying the step function with the original loss function, we obtain the skewed loss functions

$$\widetilde{\mathcal{L}}(y, \hat{y}) = \mathcal{L}(y, \hat{y}) \; s(y - \hat{y}) \tag{3.10}$$

We can further simplify $s(x)$ to have only one hyperparameter λ that controls the amount of skew by setting $\lambda_0 := \exp(-\lambda)$, $\lambda_1 := \lambda_0^{-1} = \exp(\lambda)$. Then $s(x)$ simplifies to

$$s(x) = \exp(\text{sgn}(x)\lambda) \tag{3.11}$$

where sgn : $\mathbb{R} \to \{-1, 1\}$ is the sign function and λ controls the skew.

This simplification (3.11) reduces flexibility by imposing an "inverse symmetry" constraint on λ_0 and λ_1 in (3.9), whereas it also reduces model complexity with only one parameter controlling the behaviour of the skewed loss.

To make the skewed loss function (3.10) compatible with the brain age prediction, two more adjustments are needed:

1. Skewed loss should behave differently in different age ranges. For young participants, it should assign more penalties to overpredictions than underpredictions. For elderly participants, it should assign more penalties to underpredictions than overpredictions.

2. The bias is more significant for participants with ages closer to either end of the data range than those with ages closer to the mean age (of the training dataset) [39]. Participants with age closer to either end of the data range need larger levels of skew.

Therefore, we can further modify $s(x)$ by setting λ as a function of chronological age (y). The range of λ is then constrained within $[-\lambda_{max}, +\lambda_{max}]$ where λ_{max} is a positive hyperparameter:

$$s(x, y) = \exp(\text{sgn}(x)\lambda(y)) \tag{3.12}$$

Based on (3.12), a general guide to defining $\lambda(y)$ is that a smaller y should have a negative λ while a larger y should have a positive λ. We propose two approaches to define $\lambda(y)$ using linear functions for simplicity.

- Approach 1: $\lambda(y)$ can be defined as a linear function of y.

$$\lambda(y) = g(y) \times \lambda_{max} + (1 - g(y)) \times (-\lambda_{max}) \tag{3.13}$$

$$g(y) = (y - y_{min})/(y_{max} - y_{min}) \tag{3.14}$$

where y_{min} and y_{max} represent the minimum and maximum age of the dataset.

- Approach 2: $\lambda(y)$ can be defined as a piecewise linear function of y by setting the median value of y in the dataset as a midpoint.

$$\lambda(y) = \begin{cases} (1 - g(y)) \times (-\lambda_{max}) & \text{if } y \leq y_{med} \\ g(y) \times \lambda_{max} & \text{otherwise} \end{cases} \tag{3.15}$$

$$g(y) = \begin{cases} (y - y_{min})/(y_{med} - y_{min}) & \text{if } y \leq y_{med} \\ (y - y_{med})/(y_{max} - y_{med}) & \text{otherwise} \end{cases} \tag{3.16}$$

where y_{min}, y_{max}, and y_{med} represent the minimum, maximum, and median age of the dataset.

The difference between these two approaches of defining $\lambda(y)$ is that the first one assigns the middle value of y (in the training set) to have a zero λ value, whereas the second one uses the median value. For datasets with highly skewed age distributions, the second approach results in a roughly equal number of participants having positive λ values and negative λ values, which improves stability in practice.

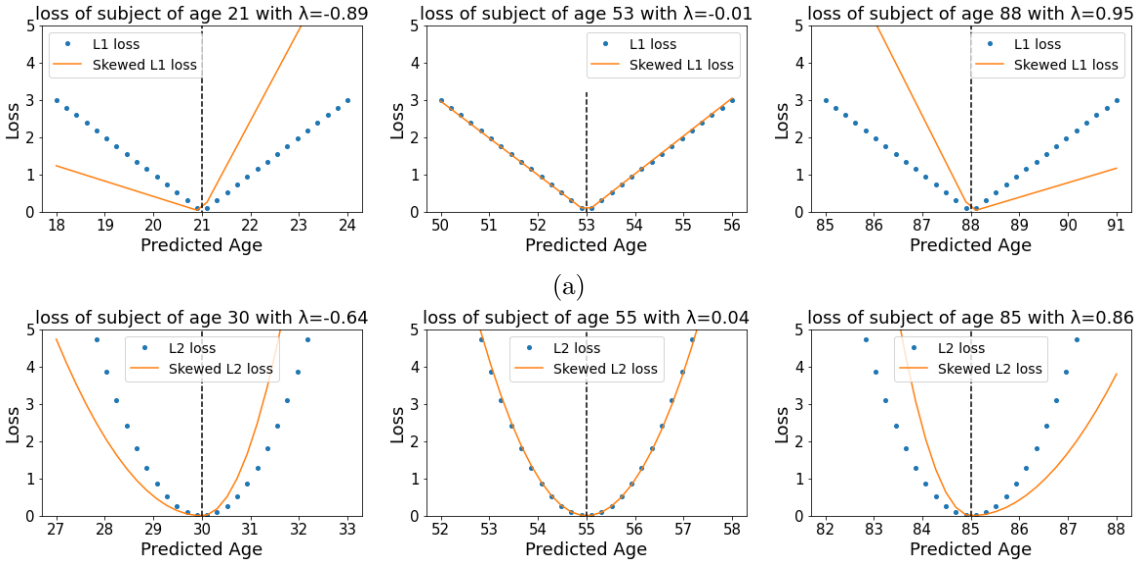
By combining (3.10), (3.12), (3.15), and (3.16), we formally propose the skewed loss function for brain age prediction

$$\widetilde{\mathcal{L}}(y, \hat{y}) = \mathcal{L}(y, \hat{y})\ s(y - \hat{y}, y) \tag{3.17}$$

$$s(y - \hat{y}, y) = \exp(\text{sgn}(y - \hat{y})\lambda(y)) \tag{3.18}$$

where $\lambda(y)$ is defined using (3.15) and (3.16).

As an example, L1 and L2 skewed loss functions are illustrated in Figure 3.2.



(a)

(b)

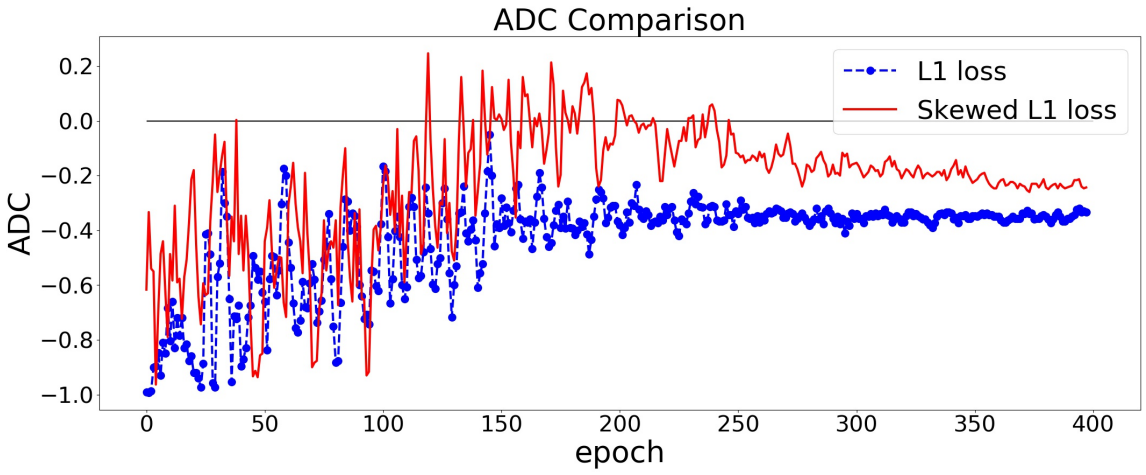
Figure 3.2: L1 and L2 skewed loss function illustrations. For each plot, the dotted and solid lines represent the normal symmetric loss function and the skewed loss function respectively. The horizontal axis means possible model predictions and the vertical axis means the corresponding loss term for each prediction. For each plot, from left to right, the skewed loss behaves differently in different age ranges and different ages are assigned with different levels of skew using (3.15) and (3.16). λ_{max} is set to 1 in both plots.

3.2.3 Effect of the skewed loss

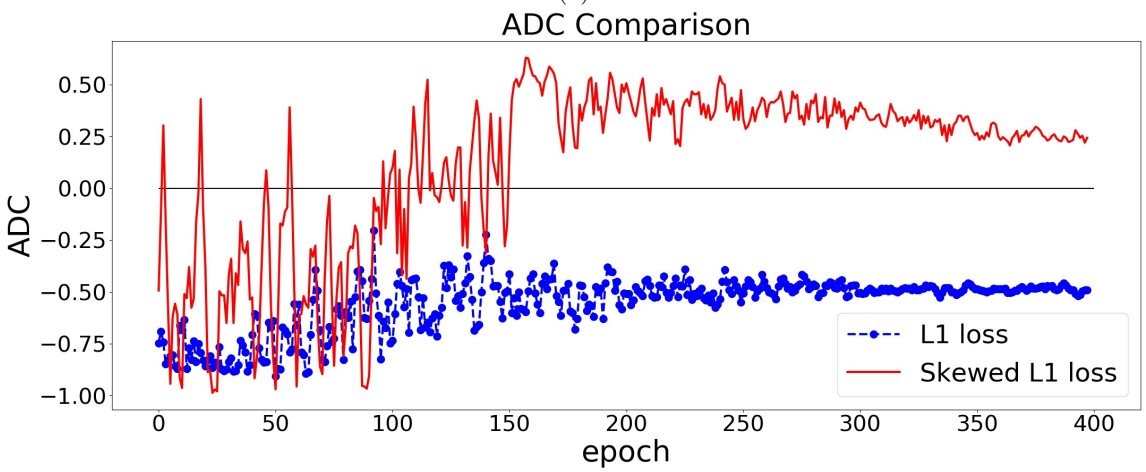
The idea of employing a skewed loss function is to assign different loss values depending on whether the model makes overestimations or underestimations. Because the regression bias manifests as an overestimation for younger individuals, we assign more penalties when the model overestimates ages for young participants to push the model to make fewer overestimations. This idea is reversed on elderly individuals and we then penalize the model more when it makes underestimations.

The effect of using the skewed loss in practice is illustrated in Figure 3.3. We can observe that models using ordinary L1 loss have negative ADC values at the end of the training, which is in line with previous studies, whereas models using skewed L1 loss end up with larger ADC values. Figure 3.3 shows that by applying the skewed loss functions, the model tends to make fewer overestimations for young individuals and fewer underestimations

for elderly individuals so that the effect of a negative ADC is reduced.



(a)



(b)

Figure 3.3: ADC Comparisons between L1 and skewed L1 loss. The model is trained separately using L1 and skewed L1 loss respectively. For each plot, the dashed and solid lines represent the changes of ADC (measured in Pearson's r) on the validation set in the training process using L1 and skewed L1 loss respectively. λ_{max} is set as 1 in (a) and 2 in (b).

3.3 Dynamic lambda training strategy

3.3.1 The necessity of dynamic lambda strategy

From Section 3.2.2, λ_{max} is a hyperparameter in the skewed loss which controls the skew of the loss function. From Figure 3.3, we can observe that setting λ_{max} as 1 results in a small negative ADC while setting λ_{max} as 2 ends up with a small positive ADC. That

indicates there should be an optimal value between 1 and 2 for λ_{max} so that the ADC can approach zero at the end of the training process.

However, due to the randomness of the network training process, different datasets used for training, and different model architectures being used, it is not realistic to foresee the optimal λ_{max} that can push the ADC to zero before the training. To solve this problem, we propose a dynamic lambda training strategy to tune λ_{max} in the training process to find the optimal value that is suitable for the current model. By dynamically changing λ_{max} in the training process, we can enable the model predictions to have (near) zero ADC at the end of training.

3.3.2 Implementations of dynamic lambda strategy

The dynamic lambda strategy is formally presented in Algorithm 1. Some notations in Algorithm 1 are defined in Table 3.1.

It should be noted that in a machine learning framework, independent test set should not be adopted in the training phase. Therefore, validation set, split from the whole dataset, is adopted to guide the model training process. Skewed loss function and other bias correction approaches rely on an underlying assumption:

Assumption 3. *A zero correlation on the validation dataset may indicate a near-zero correlation on the test set*

However, it should also be stressed that such assumption may not hold if there is a distribution shift between the validation and test set.

It should be noted that hyperparameters in Table 3.1 are insensitive to different values based on our experiment testing and the "Range" column in Table 3.1 aims to give a general guide on how to set them. The suggested value for hyperparameters α and β should be rounded to integer if needed.

In Table 3.1, hyperparameters α and β separate the whole training process into three stages, all of which should have a similar number of epochs to ensure the model is prop-

Table 3.1: Notations for Dynamic Lambda Strategy

Notation	Type	Description	Range
$\mathcal{D}_{validation}$	Dataset	validation set	
γ	hyperparameter	total number of epochs	
α	hyperparameter	stage-2 start epoch	$\gamma/3$
β	hyperparameter	stage-3 start epoch	$2\gamma/3$
ϵ	hyperparameter	update interval of λ_{max}	[5, 10]
ADC	variable	age-delta correlation	
M	variable	an array to save λ_{max}	
N	variable	an array to save ADC	
θ	hyperparameter	correlation threshold	[0.1, 0.2]
η	hyperparameter	a positive multiplier	(1, 2)
$\mathcal{L}_s(\lambda_{max})$	function	skewed loss controlled by λ_{max}	

erly trained in all stages. Hyperparameter ϵ determines how frequently we update λ_{max}. A too-small value of ϵ may prevent the model from training properly toward the newly updated loss function, whereas a too-large value may prevent the model from finding the adequate value of λ_{max} due to fewer updates. Hyperparameter θ acts as a positive threshold determining whether to adjust λ_{max} according to the current ADC value. We will change λ_{max} in the second stage only when the ADC falls outside of $[-\theta, \theta]$. Hyperparameter η determines how to change λ_{max} in each update in the second stage.

3.3.3 Explanations for dynamic lambda strategy

Algorithm 1 divides the whole training process into three stages:

1. Stage-1: The aim is to pretrain the model with a fixed λ_{max} to achieve relatively high accuracy before any modifications are made. This is done because due to the random initialization of the weights, ADC estimates in the first few epochs may be uninformative.

2. Stage-2: From this stage, we are trying to find the optimal λ_{max} that results in a zero ADC on the validation set. It is generally assumed that the validation and test set should be derived from similar distributions, so we assume a zero ADC on the validation set should be generalizable to the test set. In this stage, we apply a heuristic method to try different λ_{max} values to make the ADC approach toward

Algorithm 1: Dynamic lambda training strategy

Init: $\lambda_{max} \leftarrow 1$; skewed loss $\mathcal{L}_s(\lambda_{max})$
for $i \leftarrow 1$ **to** γ **do**
  if $i < \alpha$ **then** // stage-1
    Train the network using $\mathcal{L}_s(\lambda_{max})$
  else if $i < \beta$ **then** // stage-2
    Train the network using $\mathcal{L}_s(\lambda_{max})$
    if $(i - \alpha) \bmod \epsilon = 0$ **then**
      Calculate the current ADC on the validation set $\mathcal{D}_{validation}$
      Save λ_{max} to M and ADC to N
      if $ADC > \theta$ **then**
        $\lambda_{max} \leftarrow \lambda_{max}/\eta$
      else if $ADC < -\theta$ **then**
        $\lambda_{max} \leftarrow \lambda_{max} \times \eta$
      end
      Reinitialize $\mathcal{L}_s(\lambda_{max})$ with updated λ_{max}
  else // stage-3
    Train the network using $\mathcal{L}_s(\lambda_{max})$
    if $(i - \beta) \bmod \epsilon = 0$ **then**
      Calculate the current ADC on the validation set $\mathcal{D}_{validation}$
      Save λ_{max} to M and ADC to N
      Fit a linear regression F between M and N
      Select λ_{max} resulting in a zero ADC using F as the optimal λ_{max}
      Reinitialize $\mathcal{L}_s(\lambda_{max})$ with optimal λ_{max}
  end
end

zero. We also save λ_{max} and the corresponding ADC for Stage 3.

3. Stage-3: Using the previously saved λ_{max} and ADC pairs from stage-2, we can fit a linear regression model to find the optimal λ_{max} that results in a zero ADC and set it as the updated λ_{max}. This method is repeated multiple times until the end of training so that λ_{max} is tuned iteratively. We use linear regression for its simplicity and it has been proven to be effective in finding the optimal λ_{max} in practice.

There is a slight increase in training time when using the skewed loss compared with using a symmetric loss function because of the training of linear models in Stage-3. However, it is usually negligible compared with training a network.

Throughout the training process, our approach tries to control ADC explicitly by iteratively tuning λ_{max}. Therefore, by measuring ADC with different types of correlation,

our approach can find the optimal λ_{max} resulting in the specified correlation approaching zero, which is not feasible for the two-stage approach.

The effect of using the dynamic lambda strategy is illustrated in Figure 3.4. In Figure 3.4, we train our model twice using symmetric L1 loss and skewed L1 loss with dynamic lambda strategy. It can be easily observed that at the end of the training, the model using skewed loss ends up with a (near) zero ADC on the validation set.

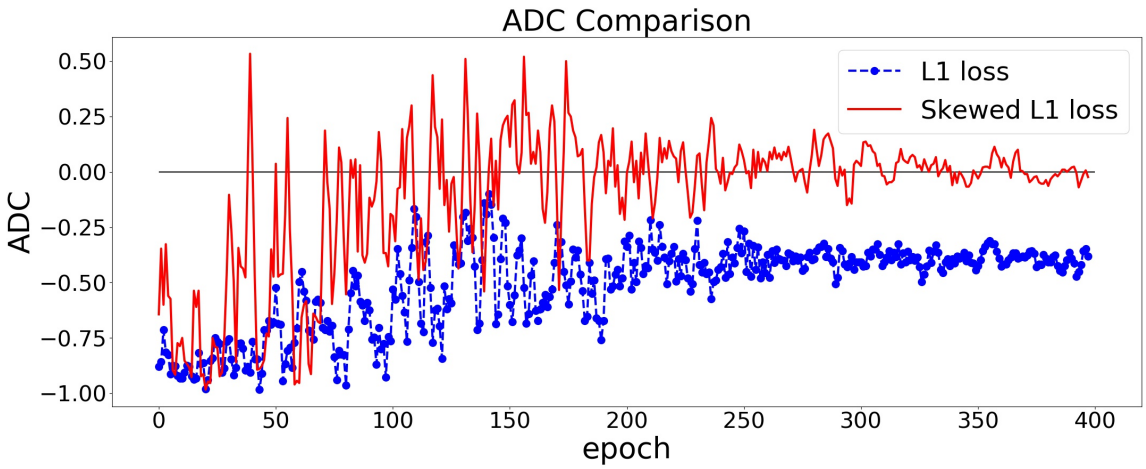


Figure 3.4: ADC Comparisons between L1 and skewed L1 loss. Dynamic lambda strategy is applied for the skewed loss. The model is trained separately using L1 loss and skewed L1 loss. The dashed and solid lines represent the changes of ADC (measured in Pearson's r) on the validation set in the training process using L1 and skewed L1 loss respectively.

3.4 Experiments

To empirically investigate our approach, we conducted experiments using two publicly available ageing datasets and several classical neural network architectures. The models were trained using the normal symmetric loss and the skewed loss respectively. The two-stage approach was applied to models using the normal symmetric loss as the contrast.

3.4.1 Datasets

To validate the robustness of the skewed loss, we selected two public neuroimaging datasets, Cam-CAN [122], [123] and ABIDE [124], with preprocessed 3D structural T1-

weighted MR brain images. The dataset descriptions and the specific preprocessing pipeline for each dataset can be found at the Cam-CAN[2] and ABIDE[3] website. The data from the Cam-CAN are preprocessed using the Automatic Analysis pipeline [125] and we selected the gray-matter density maps from it. The data from the ABIDE derives from cortical thickness measures using ANTs pipeline [109] and we selected the 3D volume containing voxel-wise measures of cortical thickness from it. It should be stressed that neither the two-stage approach nor our approach relies on a specific dataset or a specific type of data. We focus on these two specific datasets because these images are easier to access and do not require any further preprocessing, which provides a fairer comparison set-up between methods. The 3D image resolutions are $(96 \times 112 \times 96)$ for the Cam-CAN dataset and $(141 \times 120 \times 178)$ for the ABIDE dataset. To further compare our approach and the two-stage approach, we randomly removed some participants from both datasets to create two more modified datasets. We randomly removed some participants whose ages were smaller than 60 for the Cam-CAN dataset to create the skewed Can-CAN dataset and removed all participants whose ages were larger than 20 for the ABIDE dataset to create a symmetric ABIDE dataset. For CamCAN dataset, we would like to evaluate the model performance on a skewed age distribution dataset so that we chose age 60 as the split point. For ABIDE dataset, we would like to evaluate the model performance on a more symmetric age distribution dataset so that we chose age 20 as the split point to remove the long tails for ABIDE dataset.

Our approach and the two-stage approach will then be compared on all four datasets. The description of each dataset is provided below and Figure 3.5 shows their chronological age distributions:

- Cam-CAN dataset:

  The Cam-CAN dataset contains 653 cognitively normal participants (mean age 54.3 years, standard deviation 18.5 years, range 18-88 years).

- Skewed Cam-CAN dataset:

[2] https://www.cam-can.org
[3] http://preprocessed-connectomes-project.org/abide/download.html

The Cam-CAN dataset has a roughly balanced age distribution with slightly more elderly individuals. We removed 70% of the participants with ages smaller than 40 years and 50% of the participants with ages ranging from 40 years to 60 years to create a skewed age distribution.

The skewed Cam-CAN dataset contains 423 cognitively normal participants (mean age 62.2 years, standard deviation 16.9 years, range 18-88 years).

- ABIDE dataset:

  The ABIDE dataset contains 571 cognitively normal participants (mean age 17.1 years, standard deviation 7.7 years, range 6-56 years).

- Symmetric ABIDE dataset:

  The ABIDE dataset has a highly skewed age distribution toward young ages. We removed the participants whose ages are larger than 20 years to make this dataset have a more symmetric age distribution.

  The symmetric ABIDE dataset contains 415 cognitively normal individuals (mean age 13.2 years, standard deviation 3.1 years, range 6-20 years).

3.4.2 Models

To validate that the skewed loss is robust to different network architectures, we implemented three networks roughly based on ResNet [70], VGG [121], and GoogleNet [41] architectures. We replaced the 2D convolution layers with 3D convolution layers [126] and the specific implementations can be found in the GitHub repository mentioned in Section 3.1. The ResNet contains six consecutive blocks in total. The first five blocks adopt the same architecture design, which contains a convolution layer, a batch norm layer, ELU activation, another convolution layer, another batch norm layer, and another ELU activation in order, followed by a max pooling layer. The initial number of filter is set to 8 and is doubled after each max pooling layer, except for the last one. The 6th block consists of a flatten layer, a dropout layer, a fully connected layer, ELU activation, and another fully connected layer in order to derive the final prediction. The VGG also

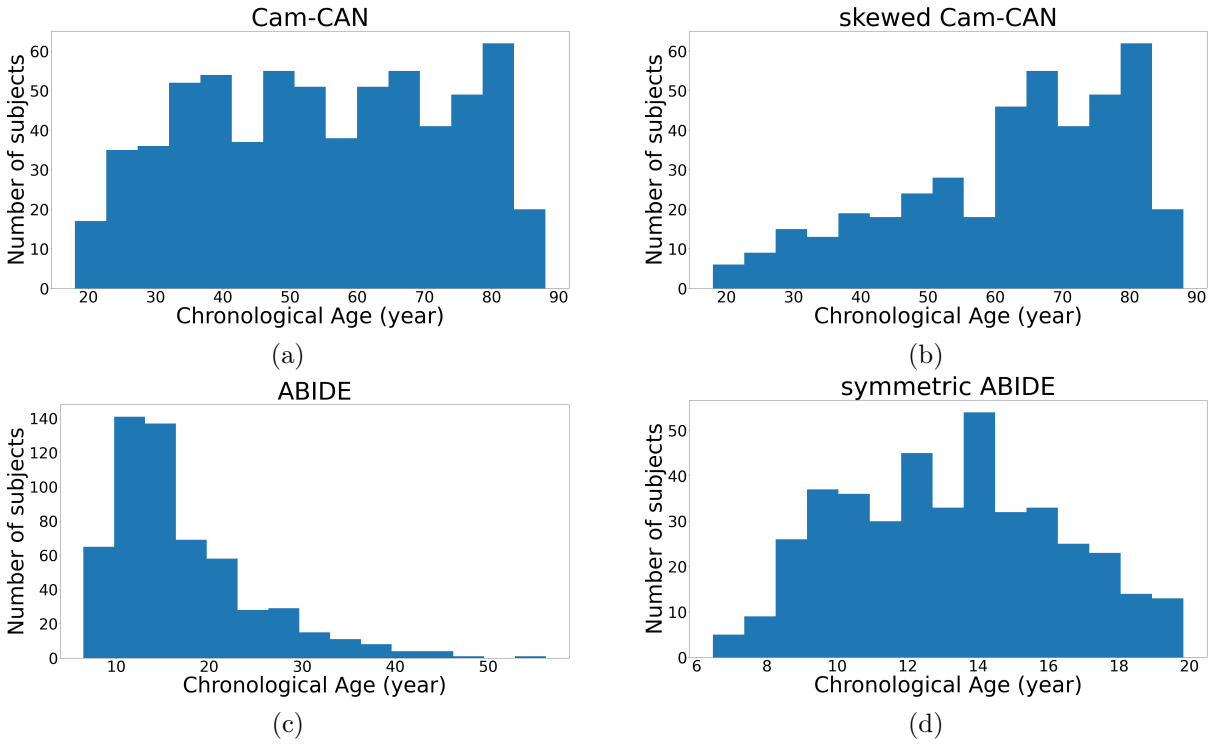


Figure 3.5: Age distribution for the Cam-CAN and ABIDE datasets and their skewed versions

contains six consecutive blocks in total. The first five blocks adopt the same architecture design, which contains a convolution layer, a batch norm layer, ReLU activation, another convolution layer, another batch norm layer, and another ReLU activation in order, followed by a max pooling layer. The initial number of filter is set to 8 and is doubled after each max pooling layer, except for the last two max pooling layers. The 6th block consists of a flatten layer, a dropout layer, a fully connected layer, ReLU activation, and another fully connected layer in order to derive the final prediction. The GoogleNet adopts a slightly different design. It contains seven consecutive blocks in total. The first block contains a convolution layer, a batch norm layer, ReLU activation, a max pooling layer, another convolution layer, another batch norm layer, another ELU activation and another max pooling layer in order. Block 2, 3, 4, 5, and 6 are inception blocks [41] with number of filters set to 16, 32, 64, 126, and 192 respectively. In each inception block, we follow the original architecture design using four different branches with 1×1 convolution, 3×3 convolution, 5×5 convolution, and 3×3 max pooling respectively, followed

by a concatenation of the four different branches. The last block contains a average pooling layer, a flatten layer, a dropout layer, a fully connected layer, ReLU activation, and another fully connected layer in order to derive the final prediction. The number of trainable parameters for each model architecture is demonstrated in Table 3.2.

