Towards Explainable AI Modelling on Brain Ageing

Zhaonian Zhang
MSc

A dissertation submitted for the degree of
PhD of Science in Computer Science

Supervised by *Dr. Richard Jiang and Dr. Bryan Williams*

School of Computing and Communications
Lancaster University

October, 2025

Statement

This dissertation is based on the following published papers:

1. Zhaonian Zhang, Richard Jiang, Ce Zhang, Bryan Williams, Ziping Jiang, Chang-Tsun Li, Paul Chazot, Nicola Pavese, Ahmed Bouridane, Azeddine Beghdadi. *Robust brain age estimation based on SMRI via nonlinear age-adaptive ensemble learning.* **IEEE Transactions on Neural Systems and Rehabilitation Engineering** (Z. Zhang, R. Jiang, et al., 2022)

2. Zhaonian Zhang, R Jiang. *User-Centric Democratization towards Social Value Aligned Medical AI Services.* **IJCAI** (Z. Zhang and R. Jiang, 2023)

3. Zhaonian Zhang, R Jiang. *Modeling Brain Aging with Explainable Triamese ViT: Towards Deeper Insights into Autism Disorder.* **IEEE Journal of Biomedical and Health Informatics** (Z. Zhang, Aggarwal, et al., 2025)

I was the lead author and was primarily responsible for the conception, implementation, data analysis, and writing of these works. Co-authors contributed to supervision, editing, and general guidance.

Declaration

I declare that the work presented in this dissertation is, to the best of my knowledge and belief, original and my own work. The material has not been submitted, either in whole or in part, for a degree at this, or any other university. Estimated word count is: **43262**

Name: **Zhaonian Zhang**
Date: **October, 2025**

Towards Explainable AI Modelling on Brain Ageing

Zhaonian Zhang, MSc.
School of Computing and Communications, Lancaster University

A dissertation submitted for the degree of *PhD of Science* in Computer Science.

October, 2025

Abstract

Machine learning, when combined with advanced neuroimaging such as three-dimensional Magnetic Resonance Imaging (MRI), has opened new possibilities for understanding brain health and disease. Among MRI modalities—structural MRI (sMRI), functional MRI (fMRI), and Diffusion Tensor Imaging (DTI)—sMRI is most widely applied in machine learning research, as it provides detailed measures of cortical thickness, gray matter volume, and subcortical anatomy.

Despite significant progress, existing brain age estimation methods face three persistent challenges: limited predictive accuracy across diverse age groups, insufficient interpretability of model predictions, and a lack of fairness in mitigating demographic biases such as age-related bias. These gaps restrict the utility of brain age as a reliable biomarker in both research and clinical settings.

This thesis addresses these limitations by developing new approaches for brain age estimation from sMRI, aiming to improve accuracy, enhance interpretability, and incorporate fairness. To this end, I compiled several large-scale sMRI datasets and proposed three models: the Nonlinear Age-Adaptive Ensemble (nl-AAE), the Triamese Vision Transformer (Triamese-ViT), and the Democratic AI framework (u-DemAI).

The nl-AAE improves predictive accuracy by dynamically weighting multiple base learners according to age groups, achieving a mean absolute error (MAE) of 3.19 years ($r = 0.95$). The Triamese-ViT leverages three orthogonal MRI views and integrates built-in interpretability, meaning that its attention mechanisms generate explanatory maps directly during the prediction process. These intrinsic explanations, validated against conventional explainable AI (XAI) techniques, highlight age-related and ASD-related brain regions consistent with established clinical findings. The u-DemAI framework extends beyond predictive performance by incorporating user personalization into the framework, enabling community-driven model updates and explicitly addressing fairness—particularly reducing age-related bias (ageism) in predictions.

Taken together, these contributions advance the state of the art in brain age estimation by combining accuracy, interpretability, and fairness. More broadly, this work demonstrates how democratic principles can be embedded into machine learning frameworks to promote equitable, transparent, and socially responsible applications in neuroscience and clinical practice.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Richard Jiang, for his invaluable supervision, insightful guidance, and continuous support throughout my PhD journey. His expertise, encouragement, and patience have been instrumental in shaping both this dissertation and my growth as a researcher.

I would also like to extend my heartfelt thanks to my parents, Jizhong Zhang and Jianhong Xiong. Their unwavering love, encouragement, and support have been the foundation of my perseverance and achievements. This work would not have been possible without their belief in me.

Thank you all for being part of this journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The global increase in aging populations poses significant challenges across medical, economic, and societal spheres. Notably, aging is closely associated with declines in cognitive functions and heightened prevalence of neurodegenerative disorders, which together represent considerable burdens for both healthcare systems and affected individuals (Reeve, Simcox, and Turnbull, 2014; R. Jiang, P. Chazot, et al., 2022). Consequently, accurate understanding and prediction of brain aging processes have emerged as vital research priorities within life sciences and biomedicine, holding substantial implications for early disease detection, risk assessment, and interventions aimed at reducing cognitive decline (Cole and Franke, 2017).

Biological aging is marked by the gradual accumulation of adverse biological changes, leading to progressive deterioration in physiological functions. Brain aging, in particular, is associated with structural and functional alterations that impact cognition and mental health. Studies have shown that age-related changes include reductions in brain volume, particularly in the prefrontal cortex, hippocampus, and insular cortex—regions essential for memory, planning, and decision-making (Groves et al., 2012; Storsve et al., 2014; Fjell, Walhovd, et al., 2009). Concurrently, the degradation of white matter integrity (Raz and Rodrigue, 2006) and the increase in the volume of the ventricular system and intracranial cerebrospinal fluid further contribute to cognitive decline (Courchesne et al., 2000; Good et al., 2001; Raz and Rodrigue, 2006). In some cases, neurodegenerative diseases such as Alzheimer's are characterized by abnormal amyloid-beta plaques (Sadigh-Eteghad et al., 2015) and tau protein tangles (Binder et al., 2005), accelerating neuronal degeneration.

In response to the growing incidence of disabling, albeit non-fatal conditions such as dementia and cognitive deterioration associated with population aging, there is an urgent need to elucidate the underlying relationships between the mechanisms of brain aging and the progression of neurodegenerative diseases. Effective methodologies must be developed for early identification of individuals at heightened risk of age-associated neurological

deterioration, continuous monitoring of disease progression, and implementation of targeted therapeutic interventions.

Age-related structural and functional brain changes significantly contribute to the etiology of various neurological conditions. The divergence between biological brain age and chronological age offers a potential biomarker for assessing vulnerability to health complications across different life stages. One prominent approach in brain aging research is brain age estimation (Figure 1.1), a technique that uses neuroimaging data to predict an individual's brain age based on characteristic age-related patterns. The deviation between predicted brain age and chronological age, commonly referred to as the brain age gap (BAG), has emerged as a significant biomarker for neurological and psychiatric disorders, including Alzheimer's disease (Beheshti, Maikusa, and Matsuda, 2018), psychosis (Chung et al., 2018), mild cognitive impairment (Gaser et al., 2013), and depression (Han et al., 2021). An elevated brain age gap typically indicates a higher likelihood of neurodegenerative disorders and increased mortality, underscoring its diagnostic and prognostic value (Cole and Franke, 2017; Cole, Ritchie, et al., 2018). In recent decades, advances in data-driven methodologies—particularly machine learning models applied to magnetic resonance imaging (MRI) scans—have greatly enhanced brain age estimation. These models commonly employ supervised regression trained on neuroimaging data from cognitively healthy individuals to map brain-derived features to chronological age, and then predict brain age in unseen cases. Deviations from normative trajectories serve as indicators of brain health: a positive BAG reflects an older-appearing brain and is associated with pathological alterations and mortality risk, whereas a negative BAG suggests comparatively preserved cognitive function (Cole and Franke, 2017). Beyond its diagnostic utility, brain age estimation has also shed light on lifestyle and environmental influences on cognitive aging, demonstrating protective effects of higher education and physical activity (Steffener et al., 2016), as well as practices such as meditation, in maintaining cognitive resilience (Luders, Cherbuin, and Gaser, 2016). Together, these findings establish brain age estimation as a valuable framework for both disease detection and preventive strategies that promote healthy aging.

Brain age estimation also plays a significant role in medicine research and development, particularly in clinical trials, which are fundamental to clinical science (Bzdok, Varoquaux, and Steyerberg, 2021; J. A. Sidey-Gibbons and C. J. Sidey-Gibbons, 2019). Many pharmaceutical companies worldwide are actively engaged in the development of medications for age-related diseases. However, the therapeutic effects of these treatments are often not immediately discernible, making it challenging for clinicians to assess their efficacy in the short term. Even experienced physicians may find it difficult to determine whether a drug has yielded the desired effects, as its impact on the aging process may take years to manifest. This prolonged evaluation period complicates data collection for pharmaceutical companies, ultimately hindering progress in the development of therapeutics for age-related conditions (J. A. Sidey-Gibbons and C. J. Sidey-Gibbons, 2019).

Brain age estimation provides an alternative approach to addressing this challenge by

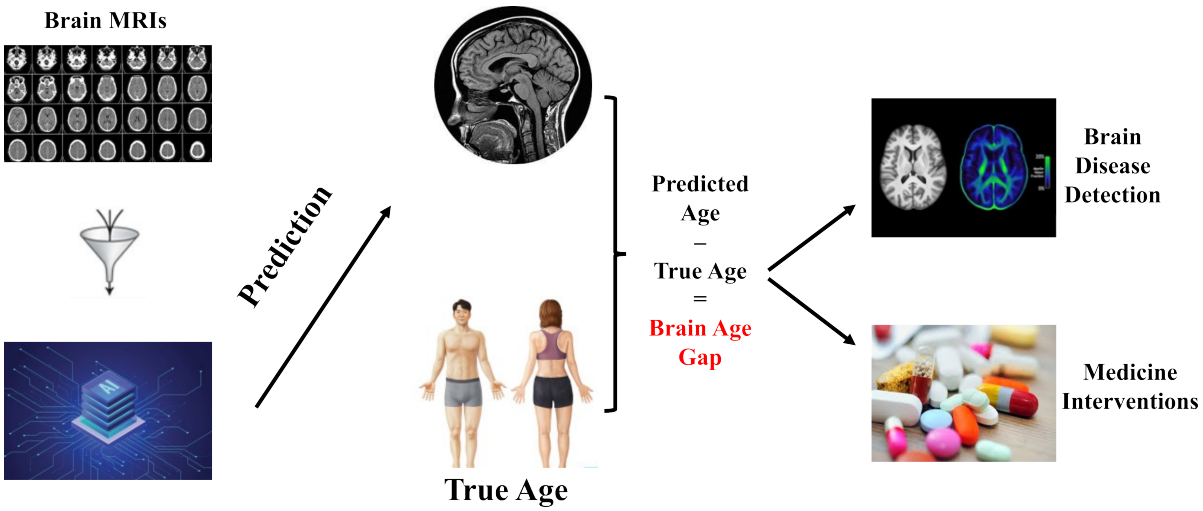


Figure 1.1: The brain age estimation process. MRIs serve as input to the deep learning models, which then predict the subjects' ages based on these images. These predicted ages are compared with the subjects' actual chronological ages to calculate key indicators, notably the brain age gap (predicted age minus chronological age). This brain age gap can be instrumental in detecting various brain diseases and assessing medicine interventions.

facilitating the continuous monitoring of drug effects through changes in predicted brain age over time (J. A. Sidey-Gibbons and C. J. Sidey-Gibbons, 2019). It leverages hierarchical feature representations in an end-to-end manner to capture subtle neuroanatomical changes (Cole and Franke, 2017). Empirical studies have shown that the discrepancy between predicted brain age and chronological age is minimal in cognitively healthy individuals (Cole and Franke, 2017; Luders, Cherbuin, and Gaser, 2016; Cole, Ritchie, et al., 2018). Brain age estimation enables pharmaceutical companies to conduct real-time follow-ups from the onset of treatment, allowing for timely assessments of drug efficacy and expediting the collection of patient data.

It is important to emphasize that the validity of interpreting brain age findings fundamentally depends on the robustness of the employed brain age estimation framework. Indeed, a highly precise brain age estimation framework can produce more reliable and clinically meaningful results. Consequently, developing increasingly accurate frameworks for estimating brain age is crucial, prompting numerous research groups to pursue enhancements by leveraging various machine learning approaches. Alongside conventional machine learning methods utilized in brain age estimation (Beheshti, Ganaie, et al., 2021; Ganaie, Tanveer, and Beheshti, 2024; Ganaie, Muhammad Tanveer, and Beheshti, 2022), deep learning has recently emerged as a prevalent methodology within the neuroimaging domain, being extensively applied to diverse tasks such as segmentation, lesion detection, and classification (Sajedi and Pardakhti, 2019). A significant advantage of deep learning techniques lies in their inherent

capability to integrate feature extraction, dimensionality reduction, and predictive modeling within a unified computational architecture, enabling superior performance compared to traditional machine learning models—particularly when analyzing highly complex datasets. Accordingly, deep learning has increasingly become the method of choice in brain imaging research, and the quantity of deep learning-driven neuroimaging investigations has exhibited substantial and consistent growth over the past decade (Sajedi and Pardakhti, 2019).

1.2 Scope and Boundaries of the Research

1.2.1 Research Questions and Hypotheses

This research is guided by three main questions:

1. Can novel deep learning models improve the accuracy of brain age estimation from sMRI compared with existing state-of-the-art approaches?

2. How can model interpretability be enhanced to identify brain regions associated with normal aging and to analyze differences between individuals with ASD and healthy controls?

3. Fairness in model prediction refers to ensuring that the model's predictive performance remains consistent across different sub-populations, thereby preventing unequal treatment of particular groups. In the context of brain age prediction, fairness is most directly related to mitigating ageism—that is, avoiding systematic overestimation or underestimation of brain age for specific chronological age groups. So, can fairness be incorporated into brain age estimation?

4. Can a democratized AI framework be developed, which enables community involvement in model optimization while embedding fairness and promoting social values?

Based on these questions, the research advances the following hypotheses:

- Deep learning models with novel architectures will achieve lower prediction errors than current leading models.

- Built-in interpretability mechanisms in the model will provide reliable insights into regional brain changes associated with aging and ASD diagnosis.

- By considering fairness in the models, they will reduce performance disparities across demographic subgroups.

- A democratized AI system can enhance accessibility, fairness, and societal impact by integrating user participation into the continuous optimization process.

1.2.2 Research Objectives

The aim of this dissertation is to advance brain age estimation through methodological innovation, interpretability, fairness, and social applicability. The specific objectives are as follows:

1. To design and implement novel deep learning models for brain age prediction based on sMRI, achieving higher accuracy than existing models.

2. To develop a built-in interpretability approach within the model, enabling direct analysis of brain regions contributing to normal aging and brain differences between healthy individuals and ASD patients.

3. To evaluate and improve fairness in model predictions, ensuring consistent performance across demographic factors, such as sex, ethnicity, and age groups. In this dissertation, we only discussed the age groups.

4. To propose a Democratized AI System that integrates fairness principles and involves relevant stakeholders in model optimization, thereby promoting social value and community benefit.

1.2.3 Scope of the Research

This research is focused on brain age estimation using sMRI data. Compared with other aging biomarkers such as telomere length or physiological measures, brain age estimation poses unique challenges. First, it relies on high-dimensional neuroimaging data, where complex spatial patterns must be captured and interpreted, making prediction highly sensitive to model design and data quality. Second, interpretability remains a central challenge: while brain age can indicate accelerated or decelerated aging, linking these deviations to specific neural mechanisms or clinical outcomes requires careful validation. Finally, issues of fairness and generalizability are particularly salient, as differences in demographic distributions—such as sex, ethnicity, or age range—may bias model performance. These challenges highlight both the complexity and the potential of brain age as a biomarker, underscoring the need for methodological innovation in accuracy, interpretability, and fairness.

The work is limited to sMRI modalities and does not extend to other imaging types such as functional MRI (fMRI), diffusion tensor imaging (DTI), or Computed Tomography (CT). Compared with fMRI, which captures transient neural activity, and DTI, which focuses on white matter connectivity, sMRI provides stable and high-resolution measurements of brain anatomy, such as cortical thickness, gray matter volume, and subcortical structures, which are closely linked to age-related changes. Unlike CT, which is widely used in clinical practice, sMRI is non-invasive and better suited for longitudinal studies in healthy and clinical populations.

The scope covers methodological development, model interpretability, and fairness evaluation, as well as the conceptual design of a democratized AI framework. While the models are tested on large-scale research datasets, they are not intended for direct clinical deployment. Instead, the emphasis lies on advancing technical performance, interpretability, and fairness in brain age estimation, and on exploring how such models can contribute to more equitable and socially responsible AI systems.

1.3 Contribution

A central contribution of this dissertation is the development of two novel model algorithms for brain age estimation: the Nonlinear Age-Adaptive Ensemble model (nl-AAE) (Z. Zhang, R. Jiang, et al., 2022) and the Explainable Triamese ViT (Z. Zhang, Aggarwal, et al., 2025). In addition, this work introduces a new perspective on Democratic AI (Z. Zhang and R. Jiang, 2023), which combines democratic participation with fairness in predictions while maintaining high accuracy.

The nl-AAE model integrates multiple independent predictors within a nonlinear, age-adaptive ensemble framework. By leveraging the complementary strengths of GoogLeNet, ResNet, Support Vector Regression (SVR), and a custom-designed Convolutional Neural Network (CNN), the model adapts dynamically to age-related variations. Its nonlinear weighting mechanism adjusts contributions from each constituent model based on the chronological age of the input, thereby capturing distinct aging patterns across the lifespan and enhancing predictive performance.

The nl-AAE was evaluated using the PAC 2019 competition dataset and benchmarked against its four constituent models. Experimental results demonstrate that the ensemble substantially improves predictive accuracy, achieving a mean absolute error (MAE) of 3.19 years and a Spearman correlation of 0.95, outperforming conventional approaches in brain age estimation. These findings highlight the potential of nl-AAE for applications such as early detection of Alzheimer's disease, assessment of traumatic brain injury, schizophrenia diagnosis, and evaluation of neuroprotective interventions in clinical trials.

Triamese-ViT is a deep learning model designed to achieve both high predictive accuracy and intrinsic interpretability in brain age estimation and the study of neurological disorders. The model was trained on a diverse cohort of 1,351 cognitively healthy individuals, aged 6 to 80, by integrating data from the IXI and ABIDE datasets to establish normative brain aging trajectories.

The architecture processes sMRI scans from three distinct anatomical orientations using Vision Transformers (ViTs). Features extracted from these perspectives are then combined through a tri-MLP framework to generate age predictions. This tri-view design, introduced for the first time in this work, enables the model to capture complementary structural information from different orientations. Triamese-ViT achieved a mean absolute error (MAE) of 3.85 years, a Spearman correlation of 0.94, and a correlation of -0.3 between chronological

age and the brain age gap, surpassing existing state-of-the-art methods in terms of predictive accuracy, fairness, and interpretability.

Beyond performance, the model offers a built-in interpretability mechanism that generates three-dimensional attention maps by integrating information from multiple views. These maps provide direct insight into structural correlates of aging and neurological conditions without requiring post-hoc explainability methods. Compared with recent approaches (Tanveer et al., 2023; L. Chen and Luo, 2023), Triamese-ViT demonstrates superior accuracy and fairness, while also offering enhanced interpretability. In contrast to 3D ViT models such as that proposed in Singla et al. (2022), Triamese-ViT provides additional advantages:

- Computational efficiency: the multi-view processing strategy reduces complexity relative to volumetric 3D ViTs;
- Lower memory requirements: the architecture is lightweight and more practical for large-scale training and deployment;
- Simplified implementation: the model achieves high accuracy while avoiding the heavy computational burden of full 3D ViTs.

Leveraging its interpretability, Triamese-ViT was applied to investigate normal brain aging and Autism Spectrum Disorder (ASD). Through attention map analysis, we identified age-specific structural changes in regions such as the Rolandic Operculum, Cingulum, Thalamus, and Vermis, which are strongly associated with common neurological conditions. In ASD patients, the model highlighted the Thalamus and Caudate Nucleus, underscoring their relevance in the disorder's pathology. These findings were further validated by conventional occlusion analysis, which confirmed the alignment between built-in interpretability and established explainable AI techniques.

In summary, the contributions of Triamese-ViT are as follows:

- Proposes a novel tri-view ViT framework for brain age estimation, achieving higher predictive accuracy and fairness compared to state-of-the-art models;
- Demonstrates improved computational efficiency, reduced memory demand, and scalability relative to high-accuracy 3D ViT models;
- Provides intrinsic interpretability through attention maps, empirically validated with occlusion analysis;
- Enables the identification of brain regions associated with normal aging, offering machine learning–based insights into neurobiological changes;
- Highlights key regions relevant to ASD, demonstrating the model's potential for clinical research applications.

The third contribution of this dissertation is the proposal of u-DemAI, a framework that operationalizes the concept of Democratic AI by addressing both societal and technological challenges. Here, democratic AI refers to an implementation in which relevant stakeholders are directly involved in optimizing AI services, thereby promoting social values and benefiting user communities. The u-DemAI system enables continual self-improvement through user interaction and supports model personalization according to individual requirements. Although it is constructed from a set of basic models that can be locally trained by non-expert users, its performance rivals or even surpasses advanced expert-driven models. Importantly, the framework maintains both high predictive accuracy and fairness, where fairness is defined as consistent predictive performance across demographic subgroups.

To validate the framework, we conducted a case study on brain age estimation using the PAC2019 dataset. The u-DemAI achieved strong results with a mean absolute error (MAE) of 2.67, a standard deviation of absolute error of 2.67, and a Pearson correlation of 0.01 between the brain age gap and chronological age, indicating both high accuracy and fairness. These findings serve as a proof-of-concept implementation of Democratic AI, demonstrating the feasibility of engaging non-experts in AI optimization and highlighting the advantages of this approach over conventional expert-dominated AI services.

1.4 Overview of the Dissertation

The remainder of this dissertation is organized as follows. In the next chapter, a comprehensive literature review of brain age estimation is presented, including an introduction to neuroimaging analysis, the development of AI models in this field, and a discussion of popular deep learning architectures applied to medical image analysis.

Subsequently, Chapter 3 introduces the Nonlinear Age-Adaptive Ensemble Model (nl-AAE) in detail. This includes dataset preprocessing, model architecture, and a performance comparison with state-of-the-art algorithms, followed by a discussion of its implications for brain aging analysis.

Chapter 4 presents the Explainable Triamese ViT. We describe the preprocessing pipeline, model structure, and comparative performance evaluation. The chapter further emphasizes the built-in interpretability mechanism, demonstrating its utility in analyzing normal brain aging as well as differences between healthy individuals and patients with ASD.

Following this, Chapter 5 introduces the user-centric Democratic AI (u-DemAI) framework. We begin with a conceptual discussion of democratized AI and our proposed definition, then describe the system's structure and application to brain age estimation, along with experimental comparisons against existing methods.

Finally, Chapter 6 concludes the dissertation with a synthesis of the key contributions, a reflection on the significance and limitations of the work, and a discussion of potential directions for future research.

Chapter 2

Literature Review

2.1 Background

Neuroimaging has become a useful tool in the diagnosis and study of neurodegenerative diseases, offering non-invasive methods to capture both structural and functional changes in the brain. Among the commonly used imaging modalities, structural Magnetic Resonance Imaging (sMRI) provides high-resolution anatomical information, such as cortical thickness, gray matter volume, and hippocampal atrophy, which are widely used as biomarkers for conditions like Alzheimer's disease and Parkinson's disease (Frisoni et al., 2010). Functional Magnetic Resonance Imaging (fMRI) measures blood-oxygen-level-dependent (BOLD) signals and is employed to assess alterations in brain activity and connectivity patterns, offering valuable insights into disrupted neural networks associated with disorders such as dementia and Huntington's disease (Greicius et al., 2004). Diffusion Tensor Imaging (DTI) can capture the microstructural integrity of white matter tracts, making it particularly relevant for detecting connectivity disruptions in multiple sclerosis and other neurodegenerative conditions (Pierpaoli et al., 1996). In addition, Computed Tomography (CT) remains clinically important for its rapid acquisition and ability to detect structural abnormalities, including ischemic lesions and vascular pathologies that may contribute to cognitive decline (Jack Jr et al., 2008).

Since sMRI provides stable and high-resolution measurements of brain anatomy and has more shared public datasets to analyze, it is always a suitable modality to investigate both normal brain aging and neurodevelopmental conditions such as Autism Spectrum Disorder (ASD), offering quantitative measures of cortical thickness, surface area, and subcortical volumes.

In the context of normal aging, longitudinal and cross-sectional sMRI studies consistently report regionally specific cortical thinning and volumetric decline. For instance, Raz and Rodrigue (2006) followed 55 healthy adults longitudinally and found pronounced shrinkage in the prefrontal cortex and hippocampus, regions particularly vulnerable to

aging. Prefrontal Cortex is responsible for higher-order cognitive functions, such as decision-making, planning, and social behavior regulation. It plays a crucial role in executive control and emotional regulation. The hippocampus is central to memory formation and spatial navigation. Age-related hippocampal atrophy is strongly associated with cognitive decline and neurodegenerative disorders such as Alzheimer's disease.

Expanding to a larger cohort, Fjell, Westlye, et al. (2009) analyzed over 880 cross-sectional scans and observed widespread cortical thinning with advancing age, especially in association cortices. The association cortices integrate information from multiple sensory modalities and are responsible for higher cognitive processes such as language, attention, and abstract reasoning. They link perception with memory and decision-making, allowing the brain to synthesize complex information. Longitudinal experiments further clarify these patterns: Storsve et al. (2014) studied 207 adults across a 4-year follow-up and showed that cortical volume decline is primarily driven by thinning rather than surface area reduction, with accelerated changes in temporal and occipital regions. Temporal includes processing centers and regions important for language comprehension and semantic memory. And Occipital Lobe is the primary center for visual processing, responsible for interpreting visual stimuli and spatial orientation.

Meta-analyses have confirmed the robustness of these findings: Hedman et al. (2012), synthesizing 56 longitudinal sMRI studies, concluded that brain atrophy occurs progressively across the lifespan, with regional trajectories varying by lobe. Together, these studies establish cortical thinning and subcortical atrophy, as measured by sMRI, as reliable biomarkers of normal aging.

In ASD research, sMRI has provided insights into both cortical and subcortical alterations. Early small-sample studies, such as Ecker et al. (2013), analyzed 168 adult males and reported increased cortical thickness in frontal regions and reduced surface area in orbitofrontal and posterior cingulate cortices, highlighting distinct developmental pathways underlying ASD morphology. Frontal Lobe governs voluntary movement, problem-solving, and personality expression. Its progressive volume decline with age. As for Orbitofrontal Cortex, it involved in reward processing, impulse control, and social cognition. And Cingulate Cortex integrates emotional and cognitive information, mediating decision-making and adaptive behavior. It forms part of the limbic system and is sensitive to both aging and neurodevelopmental abnormalities.

Larger consortia datasets have since clarified the picture. Haar et al. (2016), using the Autism Brain Imaging Data Exchange (ABIDE) dataset with over 1,100 participants, found only limited structural differences—namely, enlarged ventricles and smaller corpus callosum—while failing to replicate several previously reported findings, underscoring the role of sample heterogeneity. Corpus Callosum is the largest white matter structure connecting the two hemispheres of the brain, enabling efficient interhemispheric communication. Its deterioration with age is linked to slowed information transfer and cognitive decline. Van Rooij et al. (2018) pooled sMRI data from over 3,000 individuals, identified subtle

but consistent alterations in cortical thickness and subcortical volumes, with modest effect sizes that varied with age.

While prior studies have provided valuable insights into normal aging and ASD using sMRI, most of them relied on region-of-interest (ROI) analyses, voxel-based morphometry (VBM), or surface-based morphometry (SBM). These approaches often focused on specific structures or tested group-level differences, which may overlook subtle and distributed brain-wide patterns. Moreover, many earlier studies were constrained by relatively small sample sizes or single-site cohorts, limiting the generalizability of their findings. In contrast, the present work employs a whole-brain deep learning approach applied directly to sMRI, allowing the model to capture complex, distributed features of brain structure without predefined assumptions about specific regions. In Chapter 4, the study on the Triamese-ViT model (Z. Zhang, Aggarwal, et al., 2025) leverages a comparatively large dataset of 1,349 individuals spanning a wide age range, providing sufficient statistical power to model normative brain aging trajectories and to identify deviations associated with ASD. By integrating large-scale sMRI with advanced deep learning methods, this dissertation contributes a more comprehensive and scalable framework for understanding both normal aging and brain disease.

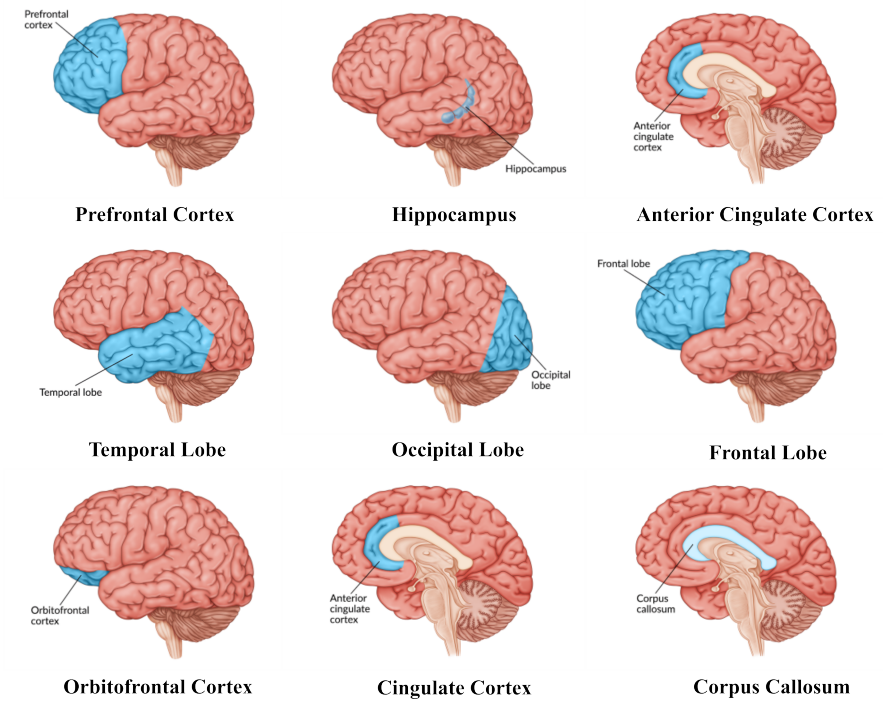


Figure 2.1: Figures adapted from Flint Rehab (https://www.flintrehab.com/corpus-callosum-injury/), accessed 2025. These illustrations highlight major cortical and subcortical structures relevant to both normal aging and Autism Spectrum Disorder (ASD) analysis.

2.2 Data Preprocessing

The preprocessing of neuroimaging data represents an essential phase in the development of accurate brain age estimation frameworks. The choice of preprocessing techniques typically varies based on the neuroimaging modality utilized. Most prior deep learning studies in brain age estimation have relied on T1-weighted MRI (T1w-MRI) data, primarily due to its greater accessibility compared to other imaging modalities, I also used this type of data in my project. Within T1w-MRI preprocessing, VBM, commonly implemented using statistical parametric mapping (SPM), has emerged as a prevalent and effective method. VBM is particularly valued for its ability to detect morphological brain changes and age-related variations at the voxel level (Beheshti, Sone, et al., 2018; Farokhian, C. Yang, et al., 2017). A key advantage of VBM-based preprocessing lies in its generation of grey matter (GM) and white matter (WM) maps, which can subsequently serve as inputs for both three-dimensional voxel-based and two-dimensional slice-based prediction models. Comprehensive technical descriptions of the VBM preprocessing pipeline are available in prior research (Farokhian, Beheshti, et al., 2017).

Alternatively, region-based preprocessing techniques, such as those provided by FreeSurfer, are also frequently employed for processing T1w-MRI data. These methods facilitate the extraction of morphological characteristics from cortical and subcortical brain regions, including surface-based morphometry and measurements of cortical thickness. Beyond T1w-MRI, studies have also utilized fluorodeoxyglucose positron emission tomography (FDG-PET) imaging to estimate brain age through deep learning approaches (J. Lee et al., 2022). FDG-PET offers complementary neuroimaging information related to brain glucose metabolism.

2.3 Popular Deep Learning Architectures

After pre-processing the data, the next step is to feed the data to the model. In this section, we will review the popular deep learning architectures used in brain age estimation. We present the models in order of increasing complexity, beginning with the CNN model, followed by VGG, ResNet, the Transformer framework, Ensemble Learning, and concluding with other popular used architectures.

2.3.1 Convolutional Neural Networks

Since 2017, convolutional neural networks (CNNs) have increasingly attracted attention from researchers in the field of brain age estimation. CNNs have gained popularity primarily due to their capability to automatically extract relevant features and their superior predictive performance compared to traditional methods.

In the context of slice-based CNN methodologies, 2D CNNs are frequently trained using individual 2D MRI slices. Previous work (P. K. Lam et al., 2020) has addressed certain limitations inherent to 3D CNNs when applied to brain age prediction, particularly noting the substantial number of parameters required and the computational complexity involved during training. Consequently, the authors introduced a hybrid architecture combining a 2D CNN and a recurrent neural network (RNN) to improve brain age prediction. In their proposed method, the 2D CNN component extracts essential intra-slice features, while the RNN captures inter-slice dependencies across sequential sagittal MRI slices. Model parameters were initialized using the Kaiming initialization strategy, and the training utilized the Adam optimizer. The primary benefit of this hybrid CNN-RNN approach lies in its significant reduction in the number of parameters—approximately 1,068,065 compared to 2,018,000 parameters in traditional 3D CNNs—while maintaining improved accuracy. However, the primary limitation noted in this work was the absence of analysis regarding the association between predicted brain age and clinical or neurological health outcomes.

Similarly, another slice-based CNN approach was proposed by Amoroso et al. (2019), which involved segmenting T1-weighted MRI scans into smaller patches. The authors utilized Pearson correlation coefficients to quantify pairwise similarity among patches, subsequently employing these similarity metrics to construct a complex network in which patches represented nodes, and correlation values were treated as weighted connections. Furthermore, the study employed the Gedeon method to identify brain regions most influential for age prediction, highlighting the ten most significant age-related features localized predominantly in the left hemisphere. Nevertheless, the principal limitation of their study was the relatively small sample size, comprising only 488 subjects, potentially limiting the generalizability of their findings.

Expanding upon slice-based methods, recent work (Dular, Špiclin, and Alzheimer's Disease Neuroimaging Initiative, 2021) compared four CNN architectures specifically designed for brain age estimation, incorporating techniques such as transfer learning with domain adaptation and bias correction to improve generalization to unseen MRI data. Of these four models, the first and fourth were trained using complete, full-resolution 3D T1-weighted MRI images, while the second model utilized a 2D CNN trained on fifteen extracted axial slices per subject. The third model, however, trained a 3D CNN on downsampled MRI data. This variation in input data type and resolution across models potentially influenced the overall predictive performance and comparability of the proposed CNN architectures.

Voxel-based CNN approaches involve training convolutional neural networks on three-dimensional volumetric MRI scans. For example, a voxel-based CNN strategy described in Cole and Franke (2017) employed segmentation of T1-weighted MRI images into GM and WM using the SPM12 toolbox. Their results indicated that grey matter volume alone yielded superior predictive performance compared to using WM separately, raw images, or combinations of GM and WM. Additionally, this study explored the heritability of brain age predictions based on GM, WM, and their combined usage. However, that study was

limited by the use of data from only two MRI scanners, restricting broader validation of the model's reliability across scanners. In addition, the authors did not examine how predicted brain age gaps relate to specific neuroanatomical features, nor did they consider the impact of in-scanner motion artifacts that may affect practical model performance. In contrast, our Triamese-ViT was trained and tested on datasets acquired from multiple scanners, and ComBat harmonization was applied to correct for scanner- and site-specific effects while preserving meaningful biological variability. Furthermore, we combined conventional XAI methods with the model's built-in interpretability function to investigate the association between brain regions and predicted brain age gaps. We also systematically analyzed the sources of artifacts observed in the interpretability maps to ensure more reliable regional inference.

Related research by J. Wang et al. (2019) segmented 1.5T MRI data into GM, WM, and cerebrospinal fluid, selecting GM volumes as inputs for a 3D CNN. Logistic regression and Cox proportional hazards models were subsequently utilized to explore the association between predicted brain-age gaps and dementia incidence. Through gradient-weighted activation mapping, attention maps highlighted alterations in GM intensity around the hippocampus and amygdala associated with aging. Nevertheless, the primary drawback of their model is its limited generalizability, as it was incapable of effectively handling images sourced from diverse datasets. Furthermore, the authors explicitly excluded dementia and stroke patients from their training set, potentially limiting the practical applicability of the developed CNN model.

And the research conducted in Bellantuono et al. (2021) introduced a complex graph-based framework using structural connectivity models to address brain age prediction. Specifically, T1-weighted MRI images were partitioned into smaller patches, and the similarity between each pair of patches was quantified using Pearson correlation coefficients. Subsequently, these patch pairs and their correlation metrics were integrated into a deep learning model to facilitate prediction. The proposed approach effectively captured age-related structural connectivity changes across various brain regions while achieving robust performance with minimal preprocessing—primarily brain extraction and linear image registration—to manage computational complexity. The authors also applied the Gedeon method to determine feature importance, identifying twelve key brain regions that were significantly informative regarding aging, such as the medial frontal gyrus, caudate nucleus, paracentral lobule, putamen, cingulate gyrus, brainstem, and sub-gyral regions. Nevertheless, a notable limitation of their work is that the analysis was conducted exclusively on a dataset comprising Autism Spectrum Disorder subjects, potentially restricting the generalizability of their findings to other neurological or psychiatric conditions.

Hybrid modeling strategies have also emerged within voxel-based CNN research for brain age estimation. For instance, the study presented in Pardakhti and Sajedi (2020) proposed integrating a 3D CNN with support vector regression (SVR) and Gaussian process regression to enhance predictive accuracy and generalization. The researchers trained their hybrid

model on healthy participants from the IXI dataset, subsequently validating it using 47 cognitively healthy subjects and 22 Alzheimer's disease (AD) patients from the ADNI dataset. Despite demonstrating good generalization potential, the authors did not perform any regional neuroanatomical analyses to identify specific brain areas associated with Alzheimer's pathology.

Finally, Hong et al. (2020), who developed a 3D CNN-based model for predicting brain age in pediatric populations using routine MRI scans. To enhance the dataset size for training, they employed data augmentation strategies. Additionally, they investigated slice-based 2D CNN modeling by converting volumetric MRI data into multiple 2D slices. However, the predictive performance of the 2D CNN was inferior to that of the 3D CNN, underscoring the critical importance of spatial correlations among adjacent slices in brain age estimation. Interestingly, their analysis revealed higher accuracy in predicting brain age for children younger than two years compared to older children.