It should be also noted that different choices of the model-dependent parameters do not affect the final ADC value as neither correction approach focuses on the model architectures. We denote our models as ResNet, VGG, and GoogleNet respectively in the rest of this paper. Compared with the original network architectures, our models have fewer layers due to the size of the dataset and the number of parameters is reduced below 1 million. All models only take the raw images as input and demographic variables such as gender are not included to follow the conventions of brain age prediction problem [33], [34].

Table 3.2: Number of trainable parameters

Model architecture	**Number of parameters**
ResNet	960,769
VGG	474,857
GoogleNet	636,805

Also, we would like to stress that the brain age prediction model is a regression model and hence the bias always exists regardless of different model architectures. The skewed loss and the two-stage approach can both be applied to any model architectures and we only select ResNet, VGG and GoogleNet architecture because they are still widely used in practice [1], [31]–[33] and form the building blocks for more complex models.

3.4.3 Training and testing

The skewed loss approach is generic and can be applied to any symmetric loss function. However, in the brain age prediction literature, MAE (i.e. average L1 loss) is the most commonly used metric. Therefore, we focus our analyses on L1 loss.

Because the size of the datasets we used is relatively small, to prevent models from overfitting, we applied data augmentation in the training process using TorchIO [127]. In

each training iteration, every input image had a probability of 50% being flipped around the horizontal plane. Also, the L2 weight decay coefficient was set to 0.001.

During the training process, the Adam optimizer [128] was used as the default optimizer for all models. The initial learning rate was set to 0.01 and then multiplied by 0.5 every 50 epochs. The batch size was set to 16 as default and the total number of epochs was set to 400.

When we trained the model using the skewed loss, the hyperparameters used in Table 3.1 are set the same for all models. We set hyperparameter α to 150, β to 300, ϵ to 5, θ to 0.15, and η to 1.5. It should be noted that the hyperparameter settings mentioned above, such as L2 weight decay coefficient and skewed loss relevant parameters, are based on our experiment testing.

Datasets were split using a stratified split strategy of which 80% was used for training, 10% for validation, and 10% for testing. To make full use of the whole dataset, we applied a split strategy similar to cross-validation. We randomly split each dataset in 20 different ways so that no two train/validation/test sets are identical. Also, to minimize the fluctuations of the model results due to random initialization of the weights, we trained our model on each train/validation/test split 5 times. In that way, we ended up training a specific model on a particular dataset 100 times.

3.5 Results

The performance comparisons between the two-stage approach and the skewed loss across four datasets described in Section 3.4.1 are listed in the following subsections. We also add the performances using the normal loss function before the correction stage as comparisons. The ADC is measured using Pearson's r unless specified otherwise. In Section 3.5.3, we provide the performance using Spearman's rank correlation coefficient to measure ADC to illustrate the flexibility of our approach.

It should also be noted that, unlike normal loss functions, correction approaches try

to sacrifice model accuracy for more unbiased model predictions. Thus, all correction approaches will result in an increase in MAE compared with normal loss. In our experiments, model performance using normal loss could be regarded as the lower bound of model errors.

To save space, Table 3.3 lists some notations used in the following subsections.

Table 3.3: Notations for Performance Comparisons

Notations	**Descriptions**
Normal loss	models using normal loss without two-stage approach
Two-stage	models using normal loss with two-stage approach
Skewed loss	models using the skewed loss with dynamic lambda strategy
Wilcoxon	Wilcoxon signed rank test
Paired-t	paired t-test

3.5.1 Model performance

From Section 3.4.3, we split each dataset using 20 different ways and for each split, we train the network 5 times. At the end of the training, we calculate the average value of the MAE and ADC from the 5 repeated runs to reduce fluctuations in model performances. In that way, we end up with 20 pairs (one split for one pair) of averaged MAE and average ADC between the two-stage approach and skewed loss. The significance tests are performed by comparing the averaged MAE and ADC values across the 20 splits. The model performance across four datasets is shown in Table 3.4 and the significance test results between the two-stage approach and the skewed loss are shown in Table 3.5.

From Table 3.4, the skewed loss achieves a lower MAE compared with the two-stage approach for most models. The averaged ADC of the skewed loss almost always falls within -0.1 to +0.1, which indicates the bias in the brain-age delta has been significantly reduced compared with normal loss. On average, the skewed loss increases the MAE by 0.2 years for the Cam-CAN datasets and 0.97 years for the ABIDE datasets, whereas the two-stage approach increases the MAE by 0.28 years and 1.12 years.

From Table 3.5, some models using the skewed loss achieve significantly lower MAE com-

Table 3.4: Model Performances on Four Datasets (mean ± standard deviation)

Dataset	**Model**	**Normal loss**		**Two-stage**		**Skewed loss**	
		MAE	ADC	MAE	ADC	MAE	ADC
Cam-CAN	ResNet	4.80±0.42	-0.28±0.12	5.06±0.49	-0.03±0.14	**4.96±0.50**	-0.03±0.13
	VGG	5.09±0.44	-0.23±0.11	5.35±0.56	-0.02±0.14	**5.22±0.59**	-0.05±0.12
	GoogleNet	5.10±0.54	-0.25±0.13	5.44±0.68	-0.03±0.14	**5.22±0.60**	-0.05±0.15
Skewed Cam-CAN	ResNet	5.02±0.70	-0.35±0.19	5.31±0.71	-0.04±0.20	**5.21±0.76**	-0.07±0.20
	VGG	5.23±0.66	-0.34±0.17	**5.42±0.71**	-0.06±0.19	5.43±0.75	-0.06±0.17
	GoogleNet	5.28±0.71	-0.38±0.15	**5.62±0.79**	-0.05±0.19	5.67±0.76	-0.06±0.18
ABIDE	ResNet	3.40±0.43	-0.62±0.15	4.71±0.97	-0.04±0.28	**4.54±0.66**	-0.08±0.26
	VGG	3.33±0.32	-0.6±0.15	4.35±0.89	-0.06±0.24	**4.19±0.69**	-0.09±0.25
	GoogleNet	3.37±0.37	-0.60±0.15	4.50±0.75	-0.08±0.25	**4.36±0.65**	-0.12±0.27
Symmetric ABIDE	ResNet	2.08±0.27	-0.73±0.08	3.55±0.88	-0.05±0.20	**3.28±0.81**	-0.09±0.23
	VGG	1.92±0.26	-0.64±0.11	2.81±0.60	-0.07±0.21	**2.81±0.57**	-0.03±0.24
	GoogleNet	2.03±0.53	-0.64±0.11	2.93±0.73	-0.05±0.24	**2.76±0.61**	-0.07±0.23

Table 3.5: Significance Test between Two-stage and Skewed Loss with p-values

Dataset	**Model**	**Wilcoxon**		**Paired-t**	
		MAE	ADC	MAE	ADC
Cam-CAN	ResNet	0.09	0.77	0.06	0.76
	VGG	**0.06**	0.16	**0.06**	0.21
	GoogleNet	**<0.01**	0.19	**<0.01**	0.12
Skewed Cam-CAN	ResNet	0.10	0.14	0.28	0.14
	VGG	0.50	0.91	0.85	0.98
	GoogleNet	0.85	0.91	0.72	0.66
ABIDE	ResNet	0.70	<0.01	0.19	<0.01
	VGG	0.09	0.16	0.12	0.05
	GoogleNet	0.18	0.19	0.13	0.17
Symmetric ABIDE	ResNet	**<0.01**	0.07	**<0.01**	0.05
	VGG	0.82	0.10	0.99	0.09
	GoogleNet	**0.01**	0.45	**<0.01**	0.41

pared with the two-stage approach, for example, GoogleNet on Cam-CAN and symmetric ABIDE datasets; ResNet on symmetric ABIDE dataset and VGG on Cam-CAN dataset. For the rest models, the skewed loss is likely to achieve lower MAE compared with the two-stage approach whereas the improvement is not significant enough.

To sum up, both the two-stage approach and the skewed loss correct the model to have a near-zero ADC at the cost of an increase in MAE, whereas using the skewed loss achieves comparable or significantly lower MAE compared with the two-stage approach.

3.5.2 Model performance of ensemble models

In Section 3.5.1, we calculate the MAE of each model and take the average of MAE from the 5 repeated runs. Alternatively, we can create an ensemble model by averaging the model predictions of 5 repeated runs and then calculate its MAE and ADC. Because we split the dataset in 20 different ways, we also end up with 20 ensemble models and thus 20 pairs of MAE and ADC for each model architecture. The ensemble model performance across four datasets is shown in Table 3.6 and the significance test results between the two-stage approach and the skewed loss are shown in Table 3.7.

Table 3.6: Ensemble Model Performances on Four Datasets (mean ± standard deviation)

Dataset	**Model**	**Normal loss**		**Two-stage**		**Skewed loss**	
		MAE	ADC	MAE	ADC	MAE	ADC
Cam-CAN	ResNet	4.00±0.24	-0.34±0.11	4.12±0.29	-0.03±0.12	**4.07±0.33**	-0.04±0.12
	VGG	4.12±0.31	-0.29±0.10	4.26±0.39	-0.03±0.11	**4.21±0.36**	-0.06±0.11
	GoogleNet	4.06±0.32	-0.32±0.13	4.20±0.48	-0.03±0.14	**4.09±0.35**	-0.05±0.14
Skewed Cam-CAN	ResNet	4.33±0.61	-0.4±0.15	4.49±0.57	-0.04±0.21	**4.38±0.66**	-0.07±0.20
	VGG	4.45±0.57	-0.39±0.15	4.50±0.56	-0.06±0.19	**4.45±0.58**	-0.07±0.17
	GoogleNet	4.30±0.49	-0.46±0.12	**4.38±0.48**	-0.06±0.19	4.55±0.48	-0.07±0.17
ABIDE	ResNet	3.02±0.35	-0.67±0.11	4.04±0.79	-0.04±0.30	**3.66±0.43**	-0.08±0.27
	VGG	3.00±0.25	-0.64±0.13	3.79±0.76	-0.06±0.25	**3.57±0.51**	-0.09±0.25
	GoogleNet	3.00±0.34	-0.66±0.12	3.89±0.60	-0.08±0.25	**3.65±0.47**	-0.12±0.27
Symmetric ABIDE	ResNet	1.85±0.19	-0.82±0.04	2.67±0.57	-0.05±0.20	**2.52±0.42**	-0.09±0.24
	VGG	1.67±0.19	-0.73±0.06	**2.20±0.36**	-0.07±0.21	2.30±0.39	-0.03±0.27
	GoogleNet	1.79±0.50	-0.72±0.07	2.39±0.57	-0.05±0.26	**2.20±0.47**	-0.08±0.24

From Table 3.6, the skewed loss also achieves a lower MAE compared with the two-stage approach for most models. The ADC of the ensemble model using the skewed loss almost always falls within -0.1 to +0.1, which also indicates the bias has been significantly reduced compared with normal loss. On average, the skewed loss increases the MAE by 0.08 years for the Cam-CAN datasets and 0.6 years for the ABIDE datasets, whereas the

Table 3.7: Ensemble Significance Test between Two-stage and Skewed Loss with p-values

Dataset	**Model**	**Wilcoxon**		**Paired-t**	
		MAE	ADC	MAE	ADC
Cam-CAN	ResNet	0.19	0.68	0.31	0.73
	VGG	0.46	0.18	0.44	0.20
	GoogleNet	0.14	0.16	0.10	0.13
Skewed Cam-CAN	ResNet	0.28	0.13	0.31	0.13
	VGG	0.41	0.91	0.49	0.78
	GoogleNet	0.14	0.97	0.09	0.65
ABIDE	ResNet	**<0.01**	0.02	**<0.01**	0.02
	VGG	**0.04**	0.26	**0.05**	0.09
	GoogleNet	**<0.01**	0.30	**0.01**	0.28
Symmetric ABIDE	ResNet	0.15	0.17	0.07	0.24
	VGG	0.26	0.18	0.24	0.17
	GoogleNet	**<0.01**	0.39	**<0.01**	0.40

two-stage approach increases the MAE by 0.12 years and 0.78 years.

From Table 3.7, ResNet on ABIDE dataset, VGG on ABIDE dataset, and GoogleNet on ABIDE and symmetric ABIDE datasets achieve significantly lower MAE using the skewed loss compared with the two-stage approach. For the rest models, the skewed loss tends to achieve lower MAE compared with the two-stage approach whereas the improvement is not significant.

To sum up, for ensemble models, the skewed loss can achieve comparable or significantly better performances compared with the two-stage approach. Using the skewed loss in ensemble models can also remove the bias in the brain age delta.

3.5.3 Performances using Spearman's rank correlation as ADC

A further experiment is carried out to evaluate how our approach performs by changing the ADC from Pearson's r to other correlation metrics. Here, we demonstrated the performances using ResNet on the Cam-CAN dataset with ADC measured in Spearman's rank correlation in Table 3.8 and Figure 3.6.

In Figure 3.6, we again train our model using L1 loss and skewed L1 loss with dynamic

Table 3.8: ResNet Performance on Cam-CAN Using Spearman's Rank Correlation as ADC (mean ± standard deviation)

Model	Metric	Normal loss	Two-stage	Skewed loss
ResNet	MAE	4.79±0.55	5.08±0.60	**4.95±0.56**
	ADC	-0.33±0.18	-0.02±0.20	-0.05±0.20

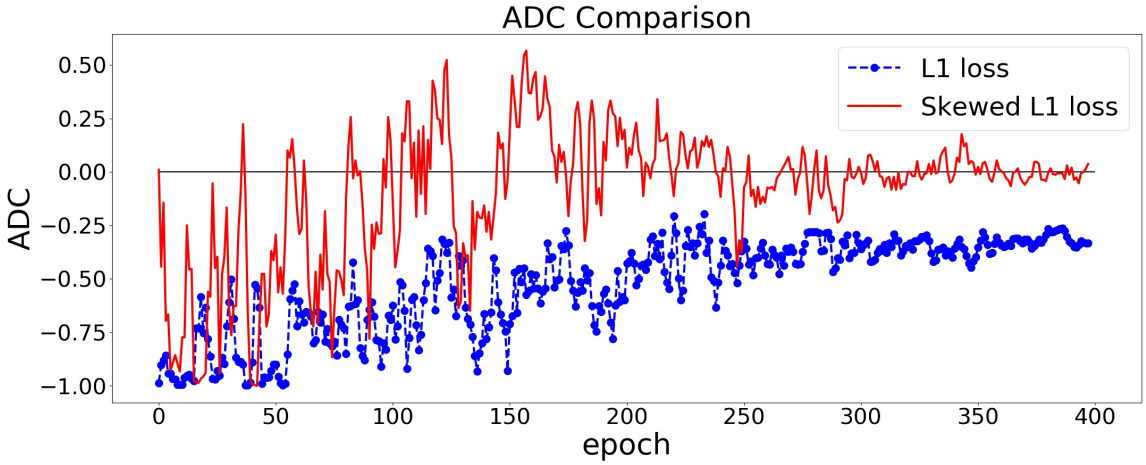


Figure 3.6: ADC Comparisons between L1 and skewed L1 loss. Dynamic lambda strategy is applied for the skewed loss. The model is trained twice using L1 loss and skewed L1 loss respectively. The dashed and solid lines represent the changes of ADC (measured in Spearman's rank correlation) on the validation set in the training process using L1 and skewed L1 loss respectively.

lambda strategy. We can observe that switching to different types of correlation metrics results in the specified metric approaching zero. It shows the potential to be generalized to other types of correlation metrics.

3.5.4 Consistency of correlation trends

Figure 3.3, Figure 3.4, and Figure 3.6 show the changes of correlation of a single model. To further validate the stability of our approach, we can examine how the variance of ADC changes in the training process. A large variance of ADC indicates large oscillations in ADC and vice versa.

In Section 3.4.3, we train each model 5 times to reduce the fluctuations. Thus, we group these 5 runs to calculate the mean and standard deviation of ADC. Figure 3.7 shows the

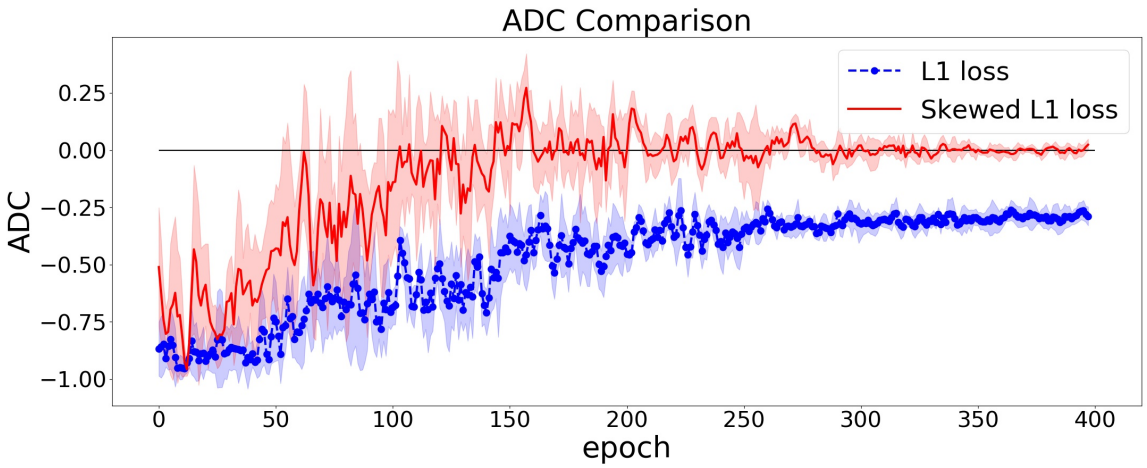


Figure 3.7: Averaged ADC Comparisons between L1 and skewed L1 loss. The dashed and solid lines represent the changes of ADC on the validation set in the training process using L1 and skewed L1 loss respectively. The shaded area represents the mean ADC plus or minus one standard deviation.

trend of averaged ADC of the 5 repeated runs.

From Figure 3.7, we can observe that the variance of ADC decreases in the training process and arrives at a small value at the end of training. This indicates the ADC can almost always arrive at (near) zero using the skewed loss on the validation set, which further validates the stability of the skewed loss.

3.5.5 Robustness to data distribution shift

Machine learning models typically require that the training and test set should share some similar properties like data distributions and the models are expected to learn these similarities during the training process. However, due to the variability of the unseen test data, these similarities may not always hold, which inevitably causes performance degradation in most cases.

Although learning with distribution shift has been extensively studied in recent years, it remains a challenging and ongoing topic in machine learning. In Section 3.5.1, we have discussed the model performance of the skewed loss without distribution shift. Therefore, it is crucial to investigate whether the skewed loss is robust to distribution shift or not.

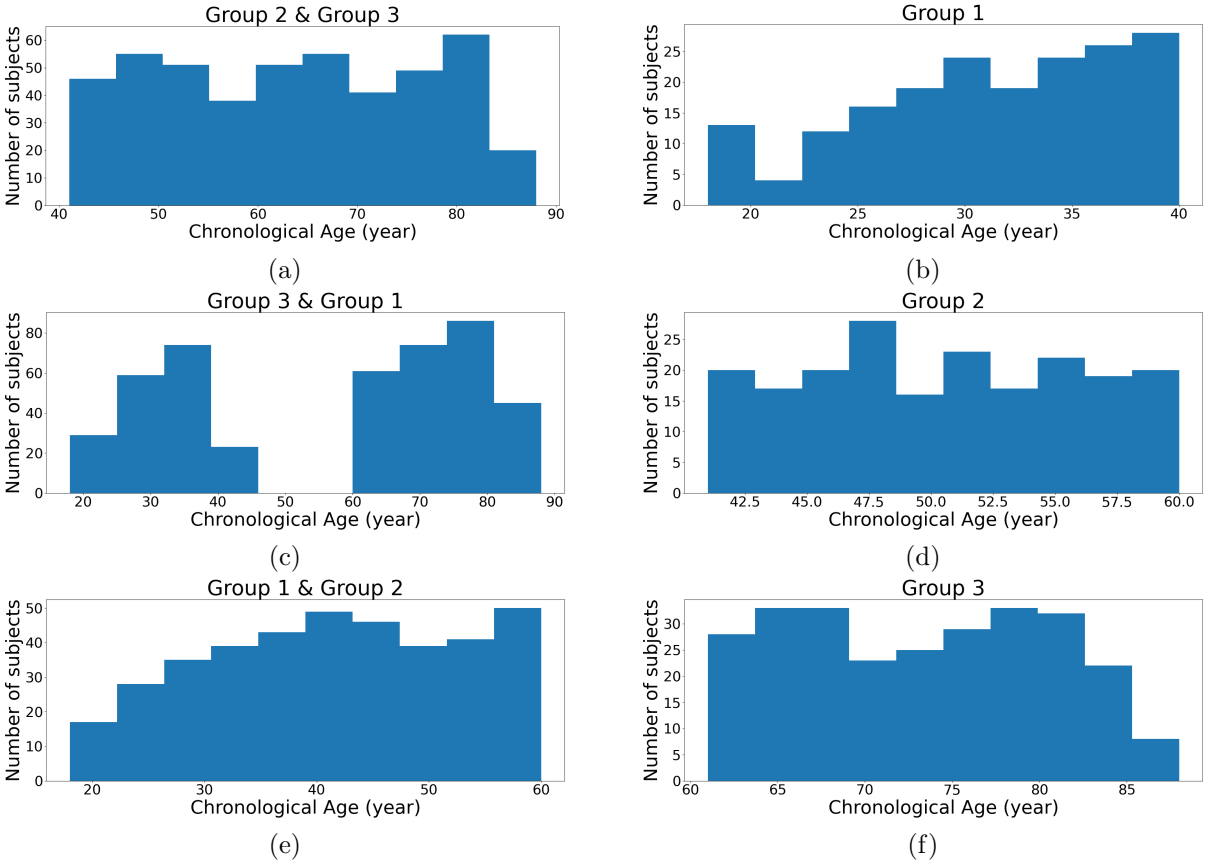


Figure 3.8: Age distributions of three types of train-test split. Each row represents a different way of creating the training and test set. The left three plots represent the three training sets and the right ones represent the corresponding test sets.

3.5.5.1 Experiment settings

To make the test set significantly different from the training set, we manually split our dataset based on the participants' chronological age. For illustration purposes, we only selected the Cam-CAN dataset and ResNet architecture to discuss the distribution shift problem. The Cam-CAN dataset was split into three groups. Group 1 consists of participants with ages below 40 years. Group 2 consists of participants with ages above 40 years and below 60 years. Group 3 consists of participants with ages above 60 years.

The model was then trained three times using each correction approach. For each time, one group was selected as the test set and the remaining two groups were combined as the training set. The chronological age distribution of the training and test sets of the three train-test splits are illustrated in Figure 3.8.

3.5.5.2 Results

The model performance comparisons on three different train-test splits are shown in Table 3.9. The correlation is measured in Pearson's r.