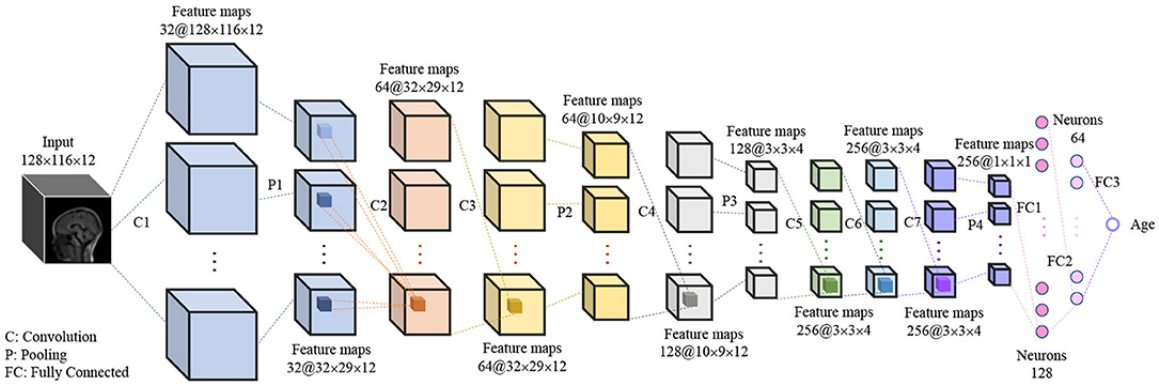


Figure 2.2: 3D CNN Architecture for brain age estimation used in Hong et al. (2020). (figure from Hong et al. (2020))

2.3.2 VGG

Given the increasing popularity of DL techniques, researchers have increasingly employed pretrained DL architectures to enhance performance in brain age estimation tasks. One prominent pretrained DL network is VGGNet (T.-W. Huang et al., 2017), characterized by convolutional layers of fixed dimensions, allowing it to effectively capture fine-grained details within input images. Additionally, VGGNet progressively doubles the number of convolutional filters in subsequent convolutional stacks, enabling the network to learn increasingly complex hierarchical features.

The study presented in Yao et al. (2023) explored the efficacy of a 2D VGGNet architecture. Specifically, the authors utilized VGGNet to identify and extract 2D slices with maximal mutual information from 3D volumetric MRI scans, aiming to enhance

computational efficiency and prediction speed. However, the study exhibited several notable limitations. Firstly, preprocessing of MRI data was restricted exclusively to channel normalization, neglecting critical preprocessing steps such as motion artifact correction, linear registration, and bias-field correction, potentially compromising the accuracy of results. Additionally, the authors did not provide a comparative performance assessment of VGGNet against alternative pretrained deep learning architectures.

In contrast to slice-based approaches, voxel-based VGGNet models have received greater attention in studies. A study utilizing voxel-based VGGNet models was presented in Dinsdale et al. (2021), employing a 3D VGGNet trained on T1-weighted images from the UK Biobank dataset. The authors investigated correlations between estimated brain age differences and an extensive set of 8,787 non-imaging variables, including lifestyle, physiological and medical history, and self-reported mental health conditions. Their results revealed modest correlations between brain age differences and various image-derived phenotypes (IDPs) across multiple imaging modalities. Additionally, the study explored the utility of attention mechanisms applied to both linearly and nonlinearly registered MRI scans, identifying subtle cortical changes primarily visible in linear registrations. A notable drawback of their approach, however, was the reliance on IDPs—which compress voxel-level information—to explore relationships between brain age differences and other imaging modalities, potentially limiting sensitivity compared to analyses performed directly on T1-weighted 3D images.

The research by H. Jiang et al. (2020) employed a voxel-based VGGNet model after segmenting T1-weighted MRI data into seven cortical regions using the CorticalParcellation-Yeo2011 atlas. Individual 3D VGGNet models trained separately on these cortical networks demonstrated the lowest mean absolute errors in age estimation for the frontoparietal, dorsal attention, and default mode networks. Additionally, the authors examined the relationship between grey matter volumes and brain aging using Pearson correlation analysis. Nonetheless, this approach was limited by considering only seven broad cortical networks, thereby neglecting finer-scale age-related effects observable within subcortical structures and smaller cortical subnetworks.

A noteworthy alternative approach was proposed by Peng et al. (2021), who introduced a simplified, fully-connected convolutional network (SFCN) architecture inspired by the VGGNet design. Their lightweight model incorporated regularization methods, such as dropout, voxel shifting, and mirroring, to enhance predictive accuracy. However, the improved performance came at the expense of significantly increased training durations, exceeding 50 hours using two dedicated NVIDIA P100 GPUs. Moreover, this study did not explore potential associations between predicted brain age discrepancies and specific health outcomes or neurological conditions.

Last, X. Feng et al. (2020) utilized a five-layer 3D VGGNet architecture to analyze a large-scale, heterogeneous MRI dataset. In their analysis, the authors performed regional brain assessments and identified a negative correlation between cortical thickness measures in frontal regions and estimated brain age differences. Additionally, the predicted brain age

differences were correlated with cognitive performance using the Benton Face Recognition Test (BFRT), a neuropsychological measure assessing baseline visual memory. While their study provided valuable insights through quantitative and region-specific analyses, limitations included the exclusive reliance on the BFRT to associate age deviation with cognitive impairment, rather than linking it explicitly to disease-specific cognitive decline. Furthermore, the dataset had limited representation from older adults and entirely excluded participants younger than 18, thereby potentially restricting generalizability.

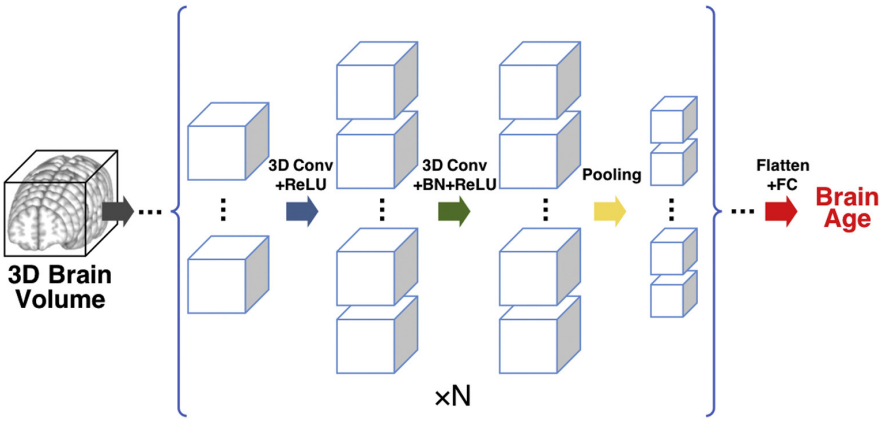


Figure 2.3: 3D VGGNet Architecture for brain age estimation used in X. Feng et al. (2020). (figure from X. Feng et al. (2020))

2.3.3 ResNet

ResNet is a specialized convolutional neural network architecture characterized by its hierarchical composition of residual blocks, which effectively mitigates training difficulties associated with vanishing gradient problems. This architecture utilizes skip connections that facilitate smoother gradient flow, enabling the successful training of deeper networks.

W. Shi et al. (2020) investigated fetal brain age estimation by implementing a convolutional neural network utilizing the ResNet architecture trained on T2-weighted MRI images. Additionally, attention maps were derived from the final layer of the CNN model, enabling identification of brain regions associated with predicted age discrepancies. Their results demonstrated significant correlations between predicted age differences and specific fetal abnormalities, including reduced head circumference and structural malformations.

The study in Fisch et al. (2021) introduced a simplified two-layer 3D CNN based on a ResNet architecture, requiring minimal preprocessing (brain extraction and image cropping) of T1-weighted MRI images. Transfer learning techniques were utilized to extract robust features, with the model trained on the GNC dataset and subsequently validated on the BiDirect, FOR2107/MACS, and IXI datasets to assess its generalization

capabilities. However, since this research focused solely on the developed 3D CNN model, it lacked comparative analyses against other established pretrained deep learning networks. Additionally, the relationship between the predicted age differences and specific neuroanatomical brain regions was not examined.

The seminal study conducted by Jónsson et al. (2019) proposed a multimodal 3D ResNet architecture for brain age estimation, leveraging inputs from T1-weighted MRI, GM, WM, and Jacobian maps. Additionally, sex and MRI scanner information were incorporated into the final layers to enhance predictive accuracy. To avoid random initialization of the CNN parameters, the authors initially trained an ensemble model on the Icelandic dataset, subsequently fine-tuning it using transfer learning on the IXI dataset. The performance of the trained model was further validated on the UK Biobank dataset, with final predictions derived through majority voting across multiple trained networks. The authors also conducted a genome-wide association analysis, identifying significant relationships between the estimated brain age difference and two genetic variants, rs1452628-T and rs2435204. Furthermore, a negative correlation between the predicted age difference and neuropsychological test performance was observed. Nevertheless, the primary limitation of this approach is the substantial computational complexity resulting from the use of multiple 3D CNN models as an ensemble.

Subsequently, Kolbeinsson et al. (2020) adopted a voxel-based methodology to quantify the contributions of specific brain regions—namely, the left amygdala, right hippocampus, left cerebellum, left insular cortex, left crus, and vermis—to brain aging. The authors employed permutation-based feature importance analysis to investigate associations between predicted brain age differences and clinical conditions such as hypertension, multiple sclerosis, systolic and diastolic blood pressure, and diabetes types I and II.

Another recent voxel-based study by (Ning, Duffy, et al., 2021) applied ResNet-based deep learning for identifying genetic correlates of brain aging. Employing a genome-wide association study approach, the research identified four genomic loci containing single nucleotide polymorphisms significantly linked to variations in predicted brain age. Despite this contribution, the investigation narrowly focused on genetic determinants, overlooking other critical influences such as lifestyle behaviors and health conditions known to accelerate brain aging. For instance, habits such as heavy smoking or excessive alcohol consumption significantly influence brain aging processes (Ning, L. Zhao, et al., 2020). Similarly, disorders like diabetes and schizophrenia have been associated with accelerated brain aging (Franke, Gaser, et al., 2013; Schnack et al., 2016), whereas physical exercise has demonstrated protective effects against age-related neural deterioration (A. F. Kramer, Erickson, and Colcombe, 2006; Larson et al., 2006).

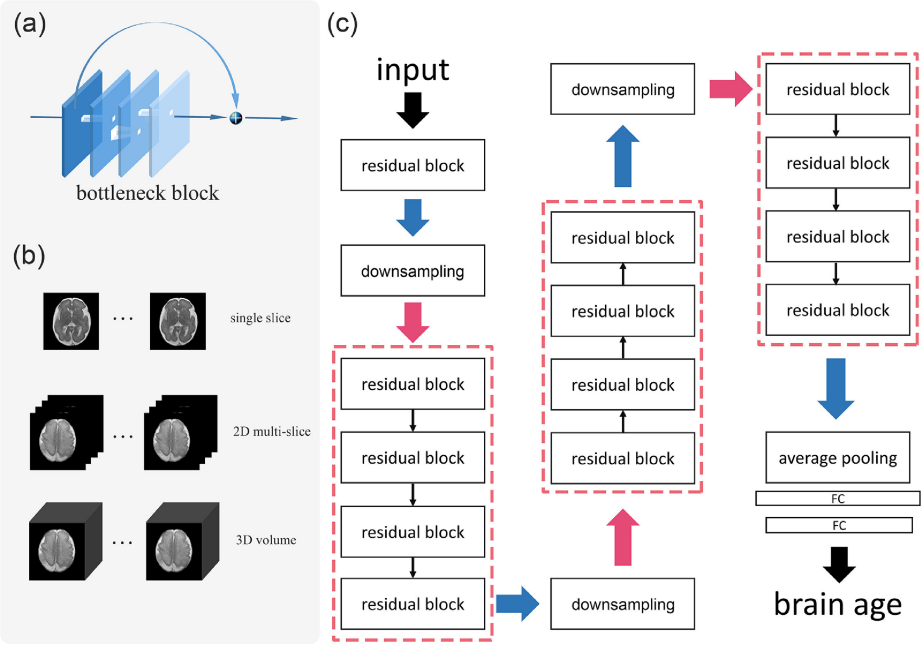


Figure 2.4: ResNet Architecture for brain age estimation used in W. Shi et al. (2020). (figure from W. Shi et al. (2020))

2.3.4 Transformer Frameworks

Although CNNs have been extensively utilized across various applications, they rely primarily on convolutional operations that extract features from small, localized neighborhoods, limiting their ability to represent global contextual relationships within the data (Y. Hu, H. Wang, and B. Li, 2022). Recently, Transformer-based architectures have emerged, capable of effectively modeling long-range dependencies within input sequences. Transformers have demonstrated remarkable performance, particularly in natural language processing and computer vision tasks, owing largely to their self-attention mechanism.

Inspired by these advantages, researchers introduced the Global–Local Transformer (GLT) (He, Grant, and Ou, 2021), a framework specifically designed to integrate both local and global feature information for brain age estimation tasks. The GLT framework seeks to unify detailed local context with broader global dependencies. Nonetheless, as highlighted by Y. Hu, H. Wang, and B. Li (2022), the global–local feature representations captured by GLT may represent only a subset of deeper local–local contextual information. To address this limitation and further enhance the representational power and generalization capabilities, Y. Hu, H. Wang, and B. Li (2022) proposed combining a pyramid-structured architecture with squeeze-and-excitation modules and self-attention mechanisms, thereby capturing both fine-grained local patterns and richer local–local interactions.

Building upon these developments, Cai, Yue Gao, and Liu (2022) adopted a multimodal

graph transformer approach to leverage comprehensive global and local multimodal features for more accurate brain age estimation. Their method incorporated cross-modal interactions, hierarchical multimodal feature fusion, and geometric learning-based feature aggregation, leading to improved estimation performance and more precise predictions of brain age. Turning to volumetric image models, Roibu (2023) examined 3D Swin Transformers alongside 3D CNNs on large, multi-map MRI representations, reporting that transformer backbones capture distributed age-related structure and that ensembling across maps improves accuracy and robustness.

Siegel et al. (2025) conducted a large-scale comparison on tens of thousands of UK Biobank T1w scans, adapting simple ViT and Swin Transformer backbones and benchmarking them against a strong ResNet baseline. The results suggested broadly comparable accuracy under matched training, with indications that transformer performance benefits disproportionately from further data scaling.

H. Zhao, Cai, and Liu (2024) proposes a multi-modal deep learning framework that learns modality-specific features from T2-weighted structural MRI and diffusion MRI in two parallel streams, then fuses them with a transformer module. The model targets chronological age estimation across the lifespan and also supports preterm and term classification, showing that attention-based fusion improves age prediction over single-modality baselines.

Wood, Townend, et al. (2024) trains five brain-age predictors on large routine clinical MRI covering T1-weighted, T2-weighted, T2-FLAIR, diffusion-weighted, and GRE T2* sequences, achieving strong accuracy. Crucially, the work demonstrates transfer learning to new sites, orientations, and sequences with limited fine-tuning data.

And Alp et al. (2024) presents a ViT-based pipeline tailored to 3D brain MRI for Alzheimer's disease classification, comparing transformer variants against CNN/LSTM-style baselines and detailing practical training choices that make ViTs competitive on routine structural MRI. Results show that transformer encoders can capture discriminative 3D patterns for neurodegenerative disease, with reproducible implementation details valuable for adapting to other 3D MRI tasks.

A two-stage ViT pretraining strategy was proposed for 3D neuroimaging (Cox et al., 2024): stage-1 learns anatomy-centric features on large unlabeled healthy-brain MRIs, and stage-2 refines spatial/context encoding—yielding a general 3D backbone. Evaluated on BraTS and ATLAS-v2, the pretrained model delivers sizable gains over fully supervised baselines and demonstrates strong sample efficiency, positioning it as a transferable representation for other MRI tasks beyond segmentation (such as brain-age).

Beyond adult cohorts, transformer modules have also been leveraged for neonatal brain-age estimation by integrating T2-sMRI and DTI, where the self-attention block serves as the fusion mechanism and yields precise age prediction together with development-related classification (H. Zhao, Cai, and Liu, 2024).

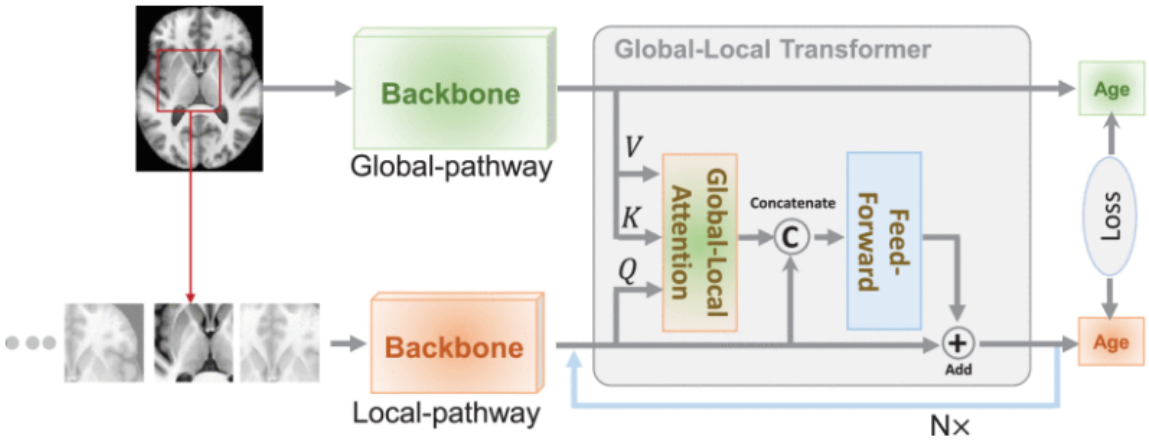


Figure 2.5: Global–Local Transformer Architecture in He, Grant, and Ou (2021). (figure from He, Grant, and Ou (2021))

2.3.5 Ensemble Learning

Ensemble learning integrates predictions from multiple neural network models to enhance robustness and predictive accuracy, albeit typically at the expense of increased computational complexity during training. Recently, ensemble deep learning approaches have gained traction among researchers seeking improved predictive performance in brain age estimation tasks.

Firstly, Hwang et al. (2021) developed an ensemble CNN model for estimating brain age using routine clinical T2-weighted spin-echo MRI scans. However, this work exhibited some notable shortcomings (Wood, Kafiabadi, et al., 2022). First, it computed final predictions by averaging across all slices without weighting, potentially reducing overall performance since not every slice contributes equally to predictive accuracy. Second, treating individual slices independently disregards potential nonlinear dependencies among spatially adjacent slices.

And Ballester et al. (2021) presented a slice-level brain age estimation approach combining convolutional neural networks (CNNs) and linear regression. Rather than processing complete volumetric brain images, this method utilized individual slices, allowing identification of specific brain regions contributing significantly to brain age prediction errors. Furthermore, the authors investigated how prediction accuracy was influenced by various factors, including slice orientation, position, participant age, sex, and MRI acquisition site. Their analysis demonstrated that certain MRI slices disproportionately impacted predictive errors. The authors also showed that employing stratified sampling for training and testing datasets effectively mitigated site-related biases and minimized sex-related prediction discrepancies. Notably, the overall accuracy of the predictions was sensitive to the selection of specific MRI slices.

Moreover, there are broader limitations inherent in many brain age estimation models that have not been explicitly designed for routine clinical MRI settings. Primarily, these models often rely on high-quality, isotropic or near-isotropic, volumetric T1-weighted MRI scans, which are infrequently acquired in standard clinical practice. Additionally, existing models typically rely on strictly standardized imaging protocols and selective participant criteria, contrasting significantly with the diverse and heterogeneous imaging data routinely encountered in hospitals, which involve multiple scanner manufacturers, varying protocols, and diverse patient populations. Lastly, many advanced brain age models require computationally demanding preprocessing steps such as spatial normalization, skull stripping, and bias field correction, further limiting their practical applicability in routine clinical contexts.

The initial contribution in this category was presented by Hofmann et al. (2022), a multilevel ensemble deep learning approach was proposed to improve model interpretability in neuroimaging contexts. The authors found that ensemble models consistently outperformed their individual component models, with notable contributions from voxels surrounding the ventricles, the meningeal boundary regions, and cortical sulcal structures, the latter being especially significant for older adults.

Additionally, Poloni, Ferrari, Alzheimer's Disease Neuroimaging Initiative, et al. (2022) introduced two CNN-based models specifically designed to estimate hippocampal age from T1-weighted MRI scans. By applying extensive data augmentation, the authors increased the effective training dataset size, enabling robust model training. The study further investigated associations between prediction errors and clinical conditions, comparing cognitively normal individuals with Alzheimer's disease (AD) and mild cognitive impairment (MCI) subjects. Statistical analyses revealed negative correlations between estimated age discrepancies and clinically derived measures. Importantly, the proposed model provided efficient inference, requiring only approximately 0.12 seconds per sample and maintaining an overall processing time under seven minutes.

Another notable example of voxel-based ensemble deep learning was presented by Levakov et al. (2020), who proposed combining predictions from multiple 3D CNN models trained on a substantial cohort of 10,176 participants for chronological age prediction. Their ensemble achieved a mean absolute error (MAE) of 3.07 years when evaluated on previously unseen data. Additionally, the authors identified cerebrospinal fluid cavities as biomarkers associated with brain atrophy and aging. By aggregating multiple explanation maps into a population-level visualization, the study highlighted ventricles and cisterns as key regions linked to early aging processes. Despite these strengths, the study's reliance on cross-sectional data rather than longitudinal data limited its ability to model individual trajectories of brain aging, focusing instead solely on between-subject variability.

In Kuo et al. (2021), the authors empirically examined how different combinations of input features affect the predictive accuracy of conventional machine learning frameworks for brain age estimation. Their study demonstrated that integrating multiple input features with

task-specific objective functions through an ensemble deep learning architecture significantly improved prediction accuracy.

Similarly, Couvy-Duchesne, Faouzi, et al. (2020) employed an ensemble comprising seven distinct classifiers, each trained on voxel-based GM and WM, along with vertex-level surface area metrics. Their results indicated that optimal predictive performance could be attained by combining the classifiers' predictions through linear regression-based weighting strategies. Furthermore, the integration of predictions using a random forest algorithm further enhanced model accuracy. However, a common limitation associated with such ensemble methodologies is their increased computational overhead arising from the simultaneous training of multiple independent neural networks.

Sun et al. (2023) trained multiple lightweight 3D CNNs on T1-weighted MRI and combined their predictions through bagging and weighted averaging, showing that the ensemble reduced variance and improved generalizability across sites in the UK Biobank. Similarly, X. Li et al. (2024) proposed a hybrid ensemble integrating CNNs with transformer-based backbones; tested on over 4,000 subjects, their results demonstrated lower mean absolute error (MAE) compared to any single constituent model, highlighting the complementary strengths of convolutional and attention-based architectures.

In Zeineldin et al. (2024), authors introduces TransXAI, a hybrid CNN–Transformer framework for multi-modal brain-tumor segmentation that emphasizes interpretability. Beyond competitive segmentation accuracy, the method provides surgeon-readable post-hoc heatmaps without modifying the trained network or sacrificing performance.

Beyond conventional averaging, more sophisticated fusion strategies have also emerged. Yang Gao et al. (2024) introduced a stacking framework where base models included CNNs, graph neural networks, and regression trees trained on multimodal MRI features, with a meta-learner combining outputs into final age predictions. This approach yielded not only higher accuracy but also improved fairness across sex and age subgroups, addressing a common limitation of earlier ensemble approaches.

2.3.6 Others

In addition to the previously discussed deep learning models, several researchers have explored alternative advanced approaches for brain age estimation. Here, we highlight representative contributions.

Lin et al. (2021) introduced a hybrid framework employing a pretrained 2D AlexNet model (Iandola et al., 2016) for feature extraction from GM images derived from T1-weighted MRI scans. They subsequently applied principal component analysis for feature dimensionality reduction and utilized a relevance vector machine with a polynomial kernel for the final classification step. Their experimental results on Alzheimer's disease datasets showed notable predictive age deviations, with an average discrepancy of 3.13 years in patients with mild cognitive impairment and 6.08 years in patients diagnosed with

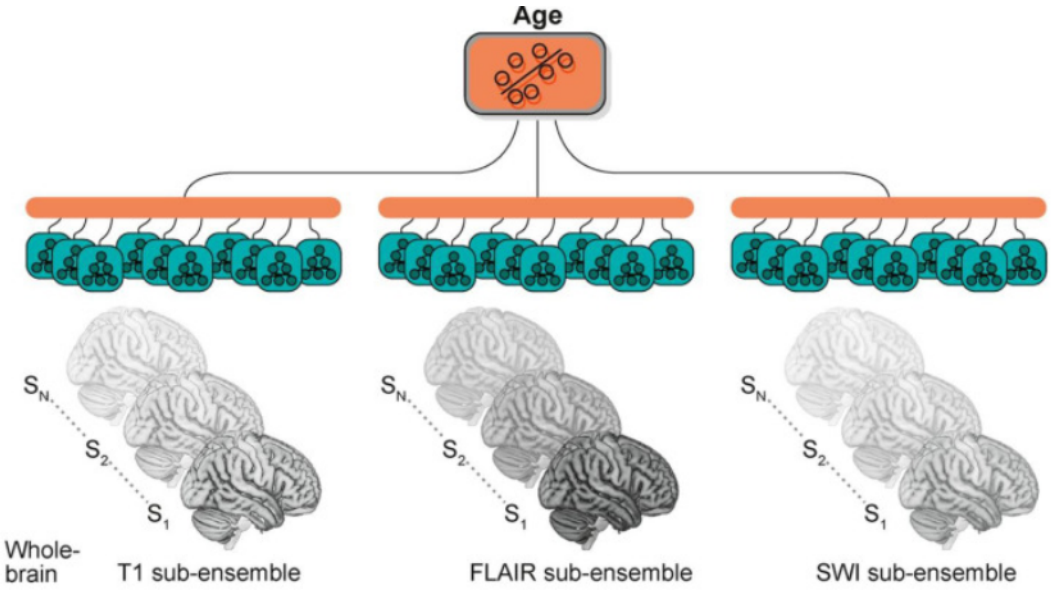


Figure 2.6: Multi-Modal Ensemble Learning in Hofmann et al. (2022). (figure from Hofmann et al. (2022))

AD. Despite the promising outcomes, the study had certain limitations. Notably, the authors exclusively evaluated performance using the pretrained AlexNet architecture, a model typically characterized as a black-box approach that lacks inherent interpretability. Additionally, comparative performance analyses involving other established pretrained CNN models were not conducted, limiting broader conclusions regarding the effectiveness of their AlexNet-based method.

A further innovative contribution was presented by Lombardi et al. (2021), who developed an explainable deep learning model focused explicitly on morphological feature extraction for brain aging. The authors explored two prominent explainability techniques—SHapley Additive exPlanations (SHAP) (Lundberg, 2017) and Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro, Singh, and Guestrin, 2018)—to identify reliable morphological biomarkers of aging. Their experiments indicated that SHAP provided superior interpretability for elucidating age-related morphological changes. Nonetheless, to more effectively leverage three-dimensional MRI data at the voxel level, alternative explainability methods such as CNN-derived saliency maps or layer-wise relevance propagation (LRP) could potentially offer improved visualization of aging-related regions within the brain. Although a correlation analysis between extracted morphological features and chronological age was performed, the authors acknowledged that further quantitative investigations are required to precisely determine the brain regions most critical to predicting biological age.

Tak et al. (2024) introduces BrainIAC, a self-supervised foundation model trained with contrastive learning on about 32,000 unlabeled brain MRIs drawn from a curated pool of

35 datasets (total 48,519 scans). As a generic 3D vision encoder for MRI, BrainIAC is adapted to multiple downstream tasks—including brain-age prediction—and consistently outperforms scratch training and prior pretrained baselines, particularly in low-data and out-of-distribution settings.

Caro et al. (2023) presents BrainLM, a foundation model trained with masked-prediction on 6,700 hours of fMRI recordings to learn temporo-spatial dynamics of brain activity. The model supports fine-tuning and zero-shot use, accurately predicting clinical variables (including age), forecasting future brain states, and generalizing to external cohorts not seen in training.

When considering 3D-based approaches, the diversity of advanced models reported in the literature is extensive. Wood, Kafiabadi, et al. (2022) introduced a voxel-based DenseNet model tailored specifically for estimating brain age from routine clinical head MRI examinations. The utilization of routine clinical imaging data ensured diverse representation across scanner types and acquisition protocols, reflecting realistic clinical conditions. Nevertheless, the authors highlighted two critical limitations: first, the model was exclusively trained on radiologically normal images, and thus its performance on brains exhibiting significant pathologies remains unknown; second, although interpretability was attributed generally to brain parenchyma, the study lacked a systematic analysis identifying specific neuroanatomical features driving model predictions.

Mouches, Wilms, Rajashekar, Sonke Langner, et al. (2021) presented a novel deterministic autoencoder framework that simultaneously achieved brain age estimation and the generation of age-specific brain templates. These age-conditioned templates illustrated age-related morphological changes such as ventricular enlargement and increased sulcal width. Expanding upon this work, the same authors (Mouches, Wilms, Rajashekar, Sönke Langner, et al., 2022) later proposed an autoencoder-based multimodal model combining structural MRI and angiographic imaging data. They employed saliency maps to elucidate contributions from cortical, subcortical, and arterial structures, concluding that integrating arterial information with brain tissue significantly enhanced predictive accuracy. Notably, structures including the lateral sulcus, fourth ventricle, and medial temporal regions emerged as critical morphological markers for age estimation.

The study conducted by He, Yanfang Feng, et al. (2022) introduced an innovative regression framework based on deep relational learning. This model learned nonlinear relationships between pairs of brain images through simultaneous relational regression and feature extraction tasks. EfficientNet was utilized for feature extraction, while transformer-based architectures modeled inter-image relationships. Evaluations encompassed various experimental conditions, such as estimating brain age from different image pairs, known-age reference comparisons, and identical-image pair analyses. Achieving a notably low MAE of 2.38, the authors argued that their deep relational learning model delivered superior generalization performance compared to contemporary state-of-the-art brain age estimation methods.

One noteworthy study by Popescu et al. (2021) employed a U-Net architecture utilizing voxel-wise MRI data to enable local predictions of brain age. This approach allowed the researchers to analyze structural distinctions between mild cognitive impairment and Alzheimer's disease patients. Additionally, the reliability of the local brain age predictions across scanners and within scanner variations was assessed. Results indicated that certain subcortical regions, including the accumbens, putamen, pallidum, hippocampus, and amygdala, were notably discriminative, exhibiting significant local predicted age differences as measured by Cohen's d values. However, a major drawback of this method was its dependence on a healthy reference population specific to each MRI site; variability due to scanner differences could result in local predicted age distributions deviating significantly from zero, complicating interpretability.

Varatharajah et al. (2018) leveraged transfer learning via a pretrained 3D Inception-V1 (Couvy-Duchesne, Strike, et al., 2020) model as a feature extractor, combined with regression and categorical (bucketed) classification tasks to estimate brain age.

Finally, J. Lee et al. (2022) adopted a modified 3D-DenseNet architecture optimized with Adam and trained using MAE loss, focusing specifically on normal aging and dementia. The authors provided age- and modality-specific interpretability maps using occlusion sensitivity analysis, which masks selected brain regions to quantify their impact on prediction accuracy. Their findings indicated that posterior regions, particularly around the posterior cingulate cortex, predominantly influenced age predictions in younger age groups (30–40 and 40–50 years). Conversely, in older groups (50–60 and 70–80 years), inferior frontal, orbitofrontal, and olfactory cortex regions played a more significant role. Additionally, their analysis demonstrated greater sensitivity of metabolic data over structural MRI for predicting brain age. However, the scope of their evaluation was limited to neurodegenerative conditions, neglecting chronic systemic or vascular diseases known to exhibit distinct brain-aging patterns.

2.4 Explainable Artificial Intelligence (XAI)

The rapid adoption of deep learning in medical imaging has raised concerns about the 'black-box' nature of predictive models, making interpretability an essential component for both scientific understanding and clinical translation. Explainable Artificial Intelligence (XAI) aims to bridge this gap by providing tools that help understand the decision-making process of complex models. In neuroimaging, XAI has been particularly important in tasks such as disease classification and brain age estimation, where identifying relevant brain regions is as crucial as obtaining accurate predictions.

Among post-hoc explanation methods, two widely used model-agnostic approaches are SHAP and LIME. SHAP (SHapley Additive exPlanations) is based on cooperative game theory and attributes the contribution of each feature to the model output based on Shapley values (Lundberg, 2017). In brain imaging, SHAP has been applied to highlight the relative

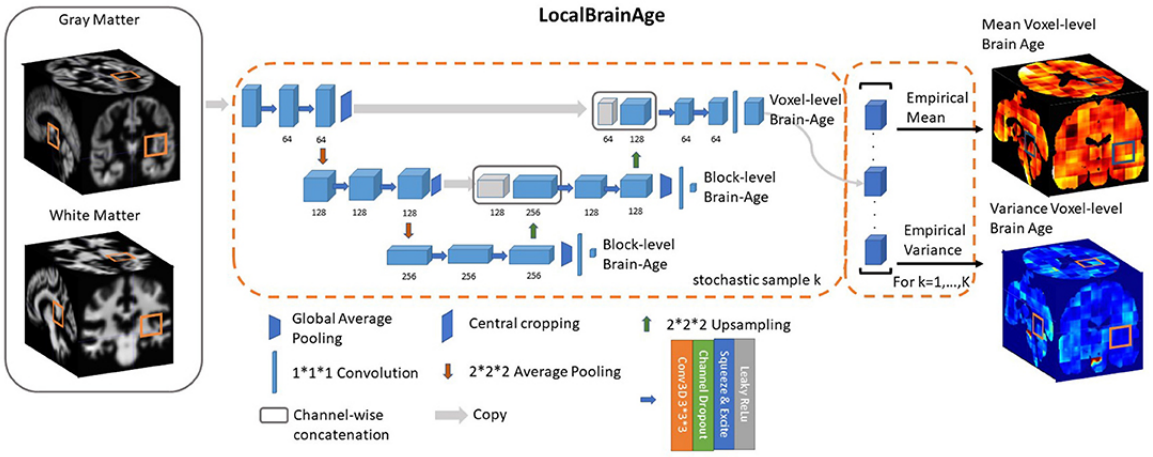


Figure 2.7: U-Net for estimating local brain age in Popescu et al. (2021). (figure from Popescu et al. (2021))

importance of different structural or functional features, thereby quantifying how specific regions influence predicted brain age or disease status. LIME (Local Interpretable Model-agnostic Explanations) approximates a complex model locally with a simpler, interpretable surrogate (such as a linear regression), providing insight into the features that most strongly influence a given prediction (Ribeiro, Singh, and Guestrin, 2016b).

For CNN applied to MRI, gradient-based saliency methods are always employed. Grad-CAM (Gradient-weighted Class Activation Mapping) visualizes class-discriminative regions by computing gradients of the output with respect to feature maps, producing heatmaps that indicate spatial importance (Selvaraju et al., 2017). This method has been adapted to regression tasks such as brain age estimation, where heatmaps can highlight cortical and subcortical regions most influential in predicting age.

Occlusion analysis represents a perturbation-based strategy: input images are systematically masked in localized regions, and the corresponding changes in prediction are measured (Zeiler and Fergus, 2014). This method directly quantifies the contribution of brain regions to prediction accuracy and has been widely used in neuroimaging to validate findings from gradient-based XAI methods. For example, occlusion has been employed in brain age studies to confirm the influence of cortical thinning or subcortical volume loss highlighted by attention-based models.

Together, these XAI methods provide complementary perspectives on model interpretability. In the context of this dissertation, occlusion analysis is employed both as a baseline and as a validation for the built-in interpretability mechanisms developed in the Triamese-ViT framework, ensuring that explanations are robust, biologically meaningful, and aligned with neuroscientific evidence.

2.5 Fairness and Bias in Brain Age Estimation

Despite recent advances in brain age estimation, concerns remain regarding fairness and potential biases in model performance across demographic groups. Many existing models have been trained on datasets with demographic imbalances, most notably with respect to sex, race, and age distribution. This imbalance often leads to systematic deviations in prediction accuracy: for example, models trained predominantly on young or middle-aged adults tend to overestimate the brain age of elderly individuals, while models biased toward one sex may yield lower accuracy for the underrepresented group (Beheshti, Nugent, et al., 2019). Similar issues have been observed in racially unbalanced datasets, where differences in brain morphology across ancestry groups may be conflated with aging effects, raising concerns about the generalizability of brain age models in diverse populations (Cole and Franke, 2017; Bashyam et al., 2020).

A number of studies have highlighted these limitations. Bashyam et al. (2020), using a large international dataset of more than 16,000 individuals across multiple sites, showed that prediction errors vary systematically by age and sex, with higher mean absolute errors in elderly cohorts and underrepresented demographic subgroups. Dibaji et al. (2023) further demonstrated sex-specific biases in brain age models, where separate models trained on male and female samples yielded different regional patterns of importance, suggesting that pooled models may mask group-specific trajectories. Importantly, these findings point to the fact that a single global model may not adequately capture heterogeneous aging processes across subgroups.

Critical evaluations also emphasize that fairness in brain age estimation is not limited to prediction error but extends to downstream interpretation. For instance, biased estimates of BAG can distort associations with clinical outcomes, leading to spurious links between accelerated aging and disease prevalence in certain populations (Beheshti, Nugent, et al., 2019). Moreover, demographic biases can undermine the use of brain age as a biomarker in clinical trials, where fair performance across sexes, ethnicities, and age ranges is essential for reliable stratification.

In conclusion, fairness and bias remain critical challenges in brain age research. Models trained on unbalanced datasets risk propagating demographic disparities, thereby limiting their reliability in cross-cohort and clinical applications. To address these challenges, we indeed require technical innovation. To respond to this issue, we introduce u-DemAI, which is a framework that has great performance on brain age estimation, not only on its high prediction accuracy, but also the fairness, which has consistent predictive performance across different age groups.

2.6 Discussion

The literature on brain age estimation demonstrates steady progress from early CNNs to

Table 2.1: Summary of representative models for brain age estimation, their advantages and disadvantages.

Model	**Data**	**Advantages**	**Disadvantages**
3D CNN (Cole and Franke, 2017)	T1w	Only used GM segmentation and achieved high accuracy.	Did not analyze the relationship between BAG and specific regions. Did not consider bias in prediction.
VGGNet (Dinsdale et al., 2021)	T1w	Investigated correlations between BAG and non-imaging variables, including lifestyle, physiological and medical history, and self-reported mental health conditions.	Analyzed correlations with image-derived phenotypes, but they compress voxel-level information, limiting sensitivity compared to direct analyses on T1w images. Did not consider bias in prediction.
ResNet (Fisch et al., 2021)	T1w	Trained on large datasets.	Did not examine the relationship between BAG and specific neuroanatomical regions. Did not consider bias in prediction.
Vision Transformer (Cai, Yue Gao, and Liu, 2022)	DTI+T1w	Leveraged multimodal data and analyzed the relationship between BAG and specific neuroanatomical brain regions.	Relied on occlusion sensitivity analysis, requiring additional processing steps. Did not consider bias in prediction.
Ensemble Model (Yang Gao et al., 2024)	T1w	Considered fairness across sex and age subgroups in prediction.	Treat each constituent model with fixed or simple weighting schemes. Did not examine the relationship between BAG and specific neuroanatomical brain regions.

more advanced architectures such as ResNet and ViTs, as well as ensemble learning strategies. Table 2.1 provides a summary of representative models in this field.