Table 3.9: Model Performance Comparisons When Facing Distribution Shift (mean ± standard deviation)

Training Set	Test Set	Metric	Normal loss	Two-stage	Skewed loss
Group 2 & Group 3	Group 1	MAE	15.4±1.51	14.54±1.22	**14.50±1.30**
		ADC	-0.85±0.05	-0.80±0.03	**-0.75±0.03**
Group 3 & Group 1	Group 2	MAE	6.9±0.31	7.75±0.25	**7.44±0.39**
		ADC	0.08±0.005	0.14±0.01	**0.11±0.01**
Group 1 & Group 2	Group 3	MAE	15.45±1.12	11.4±0.38	**10.63±1.88**
		ADC	-0.7±0.03	-0.6±0.06	**-0.52±0.06**

Group 1 contains participants whose age is below 40 years; group 2 contains participants whose age is above 40 years whereas below 60 years and group 3 contains participants whose age is above 60 years.

3.5.5.3 Discussions

From Table 3.9, it could be observed that neither the skewed loss nor the two-stage approach achieves a near-zero ADC at the end of the training, which indicates that neither approach works well when facing distribution shifts.

However, by comparing the skewed loss and the two-stage approach, it could be observed that the skewed loss suffers less in terms of both MAE and ADC in all three experiments. Moreover, when Group 1 or Group 3 is selected as the test set, both correction approaches decrease the MAE compared with the normal loss, which is not observed in Table 3.4 and Table 3.6.

The reason why the skewed loss and the two-stage approach are not robust to distribution shift is that they both heavily rely on the validation set. The two-stage approach relies on the validation set to calculate β_0 and β_1 in (2.7), whereas the skewed loss aims to achieve a near-zero ADC on the validation set. For both approaches, the validation set is used as the target for the model to optimize. If the validation set is significantly different

from the test set, both methods will suffer.

There is one more point to be noted by comparing Table 3.9 and Table 3.4, it could also be observed that MAE increases significantly. There are several reasons for this. Firstly, the size of the training set becomes much smaller compared with the one used in Section 3.4. Secondly, take the first train-test split (Group 1 as the test set) as an example. In the first train-test split, the training set only contains subjects whose age is above 40 years so that the model is also more likely to give a prediction that is above 40 years, whereas the true labels from Group 1 are all below 40 years. Therefore, when using Group 1 and Group 3 as the test set, the MAE increases significantly.

To sum up, although both the two-stage approach and the skewed loss are not robust to distribution shift, the skewed loss suffers less compared with the two-stage approach in terms of both MAE and ADC.

3.6 Generalisation of the skewed loss

We have demonstrated the ability of the skewed loss to remove the dependence of brain age delta on chronological age. We would like to stress that the skewed loss could also be generalised to other problem domains regardless of brain age prediction. We provide a simple example here to illustrate the generalisation of the skewed loss using the apparent age prediction problem. It should also be noted that in apparent age prediction, achieving a more accurate age prediction (a lower MAE) is the primary goal rather than a near-zero correlation as in brain age prediction.

3.6.1 Apparent age prediction explained

Predicting a person's real age based on a single-face image is a classic problem in the computer vision field. However, the model performance can often be affected by outliers represented by people who have an appearance that is not in line with their real age [129]. Therefore, a different approach has been developed known as apparent age prediction, which is the age perceived by humans. The problem can then be formulated as building

a regression model that takes a person's face image as input and uses the apparent age as output.

3.6.2 Experiments

To empirically investigate our approach on apparent age prediction, we conducted experiments using a publicly available face dataset. The models were trained using the normal symmetric loss and the skewed loss respectively. The two-stage approach was also applied to models using the normal symmetric loss as the contrast.

3.6.2.1 Dataset

We selected the ChaLearn Looking at People 2015 competition dataset (LAP dataset), which provided thousands of annotated images [130], [131]. The images were labelled based on web applications that averaged the opinions of 10 independent users to obtain the apparent age.

To preprocess the original dataset, we first ran a face detection program on all images to remove the useless background information. The program could be found in a GitHub repository[4] and we applied the default HOG-based approach [132] to extract the face from the whole image. The resulting face images were then resized to have a spatial resolution of 100×100 pixels. In total, we obtained 4383 face images.

3.6.2.2 Models

We adopted the ResNet [70] architecture to demonstrate the effectiveness of the skewed loss in this problem. As for the choices of model-dependent hyperparameters, we also followed the conventional design strategy discussed in Section 3.4.2. The specific implementations could also be found in the GitHub repository mentioned in Section 3.1.

[4] https://github.com/ageitgey/face_recognition

3.6.2.3 Training and testing

The general training and testing strategy stays almost the same as the brain age prediction. Because of the size of the LAP datasets, to prevent models from over-fitting, we applied data augmentation in the training process. In each training iteration, every input image had a probability of 50% being flipped horizontally. Also, the L2 weight decay coefficient was set to 0.001.

During the training process, the Adam optimizer [128] was used as the default optimizer for all models. The initial learning rate was set to 0.01 and then multiplied by 0.5 every 50 epochs. The batch size was set to 64 as default and the total number of epochs was set to 300.

When we trained the model using the skewed loss, the hyperparameters used in Table 3.1 are set the same for all models. We set hyperparameter α to 50, β to 150, ϵ to 3, θ to 0.15, and η to 1.2.

Datasets were split using a stratified split strategy of which 80% was used for training, 10% for validation, and 10% for testing. Also, to minimize the fluctuations of the model results due to the random initialization of the weights, we trained our models using skewed loss and normal loss 5 times respectively.

3.6.3 Results

The model performance comparisons on apparent age prediction are shown in Table 3.10 and Figure 3.9. The correlation is measured in Pearson's r.

It should be noted that considering the small size of this dataset and low spatial resolution (100×100), the MAE should still be considered acceptable compared with other studies [129], [133].

From Figure 3.9, We can observe that the skewed loss also achieves a near-zero correlation at the end of the training, which proves that the skewed loss could be generalized to other areas regardless of the brain age prediction.

Table 3.10: ResNet Performance on Apparent Age Prediction (mean ± standard deviation)

Model	Metric	Normal loss	Two-stage	Skewed loss
ResNet	MAE	8.32±0.22	15.42±1.22	**13.64±0.99**
	correlation	-0.75±0.06	0.02±0.04	0.01±0.06

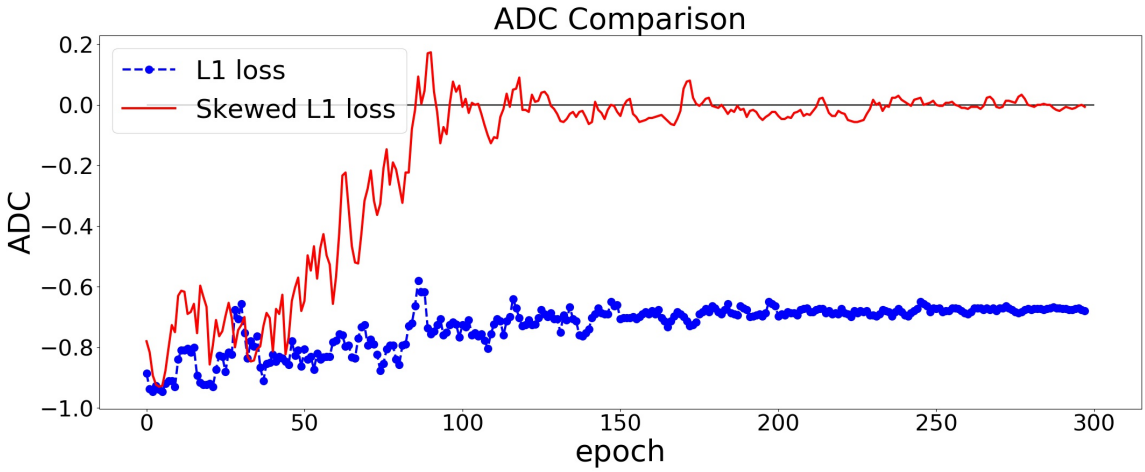


Figure 3.9: ADC Comparisons between L1 and skewed L1 loss. Dynamic lambda strategy is applied for the skewed loss. The model is trained twice using L1 loss and skewed L1 loss respectively. The dashed and solid lines represent the changes of ADC (measured in Pearson's r) on the validation set in the training process using L1 and skewed L1 loss respectively.

3.7 Discussion

In this chapter, we have proposed a skewed loss function and dynamic lambda training strategy to solve the nonzero ADC bias in brain age prediction. By debiasing predictions from brain age prediction models, we make the brain age delta robust to the confounding factor, chronological age, making it more reliable for future research and clinical studies. The skewed loss counteracts this bias by switching the normal symmetric loss function into a skewed form. The dynamic lambda strategy tunes λ_{max} iteratively to search for the optimal value that can enable the model predictions to have a near-zero ADC at the end of training.

The most significant difference between the skewed loss and the two-stage approach is

that the skewed loss controls ADC explicitly whereas the two-stage approach controls it implicitly. That means changing the way we measure ADC from Pearson's r to other types of correlation metrics, our approach could have a similar effect with the specified correlation metric approaching zero. In Section 3.5.3, we showed that using Spearman's rank correlation can also achieve similar performances.

We also demonstrated in Section 3.5 that both the skewed loss and the two-stage approach can achieve a near-zero ADC, whereas the skewed loss achieves comparable or even better model performances in terms of MAE compared with the two-stage approach. Theoretically speaking, a two-stage approach leads to a result that is strictly worse in terms of minimizing both MAE and ADC than a single integrated approach. In the two-stage approach, the model is optimized to minimize MAE. When correction is applied afterwards, there is no guarantee that the MAE will remain low. In an integrated approach, the model is optimized to minimize MAE conditioned on a desired maximum value of ADC [23]. Knowledge of the constraint allows the model to find a better local minimum to solve the task of minimizing both MAE and ADC. Also, the skewed loss only aims to encourage the model to make predictions towards the opposite side, i.e., from overestimations toward underestimations for young individuals and vice versa. In terms of MAE, there is no difference between overestimations and underestimations. In Section 3.5.5, we also evaluated whether the correction approaches are robust to distribution shifts. Although neither approach always achieves a near zero ADC, the skewed loss has been proved to suffer less in terms of both MAE and ADC from Table 3.9.

We have also demonstrated in Section 3.6 that the skewed loss, as well as the two-stage approach, could also be generalised to other areas. The observed bias is also commonly known as the regression dilution bias, which could be found in almost any regression model. However, whether to apply the skewed loss, as well as other approaches, depends on different applications. In brain age prediction, a near-zero ADC is the primary goal as a non-zero ADC could cause spurious relationships in subsequent experiments, whereas in apparent age prediction, a more accurate age estimation is more important.

It should also be noted that both the two-stage approach and the skewed loss have their advantages and shortcomings. Regarding the two-stage approach, it is easier to implement, because, in each stage, only one metric (MAE or ADC) is optimized. However, when the correction is applied afterwards, there is no guarantee that the MAE from the first stage remains low, which can be observed in Table 3.4. As for the skewed loss, it is obvious that by optimizing the two metrics (MAE and ADC) in a unified approach, the model takes care of both metrics at the same time. As for the shortcomings, a successful search of λ_{max} is crucial for the skewed loss approach. In Algorithm 1, we provide a heuristic way to search for it which proves to be robust. However, the two-stage approach does not require this additional hyperparameter searching.

There are still some aspects that require further improvements. Firstly, it would be preferred to apply the skewed loss on a larger dataset. Secondly, our approach can only be applied in regression models. Peng *et al.* [33] treated brain age prediction as a classification task and applied a weighted sum to calculate the predicted brain age. The two-stage approach can be applied to that situation while ours cannot. Also, it is useful to apply the skewed loss function in a case study to assess its sensitivity to detect pathological changes in patients. Although our approach achieves a near-zero ADC value that is similar to the two-stage approach, a case study can further confirm the relevance of our approach rather than through the ADC value.

3.8 Summary

To conclude, we developed the skewed loss function to counteract the regression bias in brain age prediction. In most cases, it achieves a better performance than the existing two-stage approach. Also, our approach has been verified in different datasets using different neural network architectures. By taking ADC explicitly into consideration in the training process, our approach also shows the potential to remove nonlinear relationships by measuring ADC using relevant correlation metrics.

In the next chapter, we turn to the second objective of this thesis to make brain age

prediction a more practical and accessible tool in clinical settings to facilitate its real-world use.

Chapter 4

Practical Brain Age Prediction - Generalisation on Low-Resolution Images via Super-Resolution

In this chapter, we turn to the second objective of this thesis to generalise brain age prediction for low-resolution MR images via super-resolution, making brain age prediction a practical and accessible application for clinical diagnosis.

We firstly propose a simple yet effective approach for the conventional patch-based training strategy for MRI super-resolution models by balancing the proportion of image patches containing high-frequency details, making the model focus more on high-frequency information in tiny regions, especially for the cerebellum. Compared with the conventional patch-based training strategy, the resultant super-resolved MR image from our approach improves significantly on the high-frequency details in the cerebellum. Secondly, we demonstrate an integrated brain age prediction workflow by incorporating MRI super-resolution. Brain age prediction model is typically trained on research-grade high-resolution images, which may face performance degradation in clinical settings, since clinical-grade MR images tend to have lower image resolutions. By adopting MRI super-resolution on clinical-grade MR images, more image details can be reconstructed, which exhibits negligible differences in predicting brain age compared with high-resolution images, after being fed into a pretrained brain age prediction model.

4.1 Introduction

High spatial resolution (HR) structural MR images contain detailed anatomical information, which is preferable in clinical diagnosis. However, HR images come at the cost of longer scan time, smaller spatial coverage, and lower signal-to-noise ratio [134], which

poses difficulties for the clinical environment due to hardware limits. One solution for this problem is to apply the single image super-resolution (SISR) technique, which only requires a low spatial resolution (LR) image to reconstruct its HR counterpart. It does not require extra scan time or high-cost scanners to generate an HR image and could be used to enhance the image quality from the low-field scanners. Nowadays, SISR has already been applied for medical image diagnosis, such as brain tumour detection [135].

Training SISR models requires paired LR and HR images so that the models can learn the mapping from the LR to HR images. Due to the rare availability of paired MR images, synthetic LR images, generated from their HR counterpart, have been widely used in the literature. To mimic the behaviours of the real-world low-field scanners, the HR images are typically transformed into the k-space using the Fast Fourier Transform (FFT), and the resolution degradation is performed on the k-space data [136].

In many previous studies, the synthetic LR images are similar to the original HR images, even if only keeping the central 25% k-space data [88], [89], [136]–[138], which may face performance degradation on the real-world LR images. In this chapter, we firstly investigate whether these previously proposed models are capable of learning more complex mappings when using more blurry LR images as input. We follow the procedures precisely in generating the LR images [88], [136], [138], while only keeping 6.25% of the k-space data. By keeping a smaller proportion of the original k-space data, there exists a more significant difference between the synthetic LR and the authentic HR images.

A more fundamental question in brain MRI super-resolution is about the patch-based training strategy. Conventionally, image patches sampled uniformly from the whole brain volume, rather than the whole images, are fed into the model as input to prevent hardware memory issues [88], [89], [136]–[138]. However, such uniform sampling introduces a data imbalance issue. The cerebrum, containing low-frequency information, generates far more patches than the cerebellum, whereas the latter contains more complex structures. It could result in the model being more focused on the information from larger regions of the brain and ignoring the fine-grained details in tiny areas. In the literature, the

reconstruction of the cerebrum has achieved almost indistinguishable performance using relatively small models [88]. However, there is still a huge gap between the super-resolved cerebellum and the authentic HR image, even using decent models [90], [91]. To solve the data-imbalance issue, we propose a simple yet effective algorithm via a non-uniform patch sampling for the conventional patch-based training strategy to treat the whole brain and cerebellum equally to derive better super-resolution quality.

Lastly, we propose an integrated brain age prediction workflow by incorporating the brain MRI super-resolution, generalising brain age prediction for low-resolution MR images. Brain age prediction models are typically trained using HR brain images so that the models can learn more detailed information about the brain structures. Performance degradation usually happens if different types of data are adopted in the testing time, for instance, images with lower resolution compared with high-resolution training images. With the help of the brain super-resolution techniques, an LR image will be super-resolved firstly and then used for predicting brain age, which makes predicting brain age more practical and accessible even in clinical environments.

Our main contributions in this chapter are:

1. We evaluate the capacity of several renowned models using more blurry LR images as input, compared with previous studies. By keeping 6.25% of the k-space data rather than 25% in the literature, the synthetic LR images look more similar to the real-world MR images generated from low field scanners, which may improve model generalization ability on real-world low-field scanners.

2. We propose a direct and effective method using a non-uniform sampling for the patch-based training strategy to improve the reconstruction quality for high-frequency details, especially for the cerebellum. To the best of our knowledge, the proposed method is the first work that treats the brain volume as two separate regions for brain MRI super-resolution tasks.

3. We integrate the brain MRI super-resolution into the brain age prediction framework, which makes it feasible to predict brain age from low-resolution images. Com-

pared with high-resolution images, low-resolution images achieve negligible differences in predicting brain age. We also provide an evaluation of the super-resolution technique and demonstrate the effectiveness of adopting super-resolved MR images for training a brain age prediction model. The super-resolved images achieve comparable predictive performance compared with HR images. The implementations are available in the GitHub repository [1].

This chapter is organised as follows. Section 4.2 proposes a non-uniform sampling approach for improving the high-frequency details for brain MRI super-resolution. Section 4.3 explains the settings of the experiments, including dataset and model architectures. Section 4.4 demonstrates quantitative experiment results and model comparisons. Section 4.5 provides a use case of the proposed super-resolution approach using brain age prediction. Section 4.6 illustrates the integrated brain age prediction workflow by incorporating brain MRI super-resolution and demonstrates performance evaluations. Section 4.7 discusses the advantages of the proposed approach and propose future works. Section 4.8 summarises the overall approach to conclude this chapter.

4.2 A non-uniform sampling for MRI super-resolution

As we previously mentioned in Section 4.1, the conventional uniform patch sampling results in a data imbalance issue between the over-represented region (cerebrum) and under-represented region (cerebellum), whereas the latter has more complex details and requires much more data to learn.

Ideally, a second model could be trained using only the cerebellum images, which is expected to achieve better results on the cerebellum. However, it requires additional training resources and might introduce problems when merging the two resultant images.

Therefore, a unified approach could be proposed by balancing the number of patches from the cerebrum and cerebellum during the training process. Such non-uniform patch sampling makes the model focus more on under-represented regions. To keep the overall

[1] https://github.com/anonymousreviewonly/MRI_SR

training time unchanged, the number of generated patches stays unchanged, whereas more patches come from the cerebellum compared with conventional patch sampling.

The non-uniform sampling approach is formally presented in Algorithm 2. Some notations in Algorithm 2 are defined in Table 4.1.

Table 4.1: Notations for Non-uniform Sampling Approach

Notation	Type	Description
$\mathcal{D}_{brain}$	constant	whole brain MRI dataset
$\mathcal{D}_{brain\_i}$	constant	whole brain MR image with index i
$\mathcal{D}_{brain\_patch}$	constant	whole brain MRI patch dataset
$\mathcal{D}_{cere}$	constant	extracted cerebellum dataset
$\mathcal{D}_{cere\_i}$	constant	extracted cerebellum image with index i
$\mathcal{D}_{cere\_patch}$	constant	extracted cerebellum patch dataset
N_{epoch}	variable	number of epoch
N_{sub}	variable	number of subject

Algorithm 2: A Non-uniform Sampling Approach

Given *Whole Brain MRI Dataset* $\mathcal{D}_{brain}$ *of length* N_{sub}
for $i \leftarrow 1$ **to** N_{sub} **do**
  Extract the cerebellum $\mathcal{D}_{cere\_i}$ from the whole brain volume $\mathcal{D}_{brain\_i}$
end
Concatenate all cerebellum volumes into a cerebellum dataset $\mathcal{D}_{cere}$
for $j \leftarrow 1$ **to** N_{epoch} **do**
  Generate patch dataset $\mathcal{D}_{brain\_patch}$ from $\mathcal{D}_{brain}$
  Randomly select 50% patches from $\mathcal{D}_{brain\_patch}$ and discard rest patches
  Generate patch dataset $\mathcal{D}_{cere\_patch}$ from $\mathcal{D}_{cere}$ with same length of $\mathcal{D}_{brain\_patch}$
  Concatenate $\mathcal{D}_{brain\_patch}$ and $\mathcal{D}_{cere\_patch}$ randomly as the training set
  Model Training
end

The intuition of the proposed approach is straightforward. Since the brainstem contains few patterns, the brain super-resolution focuses more on the cerebrum and cerebellum. Also, as the cerebrum and cerebellum contain disparate structures, it is natural to treat the whole volume super-resolution as two sub-volume super-resolution tasks. Therefore, by keeping a 50/50 data distribution of image patches, the model becomes less biased towards the cerebrum, which naturally helps the reconstruction in the cerebellum. Theoretically speaking, such non-uniform sampling might cause performance degradation in the cerebrum, as less information is used for training. However, in Section 4.4.2, we demonstrate that our modification results in negligible difference compared with conventional uniform sampling. Also, the proposed non-uniform sampling improves the recon-

struction quality for the cerebellum significantly, resulting in more high-frequency details being restored in cerebellum.

4.3 Experiments

4.3.1 Dataset and data preprocessing

To fairly compare different models and training strategies, we chose a widely used and publicly accessible brain structural MRI database, the Human Connectome Project (HCP) [139]. It contains 3D T1-weighted MR images from 1113 subjects acquired via the Siemens 3T platform using a 32-channel head coil on multiple centres. The images come in high spatial resolution as 0.7 mm isotropic in a matrix size of $320 \times 320 \times 256$. These high-quality images serve as the authentic HR images in the following experiments. We further randomly chose 800 images as the training set, 100 images as the validation set, and the rest 213 images as the test set.

We further performed the Brain Extraction Tool (BET) [140] on the original HR images to delete non-brain tissue to make our model focus only on the brain structures to improve performance.

To generate the corresponding LR images, we followed the same pipeline demonstrated in [88]. The detailed steps are as follows:

1. Applying the FFT to authentic HR images to convert the original image into the k-space data.
2. Masking (zeroing) the outer part of the 3D k-space data along two axes representing two MR phase encoding directions.
3. Applying the inverse FFT to the masked k-space data to generate the synthetic LR images.

An illustration of masking the k-space data is shown in Figure 4.1. The above LR image synthesis procedure mimics the real MR image acquisition process where a low-resolution

MRI is scanned by reducing acquisition lines in phase and slice encoding directions. It should be noted that the synthetic LR images have the same size as the HR images, avoiding the checkerboard artefacts [141].

A hyperparameter, named "scale factor", is introduced in the demonstrated image synthesis pipeline, which determines the proportion of masked k-space data. A larger scale factor masks more k-space data, resulting in more blurry images. In Figure 4.1, we set the scale factor to 2. Therefore, only the central 25% ($1/2 \times 1/2$) of the k-space data is preserved. Using a scale factor of 2 is common in many studies [88], [89], [136], whereas, in this chapter, we set the scale factor to 4, preserving the central 6.25% ($1/4 \times 1/4$) of the k-space data, to evaluate the models on more blurry LR images. An image comparison using different scale factors is shown in Figure 4.2.

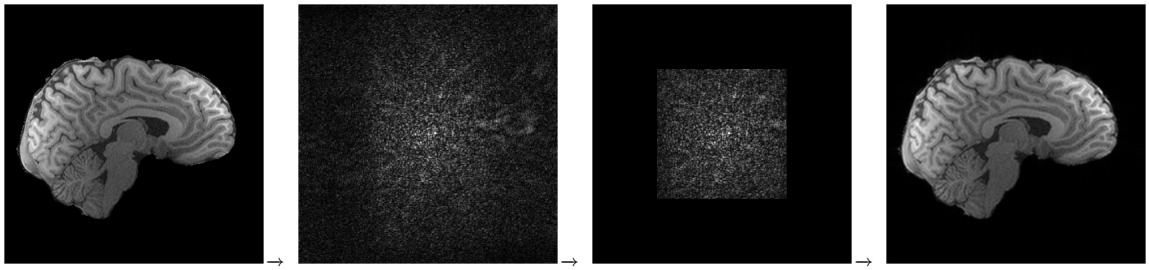
Moreover, as the proposed non-uniform sampling, described in Section 4.2, requires a separate sampling on the cerebellum, a separate dataset containing the cerebellum is created. It is achieved by using FastSurfer [142], a deep-learning-based open-sourced brain segmentation software[2]. FastSurfer automatically labels the cerebellum regions for extraction. The cerebellum extraction pipeline is demonstrated in Figure 4.3. It should be stressed that segmentation masks are generated using only the authentic HR images, and the corresponding masks will be applied on both HR and LR images to ensure both HR and LR cerebellums are also voxelwise paired. A diagram of the overall image preprocessing is shown in Figure 4.4.

4.3.2 Models

We chose two widely recognized models in this field, mDCSRN and mDCSRN-WGAN [88] to evaluate the proposed approach. We chose these two models since they both adopt the DenseNet [37] as the backbone, which is still popular in various model designs. Also, they represent two distinct model designs, CNN-based and GAN-based models.