Classical CNN-based models, including VGG and ResNet, have shown strong capability in extracting local features from sMRI and achieving reasonable prediction accuracy. However, they are limited in capturing global structural relationships, which are critical for modeling distributed patterns of brain aging. ViT-based approaches address this limitation by employing self-attention mechanisms to integrate information across the whole brain, thereby improving the modeling of global dependencies. Nevertheless, the application of attention mechanisms to XAI remains underexplored, and their value for interpretability has not been firmly established. Ensemble learning methods, in contrast, improve prediction stability and accuracy by combining diverse models.

While recent deep learning models have achieved impressive predictive accuracy, they often exhibit limited generalizability across populations and age ranges. Many approaches rely on fixed or simplistic schemes that fail to capture the nonlinear and heterogeneous nature of brain aging. In reality, age-related structural changes in the brain follow distinct trajectories across developmental and late-life stages, and the same model parameters may not adequately represent these differences. Consequently, models optimized on specific age groups tend to overfit to the dominant demographic in the training data, leading to reduced performance when applied to underrepresented cohorts such as very young or elderly individuals.

Furthermore, although these methods enhance accuracy, they frequently overlook issues of bias and fairness. Predictive performance remains sensitive to demographic imbalances—particularly in sex, ethnicity, and age distribution—which can result in systematic overestimation or underestimation of brain age in certain subgroups. These biases limit the clinical applicability of such models, as consistent and equitable performance across diverse populations is essential for reliable biomarker development. Addressing these limitations requires adaptive modeling strategies that explicitly account for demographic variability and employ fairness-aware optimization to ensure robust generalization across age groups and populations.

In response to these limitations, the present work introduces three new approaches. First, the nonlinear Age-Adaptive Ensemble (nl-AAE) advances traditional ensemble methods by dynamically adjusting the weights of constituent models according to the chronological age of the subject. This adaptive design enables the ensemble to capture age-specific structural patterns and yields higher predictive accuracy than both individual models and conventional ensembles. Second, the Triamese-ViT leverages the strengths of transformer architectures while addressing the challenge of interpretability. By extracting complementary features from three orthogonal orientations of sMRI and integrating them into a unified prediction, Triamese-ViT achieves high accuracy with built-in interpretability. Crucially, its attention-based explanations have been validated against traditional XAI methods such as occlusion analysis, ensuring that the interpretability is both intrinsic and reliable. This built-in

mechanism has also been applied to investigate the relationship between BAG and specific brain regions, providing novel insights into normal aging and ASD diagnosis. Finally, the u-DemAI framework explicitly addresses fairness and bias, dimensions often overlooked in prior models. By reducing bias in accuracy across age ranges and enabling a democratic self-updating process in which users contribute to model refinement, u-DemAI promotes both fair performance and broader accessibility.

2.7 Conclusion

In this chapter, we reviewed previous research on brain age estimation. We first introduced the background of neuroimaging analysis, followed by a discussion of popular deep learning models applied to brain age prediction, the role of XAI in this field, and work examining fairness and bias. We then highlighted the limitations of existing approaches and outlined how our proposed methods aim to address these challenges. The following three chapters present the details of our contributions: the nl-AAE, the Explainable Triamese-ViT, and the u-DemAI framework. For each model, we describe the architecture, report experimental results, and provide an analysis of its implications.

Chapter 3

Nonlinear Age-Adaptive Ensemble Learning

3.1 Introduction

Recent advancements in deep learning have transformed brain age estimation by enabling models to extract complex features from neuroimaging data with high accuracy. Deep learning approaches have outperformed traditional methods in detecting subtle structural changes in the brain and identifying deviations from normative aging trajectories. Building on these developments, this chapter introduces a novel framework, the Nonlinear Age-Adaptive Ensemble Learning model (nl-AAE), designed specifically for brain age estimation.

The key innovation of nl-AAE lies in its nonlinear age-adaptive ensemble mechanism, which integrates multiple independent models into a unified predictor. Unlike conventional ensembles that assign fixed or static weights, nl-AAE dynamically adjusts the contributions of its constituent models according to the chronological age of the input. This enables the ensemble to capture age-specific structural patterns and to model brain characteristics across the lifespan with greater accuracy. In this work, four independent models were considered: GoogLeNet, ResNet, Support Vector Regression (SVR), and a custom-designed CNN. The nonlinear weighting strategy allows the ensemble to adapt flexibly to age-related variability and to leverage the complementary strengths of each model.

We evaluate nl-AAE using the PAC 2019 competition dataset and benchmark its performance against its constituent models and other SOTA models. The results demonstrate that nl-AAE delivers superior predictive accuracy, highlighting its potential as a powerful tool for assessing brain health in clinical trials of neuroprotective therapies, identifying individuals at risk of accelerated cognitive decline, and offering insights into the downstream consequences of aging-related diseases. Its enhanced accuracy compared to existing approaches also underscores its relevance for applications such as Alzheimer's disease detection, traumatic brain injury assessment, schizophrenia diagnosis, and pharmaceutical

evaluation.

The remainder of this chapter is organized as follows. We first introduce the dataset used in our experiments. Next, we describe the independent models employed in nl-AAE, including their structure and hyperparameters. We then present the design and mechanism of the nl-AAE framework, followed by the experimental results, their analysis and discussion. Finally, we conclude the chapter by synthesizing insights, evaluating performance, identifying limitations, and suggesting directions for future improvement.

3.2 Preliminary

3.2.1 Dataset

The dataset utilized in this study is derived from (Cole and Franke, 2017) and consists of 2,641 healthy individuals' structural MRI (sMRI) scans, along with demographic attributes such as age and gender. The age range of the participants spans from 16 to 90 years, with a mean age of 35.8 years and a standard deviation of 16.2 years.

Among the participants, 53% are female and 47% are male. The mean age of female participants is 37 years, with a standard deviation of 17.2 years, whereas the mean age of male participants is 34.6 years, with a standard deviation of 14.9 years. The age distribution of the dataset is depicted in Figure 3.1.

It is worth noting that the dataset presents an imbalanced age distribution, with a relatively lower representation of older individuals and a higher proportion of younger participants. Additional details regarding the dataset and its composition are provided in (Cole and Franke, 2017). Such imbalance can negatively affect prediction accuracy, as models trained on over-represented younger samples tend to generalize less effectively to older age groups. Our proposed nl-AAE framework addresses this issue through its age-adaptive design: instead of training a single ensemble across the entire age range, nl-AAE constructs separate ensemble models for different age groups and assigns group-specific weights to the base learners. This strategy allows the model to better capture age-specific structural patterns and reduces the degradation in predictive accuracy caused by the under-representation of older individuals. Consequently, nl-AAE achieves more consistent performance across age ranges, as reflected by the reduced error disparity between younger and older groups in our results.

We selected the PAC 2019 dataset for several reasons. First, it is one of the largest publicly available datasets specifically curated for the task of brain age estimation, including over 2,600 T1w sMRI scans with chronological age labels spanning from 16 to 90 years. This wide age range makes it particularly suitable for developing and evaluating age-adaptive models. Second, the dataset has been extensively used in previous benchmark studies (Cole and Franke, 2017; Couvy-Duchesne, Faouzi, et al., 2020; Soch, 2020), enabling direct comparison of our results with existing approaches. Third, the data were preprocessed in a standardized

manner, which ensures consistency across participants and facilitates reproducibility.

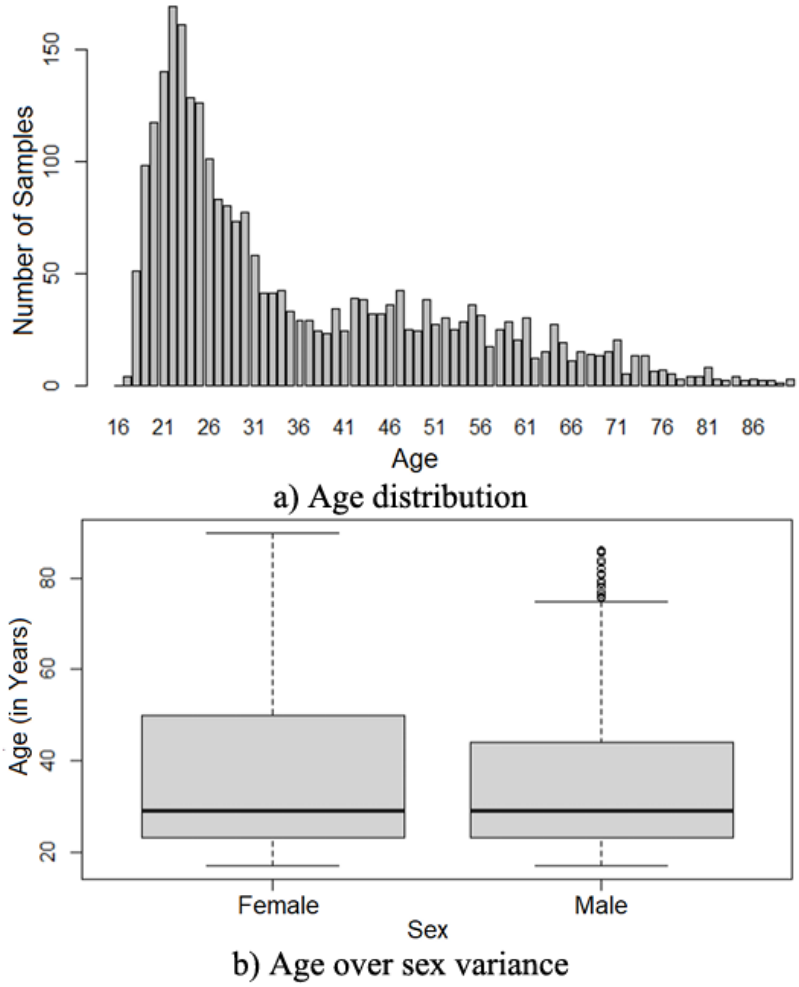


Figure 3.1: **a** presents the Age distribution of the dataset and **b** presents the sex distribution of the dataset.

3.2.2 Data Features

In this study, we utilize two distinct types of input data for our models: Gray Matter and White Matter Maps and Surface-Based Processing of Gray Matter.

The Gray Matter and White Matter Maps were provided by the Predictive Analysis Challenge (PAC) organization. Preprocessing of sMRI data involved nonlinear registration using the MNI152 space, which is a standard reference space widely used in neuroimaging to align and compare different brain scans across individuals and studies. The images were

subsequently segmented into different tissue types by a neuroimaging analysis software called SPM12 (Penny et al., 2011). Each tissue type was represented as a separate map, which was further smoothed with a 4-mm kernel. For additional details on this preprocessing method, refer to (Cole and Franke, 2017). These processed maps serve as input data for the self-defined CNN, ResNet, and GoogLeNet models used in our project.

For Surface-Based Processing of Gray Matter, we extracted vertex-wise measurements of cortical thickness and surface area from sMRIs using FreeSurfer 6.0 (Fischl, 2012). Additionally, vertex-wise features for the thickness and surface area of seven subcortical nuclei were extracted following the ENIGMA-shape protocol (Gutman, Madsen, et al., 2013; Gutman, Yalin Wang, et al., 2012). This preprocessing yielded approximately 650,000 gray matter measurements per individual. The method employed was previously validated by Baptiste Couvy-Duchesne et al. (Couvy-Duchesne, Strike, et al., 2020), who demonstrated that these processed features exhibit a strong correlation with age. So here we only used this dataset for the input of SVR. Because SVR is the only classical machine learning approach we used in this experiment. And classical machine learning approaches require explicit feature representations, and vertex-wise measures of cortical thickness and surface area provide biologically meaningful and strongly age-related predictors. In contrast, deep learning models such as CNNs, GoogLeNet, and ResNet are capable of automatically learning hierarchical representations directly from raw sMRI data, thereby obviating the need for hand-crafted surface-based features. By restricting the FreeSurfer-derived features to SVR, we ensure that each model operates on data in a manner consistent with its methodological strengths: SVR benefits from morphometric inputs, whereas deep neural networks exploit the full spatial richness of sMRI.

3.2.3 Basic Independent Models

The advancements in deep neural networks (Z. Jiang, P. L. Chazot, et al., 2019; Chiang et al., 2020; Mehboob et al., 2022) have significantly contributed to the development of medical biometrics (R. Jiang, Crookes, et al., 2022), particularly in health diagnostics and the understanding of neural function (R. Jiang and Crookes, 2019) and dysfunction (Z. Jiang, Yunpeng Wang, et al., 2022) in the human brain. In this study, we leverage deep learning techniques in combination with ensemble learning approaches to improve the accuracy of brain age estimation.

Before detailing the architecture of our ensemble model, we first introduce the fundamental components of our proposed framework. The choice of CNN, GoogLeNet, ResNet, and SVR as base learners was guided by their demonstrated utility in prior brain age studies. Shallow 3D CNN architectures have been widely applied to T1-weighted sMRI for capturing local structural patterns of gray and white matter (Cole and Franke, 2017). GoogLeNet introduces multi-scale feature extraction through its inception modules and has been successfully adapted to brain age prediction tasks (Couvy-Duchesne, Faouzi,

et al., 2020). ResNet incorporates residual connections that alleviate the vanishing-gradient problem and enable deeper architectures, which have shown competitive performance in neuroimaging-based age estimation (Fisch et al., 2021). Finally, SVR is a classical machine learning method that has consistently been used as a benchmark in brain age estimation (Couvy-Duchesne, Faouzi, et al., 2020), relying on explicitly defined morphometric features that exhibit strong correlations with chronological age. By combining these models, we integrate complementary strengths: CNNs, GoogLeNet, and ResNet automatically learn hierarchical representations directly from volumetric sMRI, while SVR leverages features with established biological relevance. This diversity provides a robust foundation for the ensemble framework.

Besides, the architecture and hyperparameters of our basic independent models were selected a priori based on prior studies of brain age estimation (Cole and Franke, 2017; Couvy-Duchesne, Faouzi, et al., 2020; Fisch et al., 2021), and these configurations achieved optimal performance in their experiments.

3.2.3.1 Convolutional Neural Networks (CNN)

The convolutional neural network (CNN) was implemented in Keras with TensorFlow as the backend. The architecture comprises seven sequential convolutional blocks, a design whose effectiveness for brain age estimation has been previously demonstrated by Couvy-Duchesne, Faouzi, et al. (2020). The seven sequential blocks includes:

- The first five blocks each include a 3D convolutional layer ($3 \times 3 \times 3$), followed by Batch Normalization, an RELU activation function, and a Max Pooling layer.
- The sixth block contains a dropout layer.
- The seventh block includes a fully connected layer.

The input to the model is a 3D volumetric image of size $121 \times 145 \times 121$ pixels, which is progressively reduced by the convolutional layers to 128 feature maps of size $4 \times 5 \times 4$. The final fully connected layer further reduces these feature maps to generate the predicted age.

The model is trained on two-channel input data, formed by concatenating gray matter and white matter maps. The training process utilizes the Mean Absolute Error (MAE) as the loss function and Adam optimizer with the following hyperparameters:

- Learning rate: 0.001
- Weight decay: 10^{-4}
- $\beta_1 = 0.9$, $\beta_2 = 0.999$

3.2.3.2 GoogLeNet (Inception V1)

The GoogLeNet (Inception V1) architecture, previously utilized for brain age estimation (Couvy-Duchesne, Faouzi, et al., 2020), is employed in this study. The model consists of:

- A stem network comprising an input layer, a convolutional filter, a max-pooling layer, two additional convolutional filters, another max-pooling layer, and an output layer.
- Two inception modules, followed by a max-pooling layer.
- Five additional inception modules, two of which are connected to an auxiliary regression head.
- Another max-pooling layer, followed by two more inception modules.
- An average pooling layer, a dropout layer, and a fully connected layer.

To adapt GoogLeNet for regression tasks, the softmax layer was replaced with a fully connected output layer. The convolutional filters in this model consist of an input layer, a convolutional layer, batch normalization, a ReLU activation, and an output layer.

The auxiliary regression heads, designed to mitigate the vanishing gradient problem, are composed of:

- An input layer, followed by an average pooling layer, a convolutional filter, and a fully connected layer.
- A ReLU activation layer, a dropout layer, and another fully connected layer leading to the output.

The input data for this model consists of 3D maps of gray matter density with dimensions $121 \times 145 \times 121$ pixels, while the output represents the predicted age. The model is trained using the MAE loss function and Adam optimizer with the following settings:

- Learning rate: 0.001
- Batch size: 8

3.2.3.3 ResNet

The ResNet model structure was always used in the previous brain research, for example, L. Hu et al. (2023) collected 658 T1-weighted MRI scans from children aged 0–3 years and trained a deep ResNet-style residual network to predict brain age in this very early developmental window, achieving a high correlation (0.91) between predicted and chronological age. Their model is specialized for infants and toddlers and focuses on raw MRI input with minimal preprocessing. However, this approach has limitations when scaled

to lifespan brain age estimation: it is optimized for a narrow age range and thus may not generalize across older populations; it does not explicitly adapt for age-group heterogeneity; it lacks integrated fairness control (for example, adjusting biases across age brackets).

The ResNet structure we used here is based on (Fisch et al., 2021), it got great performance in their research:

- Five residual blocks, each followed by a max pooling layer with a $3 \times 3 \times 3$ kernel size and a stride of $2 \times 2 \times 2$.
- A fully connected block for final prediction.

Each residual block consists of a 3D convolutional layer with a stride of $1 \times 1 \times 1$ and a kernel size of $3 \times 3 \times 3$, followed by batch renormalization and an ELU activation function. Additionally, the input signal to the residual block is directly added to the output of a later layer within the block, forming the characteristic residual connection.

The fully connected block is a Multilayer Perceptron (MLP) with:

- An input layer with $128 \times 4 \times 5 \times 4 = 10,240$ neurons.
- A hidden layer (FC1) with 256 neurons, using an RELU activation function.
- A dropout layer with a keep rate of 0.8, following the hidden layer.
- A final output layer (FC2) that performs linear regression on the hidden layer features.

The model is trained using 3D maps of gray matter density as input and Mean Absolute Error (MAE) as the loss function. The training process is optimized using Adam optimizer with the following hyperparameters:

- Learning rate: 0.001
- Weight decay: 10^{-4}
- $\beta_1 = 0.9$, $\beta_2 = 0.999$

3.2.3.4 Support Vector Regression (SVR)

Although deep neural networks have recently dominated brain age estimation, SVR remains an important and complementary component in our ensemble. First, SVR relies on explicit morphometric features, such as cortical thickness and surface area extracted by FreeSurfer, which have strong biological validity and established associations with aging (Couvy-Duchesne, Strike, et al., 2020). This contrasts with deep learning models, which learn hierarchical features from volumetric data but may obscure direct neuroanatomical

interpretations. Second, classical machine learning approaches such as SVR are less data-hungry and generally more robust in scenarios with limited sample sizes or imbalanced data distributions, which are common challenges in neuroimaging studies. Third, SVR has historically been used as a benchmark in brain age estimation research (Franke, Ziegler, et al., 2010), for example, J. Li, L. C. W. Lam, and Lu (2024) used SVR to train a brain age model on T1-MRI and further used mutual information analysis to explain the relationship between regional morphological features and estimated brain age. While his explanatory analysis was insightful, his predictive model was relatively limited, and his training objectives did not account for age bias and fairness across subgroups. Therefore, including it in our research is valuable for comparison with prior work. By integrating SVR with deep learning models, the ensemble benefits from both biologically grounded, handcrafted features and automatically learned representations, thereby enhancing robustness and interpretability across diverse age groups.

In this study, we employ SVR with a radial basis function (RBF) kernel to address the brain age estimation task. The input data consists of Surface-Based Processing of Gray Matter, containing approximately 650,000 gray matter measurements per individual, while the output represents the predicted age.

The SVR model is implemented using the Scikit-Learn package in Python. To ensure high predictive accuracy, the model is trained for over 300 epochs.

Table 3.1: Summary of independent models used in nl-AAE.

Model	**Input Type**	**Hyperparameters**
CNN	sMRI	Learning rate = 0.001, batch size 8 weight decay = 10^{-4}, optimizer = Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$), dropout rate = 0.2, epochs=300
GoogLeNet	sMRI	Learning rate = 0.001, batch size 8, optimizer = Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$), dropout rate = 0.2, epochs=300
ResNet	sMRI	Learning rate = 0.001, batch size 8, weight decay = 10^{-4}, optimizer = Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$), dropout rate = 0.2, epochs= 300
SVR	Surface-based features	Kernel = RBF, hyperparameters (C, γ, ϵ) tuned via cross-validated grid search

3.3 Age-Adaptive Ensemble Model

3.3.1 Fundamentals of Ensemble Learning

Ensemble learning enhances predictive performance by constructing and integrating multiple individual models, often referred to as multi-classifier systems or committee-based learning. The general framework of ensemble learning involves generating a set of individual learners and subsequently combining their outputs using a specific fusion strategy. Empirical studies have demonstrated that ensemble learning generally achieves superior generalization performance compared to individual models (Kuncheva and Whitaker, 2003; Sollich and Krogh, 1995).

There are six widely recognized types of ensemble learning models: Bayes Optimal Classifier, Boosting, Bootstrap Aggregating (Bagging), Bayesian Model Averaging (BMA), Bayesian Model Combination (BMC), and Stacking.

The Bayes Optimal Classifier is based on Bayesian decision theory and constructs an ensemble of all possible hypotheses within the hypothesis space (Friedman, Geiger, and Goldszmidt, 1997). It remains a widely used supervised learning approach, particularly for classification tasks.

AdaBoost is an algorithm designed to enhance weak learners into strong classifiers (Y. Freund and Schapire, 1997). It begins by training a base learner on the original training set and subsequently adjusts the sample distribution based on the learner's performance. Misclassified samples receive higher weights in subsequent iterations, ensuring that later models focus more on difficult cases. This iterative process continues until the number of base learners reaches a predefined threshold.

Bagging is a well-established parallel ensemble learning technique based on bootstrap sampling (Breiman, 1996a). Given a dataset of size m, multiple subsets are created using sampling with replacement. For an ensemble with T base learners, T subsets are generated, and each subset is used to train an individual model. The final prediction is obtained by aggregating the outputs of all base learners, typically through majority voting or averaging. Random Forest is one of the most widely recognized implementations of the Bagging approach.

Bayesian Model Averaging (BMA), Bayesian Model Combination (BMC), and Stacking represent different model integration strategies. BMA (Fraley et al., 2007) employs a weighted averaging strategy, where the weight of each model is determined by its posterior probability. BMC (Monteith et al., 2011) refines BMA by sampling not from individual models but from the space of possible ensembles, leading to improved model selection. Stacking involves training primary learners on the initial dataset and then constructing a secondary learner using the outputs of the primary learners as input features, while maintaining the original labels (Breiman, 1996b). Typically, logistic regression is used as the secondary learner. Stacking is considered more robust than BMA and BMC, as it is less sensitive to model

approximation errors.

3.3.2 Nonlinear age-adaptive ensemble model

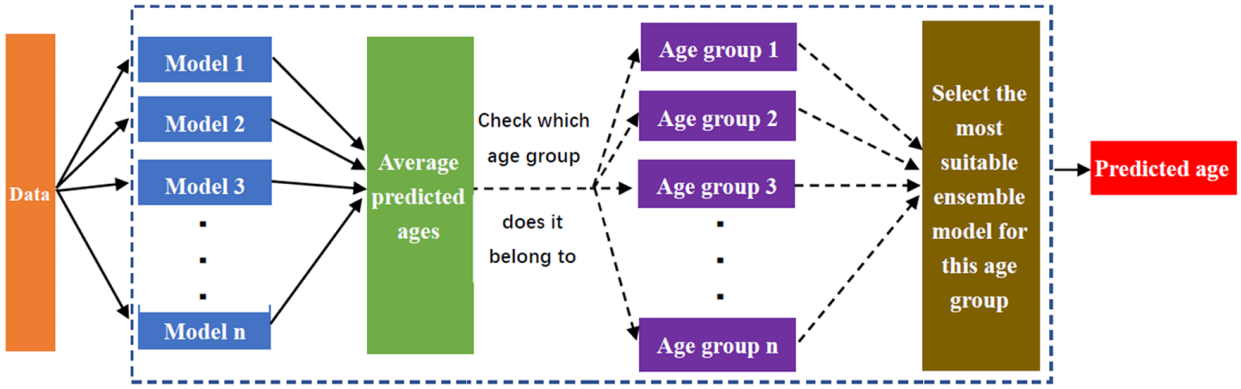


Figure 3.2: The architecture of nl-AAE. The nl-AAE model is a nonlinear age-adaptive ensemble that integrates GoogLeNet, ResNet, SVR, and a custom CNN to enhance brain age estimation. It dynamically adjusts model weights based on the average predictions by its constituent models, allowing it to adapt to age variations and capture brain aging patterns across different age groups for improved accuracy.

Extensive experimentation revealed that the performance of all models is significantly influenced by the true age of the samples (see Section **EXPERIMENTAL RESULTS**). This finding suggests that certain models are more effective at predicting brain age in younger individuals, while others perform better for older subjects. To enhance overall prediction accuracy, we developed a novel Nonlinear Age-Adaptive Ensemble Model (nl-AAE). The proposed framework is illustrated in Figure 3.2.

Initially, we employed four independent models as base learners: Support Vector Regression (SVR), ResNet, GoogLeNet, and a custom CNN model. Each model was used to generate brain age predictions, which were then recorded and utilized as input features for the ensemble model.

Subsequently, we divided the dataset into multiple age-specific groups. Within each group, a separate ensemble model was constructed, integrating the predictions from the independent models. This approach allows for age-specific adaptation, ensuring that the most suitable models contribute more significantly to the final prediction within each age range.

$$M = \sum \omega_i H_i \tag{3.1}$$

Here, M represents the prediction output of the ensemble model for a given age group. The term H_i denotes the prediction result of the i-th independent model within this age group,

while ω_i represents the weight assigned to the i-th independent model, with the constraint that the sum of all weights satisfies $\sum \omega_i = 1$.

To determine the optimal weights for the independent models, we use least squares method here. Specifically, we define a loss function within each age group, which is formulated as follows:

$$J(\omega) = \frac{1}{2}(H\omega - Y)^T(H\omega - Y) \tag{3.2}$$

H is an $m \times n$ dimensional matrix, where m represents the number of samples, and n denotes the number of independent models. The weight vector ω is an $n \times 1$ dimensional column vector, represented as $\omega = (\omega_1, \omega_2, \ldots, \omega_n)^T$, where ω_i corresponds to the weight of the i-th independent model. Similarly, Y is an $m \times 1$ dimensional column vector, expressed as $Y = (y_1, y_2, \ldots, y_n)$, where y_i denotes the real age of the i-th sample.

In this study, we evaluate two optimization methods for minimizing the loss function. The gradient descent algorithm iteratively updates the weights by following the steepest descent direction, formulated as follows:

$$\omega = \omega - \alpha H^T(H\omega - Y) \tag{3.3}$$

where α represents the learning rate.

The Ordinary Least Squares (OLS) method can also be utilized to achieve this objective. It is formulated as follows:

$$\omega = (H^T H)^{-1} H^T Y \tag{3.4}$$

It is important to note that both methods yield identical results in our experimental evaluations.

For each age group, we determine a set of optimal weights for the independent models, enabling the ensemble model to adaptively integrate the predictions from base models across different age groups. Formally, the age-adaptive model is expressed as:

$$F(x) = \sum_{A \in age} \omega_A H(x, p_A) \tag{3.5}$$

Here, age denotes the set of distinct age ranges, x represents the input data, ω_A corresponds to the value of ω at age A, and p_A denotes the parameters of the independent models at age A.

The process of predicting the brain age of a given sample proceeds as follows. First, each independent model generates a brain age prediction, which is recorded as $(H_1, H_2, \ldots, H_n)$. Once all predictions are obtained, we compute their mean value, denoted as H_{ave}, by averaging all predicted ages.

Next, we determine the age group to which H_{ave} belongs. For each age group $G_i \in (G_1, G_2, \ldots, G_n)$, there exists an ensemble model $M_i \in (M_1, M_2, \ldots, M_n)$ that adaptively

integrates the predictions from the base models. Specifically, if $H_{ave} \in G_i$, then we select M_i as the final ensemble model, which is then used to predict the brain age of the sample.

In summary, the ensemble weights are not fixed across the entire lifespan but are learned separately for different age ranges in nl-AAE. Specifically, the training data are divided into different tiny age groups, and for each group, least squares is applied to estimate the optimal weights of the base learners. This yields a distinct weight vector ω_A for every age group A, reflecting which base models are most informative within that range.

At prediction, each base learner produces an initial prediction, and their mean value H_{ave} is used to determine the most likely age group of the sample. The ensemble then applies the corresponding group-specific weight vector ω_A to combine the base model predictions. In this way, the contribution of each base model adapts dynamically with age.

Figure 3.2 illustrates the estimation process of our ensemble model, and Algorithm 1 provides the pseudocode of our method.

The nl-AAE divides the dataset into age-specific groups and trains ensemble models separately within these groups, which helps address the imbalance between younger and older participants. This strategy is not entirely new, but an adaptation of general class imbalance techniques, where dividing data into balanced subsets and training specialized classifiers has been shown to improve performance on underrepresented groups (T. G. Dietterich, 2000; Krawczyk, 2016). In our setting, the approach mitigates the overrepresentation of younger participants by ensuring that each age bracket has a dedicated ensemble model, thereby improving predictive robustness across the entire lifespan.

Algorithm 1 List of Pseudocode on Our Brain Age Estimation

Begin
Input brain sMRI as x
$H_1 = ResNet(x)$
$H_2 = GoogLeNet(x)$
$H_3 = CNN(x)$
$H_4 = SVR(x)$
$H_{ave} = \text{mean}(H_1 + H_2 + H_3 + H_4)$
for $i = 1$ to n **do**
    if $H_{ave} \in G_i$ **then**
        Age-adaptive ensemble model = M_i
    end if
end for
Final result = Age-adaptive ensemble M(x)
Output Final result
End

3.4 Experimental Results

3.4.1 Experimental Results

Table 3.3: The Details of MAE for Each Model in 5-Fold Cross-Validation

	ResNet	**6-layer CNN**	**GoogLe Net**	**SVR**	**MeanE**	**Median E**	**OE**	**nl-AAE -2**	**nl-AAE -6**	**nl-AAE -c**
Min	3.87	4.02	3.75	4.84	3.46	3.59	3.40	3.28	3.21	**2.98**
Max	4.11	4.54	4.00	5.34	3.71	3.92	3.69	3.65	3.55	**3.35**
Mean	3.99	4.33	3.88	5.15	3.57	3.72	3.52	3.42	3.39	**3.19**
Std	0.08	0.22	0.11	0.21	0.10	0.13	0.13	0.12	0.11	0.12

The evaluation follows a 5-fold cross-validation strategy, where the final results are reported as the mean of the Mean Absolute Error (MAE) and Spearman correlation coefficient between the predicted and actual ages.

The changes of nl-AAE's training loss for the last training are shown in Figure 3.3.

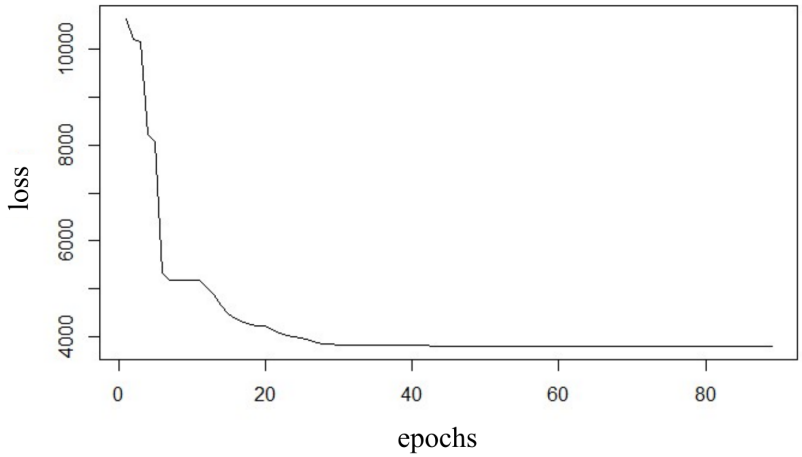


Figure 3.3: Changes of nl-AAE's training loss for the last training.

The detailed MAE results for each model in the 5-fold cross-validation are presented in Table 3.3. In the table:

- Min represents the minimum MAE.

- Max denotes the maximum MAE.
- Mean corresponds to the average MAE across five folds.
- Std refers to the standard deviation of the MAE, which quantifies the dispersion of the model's results.
- OE is a single linearly approximated ensemble model applied across all age groups, effectively testing the performance of the ensemble model without age adaptation.
- MedianE is an ensemble model applied across all age groups, which combines the outputs of the base learners by taking the median of their predictions.
- MeanE is an ensemble model applied across all age groups, which takes the prediction of base learners together using the mean of their predictions.
- nl-AAE-2 is a nonlinear age-adaptive ensemble model trained separately on two broad age groups (below 40 and 40+ years) and then integrated to improve prediction accuracy.
- nl-AAE-6 is a nonlinear age-adaptive ensemble model that divides the dataset into six distinct age ranges (10–20, 20–30, 30–40, 40–50, 50–60, and 60–90 years) for group-specific training and ensemble integration.
- nl-AAE-c is our continuous age-adaptive ensemble model that employs finer-grained grouping (year-wise from 17–30, five-year intervals from 30–60, and broader bins above 60), achieving the best overall predictive performance.

Table 3.3 indicates that the 6-layer CNN exhibits the highest standard deviation (0.22), followed by SVR with 0.21. In contrast, ResNet has the lowest standard deviation (0.08), indicating that its predictions are more stable compared to other models. The standard deviation values for other models are as follows: OE and MedianE (0.13), nl-AAE-2 and nl-AAE-c (0.12), GoogLeNet and nl-AAE-6 (0.11), and MeanE (0.10). These findings suggest that the 6-layer CNN produces the most variable predictions, while ResNet provides the most consistent and stable predictions.

The test results are summarized in Figure 3.4 and Table 3.5. In Table 3.3, we first present the results of four independent models: SVR, 6-layer self-built CNN, ResNet, and GoogLeNet. The mean absolute errors (MAE) in years are 5.15, 4.33, 3.99, and 3.88, respectively, while the Spearman correlation coefficients between predicted and actual ages are 0.83, 0.89, 0.88, and 0.89, respectively.

Furthermore, we evaluate ensemble models by aggregating the predictions of the base learners using median and mean-based strategies. The results indicate that the median-based ensemble model yields a higher mean error than the mean-based ensemble model. A

potential explanation for this outcome is that the median-based ensemble only selects a single model's output at a time, disregarding the contributions of other models that do not produce the median value.

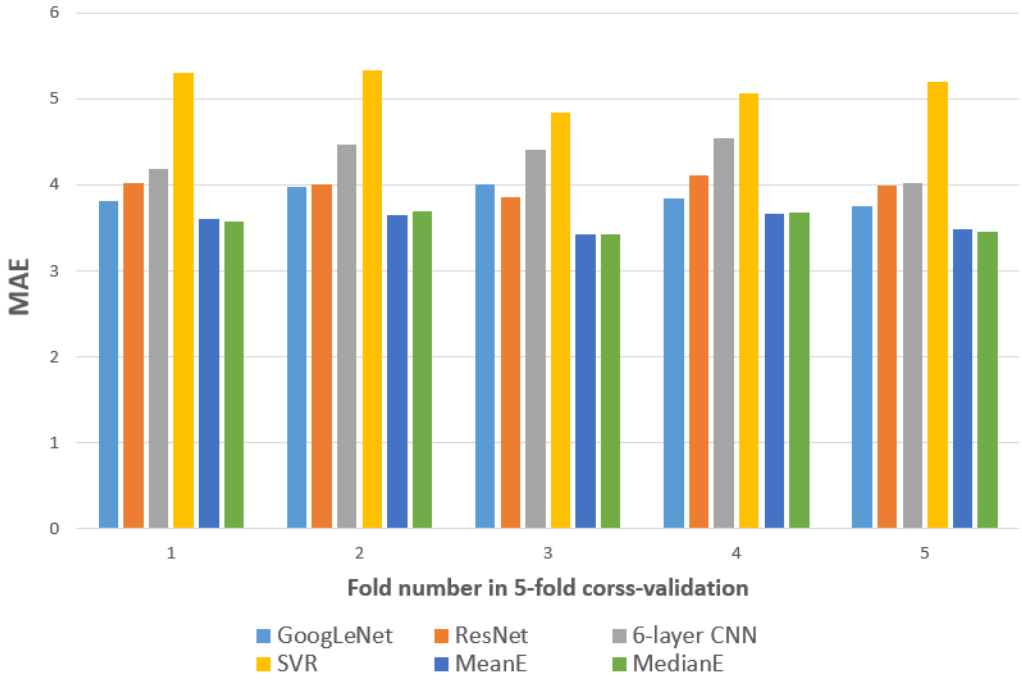


Figure 3.4: Performance of each individual model.

3.4.2 Analysis of the Nonlinear Age-Adaptive Model

Following the preliminary evaluations, we further investigate the effectiveness of the nonlinear age-adaptive ensemble model.

Initially, we employed a single linearly approximated ensemble model applied across all age groups, effectively testing the performance of the ensemble model without age adaptation. Compared to naive fusion strategies, this approach resulted in a marginal performance improvement, achieving a Mean Absolute Error (MAE) of 3.52 years and a Spearman correlation of 0.91 (OE in Table 3.5).

In the second experiment, we divided the prediction results from the four independent models into two age groups:

- Group 1: Samples aged 40 years and above.
- Group 2: Samples aged below 40 years.

Table 3.5: Results of All the Models

	Model	**Evaluation Metrics**	
		Average MAE (years)	**Spearman correlation**
Single Model	SVM (Couvy-Duchesne, Faouzi, et al., 2020)	5.15	0.83
	6-layer CNN (Couvy-Duchesne, Faouzi, et al., 2020)	4.33	0.89
	ResNet (Fisch et al., 2021)	3.99	0.88
	GoogLeNet (Couvy-Duchesne, Faouzi, et al., 2020)	3.88	0.89
	MedianE (Couvy-Duchesne, Faouzi, et al., 2020)	3.72	0.90
	MeanE (Couvy-Duchesne, Faouzi, et al., 2020)	3.57	0.91
Ensemble Method	OE (Cole and Franke, 2017)	3.52	0.91
	Our nl-AAE-2	3.45	0.89
	Our nl-AAE-6	3.39	0.95
	Our nl-AAE-c	**3.19**	**0.95**
Other Researchers' Methods	Seven algorithms combined ensemble model (Couvy-Duchesne, Faouzi, et al., 2020)	3.33	–
	Ensemble of shallow machine learning methods (Da Costa, Dafflon, and Pinaya, 2020)	3.75	–
	Distributional Transformation (Soch, 2020)	4.58	0.93

We retained the same four base models (GoogLeNet, ResNet, SVR, and CNN) and, for each age group, used their prediction outputs as input features to train a separate secondary learner. By establishing two ensemble models for different age groups and integrating them into a nonlinear ensemble model, we achieved a lower MAE of 3.45 years; however, the Spearman correlation decreased to 0.89 (nl-AAE-2 in Table 3.5).