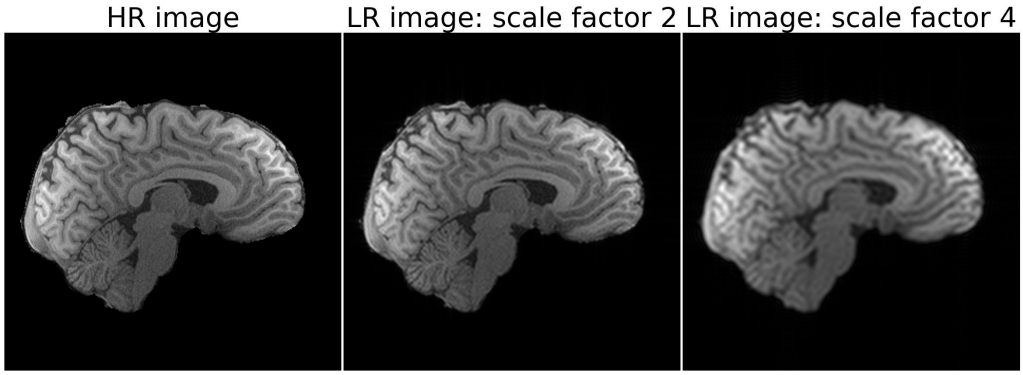
It should be noted that there are no limitations in the choice of model architectures since

[2] https://github.com/Deep-MI/FastSurfer

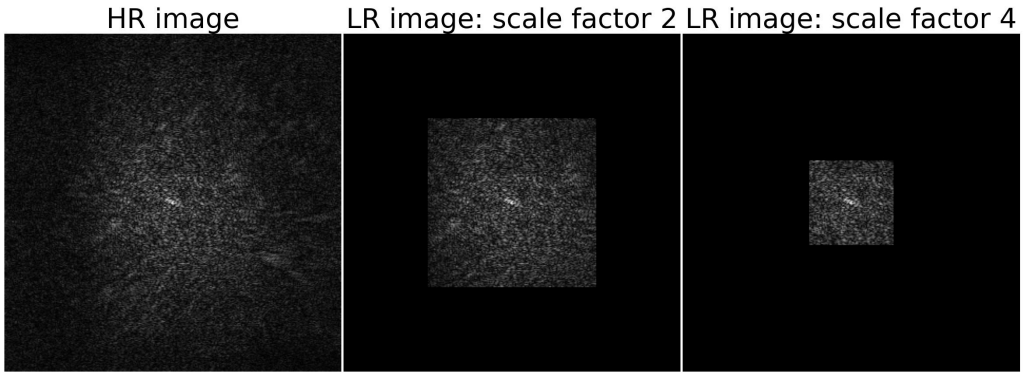


(a) (b) (c) (d)

Figure 4.1: An illustration of masking the k-space data. (a) represents the HR image. (b) is obtained by applying FFT to (a). (c) is obtained by masking (zeroing) the phase and slice encoding directions of (b) and (d) is generated by applying inverse FFT on (c). We set the scale factor to 2.



(a)



(b)

Figure 4.2: MR image and corresponding k-space data comparisons using different scale factors. In (a), from left to right, it shows the original HR image, and LR image using a scale factor of 2 and 4. (b) shows the corresponding k-space data using different scale factors. For scale factors 2 and 4, the central 25% and 6.25% of the data are preserved.

the proposed approach focuses primarily on data sampling. We chose DenseNet-based models for their stability and effectiveness. The mDCSRN model adopts a DenseNet [37]

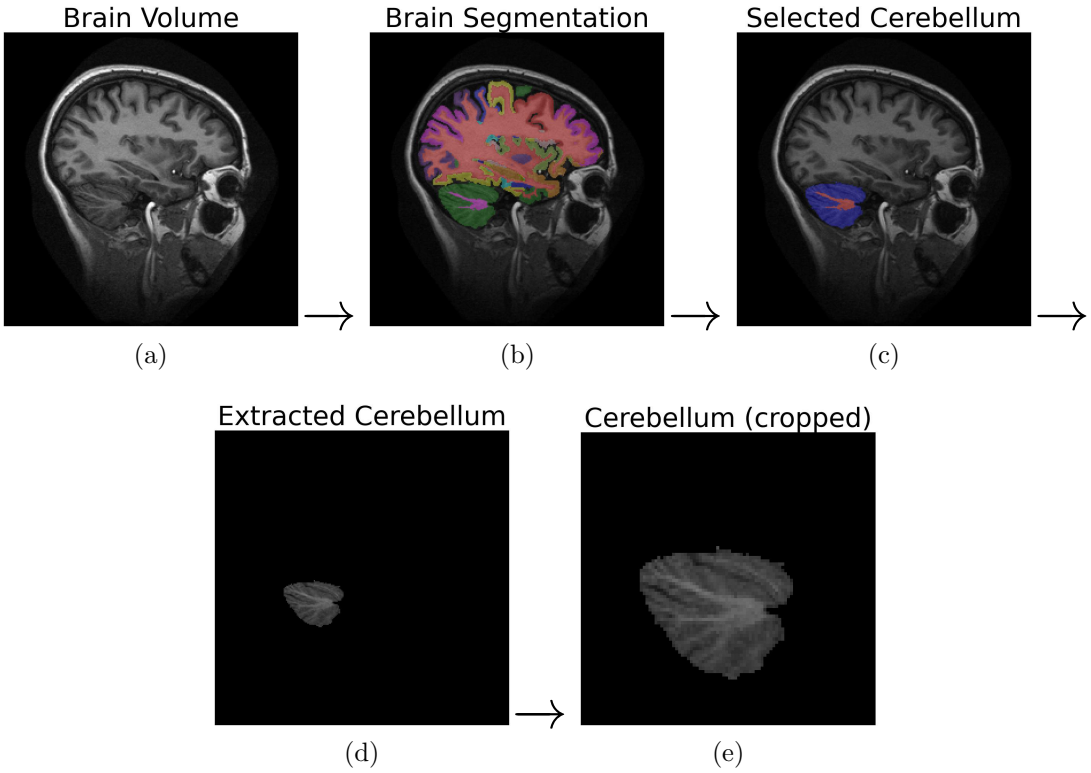


Figure 4.3: An illustration of the cerebellum extraction pipeline using FastSurfer. (a) represents the authentic HR image. (b) is obtained by applying FastSurfer to (a). (c) is obtained by selecting the cerebellum-related mask of (b) and (d) is generated by applying the cerebellum mask (c) on (a). Finally, we remove the useless background in (d) to derive (e). FastSurfer labels the cerebellum automatically for delineation.

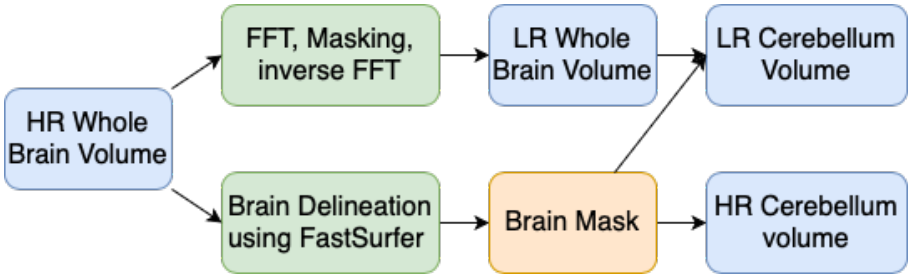


Figure 4.4: A diagram of MR image preprocessing.

architecture with 4 dense blocks. The mDCSRN-WGAN model uses the same architecture of mDCSRN as the generator and uses the Wasserstein GAN [143] to guide the model training. We adopted the identical model architecture settings from the mDCSRN and mDCSRN-WGAN mdoels in our experiments.

4.3.3 Training and testing

In the training process, Adam optimizer [128] is used as the default optimizer for all models. The initial learning rate is set to 0.001 and then multiplied by 0.5 every 10 epochs. The batch size is set to 64 as default and the total number of epochs is set to 50. We use the L1 loss for mDCSRN and the generator of the mDCSRN-WGAN model since L1 loss has been shown to generate more realistic images than L2 loss [144]. The size of image patches is set to $32 \times 32 \times 32$ for all patches. The hyperparameter settings, such as learning rate, are based on our experiment testing.

To evaluate the super-resolution performance, we used the structural similarity index (SSIM), peak signal-to-noise ratio (PSNR), and normalized root mean squared error (NRMSE) to measure the similarity between super-resolved (SR) images and HR images. SSIM uses the mean, variance, and covariance to estimate the similarity of the two images. PSNR is used to further quantify the recovered image quality using the mean squared loss. NRMSE is a more direct way to measure the pixel-wise similarity between the original and super-resolved images. In general, lower NRMSE, higher PSNR, and higher SSIM values represent better super-resolution results. Although these metrics have been criticized that a high score does not represent a better image quality [107], they are still the most commonly used metrics, and no other perceptual quality metrics for 3D MR images have been applied in the literature.

4.4 Results

In this section, we first demonstrate the SISR model performance using synthetic LR images, generated under a scale factor of 4. Then, we compare the model performance between the conventional patch sampling and the proposed non-uniform sampling. Since the CNN-based model performs similarly to the GAN-based model, to avoid unnecessary repetition, we only demonstrate sample output images from the mDCSRN model as an illustration. When conducting the significance test, the paired t-test is adopted by

default.

4.4.1 Model performance under a scale factor of 4

Figure 4.5 demonstrates the model performance of the whole brain volume using the mDCSRN model in three perspectives. It can be observed that the SR image is similar to the HR image in most regions of the brain, especially in the cerebrum. It indicates that relatively simple models, such as mDCSRN, are still capable of handling more blurry patterns, especially in the cerebrum region. The quantitative analysis is demonstrated in Table 4.2. From Table 4.2, it could be observed that both mDCSRN and mDCSRN-WGAN significantly improve the image quality in all three metrics compared with the baseline ($p < 0.01$).

Table 4.2: Model performance of the whole brain volume (mean ± standard deviation)

Model	**Image Pair**	**PSNR**	**SSIM**	**NRMSE**
None (Baseline)	HR & LR	32.33 ± 2.92	0.92 ± 0.01	0.148 ± 0.09
mDCSRN	HR & SR	34.23 ± 2.14	0.95 ± 0.01	0.119 ± 0.06
mDCSRN -WGAN	HR & SR	34.19 ± 2.11	0.95 ± 0.02	0.121 ± 0.07

4.4.2 Comparisons between conventional and proposed sampling

Table 4.3 demonstrates the quantitative analysis of model performance for mDCSRN and mDCSRN-WGAN. For each model, we compare the conventional sampling approach and the proposed approach on the whole brain volume and the cerebellum respectively. We also add the PSNR, SSIM, and NRMSE values between the HR and LR images as the baseline to demonstrate the improvement of our approach.

From Table 4.3 nad Table 4.4, when considering the reconstruction quality of the whole brain volume for both mDCSRN and mDCSRN-WGAN model, the conventional patch-based training approach achieves comparable performance with our sampling approach

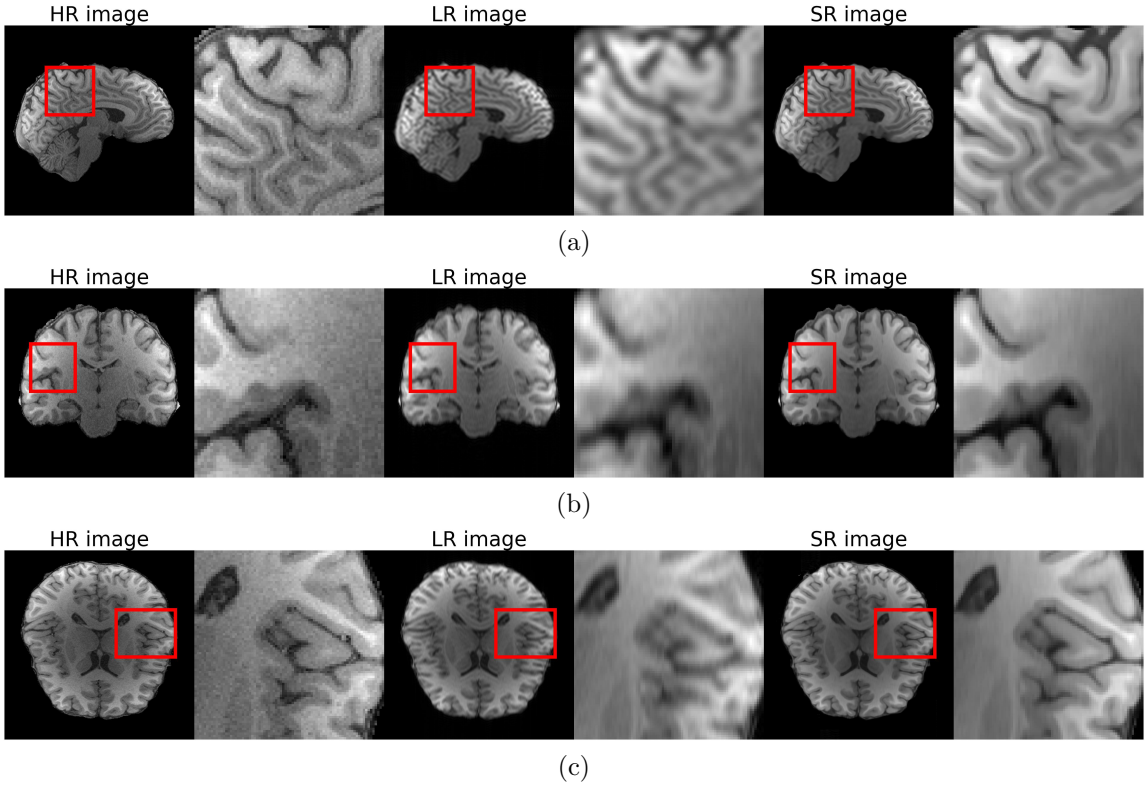


Figure 4.5: An illustration of the HR, LR, and SR brain image using the mDCSRN model. (a), (b), and (c) shows the sagittal, coronal, and axial view.

in all three metrics. There are no significant differences in all three metrics when evaluated on the whole brain volume between the two sampling approaches, whereas both approaches outperform the baseline performance significantly ($p < 0.05$). When evaluated on the cerebellum, our proposed sampling approach significantly outperforms the conventional sampling approach for both mDCSRN and mDCSRN-WGAN models in all three metrics ($p < 0.05$).

The reason for the improved performance on the cerebellum is clear, as we have increased the proportion of image patches from the cerebellum for models to learn. The potential reason for the comparable performance on the whole brain volume between the conventional and the proposed approach is that the number of patches from the whole brain is redundant. Reducing the proportion of patches from the over-represented regions does not affect model performance significantly. Also, at the beginning of each epoch, random patches will be selected until the end of training. Therefore, the model is still able to learn information from all whole brain patches.

Figure 4.6 and Figure 4.7 demonstrate the super-resolution performance between the conventional approach and the proposed approach on both whole brain volume and cerebellum respectively. It can be observed that our approach achieves almost the same performance in the whole brain volume, whereas reconstructs more details in the cerebellum.

Therefore, compared with the conventional sampling approach, the proposed non-uniform sampling approach achieves comparable performance on the whole brain volume, whereas significantly improves the reconstruction quality for the fine structures in the cerebellum.

Table 4.3: mDCSRN model performance of the whole brain and cerebellum (mean ± standard deviation)

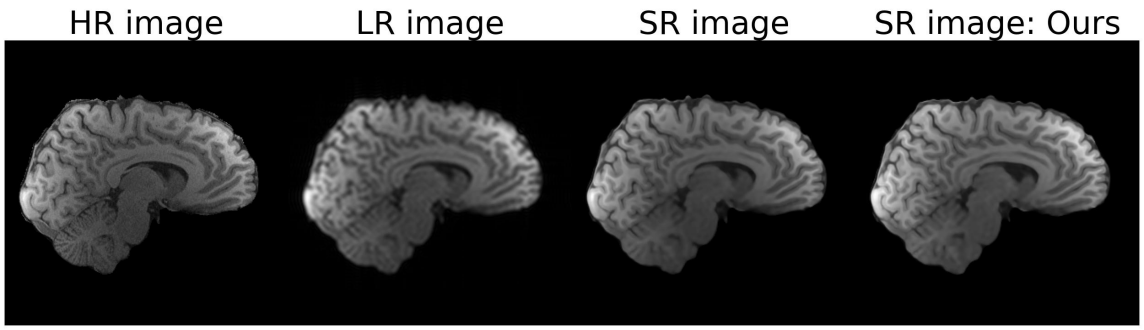
Brain Regions	Image Pair	PSNR	SSIM	NRMSE
Whole Brain	HR & LR (baseline - no model)	32.33 ± 2.92	0.92 ± 0.01	0.148 ± 0.09
	HR & SR (conventional sampling approach)	34.23 ± 2.14	0.95 ± 0.01	0.119 ± 0.06
	HR & SR (our approach)	34.18 ± 1.98	0.95 ± 0.02	0.120 ± 0.05
Cerebellum	HR & LR (baseline - no model)	28.37 ± 2.66	0.88 ± 0.01	0.15 ± 0.10
	HR & SR (conventional sampling approach)	28.89 ± 1.81	0.905 ± 0.01	0.135 ± 0.06
	HR & SR (our approach)	29.62 ± 1.55	0.93 ± 0.01	0.127 ± 0.02

Table 4.4: mDCSRN-WGAN model performance of the whole brain and cerebellum (mean ± standard deviation)

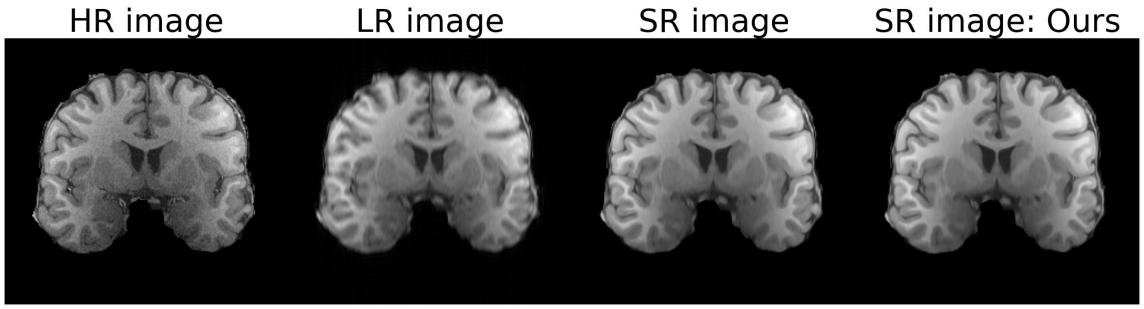
Brain Regions	Image Pair	PSNR	SSIM	NRMSE
Whole Brain	HR & LR (baseline - no model)	32.33 ± 2.92	0.92 ± 0.01	0.148 ± 0.09
	HR & SR (conventional sampling approach)	34.19 ± 2.11	0.95 ± 0.02	0.121 ± 0.07
	HR & SR (our approach)	34.13 ± 2.17	0.95 ± 0.03	0.122 ± 0.05
Cerebellum	HR & LR (baseline - no model)	28.37 ± 2.66	0.88 ± 0.01	0.15 ± 0.10
	HR & SR (conventional sampling approach)	28.91 ± 1.73	0.902 ± 0.02	0.139 ± 0.04
	HR & SR (our approach)	29.47 ± 1.81	0.918 ± 0.02	0.124 ± 0.05

4.5 Evaluation via brain age prediction

To further validate the effectiveness of the proposed super-resolution sampling approach, we provide an example use case of super-resolved MR images via brain age prediction.

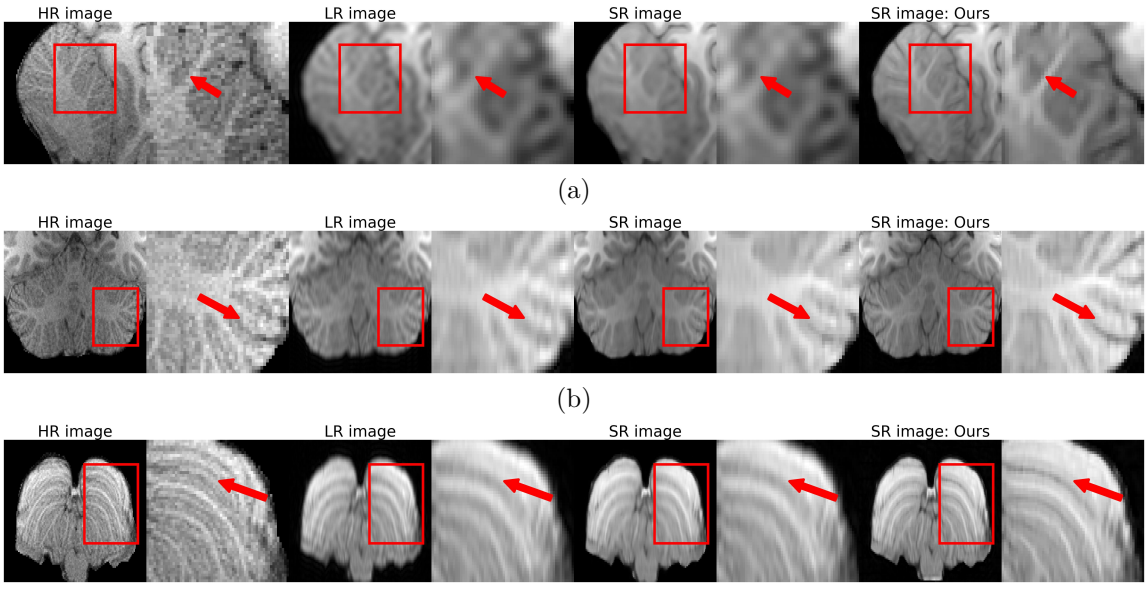


(a)



(b)

Figure 4.6: HR, LR, SR (conventional sampling), and SR (the proposed sampling) whole brain volume illustrations by mDCSRN model. (a) shows the sagittal view and (b) shows the coronal view.



(a)

(b)

(c)

Figure 4.7: An illustration of the HR, LR, SR (with conventional sampling), and SR (with the proposed sampling) cerebellum image by mDCSRN. (a), (b), and (c) show the sagittal, coronal, and axial view.

4.5.1 Experiment settings

DenseNet architecture is selected for brain age prediction considering its performance and speed and the HCP dataset is also adopted for brain age prediction. Then, the HR images are used to generate corresponding LR images, as discussed in Section 4.3.1.

Afterwards, the generated LR images are super-resolved using the mDCSRN models trained with conventional sampling and the proposed sampling approach respectively, generating two super-resolved images for each LR image. Therefore, two super-resolved MR image datasets are created. One adopts the conventional uniform sampling and the other uses the proposed non-uniform sampling.

The two SR image datasets are used to train the brain age prediction model respectively, resulting in two brain age prediction models, each dedicated to a sampling strategy. Two more brain age prediction models are also trained using original HR images and LR images respectively as the contrast. In total, four brain age prediction models are trained using HR images, LR images, SR images (conventional approach), and SR images (our approach) respectively.

To decrease the variations in the model training process, each model is trained five times. The Mean Absolute Error (MAE) is adopted as the evaluation metric. In the training process, Adam optimizer [128] is used as the default optimizer for all models. The initial learning rate is set to 0.01 and then multiplied by 0.5 every 30 epochs. The batch size is set to 64 as default and the total number of epochs is set to 300.

4.5.2 Results

Results in Table 4.5 demonstrate the model performance of brain age prediction using HR images, LR images, SR images from conventional patch sampling, and SR images from our approach respectively.

It can be observed that SR images from our approaches achieve comparable predictive accuracy compared with SR images from conventional patch sampling, and both ap-

proaches surpass LR images significantly ($p < 0.05$). The HR images show slightly better performance compared with super-resolved images from both approaches, whereas the difference is not significant.

The model performance is expected. Because, in brain age prediction, the model predicts age based on the whole brain volume rather than the cerebellum. The improvements in the cerebellum, discussed in Section 4.4, are likely to be ignored by the model. However, it also validates that our approach achieves almost the same reconstruction quality on the whole brain volume, compared with the conventional approach. Therefore, we have validated that the super-resolved MR images derived from the proposed non-uniform sampling strategy can be applied for subsequent clinical analysis and it may demonstrate more significant improvements in some cerebellum-specific analysis.

Table 4.5: Model performance of brain age prediction evaluation (mean ± standard deviation)

Model Input Images	**MAE**
HR images	**2.25 ± 0.27**
SR images (conventional sampling approach)	2.38 ± 0.31
SR images (our approach)	2.33 ± 0.34
LR images	3.01 ± 0.28

4.6 Generalising brain age prediction for LR images

As we mentioned in Section 4.1, introducing brain MRI super-resolution into brain age prediction holds many benefits, making it applicable to predicting brain age for low-resolution MR images, especially for clinical environments. To demonstrate the usability of this combination, experiments are conducted by comparing the prediction accuracy of super-resolved MR images and HR images. A detailed illustration of the integrated brain age prediction workflow with super-resolution included is demonstrated in Figure 4.8.

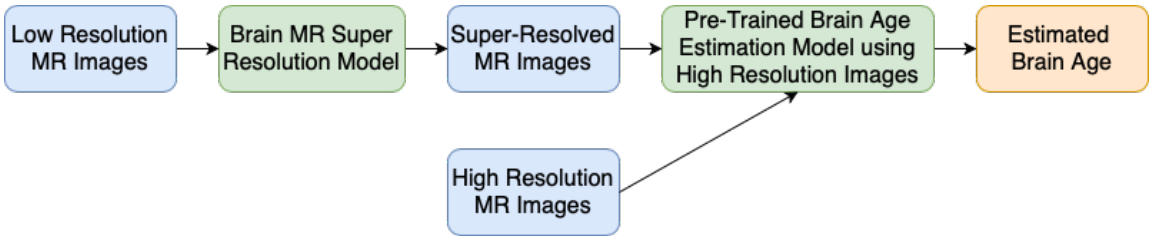


Figure 4.8: A detailed brain age prediction workflow with MRI super-resolution. A brain age prediction model is trained using HR images and it can be used to predict brain age for both high-resolution images and low-resolution images after being super-resolved.