Next, we refined the division further by categorizing samples into six age groups based on actual age:

- 10–20 years, 20–30 years, 30–40 years, 40–50 years, 50–60 years, and 60–90 years.

Using this classification, we applied the same methodology to construct the nonlinear ensemble model (nl-AAE-6). This configuration resulted in an average MAE of 3.39 years and an improved Spearman correlation of 0.95 (Table 3.5).

Finally, we implemented a more granular division by treating each individual age as a separate group where possible. Due to the limited sample size, we adopted a simplified grouping strategy:

- For samples aged 17 to 30 years, each age was treated as a separate group.
- For samples aged 30 to 60 years, we grouped every five years (e.g., 30–35, 35–40, etc.).
- For samples aged 60 to 70 years, and 70 to 90 years, data were grouped accordingly.

We refer to this finely partitioned model as the "continuous" (or year-wise) model, denoted as nl-AAE-c in Table 3.5. This age-adaptive model yielded the best overall performance, achieving a MAE of 3.19 years and maintaining a Spearman correlation of 0.95.

We compare our proposed model with prior research that has also been evaluated on the PAC 2019 dataset.

(Couvy-Duchesne, Faouzi, et al., 2020) developed an ensemble model combining seven different algorithms, achieving a Mean Absolute Error (MAE) of 3.33 years, demonstrating strong predictive performance. Similarly, (Da Costa, Dafflon, and Pinaya, 2020) constructed an ensemble of shallow machine learning methods, including Support Vector Regression (SVR) and Decision Tree-based regressors, which resulted in a MAE of 3.75 years. (Soch, 2020) proposed Distributional Transformation (DT), a method that maps predicted values to the variable's distribution within the training data to enhance decoding accuracy. Their approach achieved a MAE of 4.58 years and a Spearman correlation of 0.93 between predicted and actual age.

These prior research findings provide valuable insights. Building on this foundation, we developed the Age-Adaptive Ensemble (AAE) method, which achieves a great improvement compared to existing models.

The brain age gap is defined as the difference between the predicted age and the chronological age. Figure 3.5 illustrates the brain age gap as a function of chronological age

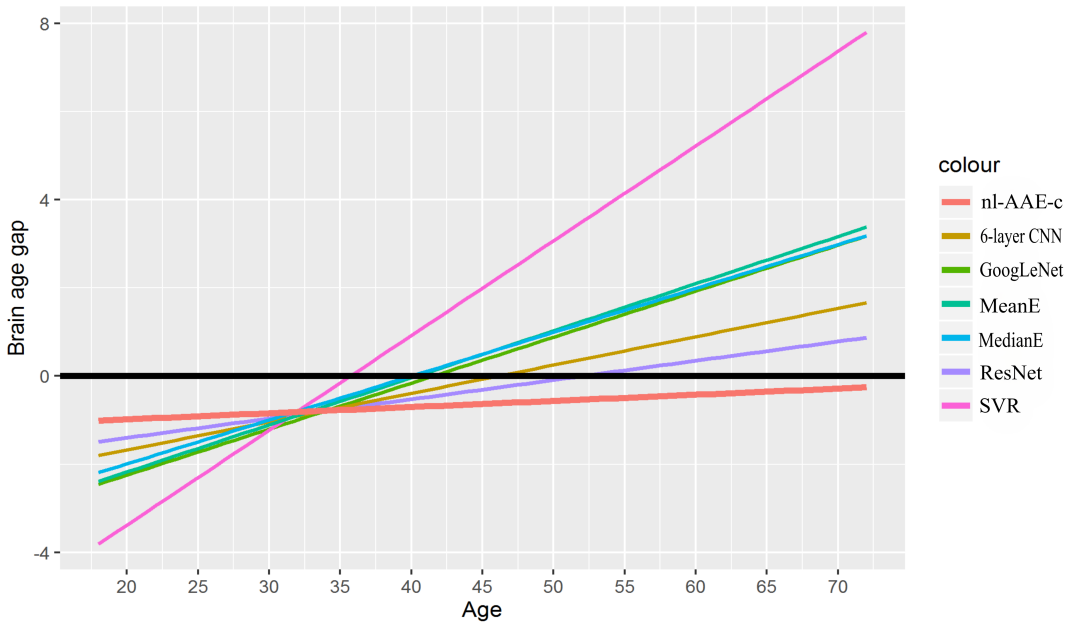


Figure 3.5: The brain age gap and age as functions of the chronological age using 7 different machine-learning methods, the horizontal black line represents 0 brain age gap.

for seven different machine learning methods. The slope of each line in Figure 3.5 quantifies the impact of age on the model's predictive accuracy. Notably:

- The AAE model exhibits the least sensitivity to aging effects, maintaining high prediction accuracy across different age groups.
- Conversely, SVR demonstrates the highest sensitivity, with predictive accuracy deteriorating more significantly with increasing age.

From the above experiments, we derive the following key insights:

1. Deep neural networks outperform traditional SVM models in brain age prediction.
2. All ensemble models achieve lower errors compared to individual discrete models, highlighting the advantage of model integration.
3. Age-adaptive ensemble models surpass non-adaptive ensemble models, demonstrating the benefits of dynamic model adaptation.
4. Finer age-based divisions lead to lower prediction errors, as evidenced by the superior performance of the nl-AAE-c model.

3.4.3 Investigation on Age-Sensitivity per Models

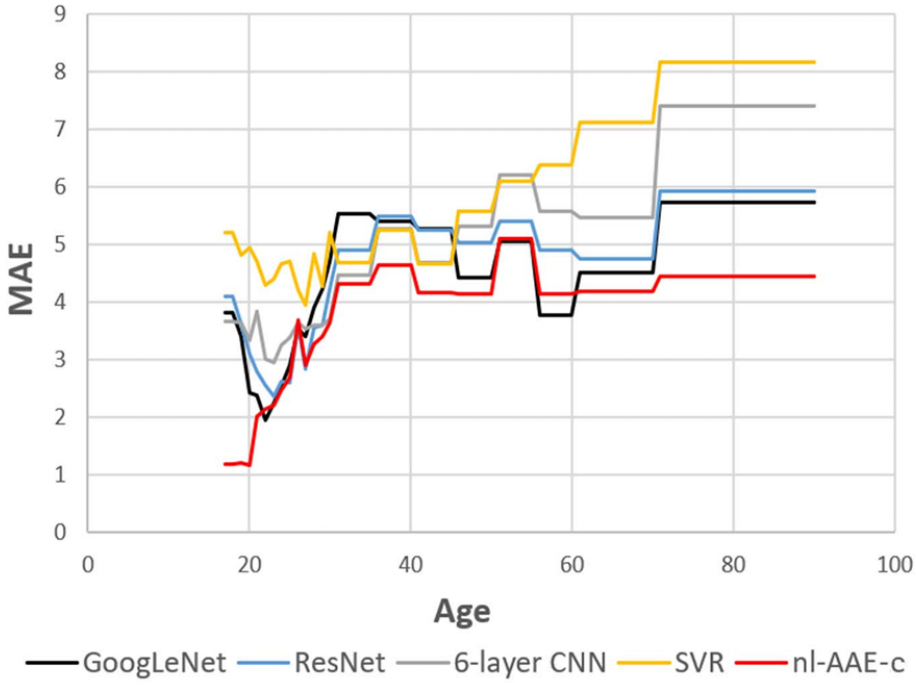


Figure 3.6: Age-sensitivity of models.

Age sensitivity describes how the Mean Absolute Error (MAE) of each model varies across different age groups. In this section, we analyze the age sensitivity of the models developed in our study.

Figure 3.6 presents the results, indicating that all models exhibit higher predictive accuracy for younger individuals but struggle with older populations, as evidenced by the increasing MAE with age. Among the independent models:

- GoogLeNet and ResNet demonstrate greater age sensitivity, with a significant rise in MAE for samples aged 20 to 30 years.
- SVR, in contrast, exhibits relatively stable performance across different age groups.
- GoogLeNet performs best for middle-aged individuals, while all models display substantial MAE increases for samples aged 70 years and older.

The nonlinear age-adaptive ensemble model exhibits a similar age-sensitivity trend but with greater stability compared to the independent models. However, a clear pattern emerges: as sample age increases, MAE progressively rises, leading to deteriorating model performance.

The worst performance is observed for individuals aged 50 to 60 years, where the MAE exceeds 5.

From a machine learning perspective, this decline in performance can be attributed to the limited availability of older samples, resulting in insufficient model training for those age groups. From a medical standpoint, we hypothesize that as individuals age, the variability in brain structures across different individuals increases, making age prediction more challenging. In contrast, younger individuals exhibit less variability, leading to higher predictive accuracy.

3.4.4 Learning the Model Weights

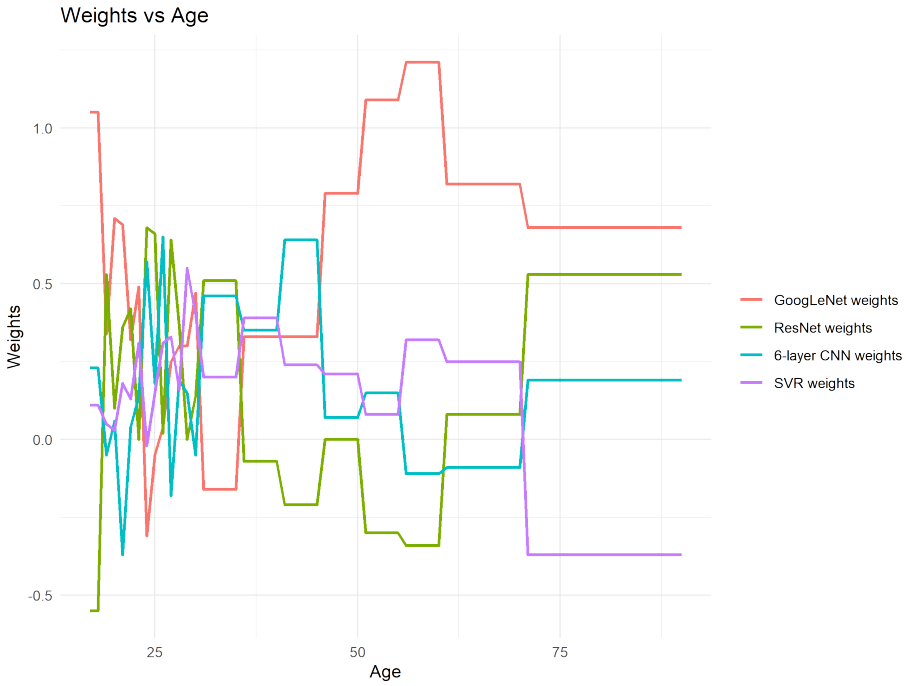


Figure 3.7: Individual models' weights changing in nl-AAE-c.

3.4.5 Analysis of Model Weights in the nl-AAE Framework

Figure 3.7 illustrates the variation in the weights assigned to each independent model within the nl-AAE across different age groups. This analysis provides insights into the relative importance of each model in contributing to the ensemble's predictions at different ages.

The SVR model maintains relatively consistent weights for samples aged 25 to 70 years. However, for younger individuals aged 10 to 25 years, SVR demonstrates poor

predictive performance, suggesting that SVR is not well-suited for age estimation in younger populations.

The CNN model plays a crucial role in predictions for individuals aged 20 to 50 years. It receives higher weights when predicting younger samples, but its contribution to the ensemble diminishes for older individuals, indicating its limited effectiveness in elderly age groups.

The GoogLeNet model becomes increasingly important in the elderly age groups. While its influence remains moderate for individuals aged 20 to 40 years, its predictions significantly impact the ensemble results in older age groups, particularly for middle-aged individuals. This suggests that GoogLeNet is highly effective for middle-aged and elderly populations.

The ResNet model maintains high weights for samples aged 10 to 35 years, highlighting its strong suitability for younger age groups. In contrast, its contribution decreases for middle-aged and elderly individuals (35–70 years), where it exhibits average performance. However, for individuals over 70 years old, ResNet demonstrates notably strong performance, suggesting that it is particularly well-suited for age estimation in elderly populations.

This weight analysis underscores the importance of an age-adaptive ensemble framework, as different models excel in different age ranges. Leveraging this diversity allows the nl-AAE model to enhance overall prediction accuracy across the lifespan.

3.5 Discussion

Our research provides several important medical insights regarding brain aging and neurodegeneration.

We observed that the performance of constituent models declines with increasing age, suggesting that younger individuals exhibit greater structural similarity in brain anatomy. Additionally, as age increases, the risk of neurodegenerative diseases also rises, leading to greater variability in brain structure and a corresponding decline in prediction accuracy. However, these findings may be influenced by biological factors, sample size limitations, or model constraints, and future studies should explore more robust validation methods to substantiate these observations.

Based on our experimental results, we propose that brain aging progresses through four distinct stages:

- 0–30 years: Significant neurodevelopmental changes occur, shaping cognitive function.
- 30–50 years: Brain structure remains relatively stable, with minor functional decline.
- 50–70 years: Noticeable cognitive decline and increased susceptibility to neurodegeneration.
- 70–80 years: The brain undergoes substantial atrophy and functional deterioration.

These transitions are determined by changes in model prediction performance, as shown in Figure 3.5.

Research findings support this classification. Studies presented at the Academy of Medical Sciences (Oxford, UK) indicate that brain maturation continues into the 30s, aligning with our findings that cognitive capacity peaks before this age. After 30 years, working memory capacity begins to decline (Guo et al., 2016). Between 30 and 50 years, brain structure remains relatively stable, but after 50 years, a significant decline becomes evident. A study published in the British Medical Journal (BMJ) (Grodstein, 2012) found that cognitive reasoning skills declined by 3.6% over 10 years among individuals initially tested at ages 45–49. Additionally, research by Peter Jones (Guo et al., 2016) indicates that overall brain volume begins to shrink in the 30s or 40s, with the rate of shrinkage accelerating between ages 60–70, a finding corroborated by our experimental results.

Given the sensitivity of brain age prediction models to neural changes, we believe that our approach can serve as a valuable tool for monitoring the efficacy of medical treatments targeting neurodegenerative disorders and aging-related conditions. Future applications may include:

- Evaluating the effectiveness of pharmaceutical interventions in clinical trials.
- Assessing the impact of lifestyle modifications on brain health.
- Supporting early diagnosis and progression tracking of neurodegenerative diseases.

Compared to our nl-AAE, traditional ensemble strategies such as stacking combine multiple base learners but assign fixed or globally optimized weights across the entire dataset. In stacking, a meta-learner is trained on the predictions of base models, and the learned weights are applied uniformly to all test samples regardless of their age distribution.

By contrast, the proposed nl-AAE framework introduces an age-adaptive mechanism that learns separate weight vectors for different age ranges. During inference, the ensemble selects the appropriate set of weights according to the estimated age group of the input, allowing the contribution of each base model to vary dynamically across the lifespan. This design is particularly important for brain age estimation, where model performance is known to vary substantially between younger and older cohorts. As a result, nl-AAE not only outperforms conventional stacking in overall accuracy, but also reduces age-related biases by tailoring model integration to the structural characteristics of each age group.

While nl-AAE demonstrates strong predictive accuracy in research settings, several limitations constrain its direct application to real-time clinical diagnosis. First, the framework integrates multiple deep networks and a classical model, which increases computational cost and inference time compared to single-model solutions. Second, the method relies on extensive preprocessing pipelines, such as nonlinear registration, tissue segmentation, and surface-based feature extraction, many of which are intensive and not practical in real clinical workflows.

In our study, model performance was primarily evaluated using widely adopted metrics in the brain age estimation literature, including mean absolute error (MAE) and Pearson correlation. These measures allow direct comparison with prior work, ensuring consistency and reproducibility across studies. However, we acknowledge that statistical significance testing of performance differences (e.g., paired hypothesis testing between models) would provide a stronger validation of the observed improvements. Incorporating such analyses represents an important direction for future work, as it would allow us to confirm whether the differences between models are not only numerically but also statistically meaningful.

3.6 Conclusion

In this chapter, we proposed the nl-AAE framework for brain age estimation. By integrating CNN, GoogLeNet, ResNet, and SVR within a nonlinear, age-adaptive ensemble, the model dynamically adjusts the contribution of its constituent learners according to the age of the subject. This design enables nl-AAE to capture age-specific structural patterns and to mitigate performance degradation caused by the imbalanced age distribution in the dataset. Experimental results on the PAC 2019 dataset demonstrate that nl-AAE achieves superior predictive accuracy compared with both individual base models and conventional ensemble methods, while offering more stable performance across age groups.

Despite these advantages, nl-AAE has limitations for real-world application. First, the framework integrates multiple deep networks and a classical model, which increases computational cost and inference time compared with single-model solutions. Second, the method relies on extensive preprocessing pipelines, many of which are computationally intensive and less practical for deployment in real clinical workflows.

Future work may focus on incorporating more advanced deep learning architectures as base models. For instance, while our current design employed Inception V1, more recent variants could further enhance predictive performance. In addition, extending the ensemble to multimodal inputs or harmonizing pipelines to reduce preprocessing demands could improve both generalizability and clinical feasibility.

Having established nl-AAE as a high-accuracy ensemble approach, the next chapter introduces the Explainable Triamese ViT framework. Unlike nl-AAE, Triamese-ViT is specifically designed with built-in interpretability and a tri-view model structure, enabling not only accurate brain age estimation but also direct analysis of the normal aging and ASD.

Chapter 4

Explainable Triamese ViT

4.1 Introduction

Brain age, estimated from neuroimaging data, has emerged as a powerful biomarker for quantifying brain health and predicting the onset of various pathologies. While deep learning models, particularly those based on CNNs, have demonstrated considerable success in this domain, their clinical translation is often hampered by significant challenges: the inherent black box nature of their decision-making processes. The lack of clear interpretability limits the trust and utility of these models in clinical settings, where understanding why a prediction is made is as crucial as the prediction itself.

In contrast, ViTs offer a compelling alternative by segmenting images into patches and employing self-attention mechanisms to capture intricate spatial relationships across these patches (Dosovitskiy et al., 2020). This approach enhances feature extraction capabilities and provides interpretability through attention maps, which highlight the regions most influential in the model's decision-making process (Khan et al., 2022). Previous ViTs are primarily designed for 2D images (Tanveer et al., 2023; Al-Hammuri et al., 2023; He, Grant, and Ou, 2021), they may not fully exploit the three-dimensional nature of MRI data, potentially leading to the loss of crucial depth-related information in brain age prediction.

As for 3D ViT, Pantelaios et al. (2024) combines a 3D CNN with a ViT. The architecture uses a dual-branch system: the CNN branch is designed to capture local, fine-grained features and spatial hierarchies from the MRI scans, while the ViT branch is used to model long-range dependencies and global contextual relationships across the entire brain volume. The features extracted from both branches are then fused to make a final, integrated prediction of brain age. Their results prove that this architecture was effective at predicting age in individual brain lobes.