4.6.1 Experiment settings

As is shown in Figure 4.8, the integrated brain age prediction workflow requires two trained models, the super-resolution model and the brain age prediction model. In Section 4.4, we have trained the mDCSRN model and mDCSRN-WGAN model for super-resolution. The two models will be used as super-resolution models, illustrated in Figure 4.8.

In terms of model architectures for brain age prediction, we choose the DenseNet [34] architecture and SFCN [33] architecture considering their performance and speed.

HCP dataset, mentioned in Section 4.3.1, is still used in brain age prediction. The HR image dataset is split using a stratified split strategy of which 80% was used for training, 10% for validation, and 10% for testing. Then, similar to Section 4.5, the testing HR images are used to generate corresponding testing LR images as discussed in Section 4.3.1. Afterwards, the generated testing LR images are super-resolved using the mDCSRN model and mDCSRN-WGAN trained with conventional patch sampling and the proposed sampling approach respectively, generating four testing super-resolved MR images for each LR testing image. Therefore, four super-resolved MR image testing datasets are created.

The HR image training dataset is used to train the brain age prediction model. The HR image testing dataset and the four super-resolved image testing datasets are tested. To decrease the variations in the model training process, each brain age prediction model is

trained five times. The Mean Absolute Error (MAE) is adopted as the evaluation metric commonly used in age prediction research. In the training process, Adam optimizer [128] is used as the default optimizer for all models. The initial learning rate is set to 0.01 and then multiplied by 0.5 every 30 epochs. The batch size is set to 64 as default and the total number of epochs is set to 300.

4.6.2 Results

Results in Table 4.6 demonstrate the model performance of brain age prediction using HR testing images, LR testing images and the four super-resolved testing images using conventional patch sampling and our approach form mDCSRN and mDCSRN-WGAN models respectively. Table 4.7 demonstrates the significance table of model performance compared with HR-image-based model. It can be observed that using the same pre-trained model for brain age prediction, four different SR testing datasets perform similarly in terms of MAE. The HR images show slightly better performance compared with super-resolved images from all four sources, whereas the difference is not significant. On the contrary, when adopting the LR images for prediction, we can observe a significant model performance degradation in terms of MAE, illustrating a clear distribution shift problem.

The performance table indicates that the testing performance of super-resolved images is comparable with HR images even if the brain age prediction models are trained using the HR images. The distribution shift problem is not significant in the refined brain age prediction framework, and the super-resolved images are capable of displaying similar information compared with the original HR images.

Therefore, we have demonstrated that the brain MRI super-resolution can be integrated into brain age prediction, which makes it feasible to predict brain age for low-resolution MR images. Such modification does not introduce a significant distribution shift problem for brain age prediction as long as the super-resolution model is properly trained.

4.6.3 Summary

The integrated brain age prediction with MRI super-resolution included demonstrates strong potential in clinical diagnosis. A sophisticated large-scale brain age prediction model can be pre-trained using research-level high-resolution MR images. Then, in clinical environments, clinical-graded MR images can be super-resolved to the target resolution used to train the age prediction model. In this way, brain age can be predicted on the super-resolved images in clinical situations, which makes brain age accessible in clinical situations.

Table 4.6: Model performance of integrated brain age prediction framework (mean ± standard deviation)

DenseNet	**MAE**	**SFCN**	**MAE**
HR images	**2.50 ± 0.27**	HR images	**2.97 ± 0.27**
SR images (mDCSRN; conventional sampling)	2.61 ± 0.28	SR images (mDCSRN; conventional sampling)	3.09 ± 0.31
SR images (mDCSRN; our approach)	2.58 ± 0.21	SR images (mDCSRN; our approach)	3.13 ± 0.29
SR images (mDCSRN-WGAN; conventional sampling)	2.65 ± 0.28	SR images (mDCSRN-WGAN; conventional sampling)	3.10 ± 0.22
SR images (mDCSRN-WGAN; our approach)	2.63 ± 0.23	SR images (mDCSRN-WGAN; our approach)	3.15 ± 0.21
LR images	4.04 ± 0.15	LR images	4.29 ± 0.20

4.7 Discussion

We have proposed a simple yet effective algorithm, via a non-uniform sampling strategy, for conventional patch-based training in brain MRI super-resolution problems. We have also proposed an integrated brain age prediction workflow, by integrating the brain MRI super-resolution, to make it feasible to predict brain age for low-resolution images in the clinical environments. We provide a different perspective by observing the brain

Table 4.7: Model performance significance table of refined brain age prediction framework compared with HR images

DenseNet	**p-values**	**SFCN**	**p-values**
SR images (mDCSRN; conventional sampling)	0.81	SR images (mDCSRN; conventional sampling)	0.73
SR images (mDCSRN; our approach)	0.86	SR images (mDCSRN; our approach)	0.58
SR images (mDCSRN-WGAN; conventional sampling)	0.65	SR images (mDCSRN-WGAN; conventional sampling)	0.64
SR images (mDCSRN-WGAN; our approach)	0.72	SR images (mDCSRN-WGAN; our approach)	0.49
LR images	**0.01**	LR images	**0.01**

structures in super-resolving brain MR images. By maintaining a 50/50 data distribution of patches, our approach achieves comparable reconstruction performance on the whole brain volume, whereas significantly surpasses the conventional approach on the cerebellum. More high-frequency details in the cerebellum have been restored from our approach.

We have demonstrated the capacity of several renowned models by using much more blurry LR images as input in Section 4.4. In the literature, a scale factor of 2 has been widely adopted to generate LR images. In this paper, we select a scale factor of 4 to generate LR images and the models are still capable of learning the mappings between the LR and HR images.

We have shown that our non-uniform sampling strategy across the whole brain volume does not result in significant reconstruction degradation in the non-cerebellum regions. The most probable reason behind this is that in the conventional approach, the number of patches from the over-represented regions is already redundant, and non-cerebellum regions contain simpler patterns. As we demonstrated in Figure 4.5 and Table 4.3, the mDCSRN model can achieve indistinguishable reconstruction quality over non-cerebellum regions.

We have also provided a use case for MRI super-resolution via brain age prediction. From Table 4.5, the super-resolved MR images demonstrate significantly better predictive accuracy in predicting brain age compared with LR images ($p < 0.05$). Although the HR images achieve the highest accuracy, the difference between HR images and super-resolved images is not significant. In terms of the two super-resolution sampling approaches, since the proposed approach focuses primarily on the cerebellum, there is no significant difference between the two approaches for age prediction using the whole brain volume.

By inserting the brain MRI super-resolution into the brain age prediction, model predictions using LR images achieve comparable performance compared with high-resolution testing images. The prediction accuracy difference between HR images and super-resolved images is negligible. It improves the flexibility of input data in brain age prediction so that the brain age prediction can handle HR and LR images. The most probable reason is that brain MRI super-resolution can achieve excellent performance in super-resolving the cerebrum and brain age prediction focuses more on the structures of the cerebrum, since it shows more apparent alterations in the ageing process and is also the largest region in the brain.

There are still some aspects that require further improvements. Firstly, although our approach focuses on data balancing and does not depend on model architectures, it might be worthwhile to apply such sampling to more advanced architectures. Also, our proposed sampling strategy is mainly based on intuitions. Given that the brain mainly contains the cerebrum, cerebellum, and brainstem, whereas the latter contains very few structures, it is reasonable to treat the whole brain volume SR into two sub-volume SR problems and apply a 50/50 data distribution for the cerebellum and non-cerebellum regions. However, more sophisticated sampling methods could be explored to evaluate the optimal combination of the two regions.

4.8 Summary

To conclude, we propose a more effective sampling approach for the conventional patch-based training strategy for brain MRI super-resolution problems. By using a non-uniform sampling across the whole brain volume, the model tends to learn more high-frequency details from the cerebellum, rather than being focused on non-cerebellum regions due to data imbalance. With the help of the proposed sampling approach, we improve the model performance on the cerebellum significantly and also achieve comparable performance on the whole brain volume. We also refine the brain age prediction workflow by inserting the brain MRI super-resolution techniques, making the brain age prediction a more practical application for clinical diagnosis and capable of handling more diverse images.

In the next chapter, we turn to the last objective of this thesis to improve the model predictive accuracy using a multi-modal approach.

Chapter 5

Accurate Brain Age Prediction - Multi-Modal Learning with structural MRI and diffusion MR measures

In this chapter, we turn to the third objective of this thesis to demonstrate a multi-modal approach for predicting brain age more accurately.

We propose a tract-wise approach for predicting brain age using diffusion MRI and quantitative MRI measures of the white matter tracts. The tract-wise approach prioritises more informative white matter tracts and discards less useful ones, providing insights into the more ageing-sensitive white matter tracts. We further compare brain age prediction that uses structural T1-weighted images to prediction that uses diffusion MR measures of the white matter tracts. By comparing the model accuracy of the two data modalities, we illustrate that diffusion MR measures are also sensitive in predicting age. A two-stream brain age prediction framework is further proposed by fusing the model predictions from diffusion MR measures and structural T1-weighted images, achieving improved performance compared with any single data modality.

5.1 Introduction

The rapid advance of deep learning techniques over the past decade makes it feasible for training deep neural networks on structural MR images to derive the predicted brain age [1], [31], [33], [34], [42]. In the literature, structural MRI and diffusion MRI are the most widely-adopted MR imaging modalities for brain age prediction [17], [19], [33], [34], [36], [117]–[119]. Structural MRI provides detailed information about the brain's anatomical structures and diffusion MRI captures the integrity and connectivity of the brain microstructure.

With the advance of deep learning models, it has become a common practice to train neural networks on MR images to predict brain age. Structural MR images, such as T1-weighted images, have demonstrated satisfactory accuracy in predicting age for healthy people [1], [33], [34], [36], [104]. However, there are few studies in directly predicting brain age using neural networks on diffusion MR images [117]. A more common practice to predict brain age using diffusion MRI data is to first extract features from diffusion MR images and then model brain age using the extracted features, rather than raw images [17], [19], [30], [110], [118], [119], [145].

To quantify the brain microstructural alterations in the ageing process, various diffusion MRI measures of white matter microstructure, such as fractional anisotropy (FA), mean diffusivity (MD), and radial diffusivity (RD), have been proposed and adopted for predicting brain age [17], [19], [30], [110], [145]. Since diffusion MRI measures can be estimated on each white matter tract, it results in a large number of diffusion MRI features and different feature selection approaches have been explored [17], [19], [110], [118], [145].

In this chapter, we propose a tract-wise approach to prioritise more informative white matter tracts for accurate brain age prediction, providing insights into the more ageing-sensitive white matter tracts. The selection of informative tracts happens in the model training process and all white matter tracts have the chance of being selected. We demonstrate that although some tracts are more likely to be prioritised, all white matter tracts are adopted for predicting brain age using the model ensemble. We also train neural networks on 3D structural T1-weighted MR images as the contrast to further investigate any possible advantages of using white matter microstructure measures.

The two data modalities mentioned above contain distinct information about the brain and therefore complement each other, providing a more complete picture of the brain and of how it ages. Therefore, we further provide a two-stream brain age prediction framework containing two separate routes, for structural images and measures of white matter microstructure respectively. The structural image pipeline adopts several renowned neural networks to make predictions, as stated in the literature [33], [34]. The microstructure

measures pipeline adopts traditional models, such as XGBoost, to make predictions as the microstructure measures come in tabular format. The two separate pipelines are later combined to create an ensemble model to produce the final age prediction, which outperforms any single pipeline. The implementations of the proposed two-stream framework are available in the GitHub repository[1].

Our contributions in this chapter are:

1. We propose a tract-wise approach for accurate brain age prediction using diffusion MRI measures of white matter microstructure. The tract-wise approach prioritises more informative white matter tracts in the model training process, providing insights into more ageing-sensitive tracts. The proposed approach serves in both dimension reduction and feature selection aspects by first separating different white matter tracts and then prioritising more ageing-sensitive ones for predictions. Unlike some schemes that perform feature selection before model training [17], [118], [145], the tract-wise approach preserves the most complete information, keeping all white matter tracts in consideration.

2. We provide a comparison between structural T1-weighted MR images and diffusion MRI measures of white matter microstructure in predicting brain age. Apart from diffusion MRI measures, quantitative MRI measures, such as myelin water fraction (MWF) and intracellular volume fraction (ICVF), are also added to model brain age. A two-stream brain age prediction framework is further proposed by combining information from structural images and white matter microstructure measures, which provides the model with a more complete picture of the brain, achieving an improved accuracy compared with any single data modality.

This chapter is organised as follows. Section 5.2 introduces our dataset and relevant preprocessing pipelines. Section 5.3 proposes a two-stream framework for structural MR images and white matter microstructure measures respectively. Section 5.4 demonstrates experiment results and model comparisons. Section 5.5 discusses the advantages of the

[1] https://github.com/hanzhiwangchn/Microstructure_Age_Prediction

multi-modal approach and proposes future works. Section 5.6 summarises the overall two-stream framework to conclude this chapter.

5.2 Dataset acquisition and preprocessing

The dataset used in this chapter contains two data modalities, 3D structural T1-weighted MR images and diffusion MRI measures of the white matter microstructure. The dataset was collected at the Cardiff University Brain Research Imaging Centre (CUBRIC), Cardiff University.

MRI data were collected using an ultra-strong magnetic field gradient (300 mT/m) 3T Connectom MRI scanner, modified from a 3T Magnetom Skyra (Siemens Healthcare, Erlangen, Germany), and a 32-channel receive-only head coil (same vendor).

5.2.1 Dataset description

The dataset contains 123 cognitively normal participants (mean age 33.1 years, standard deviation 10.4 years, age range 19-68 years). The dataset age distribution is illustrated in Figure 5.1. As is shown in Figure 5.1, our dataset has a skewed age distribution towards young ages whereas it has a relatively wide age range.

All participants were required to meet MR safety requirements for scanning. Exclusion criteria included diagnosis of any heart or breathing problem; high blood pressure; nerve issues, including carpal tunnel syndrome and nerve damage; history of stroke, brain tumour or brain injury; dizziness, palpitations or fainting; diabetes; current or previous diagnosis of psychiatric condition; use of medication known to alter breathing, blood pressure or mood; pregnancy or breast-feeding; heavy use of tobacco; frequent migraines; epilepsy; and history of concussion resulting in loss of consciousness. The study was approved by the Cardiff University School of Psychology Research Ethics Committee. All participants gave informed written consent.

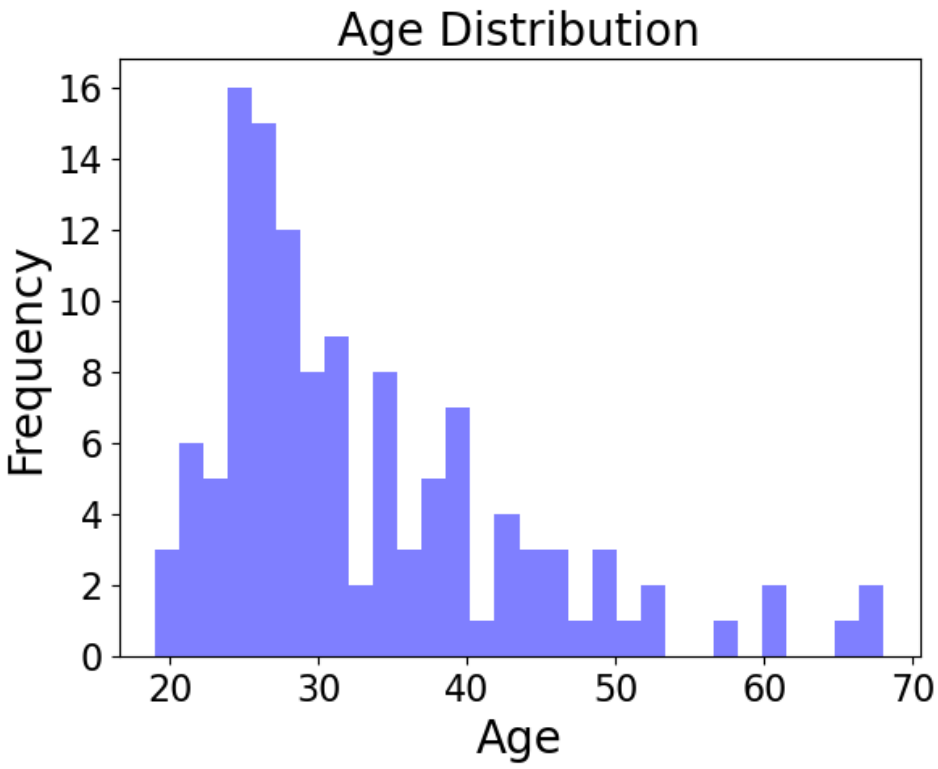


Figure 5.1: Age distribution of our dataset

5.2.2 Dataset acquisition

5.2.2.1 T1-weighted structural images

T1-weighted images were acquired using a magnetization-prepared 180-degree radio-frequency pulse and rapid gradient-echo (MPRAGE) sequence. Acquisition parameters were as follows: repetition time (TR) 2300 ms, echo time (TE) 2 ms, field of view (FOV) $256 \times 256 \times 192$ mm, matrix size $256 \times 256 \times 192$, voxel size $1 \times 1 \times 1$ mm, flip angle 9°, inversion time (TI) 857 ms, in-plane acceleration (GRAPPA) factor 2 and phase-encoding direction anterior-to-posterior (AP).

5.2.2.2 Diffusion MRI

Multi-shell diffusion-weighted MRI data were acquired using protocols described by Koller et al. [146] for the Microstructural Image Compilation with Repeated Acquisitions (MICRA) dataset. Data were acquired using a single-shot spin-echo, echo-planar imaging sequence. It was acquired in both anterior-to-posterior (AP) and posterior-to-anterior

(PA) phase-encoding directions. AP data comprised two shells of 20 diffusion encoding directions (uniformly distributed according to Jones et al. [147]) at b = 200 s/mm^2 and b = 500 s/mm^2, one shell of 30 directions at b = 1200 s/mm^2 and three shells of 61 directions each at b = 2400 s/mm^2, 4000 s/mm^2 and 6000 s/mm^2, in addition to two leading non-diffusion-weighted (b = 0 s/mm^2) images and 11 non-diffusion-weighted images dispersed throughout (33rd volume and every 20th volume thereafter). PA data comprised two leading non-diffusion-weighted images, one shell of 30 directions at b = 1200 s/mm^2, and a final non-diffusion-weighted image. Data acquisition details for all b-values were as follows: TR 3000 ms, TE 59 ms, FOV 220 × 200 mm in-plane, matrix size 110 × 110 × 66, voxel size 2 × 2 × 2 mm, with in-plane acceleration (GRAPPA) factor 2. Diffusion gradient duration and separation were 7 ms and 24 ms, respectively.

5.2.2.3 Multi-component relaxometry data

Multi-component relaxometry data were acquired using sequences implementing the McDESPOT protocol [148], including a T1-weighted 3D spoiled gradient-recalled echo sequence (SPGR) image, an inversion recovery (IR)-prepped spoiled gradient-recalled echo sequence (SPGR-IR) image and a steady-state free precession (SSFP) image. Data acquisition details were as follows: for SPGR: TR 4 ms, TE 1.9 ms, 8 flip angles (3, 4, 5, 6, 7, 9, 13, and 18°); for SPGR-IR: TR 4 ms, TE 1.9 ms, flip angle 5°, full k-space acquisition in PE and slice directions; for SSFP: TR 4.54 ms, TE 2.27 ms, 8 flip angles (10, 13.33, 16.67, 20, 23.33, 30, 43.33 and 60°). For all multi-component relaxometry data, the phase-encoding direction was AP, FOV 220 × 220 × 179 mm, matrix size 128 × 128 × 104 and voxel size 1.72 × 1.72 × 1.72 mm isotropic.

5.2.3 Structural image preprocessing

The raw structural T1-weighted MR images experienced a series of preprocessing steps before being used in the structural image stream.

Skull-stripping was first applied to the raw images to remove the non-brain tissues to delete unnecessary information for brain age prediction, using an open-sourced software,

named deepbrain [2]. Afterwards, we performed rigid registration to register the extracted brains to the MNI152 space using the linear registration tool (flirt) [149] from the FMRIB Software Library (FSL). Then, the image intensities were z-score normalized based on the individual mean and standard deviation using the *fslmath* command. Lastly, we removed the black background to reduce the overall image size to $153 \times 192 \times 161$.

5.2.4 Microstructure measures data preprocessing

5.2.4.1 Diffusion MRI preprocessing

Multi-shell diffusion-weighted MRI data were corrected for thermal noise, signal drift, susceptibility-induced distortions, motion and eddy current-induced distortions, gradient non-linearity, and Gibbs ringing artefacts.

Using the FSL brain extraction tool [140], the first non-diffusion weighted image from each phase-encoding direction was masked to exclude all non-brain data. Diffusion MRI noise level estimation and denoising was performed using a Marchenko-Pastur principal component analysis (MP-PCA)-based approach [150]–[152] in MRtrix3 [153]. Next, within-image intensity drift was corrected by fitting the diffusion-weighted MRI data to temporally interspersed b0 (non-diffusion-weighted) images. Slicewise OutLIer Detection (SOLID) [154] was then applied, employing lower and upper thresholds of 3.5 and 10, respectively, using a modified Z-score and a variance-based intensity metric. The susceptibility-induced off-resonance field was estimated from the non-diffusion-encoded data collected in AP and PA phase-encoding directions using FSL's topup [155], [156] and corrected, along with eddy current-induced distortions and subject movements, using FSL's eddy tool [157]. In-house code was then used to correct for gradient non-linearity distortions in MATLAB. Lastly, Gibbs ringing correction was performed using the local subvoxel-shifts method [158].

The fibre orientation distribution function was calculated in MRtrix3 [153] using multi-shell multi-tissue constrained spherical deconvolution [159]. These data were then used for

[2] https://github.com/iitzco/deepbrain

white matter bundle segmentation (tractography) and composite hindered and restricted model of diffusion (CHARMED) [160] analysis.

CHARMED [160] analysis was conducted using in-house code, run in MATLAB. First, a standard CHARMED model was fit using additional data gathered from the multi-shell multi-tissue constrained spherical deconvolution fiber orientation distribution to select the number of compartments in each voxel and constrain their orientation. Parameters initialized during this step were then used to generate restricted fraction maps, representing the total amount of restricted diffusion within each voxel.

5.2.4.2 Tract segmentation

White matter tracts were reconstructed from constrained spherical deconvolution peaks using a combined tract segmentation and orientation mapping approach [161], [162], available in TractSeg [163]. We reconstructed 10 major white matter tracts (consisting of 29 individual fibre bundles), including: Arcuate fasciculus (left and right), corpus callosum (rostrum, genu, rostral body, anterior midbody, posterior midbody, isthmus, and splenium), cingulum bundle (left and right), corticospinal tract (left and right), inferior fronto-occipital fasciculus (left and right), inferior longitudinal fasciculus (left and right), superior longitudinal fasciculus (sections I, II and III left and right), uncinate fasciculus (left and right), fornix (left and right), and thalamocortical (prefrontal) tract (left and right).

5.2.4.3 Multi-component relaxometry data preprocessing

Preprocessing of multi-component relaxometry data included brain extraction of T1w MPRAGE data using HD-BET [164] followed by resampling of MPRAGE and masked brain-extracted data to SPGR-IR resolution using Advanced Normalization Tools [109]. SSFP and SPGR data were corrected for motion using the FSL motion correction tool (mcflirt) [165] and SPGR-IR data were registered to the motion-corrected SPGR data using FSL's linear registration tool (flirt) [149]. These data were then used to create B1, T1 and B0 maps using functions from QUantitative Imaging Tools (QUIT) [166]. Finally,

multi-component DESPOT (mcDESPOT) using a three-compartment model [148] was used to separate SPGR and SSFP signals into multiple discrete pools, representing myelin water and intra/extra-cellular water or cerebrospinal fluid (CSF).

5.2.5 Microstructure measures data summary

After the preprocessing of the diffusion and quantitative measures of the white matter microstructure mentioned in Section 5.2.4, we ended up with eight measures of white matter microstructure, namely myelin water fraction (MWF), intracellular volume fraction (ICVF), fractional anisotropy (FA), mean diffusivity (MD), radial diffusivity (RD), axial diffusivity (AD), total restricted fraction (FRtot), and kurtosis fractional anisotropy (KFA). To extract the values of microstructural measures for each tract, we choose the median value for the tract average, as this helps to ignore large outliers from the bundle ends. Therefore, we formulated the white matter microstructure measures dataset in a matrix of shape $123 \times 29 \times 8$. The first dimension represents the number of subjects, the second represents the number of extracted tracts, and the last one represents the number of measures (MWF, ICVF, FA, MD, RD, AD, FRtot, and KFA).

The structural MR image dataset and microstructure measures dataset mentioned above are private datasets for research purposes at the Cardiff University Brain Research Imaging Centre, Cardiff University. However, it should be stressed that the proposed tract-wise approach is generic and works with any diffusion data, as long as they can be formatted in the shape of (number of subjects $\times$ number of tracts $\times$ number of microstructural measures). There are no specific requirements for the acquisition parameters, such as a specific b value in diffusion MRI. The model performance may vary when using different data to predict brain age, but the proposed tract-wise approach workflow remains consistent for different data.

5.3 A two-stream brain age prediction framework

The two-stream brain age prediction framework contains two separate routes for diffusion MRI measures of white matter microstructure and 3D structural MR images. The proposed two-stream framework is illustrated in Figure 5.2.

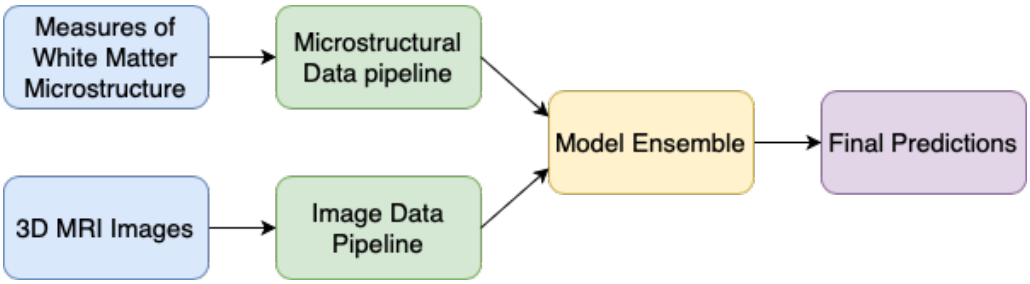


Figure 5.2: A two-stream brain age prediction framework using white matter microstructure measures and 3D structural MR images.

5.3.1 Microstructure measures pipeline

5.3.1.1 Overview

Due to distinct changes in the trajectory of different white matter tracts as age increases, we assume that some particular tracts may be more sensitive to ageing than others [167], [168]. Therefore, we propose a tract-wise training approach to enhance the model performance by prioritising more informative tracts.

The tract-wise training approach is formally presented in Algorithm 3. Some notations in Algorithm 3 are defined in Table 5.1.

The proposed tract-wise approach, illustrated in Algorithm 3 can be divided in three stages:

1. Dimension reduction: The whole microstructural dataset is divided into N_tract tract-specific sub-datasets, where the model α is fitted on each sub-dataset. This tract-wise model training reduces the number of features from $N\_tract \times N\_measures$ to $N\_measures$ for each model α.

Table 5.1: Notations for Tract-wise Training Approach

Notation	Type	Description
$\mathcal{D}$	constant	microstructural measures dataset
N_sub	constant	number of subjects
N_tract	constant	number of tracts
$N\_measures$	constant	number of microstructural measures
α	model	unfitted model
α_i	model	fitted model with index i
γ	variable	Mean Absolute Error
γ_i	variable	Mean Absolute Error with index i
N_top	variable	number of top-performing tracts
L	variable	an array to save fitted models
M	variable	an array to save mean absolute errors
N	variable	an array to save indices

Algorithm 3: A Tract-wise Training Approach

Given *Dataset* $\mathcal{D}$ *with shape* $N\_sub \times N\_tract \times N\_measures$*; Corresponding training set* $\mathcal{D}_{train}$*, validation set* $\mathcal{D}_{val}$*, test set* $\mathcal{D}_{test}$*; unfitted model* α

for $i \leftarrow 1$ **to** N_tract **do**
  Create tract-specific datasets $\mathcal{D}_{train\_i}$, $\mathcal{D}_{val\_i}$, and $\mathcal{D}_{test\_i}$ for tract with index i
  Train model α on $\mathcal{D}_{train\_i}$ and calculate MAE on $\mathcal{D}_{val\_i}$
  Save fitted model α_i in L and MAE γ_i in M
end
Select N_top lowest values from M and save their indices to N.
for $i \leftarrow 1$ **to** N_tract **do**
  if i **in** N **then**
    Test fitted model α_i on the tract-specific test set $\mathcal{D}_{test\_i}$
    Save predictions from model α_i for each subject in $\mathcal{D}_{test\_i}$
  end
end
Average all N_top predictions for each subject in $\mathcal{D}_{test\_i}$ to derive the final prediction.

2. Feature selection: After obtaining the fitted models on each sub-dataset, their performance is measured on the validation set. We then select N_top models with the smallest validation errors out of the N_tract fitted models, which are then adopted for testing.

3. Ensemble testing: The selected N_top models represent the prioritised tracts. In the test set, only the prioritised white matter tracts are adopted for the final prediction where their predictions are averaged to derive the final prediction.

We set the number of top-performing tracts to 3 to ensure a good balance between model accuracy and prediction stability. It is chosen based on our experiment testing. Generally speaking, selecting more tracts results in less variance in model performance since the final

prediction is the averaged prediction of all selected tracts. However, it may include some tracts that do not perform very well and result in decreased accuracy. And vice versa. Every tract has the chance to be prioritised by the different models using the tract-wise approach and tracts being more frequently selected indicate that they are more sensitive to ageing compared with other white matter tracts.

5.3.1.2 Pipeline details

Utilizing a tract-wise training approach transforms our dataset into a 2D matrix, allowing the application of traditional machine learning models.

To preprocess the data, we first employ the Robust Scaler [169] on each tract-specific dataset to normalize the value range within each dataset. The Robust Scaler removes the median and adjusts data based on the interquartile range (the first and third quartiles), helping the models converge faster. Afterwards, we apply the Principal Component Analysis (PCA) to the scaled data, selecting the top 6 principal components for dimensionality reduction. PCA serves not only as a dimensionality reduction technique, but it transforms different diffusion MRI measures into orthogonal features for age prediction as well. The choice of the number of principal components is based on the proportion of explained variance, and we choose 85% as the threshold. Furthermore, given our small dataset size and the skewed age distribution favouring young subjects, we perform data augmentation using the Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOGN) [170] to augment our training set, achieving a more balanced age distribution. SMOGN generates synthetic, fake samples based on the k-Nearest Neighbors (KNN) distance for regression problems, which is particularly helpful on imbalanced dataset. This data augmentation approach effectively doubles the size of our training set, which significantly stabilizes model predictions and improves model accuracy.

In terms of model selection, we opt for three distinct models, namely the Support Vector Regression (SVR) [112], XGBoost [111], and the stacking model. This model selection is based on the favourable balance between model performance and computational speed.

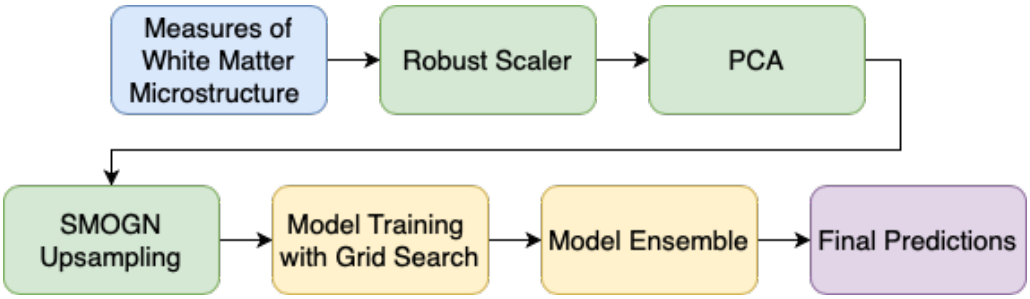


Figure 5.3: Detailed steps of the microstructure measures pipeline

The stacking model employs the SVR as its first layer model and stacks the XGBoost as the final model. Grid Search has also been applied to explore the optimal hyperparameter settings for all three models. To further enhance predictive capabilities, we apply the model ensemble by averaging predictions across SVR, XGBoost, and the stacking model to derive the final brain age estimation. The workflow of the microstructure measures pipeline is summarised in Figure 5.3.

5.3.2 Structural image pipeline

To emphasize the advantages of the proposed tract-wise approach on microstructure measures for brain age prediction, neural networks trained using 3D T1-weighted structural images are adopted as the contrast.

5.3.2.1 Overview

The structures of the structural MRI-based model pipeline adhere to standard procedures outlined in the literature [1], [33], [34], [42]. Neural networks take the preprocessed 3D structural T1-weighted MR images described in Section 5.2.3 as input, producing predictions for brain age as output. To further improve predictive accuracy, the model ensemble, described in Section 5.3.1, is adopted by averaging predictions across all model architectures to derive the final brain age estimation.

5.3.2.2 Models

Despite the vast variety of network architectures in the literature, we choose two renowned model architectures, namely SFCN [33] and DenseNet [34], in this chapter for their stabil-

ity and effectiveness. Model-dependent parameters, such as the number of filters in each layer, are adapted based on our dataset. The number of parameters is reduced below 1 million for each model.

The SFCN follows the identical architecture design from its original work [33]. It contains 6 consecutive blocks in total and number of filter for each convolution layer is set to 16. Each of the first 4 consecutive blocks consists of a $3 \times 3 \times 3$ 3D convolution layer, a Batch Norm layer, a Max Pooling layer and a ReLU activation. The 5 th block contains one $1 \times 1 \times 1$ 3D convolution layer, a Batch Norm layer and a ReLU activation. The 6 th block contains an average pooling layer, a dropout layer, an $1 \times 1 \times 1$ 3D convolution layer and a fully connected layer to derive the output. The DenseNet [34] also follows the identical architecture design from the MONAI package [171]. The initial number of filter is set to 16 and growth rate is set to 8. It contains four dense blocks, which have 4, 8, 12, and 10 dense layers respectively. A 3D adaptive average pooling layer is applied afterwards, followed by a flatten layer and a fully connected layer to derive the final prediction. The number of trainable parameters for each model architecture is demonstrated in Table 5.2. The workflow of the structural image pipeline is summarised in Figure 5.4.

Table 5.2: Number of trainable parameters

Model architecture	**Number of parameters**
SFCN [33]	471,301
DenseNet [34]	348,427

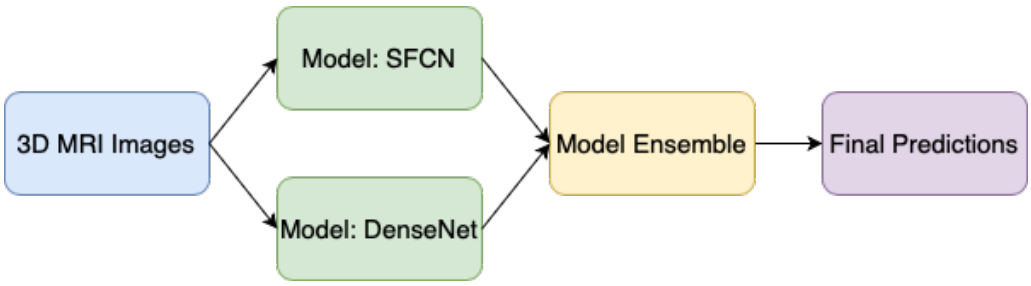


Figure 5.4: Detailed steps of the 3D structural image pipeline. The structural images are fed into model SFCN and DenseNet respectively, where predictions from both models are ensembled to derive the final predictions.

5.3.3 Training settings

The microstructure measures and structural image datasets are split using the same stratified split strategy, of which 80% is used for training, 10% for validation, and 10% for testing. The stratified split strategy is performed to maintain a similar age distribution across the train, validation and test set. To make full use of the dataset, we apply a split strategy similar to cross-validation. We randomly split each dataset in 20 different ways so that no two train/validation/test sets are identical. Moreover, to minimize the variances of the model predictions from both streams due to random initialization of the model weights, we train our models on each train/validation/test split 3 times. In that way, we end up training a specific model on a particular dataset 60 times.

When producing predictions for each participant, the two streams perform differently. In the microstructure measures pipeline, the predictions of the selected models trained on the top 3 more informative tracts are first averaged to produce the estimated brain age. In the structural image stream, the networks give the predicted brain age directly.

In terms of the network settings, we choose the Mean Absolute Error (MAE) as the loss function for all network models. The skewed loss function, demonstrated in Chapter 3 is not adopted considering the small size of our dataset. Introducing the skewed loss can debias the model predictions, whereas the small validation set does not provide a good representation of the true age-delta correlation across the whole dataset. During the training process, the Adam optimizer [128] is used for all models. The initial learning rate is set to 0.01 and is multiplied by 0.5 every 50 epochs. The batch size is set to 32 and the total number of epochs is set to 200.

Because the size of our dataset is relatively small, to prevent networks from overfitting, we apply data augmentation in the training process using TorchIO [127]. In each training iteration, every input image has a probability of 50% being flipped around the horizontal plane. The L2 weight decay coefficient is set to 0.001. It should be stressed that all hyperparameter settings, such as learning rate and batch size, mentioned above are chosen based on our experiment testing.

5.3.4 Testing settings

As we described in Section 5.3.3, each model is repeated 3 times, producing 3 brain age estimations, to prevent fluctuations in model performance. Therefore, we provide two ways to evaluate model accuracy, which will be used in Section 5.4.

On one hand, we can create an ensemble model by averaging all 3 predictions to give an averaged age prediction and MAE is calculated based on the averaged predictions for each dataset split. On the other hand, MAE can be calculated for each of the 3 predictions, resulting in 3 MAE values. These MAE values are then averaged to produce an average MAE for each dataset split. Two ways of calculating MAE discussed above are both reported in Section 5.4.

The illustration of the training and testing process on each dataset split is shown in Figure 5.5. The two streams share the same strategy.

The illustration of the training and testing process on each dataset split is shown in Figure 5.5. Figure 5.5 shows how the ensemble model predictions are calculated for each stream. Take the microstructural measures stream as an example. All models (SVR, XGBoost, and Stacking) in the microstructural measures stream are trained three times, generating three predictions from each model. The three predictions from the same model are first averaged and the resultant three averaged predictions from three different models (SVR, XGBoost, or Stacking) are further averaged to produce the final ensemble prediction. This strategy stays the same for structural images.

5.4 Results

The performance comparisons between the microstructure measures pipeline and structural image pipeline described in Section 5.3 are listed in the following subsections.

To save space, Table 5.3 lists some notations used in the following subsections.

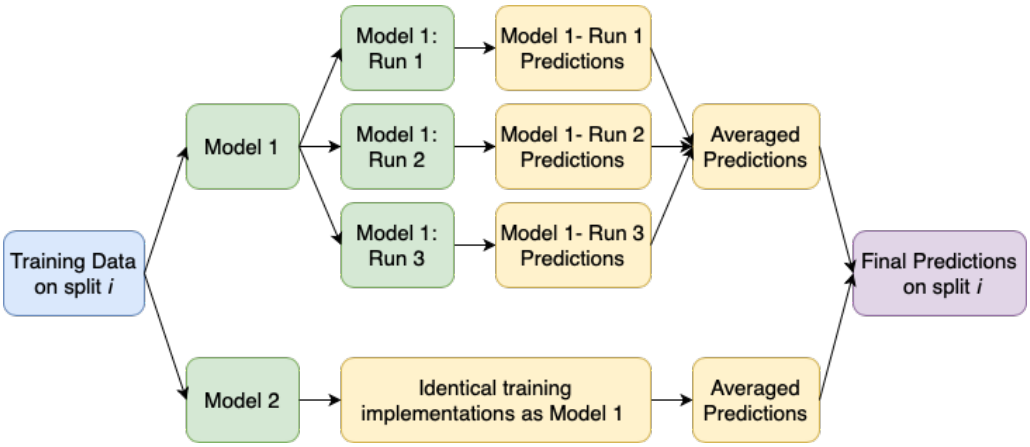


Figure 5.5: An illustration of the training and testing process on each dataset split. The two streams share the same strategy. For each split, each model from the same stream is repeated three times to prevent model performance fluctuations, producing three age estimations, which are then averaged to derive an estimation for the current model. Different models from the same stream are then averaged to produce the final estimation for the current data split. We only show two models here as an illustration.

Table 5.3: Notations for Performance Comparisons

Notations	Descriptions
Stacking	SVR-XGBoost stacking model
Averaged model	model-averaging of all trained models
Single Model Averaged MAE	Averaged MAE for single model across all splits
Ensemble Model MAE	MAE for averaged predictions across all splits

In Table 5.3, notations "Single Model Averaged MAE" and "Ensemble Model MAE" represent the two ways of calculating MAE, discussed in Section 5.3.4.

5.4.1 Performance of microstructure measures pipeline

From Section 5.3.3, we split the microstructure measures dataset using 20 different splits, and for each split, we train each model 3 times. The model performance table of the microstructure measures pipeline is demonstrated in Table 5.4.

As shown in Table 5.4, the ensemble model (model averaging across all three repeated runs) outperforms the single model for each model architecture, on accuracy and stability.

Table 5.4: Model Performances of the microstructure measures pipeline (Unit: years)

Single / Ensemble	Model	Mean	Std	Min	Max
Single Model Averaged MAE	SVR	3.61	0.62	2.42	5.02
	XGBoost	3.88	0.67	2.71	5.74
	Stacking	3.85	0.77	2.53	6.23
	Averaged model	**3.23**	0.58	2.00	4.63
Ensemble Model MAE	SVR	3.47	0.52	2.47	4.46
	XGBoost	3.42	0.49	2.54	4.71
	Stacking	3.20	0.73	2.46	5.68
	Averaged model	**3.10**	0.50	2.04	4.13

Also, the averaged model outperforms any single model by averaging different model architectures.

Therefore, we can create an ensemble model by averaging all model architectures and all three repeated runs, which achieves a mean MAE of 3.10 years for the microstructure measures pipeline.

5.4.2 Performance of structural image pipeline

The same MAE calculation used in Section 5.4.1 can also be performed on the structural image pipeline. Therefore, we also end up with 20 averaged MAE results for each neural network and 20 MAE results for each ensemble model (one split corresponds to one MAE result). The model performance table of the structural image pipeline is demonstrated in Table 5.5.

Table 5.5: Model Performances of structural image pipeline (Unit: years)

Single / Ensemble	Model	Mean	Std	Min	Max
Single Model Averaged MAE	SFCN [33]	4.01	0.87	2.7	6.37
	DenseNet [34]	3.68	0.89	2.15	6.01
	Averaged model	**3.55**	0.80	2.28	5.53
Ensemble Model MAE	SFCN [33]	3.71	0.72	2.51	5.15
	DenseNet [34]	3.45	0.93	2.02	5.51
	Averaged model	**3.34**	0.75	2.12	4.79

From Table 5.5, similar conclusions can be drawn. DenseNet shows better accuracy compared with the SFCN model. However, the difference is not significant. The ensemble model shows increased accuracy due to the averaged prediction from the three repeated runs and the averaged model still outperforms any single model architecture.

Therefore, we also create an ensemble model by averaging all model architectures and all three repeated runs, which achieves a mean MAE of 3.34 years for structural MR images.

5.4.3 Performance of the two-stream framework

From Section 5.4.1 and Section 5.4.2, it is straightforward that measures of the white matter microstructure also show predictive ability in brain age prediction. The microstructure measures pipeline has achieved similar performance compared with the structural image pipeline.

As we previously discussed in Section 5.1, white matter microstructure measures and structural images contain distinct information about brain ageing and their information complements each other, which can be combined to create another ensemble model. Therefore, another straightforward ensemble model is built by averaging the model predictions from both pipelines for each subject and each dataset split. Despite adopting a conventional fusing method (model averaging), the ensemble model has achieved an improved accuracy compared with any single pipeline. More advanced fusing methods are left as future works and will be evaluated in the future.

By combining the ensemble model performance from both pipelines, the performance table of the two-stream framework is illustrated in Table 5.6. The significance table is illustrated in Table 5.7.

Table 5.6: Ensemble Model Performances of all pipelines (Unit: years)

Pipeline	**Mean**	**Std**	**Min**	**Max**	$\mathbb{R}^2$
Microstructure measures pipeline	3.10	0.50	2.04	4.13	0.93
Structural image pipeline	3.34	0.75	2.12	4.79	0.92
Two-stream framework	**3.00**	0.49	2.08	3.87	0.93

Table 5.7: Significance test of pipeline performance with p-values

Pipelines	**Paired-t**
Microstructure Measures & Structural Images	**<0.01**
Microstructure Measures & Two-stream framework	0.15
Structural Images & Two-stream framework	**<0.01**

From Table 5.6, it can be demonstrated that the two-stream framework achieves improved accuracy in predicting brain age compared with any single data modality. From Table 5.7, by introducing information from the measures of white matter microstructure into the prediction framework, we significantly enhance the accuracy of predicting brain age compared with commonly adopted image-based models, resulting in more reliable and improved outcomes.

5.4.4 Tracts of interest

The adoption of the tract-wise approach highlights the more informative tracts for predicting brain age. A histogram of the frequency of selected tracts is illustrated in Figure 5.6. The tract frequency represents how many times is a particular tract selected for each model and each dataset split. From Figure 5.6, there are significant differences in the frequency of being selected among all white matter tracts, supporting our assumption that some particular tracts experience more age-related alterations in the ageing process than other tracts. For each model architecture and each model's repeated run, different tracts are selected. It can be observed that all white matter tracts are adopted in predicting brain age, providing a more complete picture of the brain microstructure. If feature selection is performed before model training [19], [118], [145], some tracts with a lower frequency, shown in Figure 5.6, may be removed and result in a decreased performance.

The tracts that the tract-wise approach identified as most informative are the genu, rostral body, and anterior midbody of the corpus callosum, the left fornix, and the left thalamo-prefrontal tract. These tracts have been reported in the literature to be some of the latest to complete their maturation through myelination, which can extend to the third or fourth decade of life, and to be some of the first ones to be affected by ageing.

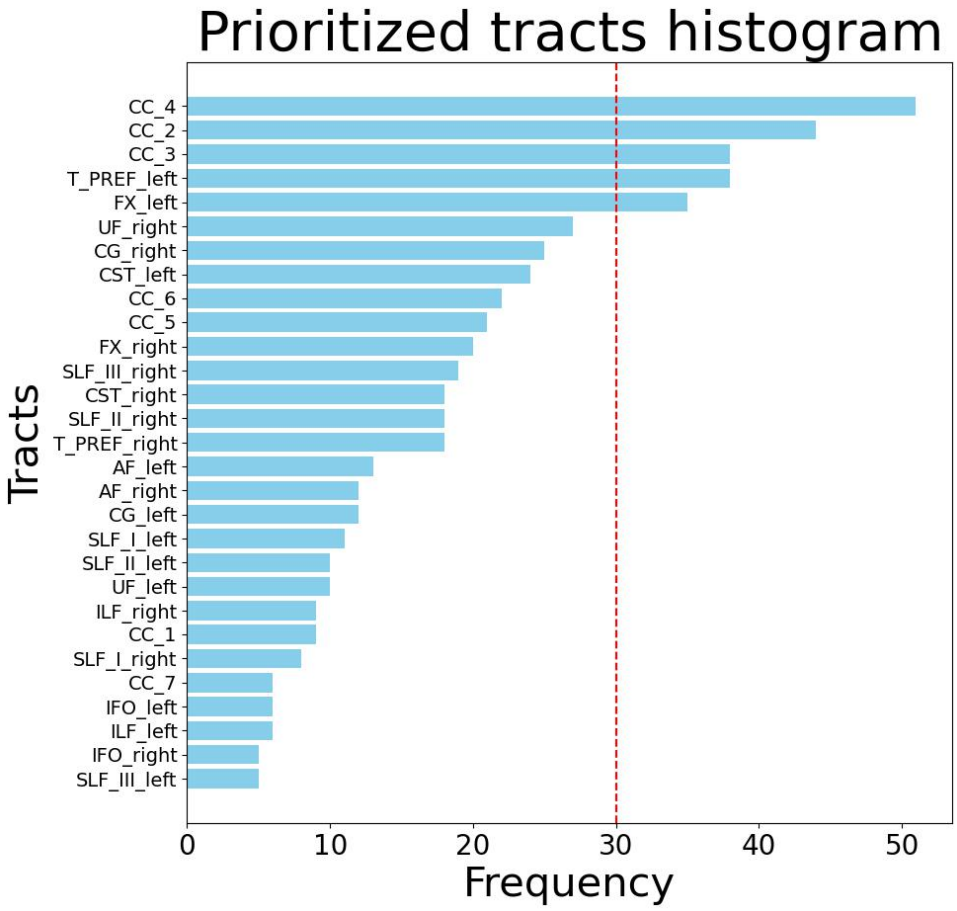


Figure 5.6: Frequency of prioritised tracts for all models.

The anterior section of the corpus callosum is more susceptible to brain ageing compared to the posterior section [172] with a steeper age decline than other tracts [103], [173], [174]. Also, selective vulnerability to ageing has been reported in late-myelinating tracts such as the genu of the corpus callosum and the fornix [175], [176]. Finally, the synaptic connectivity of thalamo-prefrontal tracts increases into at least the late teen years, with that of the left tract exhibiting a steeper increase with age than the right one [177].

5.4.5 Performance comparisons with previous studies

In the literature, the accuracy of brain age prediction models varies significantly, depending on various model architectures and datasets. The proposed image pipeline and microstructure measures pipeline both achieve a good trade-off between model accuracy and training complexity compared with previous studies.

5.4.5.1 Structural image-based models

Structural image-based brain age prediction has achieved satisfactory accuracy due to the increasing size of the dataset. Wood *et al.* used a training sample of 12,949 participants [34], which is more than two orders of magnitude than ours, to achieve an MAE of 2.1 years. Peng *et al.* achieved an MAE of around three years using a few thousand images [33], which is still more than one order of magnitude than ours. However, the main reason for achieving such high accuracy is not the adoption of more advanced models but the huge size of the dataset. Using a huge dataset not only entails a much higher cost in terms of participants and scanning time but also increases the model training complexity. Our image pipeline is generic and therefore also benefits from the adoption of larger datasets. In our image pipeline, we extensively applied model ensemble for each model architecture and each model's repeated run to aggregate average-performing models into a top-performing model, achieving an MAE of 3.34 years on average.

5.4.5.2 Microstructure measures-based models

Compared with models using structural images, models using microstructural measures for brain age prediction require significantly fewer samples. As the microstructural measures reveal age-related alterations of the brain, the selection of more informative measures becomes crucial for predicting brain age. Beck *et al.* [30] adopted a sample size of 702 and achieved MAEs between 6.99 and 9.82. Their performance is poor as they adopted each microstructure measure separately, rather than combining all measures for prediction. Cole [17] adopted a dataset with more than 2000 participants and applied LASSO to reach an MAE of 3.879 for diffusion MRI data. However, feature selection using LASSO fails to provide any complete information about which tract or measure performs better compared with other tracts or measures. Xiong *et al.* [19] trained several models on a dataset with 14,000 participants without any feature selections and achieved similar performance as ours. However, the participants' age of their dataset is older compared with ours, which makes it easier to predict brain age, as the brain alterations due to ageing become more significant compared with middle-aged participants. Cai *et al.*

[145] adopted reinforcement Q-Learning to select important features and had a dataset of 212 participants. However, similar to Xiong's work [19], their model focused on age prediction for children and adolescents, whose age range is limited.

Compared with previous studies mentioned above, our dataset focuses on middle-aged participants. The proposed tract-wise approach performs automatic feature selection based on model performances on all white matter tracts, which indicates more aging-sensitive tracts. Also, we performed another model ensemble for the selected tracts when making predictions which further improves the model accuracy to achieve an MAE of 3.10 years.

5.5 Discussion

We have proposed a tract-wise approach for accurate brain age prediction using diffusion measures of white matter microstructure. The tract-wise approach identifies more informative white matter tracts in the model training process and prioritises those tracts to produce the final brain age estimation. Every tract has the chance of being selected under different model configurations. Therefore, employing a model ensemble to further aggregate all models makes sure that information in all tracts is considered to produce the final prediction.

We have also shown that diffusion measures of white matter microstructure and structural T1-weighted images complement each other in brain age prediction. While T1-weighted imaging provides insight into the gross anatomical structure of the brain, we can obtain additional complementary information about the microstructural properties of the white matter using quantitative mapping techniques. By combining information from different aspects, the proposed two-stream brain age prediction framework obtains a more complete picture of brain ageing, resulting in improved predictive accuracy. In the literature, multi-modal models have also been applied to improve the performance, but they tend to adopt a specific type of data, such as using multi-modal images [33], [119] or using multi-modal diffusion MRI measures [17], [19], [30], [110]. Also, they tend to apply the same model

architectures for all data modalities. In contrast, the proposed two-stream framework combines more distinct data modalities and applies more diverse model architectures for age prediction compared with previous research.

We have also demonstrated the effectiveness of our microstructure measures pipeline using a dataset with a small sample size but a relatively wide age range. PCA transforms the diffusion MRI measures and quantitative MRI measures into a set of orthogonal features, avoiding the collinearity issues [113]. SMOGN data augmentation further generates synthetic samples, doubling the size of our training set, which significantly enhances the model accuracy and stability [170]. The microstructure measures pipeline achieves an MAE of 3.10 years, which is satisfactory compared with other diffusion MRI-based studies [17], [19], [30], [110], [118], [178], given the relatively small sample size.

We have also illustrated that extensively using model ensemble significantly boosts model accuracy. In this chapter, we apply the model ensemble for each model architecture, each model's repeated run, and each data modality. From Table 5.5 and Table 5.4, we can easily observe that the ensemble model significantly outperforms the single model in terms of accuracy and stability.

There are also some avenues for future work. Firstly, in the microstructure measures pipeline, we build a multi-modal ensemble model, which adopts diffusion MR measures and quantitative MR measures. However, in the structural image pipeline, we only use structural T1-weighted images as input. Including other image modalities, such as T2-weighted images, may be beneficial to improve the model stability. Secondly, we apply model averaging to fuse model predictions from different data modalities. More advanced model fusion methods will be evaluated to further boost the model performance.

5.6 Summary

To conclude, we have proposed a tract-wise approach by prioritising ageing-sensitive white matter tracts for brain age prediction. By selecting more informative tracts and discarding less useful ones in predicting brain age, our model achieves an accurate age

prediction using a relatively small sample size. The tract-wise approach makes it possible to evaluate each tract separately and more frequently selected tracts indicate more significant alterations in the microstructural measures with advancing age. The most frequently selected tracts are the genu, rostral body, and anterior midbody of the corpus callosum, which have also been reported in the literature. Also, by fusion model predictions from microstructure measures and structural images, the two-stream brain age prediction framework achieves an improved predictive accuracy compared with any single data modality.

Chapter 6

Conclusion and Future Work

This thesis presented a substantial body of work in the area of brain age prediction. This chapter provides a concluding summary of each chapter, the main contributions of the thesis as well as the proposed future work.

6.1 Brain age prediction revisit

Brain age prediction aims to evaluate the biological process of brain ageing, which shows potential in both research field and clinical applications.

In terms of research perspective, brain age prediction serves as a simplified approach to evaluate the brain ageing process. It serves as a tool to evaluate the effectiveness of some interventions in clinical trials of age-associated neurodegenerative diseases or aimed at promoting healthy brain ageing. By tracking the changes of the predicted brain age over time, the impact of lifestyle changes, pharmacological treatments or cognitive training programs on brain ageing can be evaluated [3]. In terms of clinical applications, brain age prediction could be used for early prevention and detection of neurological aspect of ageing. Using brain age as a screening tool could facilitate early detection of disorders or even their pre-clinical stages, which, in turn, would allow early intervention. Early intervention in age-related disorders is an important clinical focus. There are many studies demonstrating that people with Alzheimer's disease tend to have large positive brain age delta whereas in its earlier stage, such as mild cognitive decline, they tend to have a smaller brain age delta [27]. In clinical applications, if a patient is observed to have a relatively large positive delta, it will allow early detection of brain related diseases, before it becomes serious.

The algorithms demonstrated in Chapter 3, Chapter 4, and Chapter 5 aim to refine the brain age prediction framework in different aspects, mitigating its gap between research and clinical use. The skewed loss function demonstrated in Chapter 3 correct the regression bias in brain age prediction, which removes the effect of chronological age when evaluating the relationship between brain age delta and any cognitive diseases. The integrated brain age prediction framework demonstrated in Chapter 4 makes it possible to predict brain age in clinical settings, which improves the generalisation ability of brain age prediction, making it more easily accessible in the hospitals. The multi-modal learning approach demonstrate in Chapter 5 focuses on the predictive accuracy of brain age prediction models. By adopting multiple data modalities in predicting brain age, the brain age prediction model evaluates the brain biological ageing process more accurately.

6.2 Concluding summary

The neurological aspect of ageing is biologically complex and can be affected by various internal or external factors, such as brain-associated diseases or brain injury from accidents, resulting in accelerated or delayed brain ageing. People with accelerated brain ageing experience a higher risk of neurological diseases, as the brain appears older than healthy people with the same chronological age. Therefore, brain age prediction tries to quantify the actual brain ageing process, which provides a personalised biomarker indicating brain health. The scope of this thesis is focused on the development of novel algorithms that make brain age prediction models more reliable in research analysis, more practical in clinical diagnosis and more accurate in measuring brain ageing process.

Chapter 2 presented a thorough literature review on brain age prediction from four perspectives. Firstly, the development of brain age prediction models over the past few years is summarised and the clinical significance and potential of brain age are illustrated. Then, the commonly observed regression bias in brain age prediction models is explained and commonly adopted bias correction approaches are summarised. Afterwards, the recent advances in brain MRI super-resolution techniques are reviewed, which

demonstrates the potential to generalise brain age prediction models for clinical-grade MR images. Lastly, the MR imaging modalities for predicting brain age are studied and multi-modal brain age prediction models are summarised.

Chapter 3 presented algorithms for debiasing model predictions to make the brain age prediction model more reliable by producing an unbiased estimation of brain age. A skewed loss function is proposed to eliminate the regression bias by assigning different loss values based on whether the model over-predicts or under-predicts a participant. The proposed skewed loss replaces the conventional symmetric loss function used in the model training phase and the model can produce an unbiased estimation of brain age at the end of the training process. A dynamic training algorithm is further proposed for the skewed loss, which searches for the optimal hyperparameters using an iterative approach. By combining the skewed loss function with the dynamic training strategy, the regression bias can be eliminated at the end of the model training process, which has been proven to be robust to different datasets, data distributions, model architectures, and even problem domains.

Chapter 4 generalised brain age prediction models for low-resolution images, which makes brain age prediction a more practical and accessible application for clinical diagnosis. A straightforward and effective non-uniform sampling algorithm is proposed by prioritizing models with more high-frequency information, which improves the reconstruction quality of the cerebellum for brain MRI super-resolution. By comparing the proposed sampling approach with the conventional sampling approach, our non-uniform sampling demonstrates indistinguishable performance on the whole brain volume. In contrast, it improves significantly on the cerebellum. A refined brain age prediction workflow is then proposed by incorporating super-resolution into brain age prediction models. Given a pre-trained brain age prediction model, the low-resolution MR images are first super-resolved and then adopted to produce the predicted brain age. Performance comparison between authentic high-resolution testing images and super-resolved testing images is conducted and insignificant differences are observed. It opens the window for predicting brain age in clinical diagnosis, where high-resolution MR images are not always accessible, making

brain age prediction a more practical and accessible tool in practice.

Chapter 5 demonstrated a multi-modal approach for predicting brain age using macro- and microstructural brain MR imaging, which produces a more accurate estimation of brain age by providing a more complete picture of the brain ageing process. A novel tract-wise feature selection approach is proposed to prioritize more informative tracts for predicting brain age, providing insights into more ageing-sensitive white matter tracts. Unlike some other approaches that perform feature selection before model training, the tract-wise approach is performed in the model training process, which preserves the most complete information, keeping all white matter tracts in consideration. In addition, a performance comparison between models using structural images and models using microstructural measures is carried out. A two-stream framework is then proposed by combining information from structural images and white matter microstructure measures, which provides the model with a more complete picture of the brain, achieving an improved accuracy compared with any single data modality, which demonstrates the strength of adopting multi-modal models for predicting brain age, instead of uni-modal models.

6.3 Contributions of the thesis

This thesis has contributed one journal paper [104] and one conference paper with the second journal paper under review. The main contributions of this thesis include:

- To eliminate the regression bias in brain age prediction models to make brain age a reliable biomarker, a skewed loss function and a dynamic training algorithm are proposed for deep learning models. The proposed skewed loss tackles the observed bias by assigning different loss values based on whether the model overestimates or underestimates the participant's age. It implicitly makes the model produce fewer overestimations for young individuals and fewer underestimations for elderly participants. The proposed approach is the first work that corrects the bias in an integrated approach for deep learning models. Compared with the widely-adopted

two-stage correction approach [10], [39], [63], our approach achieves an improved accuracy in predicting brain age. At the end of the training phase, model prediction is always unbiased, without the need to perform a second stage correction. Our approach has been validated on two public neuroimaging datasets and is robust to different model architectures and correlation metrics. Apart from the brain age prediction problem, our approach can also generalise to different problem domains (Presented in Chapter 3).

- To generalise brain age prediction in the clinical environment to make brain age prediction a more practical tool for diagnosis, a modified brain age prediction workflow is demonstrated by incorporating the super-resolution techniques. The clinical-grade MR images are first super-resolved before being fed into the brain age prediction models to predict the age. Adopting image super-resolution effectively mitigates the image resolution differences between clinical and research-level images, of which the latter are used to train the brain age prediction model. The super-resolved MR images achieve indistinguishable accuracy compared with original high-resolution MR images, which makes it feasible for clinical-level brain age prediction. An effective non-sampling approach is also proposed for improving the reconstruction quality of the cerebellum for super-resolution tasks. This algorithm is the first work that treats different brain regions separately, depending on their structural differences, to improve the super-resolution quality of brain MR images. The proposed algorithm has been evaluated using CNN-based and GAN-based models on the publicly available HCP dataset (Presented in Chapter 4).

- To optimise the accuracy of the brain age prediction model, a multi-modal approach is proposed by fusing the information from structural MR images and microstructural MR measures, to predict brain age. We propose a tract-wise approach for predicting brain age using brain microstructural MR measures which demonstrates superior accuracy using a small sample dataset. The tract-wise approach prioritises more informative tracts of the brain in the model training process, providing insights into more ageing-sensitive white matter tracts. Unlike some schemes that perform

feature selection before model training, the tract-wise approach preserves the most complete information, keeping all white matter tracts in consideration. Adopting a tract-wise approach decomposes the 3D MR measures dataset into multiple 2D datasets, which reduces the dimensionality issues and makes model ensemble from different datasets a straightforward approach to predict brain age. Our approach has been validated on a brain microstructural measures dataset collected at Cardiff University and applies to other datasets as well (Presented in Chapter 5).

6.4 Future works

In this thesis, we proposed algorithms for improving brain age prediction models from three perspectives, each dedicated to a specific chapter, to make brain age prediction models more reliable, practical, and accurate. Despite those approaches, there is sufficient scope for future work:

- **More advanced model debiasing approaches**

  Model debiasing plays a crucial role in brain age prediction models since a biased estimation of brain age weakens its validity as the biomarker. Firstly, it would be crucial to thoroughly investigate under which circumstances, such as the expected size of the dataset and expected model complexity of the prediction models, the observed age-associated bias can become negligible to the predicted brain age, i.e., age delta correlation falls between -0.1 and +0.1. The main motivation behind it is that all correction approaches proposed so far [23], [39], [40], [104] sacrifice model accuracy for the unbiased predictions and correction approaches could significantly decrease the model performance. In that way, the advantages of using more advanced models can be weakened after the corrections. In the literature, there have been some studies that produce minor bias. For instance, Wood et al. [34] produced a correlation of -0.18 between brain age delta and chronological age. It is still unknown what is the expected model complexity for the brain age prediction problem. Secondly, more accurate bias correction approaches are needed. The two-

stage approach and integrated approach both focus on the group-level corrections and there is no guarantee that a near-zero ADC still holds in any subset age groups. Zhang et al. [40] proposed an age-level correction approach that tries to answer this problem, whereas their method requires the true target label, which validates the principles of the predictive framework.

- **Incorporating models from other brain-imaging fields**

  As for incorporating models from other brain-related fields to facilitate brain age prediction, brain MR super-resolution provides a more affordable way to train a brain age prediction model using low-resolution images. It also makes it feasible to predict brain age from low-resolution images, given a pre-trained brain age prediction. Some other areas may be included as well, such as brain segmentation and brain abnormality detection.

  Brain segmentation separates the whole brain volume into different regions and some weighting matrix map can be built to highlight the importance of different regions. For example, the prefrontal cortex and hippocampus are believed to be affected more by ageing, whereas the occipital shows the least influence in the ageing process [179], [180]. Therefore, the locations of the prefrontal cortex and hippocampus on the weighting map should have larger weights than the occipital locations. Such an explicit weighting map makes the model pay greater attention to more valuable and informative regions, which may improve the performance.

  Brain abnormality detection and classification may also help the model predict based on different abnormalities. Brain diseases, such as brain tumours and Alzheimer's disease, tend to result in positive brain age delta. However, different diseases may affect brain age at different levels. Combining information on different diseases might also help in brain age prediction.

- **Towards true multi-modal learning**

  Speaking of multi-modal model learning, most studies that use different data modal-

ities in brain age prediction tend to train models using a specific data modality and aggregate different models afterwards [17], [19], [30], [33]. The performance improvement after adopting such an approach comes from the use of multi-modal data as well as model ensemble. In terms of training a single model, there is no difference among different data modalities and the model can not learn any information from other data modalities. Therefore, it would be interesting to build new model architectures that learn from different modalities of data together in the model training process, i.e., a model with multiple inputs, which might improve the model performance as well.

Bibliography

[1] J. H. Cole, R. P. Poudel, D. Tsagkrasoulis, *et al.*, "Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker," *NeuroImage*, vol. 163, pp. 115–124, 2017.

[2] C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer, "The hallmarks of aging," *Cell*, vol. 153, no. 6, pp. 1194–1217, 2013.

[3] C. Gaser, P. Kalc, and J. H. Cole, "A perspective on brain-age estimation and its clinical promise," *Nature computational science*, pp. 1–8, 2024.

[4] J. H. Cole and K. Franke, "Predicting age using neuroimaging: Innovative brain ageing biomarkers," *Trends in neurosciences*, vol. 40, no. 12, pp. 681–690, 2017.

[5] K. Franke, G. Ziegler, S. Klöppel, C. Gaser, A. D. N. Initiative, *et al.*, "Estimating the age of healthy subjects from t1-weighted mri scans using kernel methods: Exploring the influence of various parameters," *Neuroimage*, vol. 50, no. 3, pp. 883–892, 2010.

[6] L. Baecker, R. Garcia-Dias, S. Vieira, C. Scarpazza, and A. Mechelli, "Machine learning for brain age prediction: Introduction to methods and clinical applications," *EBioMedicine*, vol. 72, 2021.

[7] B. Mohajer, N. Abbasi, E. Mohammadi, *et al.*, "Gray matter volume and estimated brain age gap are not linked with sleep-disordered breathing," *Human brain mapping*, vol. 41, no. 11, pp. 3034–3044, 2020.

[8] K. Franke and C. Gaser, "Longitudinal changes in individual brainage in healthy aging, mild cognitive impairment, and alzheimer's disease," *GeroPsych*, 2012.

[9] M. Ly, Z. Y. Gary, H. T. Karim, *et al.*, "Improving brain age prediction models: Incorporation of amyloid status in alzheimer's disease," *Neurobiology of aging*, vol. 87, pp. 44–48, 2020.

[10] J. H. Cole, S. J. Ritchie, M. E. Bastin, *et al.*, "Brain age predicts mortality," *Molecular psychiatry*, vol. 23, no. 5, pp. 1385–1392, 2018.

[11] K. Franke, C. Gaser, B. Manor, and V. Novak, "Advanced brainage in older adults with type 2 diabetes mellitus," *Frontiers in aging neuroscience*, vol. 5, p. 90, 2013.

[12] K. Franke, M. Ristow, C. Gaser, and A. D. N. Initiative, "Gender-specific impact of personal health parameters on individual brain aging in cognitively unimpaired elderly subjects," *Frontiers in aging neuroscience*, vol. 6, p. 94, 2014.

[13] E. Luders, N. Cherbuin, and C. Gaser, "Estimating brain age using high-resolution pattern recognition: Younger brains in long-term meditation practitioners," *Neuroimage*, vol. 134, pp. 508–513, 2016.

[14] L. Rogenmoser, J. Kernbach, G. Schlaug, and C. Gaser, "Keeping brains young with making music," *Brain Structure and Function*, vol. 223, pp. 297–305, 2018.

[15] J. Steffener, C. Habeck, D. O'Shea, Q. Razlighi, L. Bherer, and Y. Stern, "Differences between chronological and brain age are related to education and self-reported physical activity," *Neurobiology of aging*, vol. 40, pp. 138–144, 2016.

[16] L. C. Löwe, C. Gaser, K. Franke, and A. D. N. Initiative, "The effect of the apoe genotype on individual brainage in normal aging, mild cognitive impairment, and alzheimer's disease," *PloS one*, vol. 11, no. 7, e0157514, 2016.

[17] J. H. Cole, "Multimodality neuroimaging brain-age in uk biobank: Relationship to biomedical, lifestyle, and cognitive factors," *Neurobiology of aging*, vol. 92, pp. 34–42, 2020.

[18] X. Niu, F. Zhang, J. Kounios, and H. Liang, "Improved prediction of brain age using multimodal neuroimaging data," *Human brain mapping*, vol. 41, no. 6, pp. 1626–1643, 2020.

[19] M. Xiong, L. Lin, Y. Jin, W. Kang, S. Wu, and S. Sun, "Comparison of machine learning models for brain age prediction using six imaging modalities on middle-aged and older adults," *Sensors*, vol. 23, no. 7, p. 3622, 2023.

[20] K. L. Miller, F. Alfaro-Almagro, N. K. Bangerter, *et al.*, "Multimodal population brain imaging in the uk biobank prospective epidemiological study," *Nature neuroscience*, vol. 19, no. 11, pp. 1523–1536, 2016.

[21] B. Dufumier, A. Grigis, J. Victor, C. Ambroise, V. Frouin, and E. Duchesnay, "Openbhb: A large-scale multi-site brain mri data-set for age prediction and debiasing," *NeuroImage*, vol. 263, p. 119 637, 2022.

[22] H. Liang, F. Zhang, and X. Niu, "Investigating systematic bias in brain age estimation with application to post-traumatic stress disorders," *Human brain mapping*, vol. 40, no. 11, pp. 3143–3152, 2019.

[23] M. S. Treder, J. P. Shock, D. J. Stein, S. Du Plessis, S. Seedat, and K. A. Tsvetanov, "Correlation constraints for regression models: Controlling bias in brain age prediction," *Frontiers in psychiatry*, vol. 12, 2021.

[24] A. Lombardi, A. Monaco, G. Donvito, N. Amoroso, and S. Tangaro, "Brain age prediction with morphological features using deep neural networks: Results from predictive analytic competition 2019," *Frontiers in Psychiatry*, vol. 11, p. 619 629, 2021.

[25] B. Mwangi, K. M. Hasan, and J. C. Soares, "Prediction of individual subject's age across the human lifespan using diffusion tensor imaging: A machine learning approach," *Neuroimage*, vol. 75, pp. 58–67, 2013.

[26] T. T. Le, R. T. Kuplicki, B. A. McKinney, *et al.*, "A nonlinear simulation framework supports adjusting for age when analyzing brainage," *Frontiers in aging neuroscience*, vol. 10, 2018.

[27] L. Baecker, J. Dafflon, P. F. Da Costa, *et al.*, "Brain age prediction: A comparison between machine learning models using region-and voxel-based morphometric data," *Human brain mapping*, vol. 42, no. 8, pp. 2332–2346, 2021.

[28] A.-M. G. De Lange, M. Anatürk, S. Suri, *et al.*, "Multimodal brain-age prediction and cardiovascular risk: The whitehall ii mri sub-study," *NeuroImage*, vol. 222, p. 117 292, 2020.

[29] M. Anatürk, T. Kaufmann, J. H. Cole, *et al.*, "Prediction of brain age and cognitive age: Quantifying brain and cognitive maintenance in aging," *Human brain mapping*, vol. 42, no. 6, pp. 1626–1640, 2021.

[30] D. Beck, A.-M. G. de Lange, I. I. Maximov, *et al.*, "White matter microstructure across the adult lifespan: A mixed longitudinal and cross-sectional study using advanced diffusion models and brain-age prediction," *NeuroImage*, vol. 224, p. 117 441, 2021.

[31] H. Li, T. D. Satterthwaite, and Y. Fan, "Brain age prediction based on resting-state functional connectivity patterns using convolutional neural networks," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, IEEE, 2018, pp. 101–104.

[32] G. Rao, A. Li, Y. Liu, and B. Liu, "A high-powered brain age prediction model based on convolutional neural network," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2020, pp. 1915–1919.

[33] H. Peng, W. Gong, C. F. Beckmann, A. Vedaldi, and S. M. Smith, "Accurate brain age prediction with lightweight deep neural networks," *Medical Image Analysis*, vol. 68, 2021.

[34] D. A. Wood, S. Kafiabadi, A. Al Busaidi, *et al.*, "Accurate brain-age models for routine clinical mri examinations," *NeuroImage*, 2022.

[35] Y. Hu, H. Wang, and B. Li, "Sqet: Squeeze and excitation transformer for high-accuracy brain age estimation," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2022, pp. 1554–1557. DOI: 10.1109/BIBM55620.2022.9995270.

[36] S. He, P. E. Grant, and Y. Ou, "Global-local transformer for brain age estimation," *IEEE transactions on medical imaging*, vol. 41, no. 1, pp. 213–224, 2021.

[37] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[38] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[39] A.-M. G. de Lange and J. H. Cole, "Commentary: Correction procedures in brain-age prediction," *NeuroImage: Clinical*, vol. 26, 2020.

[40] B. Zhang, S. Zhang, J. Feng, and S. Zhang, "Age-level bias correction in brain age prediction," *NeuroImage: Clinical*, vol. 37, p. 103 319, 2023.

[41] C. Szegedy, W. Liu, Y. Jia, *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[42] C. Yin, P. Imms, M. Cheng, *et al.*, "Anatomically interpretable deep learning of brain age captures domain-specific cognitive impairment," *Proceedings of the National Academy of Sciences*, vol. 120, no. 2, e2214634120, 2023.

[43] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[44] I. Beheshti, N. Maikusa, and H. Matsuda, "The association between "brain-age score"(bas) and traditional neuropsychological screening tools in alzheimer's disease," *Brain and Behavior*, vol. 8, no. 8, e01020, 2018.

[45] D. P. Varikuti, S. Genon, A. Sotiras, *et al.*, "Evaluation of non-negative matrix factorization of grey matter in age prediction," *Neuroimage*, vol. 173, pp. 394–410, 2018.

[46] C. Gaser, K. Franke, S. Klöppel, N. Koutsouleris, H. Sauer, A. D. N. Initiative, *et al.*, "Brainage in mild cognitive impaired patients: Predicting the conversion to alzheimer's disease," *PloS one*, vol. 8, no. 6, 2013.

[47] R. R. Savjani, B. A. Taylor, L. Acion, E. A. Wilde, and R. E. Jorge, "Accelerated changes in cortical thickness measurements with age in military service members with traumatic brain injury," *Journal of neurotrauma*, vol. 34, no. 22, pp. 3107–3116, 2017.

[48] Y. Chung, J. Addington, C. E. Bearden, *et al.*, "Use of machine learning to determine deviance in neuroanatomical maturity associated with future psychosis in youths at clinically high risk," *JAMA psychiatry*, vol. 75, no. 9, pp. 960–968, 2018.

[49] W. H. Lee, M. Antoniades, H. G. Schnack, R. S. Kahn, and S. Frangou, "Brain age prediction in schizophrenia: Does the choice of machine learning algorithm matter?" *Psychiatry Research: Neuroimaging*, vol. 310, p. 111 270, 2021.

[50] M. Kolenic, K. Franke, J. Hlinka, *et al.*, "Obesity, dyslipidemia and brain age in first-episode psychosis," *Journal of psychiatric research*, vol. 99, pp. 151–158, 2018.

[51] N. Koutsouleris, C. Davatzikos, S. Borgwardt, *et al.*, "Accelerated brain aging in schizophrenia and beyond: A neuroanatomical marker of psychiatric disorders," *Schizophrenia bulletin*, vol. 40, no. 5, pp. 1140–1153, 2014.

[52] L. K. Han, R. Dinga, T. Hahn, *et al.*, "Brain aging in major depressive disorder: Results from the enigma major depressive disorder working group," *Molecular psychiatry*, vol. 26, no. 9, pp. 5124–5139, 2021.

[53] S. Christman, C. Bermudez, L. Hao, *et al.*, "Accelerated brain aging predicts impaired cognitive performance and greater disability in geriatric but not midlife adult depression," *Translational Psychiatry*, vol. 10, no. 1, p. 317, 2020.

[54] S. Han, Y. Chen, R. Zheng, *et al.*, "The stage-specifically accelerated brain aging in never-treated first-episode patients with depression," *Human brain mapping*, vol. 42, no. 11, pp. 3656–3666, 2021.

[55] B. Besteher, C. Gaser, and I. Nenadić, "Machine-learning based brain age estimation in major depression showing no evidence of accelerated aging," *Psychiatry Research: Neuroimaging*, vol. 290, pp. 1–4, 2019.

[56] T. Kaufmann, D. van der Meer, N. T. Doan, *et al.*, "Common brain disorders are associated with heritable patterns of apparent aging of the brain," *Nature neuroscience*, vol. 22, no. 10, pp. 1617–1623, 2019.

[57] A. Lombardi, N. Amoroso, D. Diacono, A. Monaco, S. Tangaro, and R. Bellotti, "Extensive evaluation of morphological statistical harmonization for brain age prediction," *Brain sciences*, vol. 10, no. 6, p. 364, 2020.

[58] B. G. Becker, T. Klein, C. Wachinger, A. D. N. Initiative, *et al.*, "Gaussian process uncertainty in age estimation as a measure of brain abnormality," *NeuroImage*, vol. 175, pp. 246–258, 2018.

[59] L. Ronan, A. F. Alexander-Bloch, K. Wagstyl, *et al.*, "Obesity associated with increased brain age from midlife," *Neurobiology of aging*, vol. 47, pp. 63–70, 2016.

[60] S. McWhinney, M. Kolenic, K. Franke, *et al.*, "Obesity as a risk factor for accelerated brain ageing in first-episode psychosis—a longitudinal study," *Schizophrenia Bulletin*, vol. 47, no. 6, pp. 1772–1781, 2021.

[61] H. M. Aycheh, J.-K. Seong, J.-H. Shin, *et al.*, "Biological brain age prediction using cortical thickness data: A large scale cohort study," *Frontiers in aging neuroscience*, vol. 10, 2018.

[62] S. M. Smith, D. Vidaurre, F. Alfaro-Almagro, T. E. Nichols, and K. L. Miller, "Estimation of brain age delta from brain imaging," *NeuroImage*, vol. 200, pp. 528–539, 2019.

[63] A.-M. G. de Lange, T. Kaufmann, D. van der Meer, *et al.*, "Population-based neuroimaging reveals traces of childbirth in the maternal brain," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22 341–22 346, 2019.

[64] A.-M. G. de Lange, M. Anatürk, J. Rokicki, *et al.*, "Mind the gap: Performance metric evaluation in brain-age prediction," *Human Brain Mapping*, vol. 43, no. 10, pp. 3113–3129, 2022.

[65] J. M. Elwood, *Causal relationship in medicine. A practical system for critical appraisal.* 1988.

[66] I. Beheshti, S. Nugent, O. Potvin, and S. Duchesne, "Bias-adjustment in neuroimaging-based brain age frameworks: A robust scheme," *NeuroImage: Clinical*, vol. 24, 2019.

[67] E. R. Butler, A. Chen, R. Ramadan, *et al.*, "Pitfalls in brain age analyses," Wiley Online Library, Tech. Rep., 2021.

[68] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[69] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.

[70] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[71] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 286–301.

[72] C. Ledig, L. Theis, F. Huszár, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[73] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4799–4807.

[74] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2472–2481.

[75] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, "Transformer for single image super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 457–466.

[76] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[77] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-gan," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[78] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3262–3271.

[79] W. Shi, J. Caballero, F. Huszár, *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceed-*

ings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1874–1883.

[80] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1637–1645.

[81] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Fast and accurate image super-resolution with deep laplacian pyramid networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 11, pp. 2599–2613, 2018.

[82] J. Zhu, G. Yang, and P. Lio, "How can we make gan perform better in single medical image super-resolution? a lesion focused multi-scale approach," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, IEEE, 2019, pp. 1669–1673.

[83] C. You, G. Li, Y. Zhang, *et al.*, "Ct super-resolution gan constrained by the identical, residual, and cycle learning ensemble (gan-circle)," *IEEE transactions on medical imaging*, vol. 39, no. 1, pp. 188–203, 2019.

[84] J. Du, Z. He, L. Wang, *et al.*, "Super-resolution reconstruction of single anisotropic 3d mr images using residual convolutional neural network," *Neurocomputing*, vol. 392, pp. 209–220, 2020.

[85] Q. Lyu, H. Shan, and G. Wang, "Mri super-resolution with ensemble learning and complementary priors," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 615–624, 2020.

[86] X. Zhao, Y. Zhang, T. Zhang, and X. Zou, "Channel splitting network for single mr image super-resolution," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5649–5662, 2019.

[87] C.-H. Pham, A. Ducournau, R. Fablet, and F. Rousseau, "Brain mri super-resolution using deep 3d convolutional networks," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, IEEE, 2017, pp. 197–200.

[88] Y. Chen, F. Shi, A. G. Christodoulou, Y. Xie, Z. Zhou, and D. Li, "Efficient and accurate mri super-resolution using a generative adversarial network and 3d multi-

level densely connected network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2018, pp. 91–99.

[89] Y. Li, Y. Iwamoto, L. Lin, R. Xu, R. Tong, and Y.-W. Chen, "Volumenet: A lightweight parallel network for super-resolution of mr and ct volumetric data," *IEEE Transactions on Image Processing*, vol. 30, pp. 4840–4854, 2021.

[90] Y. Zhang, K. Li, K. Li, and Y. Fu, "Mr image super-resolution with squeeze and excitation reasoning attention network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 425–13 434.

[91] G. Li, J. Lv, Y. Tian, *et al.*, "Transformer-empowered multi-scale contextual matching and aggregation for multi-contrast mri super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 636–20 645.

[92] H. Zhou, Y. Huang, Y. Li, Y. Zhou, and Y. ZhengFellow, "Blind super-resolution of 3d mri via unsupervised domain transformation," *IEEE Journal of Biomedical and Health Informatics*, 2022.

[93] C.-H. Pham, C. Tor-Díez, H. Meunier, *et al.*, "Multiscale brain mri super-resolution using deep 3d convolutional networks," *Computerized Medical Imaging and Graphics*, vol. 77, p. 101 647, 2019.

[94] S. Dong, G. Hangel, W. Bogner, *et al.*, "Multi-scale super-resolution magnetic resonance spectroscopic imaging with adjustable sharpness," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2022, pp. 410–420.

[95] D. C. Alexander, D. Zikic, A. Ghosh, *et al.*, "Image quality transfer and applications in diffusion mri," *NeuroImage*, vol. 152, pp. 283–298, 2017.

[96] R. Tanno, D. E. Worrall, E. Kaden, *et al.*, "Uncertainty modelling in deep learning for safer neuroimage enhancement: Demonstration in diffusion mri," *NeuroImage*, vol. 225, p. 117 366, 2021.

[97] H. Lin, M. Figini, F. D'Arco, *et al.*, "Low-field magnetic resonance image enhancement via stochastic image quality transfer," *Medical Image Analysis*, vol. 87, p. 102 807, 2023.

[98] V. Lau, L. Xiao, Y. Zhao, *et al.*, "Pushing the limits of low-cost ultralow-field mri by dual-acquisition deep learning 3d superresolution," *Magnetic Resonance in Medicine*, 2023.

[99] R. Chen, X. Tang, Y. Zhao, *et al.*, "Single-frame deep-learning super-resolution microscopy for intracellular dynamics imaging," *Nature Communications*, vol. 14, no. 1, p. 2854, 2023.

[100] R. Ranjan and X. Chen, "Super-resolution live cell imaging of subcellular structures," *JoVE (Journal of Visualized Experiments)*, no. 167, e61563, 2021.

[101] J. Valli, A. Garcia-Burgos, L. M. Rooney, B. V. d. M. e Oliveira, R. R. Duncan, and C. Rickman, "Seeing beyond the limit: A guide to choosing the right super-resolution microscopy technique," *Journal of Biological Chemistry*, vol. 297, no. 1, 2021.

[102] W. Yu, C. Rush, M. Tingey, S. Junod, and W. Yang, "Application of super-resolution speed microscopy in the study of cellular dynamics," *Chemical & Biomedical Imaging*, vol. 1, no. 4, pp. 356–371, 2023.

[103] E. V. Sullivan, T. Rohlfing, and A. Pfefferbaum, "Longitudinal study of callosal microstructure in the normal adult aging brain using quantitative dti fiber tracking," *Developmental neuropsychology*, vol. 35, no. 3, pp. 233–256, 2010.

[104] H. Wang, M. Treder, D. Marshall, D. Jones, and Y. Li, "A skewed loss function for correcting predictive bias in brain age prediction," *IEEE Transactions on Medical Imaging*, vol. 42, no. 6, pp. 1577–1589, 2023.

[105] L. Liu, W. Ouyang, X. Wang, *et al.*, "Deep learning for generic object detection: A survey," *International journal of computer vision*, vol. 128, pp. 261–318, 2020.

[106] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

[107] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3365–3387, 2020.

[108] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "Fsl," *Neuroimage*, vol. 62, no. 2, pp. 782–790, 2012.

[109] B. B. Avants, N. Tustison, G. Song, *et al.*, "Advanced normalization tools (ants)," *Insight j*, vol. 2, no. 365, pp. 1–35, 2009.

[110] A. Richie-Halford, J. D. Yeatman, N. Simon, and A. Rokem, "Multidimensional analysis and detection of informative features in human brain white matter," *PLoS computational biology*, vol. 17, no. 6, e1009136, 2021.

[111] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[112] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," *Advances in neural information processing systems*, vol. 9, 1996.

[113] M. Chamberland, E. P. Raven, S. Genc, *et al.*, "Dimensionality reduction of diffusion mri measures for improved tractometry of the human brain," *NeuroImage*, vol. 200, pp. 89–100, 2019.

[114] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[115] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[116] A. P. Merluzzi, D. C. Dean III, N. Adluru, *et al.*, "Age-dependent differences in brain tissue microstructure assessed with neurite orientation dispersion and density imaging," *Neurobiology of aging*, vol. 43, pp. 79–88, 2016.

[117] Y. Wang, J. Wen, J. Xin, Y. Zhang, H. Xie, and Y. Tang, "3dcnn predicting brain age using diffusion tensor imaging," *Medical & Biological Engineering & Computing*, vol. 61, no. 12, pp. 3335–3344, 2023.

[118] C.-L. Chen, Y.-C. Hsu, L.-Y. Yang, *et al.*, "Generalization of diffusion magnetic resonance imaging–based brain age prediction model through transfer learning," *NeuroImage*, vol. 217, p. 116 831, 2020.

[119] H. Zhao, H. Cai, and M. Liu, "Transformer based multi-modal mri fusion for prediction of post-menstrual age and neonatal brain development analysis," *Medical Image Analysis*, p. 103 140, 2024.

[120] A. Abbott, "Dementia: A problem for our age," *Nature*, vol. 475, no. 7355, S2–S4, 2011.

[121] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[122] M. A. Shafto, L. K. Tyler, M. Dixon, *et al.*, "The cambridge centre for ageing and neuroscience (cam-can) study protocol: A cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing," *BMC neurology*, vol. 14, no. 1, pp. 1–25, 2014.

[123] J. R. Taylor, N. Williams, R. Cusack, *et al.*, "The cambridge centre for ageing and neuroscience (cam-can) data repository: Structural and functional mri, meg, and cognitive data from a cross-sectional adult lifespan sample," *Neuroimage*, vol. 144, pp. 262–269, 2017.

[124] A. Di Martino, C.-G. Yan, Q. Li, *et al.*, "The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism," *Molecular psychiatry*, vol. 19, no. 6, pp. 659–667, 2014.

[125] R. Cusack, A. Vicente-Grabovetsky, D. J. Mitchell, *et al.*, "Automatic analysis (aa): Efficient neuroimaging workflows and parallel processing using matlab and xml," *Frontiers in neuroinformatics*, vol. 8, p. 90, 2015.

[126] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[127] F. Pérez-García, R. Sparks, and S. Ourselin, "Torchio: A python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," *Computer Methods and Programs in Biomedicine*, vol. 208, 2021.

[128] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[129] C. Miron, V. Manta, R. Timofte, A. Pasarica, and R.-I. Ciucu, "Efficient convolutional neural network for apparent age prediction," in *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*, IEEE, 2019, pp. 259–262.

[130] S. Escalera, J. Fabian, P. Pardo, *et al.*, "Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 1–9.

[131] E. Agustsson, R. Timofte, S. Escalera, X. Baro, I. Guyon, and R. Rothe, "Apparent and real age estimation in still images with deep residual regressors on appa-real database," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, IEEE, 2017, pp. 87–94.

[132] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Ieee, vol. 1, 2005, pp. 886–893.

[133] R. Rothe, R. Timofte, and L. Van Gool, "Dex: Deep expectation of apparent age from a single image," in *Proceedings of the IEEE international conference on computer vision workshops*, 2015, pp. 10–15.

[134] E. Plenge, D. H. Poot, M. Bernsen, *et al.*, "Super-resolution methods in mri: Can they improve the trade-off between resolution, signal-to-noise ratio, and acquisition time?" *Magnetic resonance in medicine*, vol. 68, no. 6, pp. 1983–1993, 2012.

[135] F. Özyurt, E. Sert, and D. Avcı, "An expert system for brain tumor detection: Fuzzy c-means with super resolution and convolutional neural network with extreme learning machine," *Medical hypotheses*, vol. 134, p. 109 433, 2020.

[136] Y. Chen, Y. Xie, Z. Zhou, F. Shi, A. G. Christodoulou, and D. Li, "Brain mri super resolution using 3d deep densely connected neural networks," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, IEEE, 2018, pp. 739–742.

[137] Q. Lyu, H. Shan, C. Steber, *et al.*, "Multi-contrast super-resolution mri through a progressive network," *IEEE transactions on medical imaging*, vol. 39, no. 9, pp. 2738–2749, 2020.

[138] J. Wang, Y. Chen, Y. Wu, J. Shi, and J. Gee, "Enhanced generative adversarial network for 3d brain mri super-resolution," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 3627–3636.

[139] D. C. Van Essen, S. M. Smith, D. M. Barch, *et al.*, "The wu-minn human connectome project: An overview," *Neuroimage*, vol. 80, pp. 62–79, 2013.

[140] S. M. Smith, "Fast robust automated brain extraction," *Human brain mapping*, vol. 17, no. 3, pp. 143–155, 2002.

[141] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, e3, 2016.

[142] L. Henschel, S. Conjeti, S. Estrada, K. Diers, B. Fischl, and M. Reuter, "Fastsurfer-a fast and accurate deep learning based neuroimaging pipeline," *NeuroImage*, vol. 219, p. 117 012, 2020.

[143] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*, PMLR, 2017, pp. 214–223.

[144] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on computational imaging*, vol. 3, no. 1, pp. 47–57, 2016.

[145] H. Cai, A. Li, G. Yu, X. Yang, and M. Liu, "Brain age prediction in developing childhood with multimodal magnetic resonance images," *Neuroinformatics*, vol. 21, no. 1, pp. 5–19, 2023.

[146] K. Koller, U. Rudrapatna, M. Chamberland, *et al.*, "Micra: Microstructural image compilation with repeated acquisitions," *Neuroimage*, vol. 225, p. 117 406, 2021.

[147] D. K. Jones, M. A. Horsfield, and A. Simmons, "Optimal strategies for measuring diffusion in anisotropic systems by magnetic resonance imaging," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 42, no. 3, pp. 515–525, 1999.

[148] S. C. Deoni, L. Matthews, and S. H. Kolind, "One component? two components? three? the effect of including a nonexchanging "free" water component in multicomponent driven equilibrium single pulse observation of t1 and t2," *Magnetic resonance in medicine*, vol. 70, no. 1, pp. 147–154, 2013.

[149] M. Jenkinson and S. Smith, "A global optimisation method for robust affine registration of brain images," *Medical image analysis*, vol. 5, no. 2, pp. 143–156, 2001.

[150] L. Cordero-Grande, D. Christiaens, J. Hutter, A. N. Price, and J. V. Hajnal, "Complex diffusion-weighted image estimation via matrix recovery under general noise models," *Neuroimage*, vol. 200, pp. 391–404, 2019.

[151] J. Veraart, D. S. Novikov, D. Christiaens, B. Ades-Aron, J. Sijbers, and E. Fieremans, "Denoising of diffusion mri using random matrix theory," *Neuroimage*, vol. 142, pp. 394–406, 2016.

[152] J. Veraart, E. Fieremans, and D. S. Novikov, "Diffusion mri noise mapping using random matrix theory," *Magnetic resonance in medicine*, vol. 76, no. 5, pp. 1582–1593, 2016.

[153] J.-D. Tournier, R. Smith, D. Raffelt, *et al.*, "Mrtrix3: A fast, flexible and open software framework for medical image processing and visualisation," *Neuroimage*, vol. 202, p. 116 137, 2019.

[154] V. Sairanen, A. Leemans, and C. M. Tax, "Fast and accurate slicewise outlier detection (solid) with informed model estimation for diffusion mri data," *Neuroimage*, vol. 181, pp. 331–346, 2018.

[155] J. L. Andersson, S. Skare, and J. Ashburner, "How to correct susceptibility distortions in spin-echo echo-planar images: Application to diffusion tensor imaging," *Neuroimage*, vol. 20, no. 2, pp. 870–888, 2003.

[156] S. M. Smith, M. Jenkinson, M. W. Woolrich, *et al.*, "Advances in functional and structural mr image analysis and implementation as fsl," *Neuroimage*, vol. 23, S208–S219, 2004.

[157] J. L. Andersson and S. N. Sotiropoulos, "An integrated approach to correction for off-resonance effects and subject movement in diffusion mr imaging," *Neuroimage*, vol. 125, pp. 1063–1078, 2016.

[158] E. Kellner, B. Dhital, V. G. Kiselev, and M. Reisert, "Gibbs-ringing artifact removal based on local subvoxel-shifts," *Magnetic resonance in medicine*, vol. 76, no. 5, pp. 1574–1581, 2016.

[159] B. Jeurissen, J.-D. Tournier, T. Dhollander, A. Connelly, and J. Sijbers, "Multi-tissue constrained spherical deconvolution for improved analysis of multi-shell diffusion mri data," *NeuroImage*, vol. 103, pp. 411–426, 2014.

[160] Y. Assaf and P. J. Basser, "Composite hindered and restricted model of diffusion (charmed) mr imaging of the human brain," *Neuroimage*, vol. 27, no. 1, pp. 48–58, 2005.

[161] J. Wasserthal, P. F. Neher, and K. H. Maier-Hein, "Tract orientation mapping for bundle-specific tractography," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11*, Springer, 2018, pp. 36–44.

[162] J. Wasserthal, P. F. Neher, D. Hirjak, and K. H. Maier-Hein, "Combined tract segmentation and orientation mapping for bundle-specific tractography," *Medical image analysis*, vol. 58, p. 101 559, 2019.

[163] J. Wasserthal, P. Neher, and K. H. Maier-Hein, "Tractseg-fast and accurate white matter tract segmentation," *NeuroImage*, vol. 183, pp. 239–253, 2018.

[164] F. Isensee, M. Schell, I. Pflueger, *et al.*, "Automated brain extraction of multisequence mri using artificial neural networks," *Human brain mapping*, vol. 40, no. 17, pp. 4952–4964, 2019.

[165] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," *Neuroimage*, vol. 17, no. 2, pp. 825–841, 2002.

[166] T. C Wood, "Quit: Quantitative imaging tools," *Journal of Open Source Software*, vol. 3, no. 26, p. 656, 2018.

[167] C. Lebel, L. Walker, A. Leemans, L. Phillips, and C. Beaulieu, "Microstructural maturation of the human brain from childhood to adulthood," *Neuroimage*, vol. 40, no. 3, pp. 1044–1055, 2008.

[168] S. W. Davis, N. A. Dennis, N. G. Buchler, L. E. White, D. J. Madden, and R. Cabeza, "Assessing the effects of age on long white matter tracts using diffusion tensor tractography," *Neuroimage*, vol. 46, no. 2, pp. 530–541, 2009.

[169] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[170] P. Branco, L. Torgo, and R. P. Ribeiro, "Smogn: A pre-processing approach for imbalanced regression," in *First international workshop on learning with imbalanced domains: Theory and applications*, PMLR, 2017, pp. 36–50.

[171] M. J. Cardoso, W. Li, R. Brown, *et al.*, "Monai: An open-source framework for deep learning in healthcare," *arXiv preprint arXiv:2211.02701*, 2022.

[172] A. Salami, J. Eriksson, L.-G. Nilsson, and L. Nyberg, "Age-related white matter microstructural differences partly mediate age-related decline in processing speed but not cognition," *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1822, no. 3, pp. 408–415, 2012.

[173] D. Head, R. L. Buckner, J. S. Shimony, *et al.*, "Differential vulnerability of anterior white matter in nondemented aging with minimal acceleration in dementia of the

alzheimer type: Evidence from diffusion tensor imaging," *Cerebral cortex*, vol. 14, no. 4, pp. 410–423, 2004.

[174] D. Salat, D. Tuch, D. Greve, *et al.*, "Age-related alterations in white matter microstructure measured by diffusion tensor imaging," *Neurobiology of aging*, vol. 26, no. 8, pp. 1215–1227, 2005.

[175] T. R. Barrick, R. A. Charlton, C. A. Clark, and H. S. Markus, "White matter structural decline in normal ageing: A prospective longitudinal study using tract-based spatial statistics," *Neuroimage*, vol. 51, no. 2, pp. 565–577, 2010.

[176] A. R. Bender, M. C. Völkle, and N. Raz, "Differential aging of cerebral white matter in middle-aged and older adults: A seven-year follow-up," *Neuroimage*, vol. 125, pp. 74–83, 2016.

[177] B. Alkonyi, C. Juhász, O. Muzik, M. E. Behen, J.-W. Jeong, and H. T. Chugani, "Thalamocortical connectivity in healthy children: Asymmetries and robust developmental changes between ages 8 and 17 years," *American Journal of Neuroradiology*, vol. 32, no. 5, pp. 962–969, 2011.

[178] Y. Yang, A. Sathe, K. Schilling, *et al.*, "A deep neural network estimation of brain age is sensitive to cognitive impairment and decline," in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024*, World Scientific, 2023, pp. 148–162.

[179] B. H. Anderton, "Ageing of the brain," *Mechanisms of ageing and development*, vol. 123, no. 7, pp. 811–817, 2002.

[180] M. M. Esiri, "Ageing and the brain," *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, vol. 211, no. 2, pp. 181–187, 2007.