The Swin Transformer (J. Kim, M. Kim, and Park, 2024) is an advanced hierarchical Vision Transformer, was adapted for 3D medical imaging. Unlike the standard ViT which has high computational complexity with high-resolution images, the Swin Transformer uses

a shifted window mechanism for self-attention. This allows it to compute self-attention locally within non-overlapping windows that are shifted between layers, capturing features at various scales more efficiently. Their pre-trained model outperformed existing supervised and self-supervised methods on several tasks, including brain age prediction, Alzheimer's classification, and Parkinson's classification.

Model interpretability is a critical aspect of brain age estimation, as it enables the identification of key brain regions associated with aging, thereby facilitating both neuroscientific research and clinical diagnosis. Explainability in machine learning models generally falls into two categories: post-hoc explainability and inherently interpretable models. Post-hoc methods aim to provide explanations for black-box model predictions, either at an individual-instance level or globally across datasets, with feature attribution being the most commonly used technique. These approaches assess feature importance using perturbation-based methods (Ribeiro, Singh, and Guestrin, 2016a; Lundberg, 2017) or input gradients (Srinivas and Fleuret, 2019; Selvaraju et al., 2017). Despite their widespread use in computer vision, perturbation-based methods often produce unreliable attributions due to their underlying assumptions regarding feature removal (Bhalla, Srinivas, and Lakkaraju, 2024).

Conversely, inherently interpretable models are designed to be transparent in their structure or parameterization, offering a more direct and accurate understanding of their decision-making process compared to post-hoc techniques. Examples of such models include linear regression, decision trees, generalized linear models (GLMs), generalized additive models (GAMs)(Hastie, 2017), joint additive models (JAMs)(J. Chen et al., 2018), prototype- and concept-based models (C. Chen et al., 2019), and weight-input aligned models (Böhle, Fritz, and Schiele, 2022). However, while these models provide superior interpretability, they often exhibit lower predictive performance compared to deep learning-based black-box models, creating a trade-off between transparency and accuracy.

To address these limitations, this chapter introduces a novel architecture, the Explainable Triamese ViT, designed for accurate and interpretable brain age estimation. Moving beyond conventional 3D networks, Triamese-ViT leverages the power of ViTs by decomposing 3D MRI scans into three orthogonal 2D views. This approach not only mitigates the prohibitive computational cost associated with volumetric analysis but also captures a rich, multi-faceted representation of brain anatomy. By integrating features from these distinct perspectives, the model develops a comprehensive understanding of complex brain structures, enhancing predictive performance.

A key innovation of this work lies in the model's inherent explainability. The attention mechanisms from the Transformer architecture are harnessed to generate detailed, 3D-like attention maps, offering unprecedented insight into the specific brain regions that drive age prediction. This chapter details the systematic evaluation of Triamese-ViT, demonstrating its superior accuracy and fairness compared to a suite of state-of-the-art algorithms. Furthermore, it validates the model's interpretability through occlusion sensitivity analyses.

Finally, the clinical utility of Triamese-ViT is explored through two primary applications. First, the model is employed to map the neuroanatomical changes associated with the normal aging process across the human lifespan. Second, it is applied to a cohort of individuals with Autism Spectrum Disorder (ASD) to identify the key neural signatures associated with the condition. These analyses are further stratified by gender to investigate sex-specific differences in both healthy aging and ASD pathology, underscoring the model's potential as a versatile tool for advancing neuroimaging research.

4.2 Method

4.2.1 Data and Code Availability

In this study, we utilized MRI scans from the IXI[1] and ABIDE[2] datasets. Specifically, we compiled a dataset of healthy individuals to train the model and examine normal brain aging, as well as a dataset of individuals with Autism Spectrum Disorder (ASD) to identify key brain regions associated with ASD detection. All MRI scans were T1-weighted.

The dataset of healthy participants comprises 1351 scans from individuals aged 6 to 80 years, with a mean age of 30.5 years and a standard deviation of 19.95 years. This cohort includes 872 males and 479 females. Given that previous studies have suggested that gender does not have a significant impact on brain age estimation (Couvy-Duchesne, Faouzi, et al., 2020), gender-specific analyses were not conducted in this study.

The age distribution across different subgroups within the healthy population is presented in Table 4.1.

Age	0s	10s	20s	30s	40s	50s	60s	70s
Samples	142	420	257	138	112	104	120	58

Table 4.1: Healthy participants' dataset age distribution.

For the healthy cohort, the dataset was stratified into eight age groups: 0s, 10s, 20s, 30s, 40s, 50s, 60s, and 70s. Within each age group, 70% of the samples were assigned to the training set, 15% to the validation set, and 15% to the test set, ensuring a balanced and representative distribution across all subsets.

For the ASD cohort, the dataset comprises 280 samples, with participants ranging in age from 6 to 62 years (mean = 18.8 years, standard deviation = 13.78 years). The detailed age distribution is provided in Table 4.2. For these two datasets, we didn't stratify them by age groups when training like nl-AAE.

[1] `https://brain-development.org/ixi-dataset/`

[2] `https://fcon_1000.projects.nitrc.org/indi/abide/`

It is important to note that the ASD dataset in this study was not used for model training or validation but exclusively for interpretability analysis with the pretrained Triamese-ViT model. Consequently, stratification was not required.

Age	0s	10s	20s	30s	40s	50s	60s
Samples	82	112	48	6	12	18	2

Table 4.2: ASD participants' dataset age distribution.

To ensure compatibility and mitigate the potential impact of protocol variability across different datasets, we applied a standardized preprocessing pipeline using FSL 5.10 (Jenkinson et al., 2012), a comprehensive library of analysis tools for FMRI, MRI, and DTI brain imaging data. This preprocessing procedure consisted of several steps: brain extraction (Smith, 2002), bias field correction, nonlinear registration to the MNI standard space, and voxel-wise intensity normalization within the brain region by subtracting the mean and dividing by the standard deviation. Additionally, ComBat statistical harmonization was employed to adjust for scanner- and site-specific effects while preserving biological variability.

Following preprocessing, all MRI scans were resampled to a voxel resolution of $91 \times 109 \times 91$ with an isotropic spatial resolution of 2 mm.

To assess the effectiveness of harmonization, we visualized the voxel intensity distributions from two different imaging sites in our dataset—Trinity College Dublin and Georgetown University—before and after applying ComBat harmonization (Figure 4.1).

Prior to harmonization, substantial differences in intensity distributions were observed between datasets, reflecting scanner- and site-specific variability. However, after applying ComBat harmonization, the intensity distributions became well-aligned, indicating that the method successfully mitigates unwanted scanner-induced variability while preserving biologically relevant variations. These results highlight the effectiveness of ComBat in standardizing multi-site neuroimaging data, ensuring greater consistency across datasets.

The code used in this study has been made publicly available on GitHub[3].

4.2.2 Proposed Triamese-ViT

In this section, we present our novel architecture named Triamese-ViT. Our approach is inspired by (R. Jiang, Ho, et al., 2017), which highlights that different views of a 3D image contain unique and independent information that can be leveraged in machine learning models. As illustrated in Figure 4.2, the structure of Triamese-ViT is based on the Vision Transformer (ViT) (Dosovitskiy et al., 2020). Triamese-ViT processes 3D MRIs, denoted as $M \in \mathbb{R}^{H \times W \times C}$, where H, W, and C represent the height, width, and the slice number, respectively. The MRI M is then reshaped into three distinct viewpoints, represented as

[3] `https://github.com/zhangz59/Triamese-ViT`

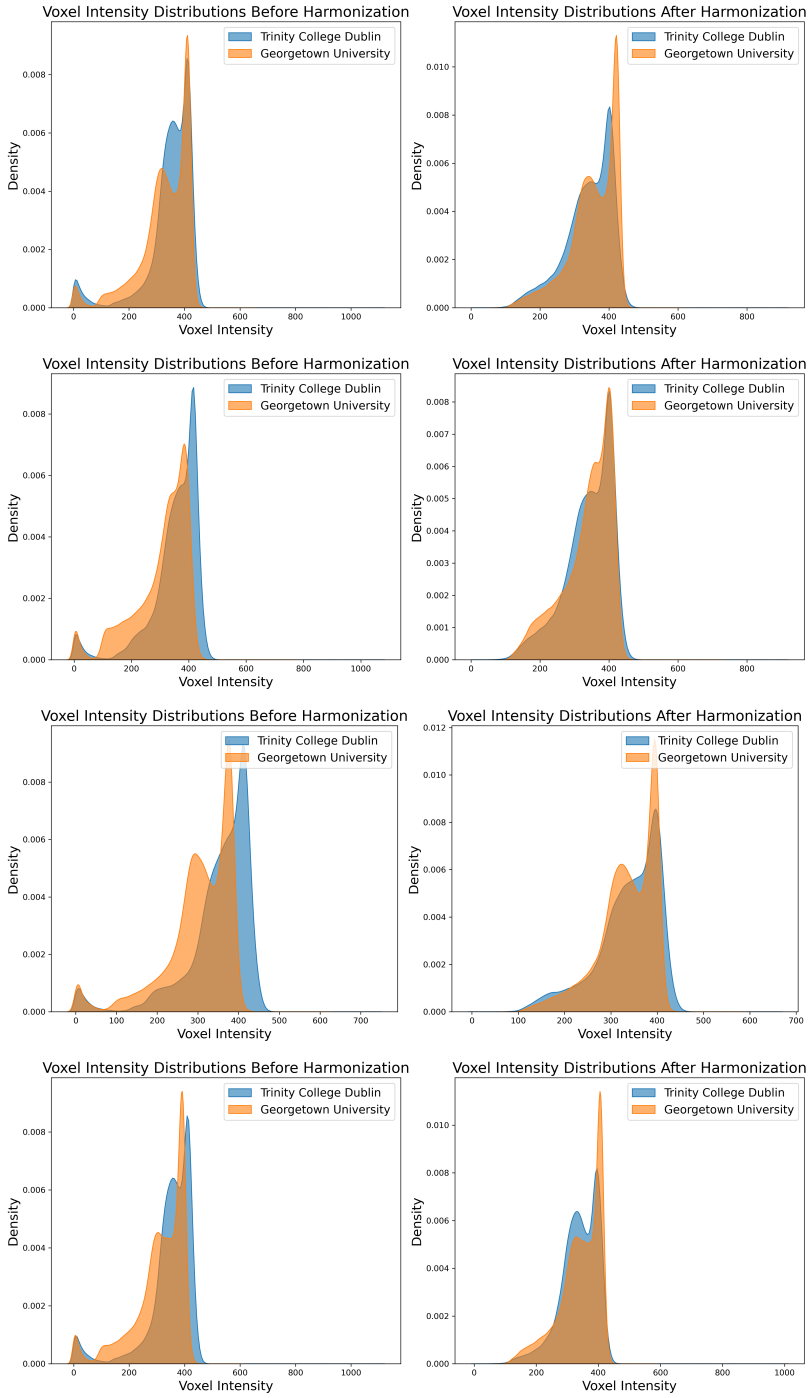


Figure 4.1: The effect of harmonization, we visualized the voxel intensity distributions from two different sites within our dataset—Trinity College Dublin and Georgetown University—before and after applying ComBat harmonization.

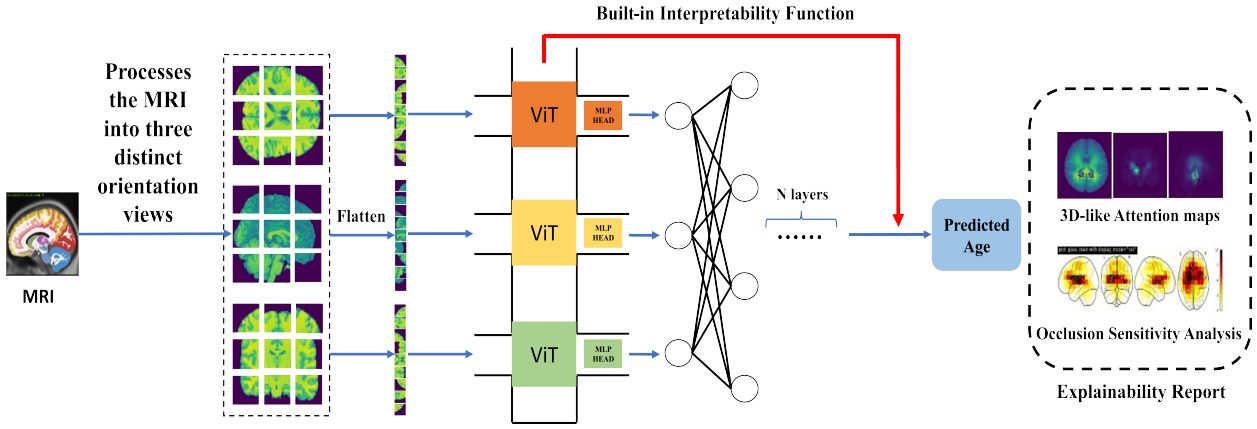


Figure 4.2: The architecture of Triamese-ViT. This model processes brain MRI images from three distinct perspectives utilizing the Vision Transformer (ViT) to extract unique features. These features are then integrated within a Tri Multi-Layer Perceptron (MLP) framework to generate age predictions. And built-in interpretability function generates 3D-like images to explain different brain regions influence during prediction.

$M \rightarrow (M_x, M_y, M_z)$, with $M_x \in \mathbb{R}^{H \times W}$ (C channels), $M_y \in \mathbb{R}^{H \times C}$ (W channels), and $M_z \in \mathbb{R}^{W \times C}$ (H channels).

Focusing initially on M_x, the MRI is divided into a sequence of flattened 2D squares, denoted as $M_{x,s} \in \mathbb{R}^{N \times (S^2 \cdot C)}$, where the side length of the square is S, and the number of squares is $N = \frac{H \times W}{S^2}$.

In the transformer encoder layers, the vectors processed are of dimension D. Thus, M_x needs to be mapped to D dimensions using a trainable linear projection. The process is formulated as follows:

$$t_{x,0} = \text{Concat}(M_{x,class}; M^1_{x,s}E; M^2_{x,s}E; \ldots; M^N_{x,s}E) + E_{pos} \tag{4.1}$$

In Equation 4.1, $M_{x,class}$ is a learnable token (or class token) added to ViT, akin to the method used in Devlin et al., 2018. This class token, $M_{x,class}$, is eventually output from the Transformer Encoder as $t^0_{x,L}$, representing the image representation P (Equation 4.7). Here, $E \in \mathbb{R}^{(S^2 \cdot C) \times D}$ is the linear projection matrix, Concat denotes token concatenation, and $E_{pos} \in \mathbb{R}^{(N+1) \times D}$ is the positional encoding added to each token embedding. $t_{x,0}$ represents the input sequence to the 0-th (first) Transformer encoder layer. The same preprocessing steps are applied to M_y and M_z, resulting in $t_{y,0}$ and $t_{z,0}$.

The transformed matrices $t_{x,0}$, $t_{y,0}$, and $t_{z,0} \in \mathbb{R}^{(N+1) \times D}$ are fed into the transformer encoder. Each encoder consists of multiple layers, where each layer sequentially processes the input through Layer Normalization (LN), Multi-Head Attention (MSA), another Layer Normalization, and a Multi-Layer Perceptron (MLP). The MSA performs parallel attention calculations across multiple heads, allowing for diverse representation and richer

understanding of the input data.

$$[Q, K, V] = FC(t_{x,0}) \tag{4.2}$$

Here, $Q \in \mathbb{R}^{(N+1)\times d}$, $K \in \mathbb{R}^{(N+1)\times d}$, and $V \in \mathbb{R}^{(N+1)\times d}$ represent the Query, Key, and Value matrices, respectively. Assuming the MSA has n heads and $D = n \times d$, each head independently processes the input:

$$\text{head}_i = \text{softmax}\left(\frac{Q_i K_i^{\mathsf{T}}}{\sqrt{d}}\right) V_i \tag{4.3}$$

$$\text{MSA}(z_{x,0}) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_n) \tag{4.4}$$

Let $t_{x,0}$ be the input to the first layer of the Transformer Encoder. The feedforward calculations in the encoder are given by:

$$t^{'}_{x,l} = \text{MSA}(\text{LN}(t_{x,l-1})) + t_{x,l-1} \tag{4.5}$$

$$t_{x,l} = \text{MLP}(\text{LN}(t^{'}_{x,l})) + t^{'}_{x,l} \tag{4.6}$$

where $l \in [1, 2, \ldots, L]$. The outputs from each Transformer Encoder are then passed to an MLP head, consisting of a hidden layer and an output layer, to generate the final prediction for each view. The prediction from the first view, M_x, is denoted as P_x. By applying the same procedure to M_y and M_z, we obtain two additional predictions, P_y and P_z.

In the final stage, these three view-based predictions (P_x, P_y, and P_z) are fed into the MLP, which integrates the information from all three views to produce the final comprehensive prediction:

$$P_{Tri} = \text{MLP}(P_x, P_y, P_z) \tag{4.7}$$

Here, P_{Tri} denotes the final prediction.

The pseudocode of Triamese-ViT is shown in Algorithm 2.

The decision to adopt an axis-wise Vision Transformer (ViT) instead of a full 3D ViT for brain age estimation is motivated by several key advantages:

- Lower Computational Cost: Triamese-ViT avoids the high computational burden of processing entire 3D volumes by decomposing them into three orthogonal 2D views. This approach significantly reduces computational complexity, as each view is treated as a 2D input to a standard ViT, which scales linearly with input size. Consequently, Triamese-ViT enables faster training and requires substantially less GPU memory compared to full 3D ViTs, making it more suitable for large-scale 3D medical imaging datasets.

Algorithm 2 Tri-View Transformer-based Prediction Algorithm

Input: 3D image I
Output: Predicted result y
Extract views I_x, I_y, I_z from I
▷ Process viewpoint I_x
Flatten I_x into 2D slices $\{x_1, x_2, \ldots, x_n\}$
Map each x_i into patch embeddings $\{p_1^x, p_2^x, \ldots, p_n^x\}$
▷ Repeat the same process for I_y and I_z
Flatten I_y into 2D slices $\{y_1, y_2, \ldots, y_n\}$
Map each y_i into patch embeddings $\{p_1^y, p_2^y, \ldots, p_n^y\}$
Flatten I_z into 2D slices $\{z_1, z_2, \ldots, z_n\}$
Map each z_i into patch embeddings $\{p_1^z, p_2^z, \ldots, p_n^z\}$
▷ Feed the processed matrices into the Transformer Encoder
for all view v in $\{x, y, z\}$ **do**
    for $i = 1$ to L **do**
        for all patch embedding p_i^v **do**
            $p_i^v \leftarrow \text{MultiHeadSelfAttention}(p_i^v)$
            $p_i^v \leftarrow \text{FeedForward}(p_i^v)$
        end for
    end for
end for
▷ Integrate the predictions from all three views
Concatenate final outputs $\{o_x, o_y, o_z\}$
Predict y using MLP on concatenated output

- Model Simplicity and Implementation: By leveraging the well-established 2D Vision Transformer framework, Triamese-ViT maintains a straightforward implementation that requires minimal adaptation. This simplifies model design, debugging, and fine-tuning while enabling the integration of pre-trained weights and existing tools developed for 2D ViTs. In contrast, 3D ViTs necessitate extensive architectural modifications, including 3D tokenization and positional encoding, which introduce additional computational and technical complexities.

- Higher Predictive Accuracy: Empirical evaluations indicate that Triamese-ViT outperforms 3D ViTs in predictive accuracy. This improvement arises from its ability to integrate multiple 2D views, effectively capturing diverse and complementary spatial features from different anatomical perspectives. By enhancing spatial representation learning, Triamese-ViT improves the robustness and precision of brain age estimation.

Compared to other multi-view or pseudo-3D ViT architectures, our proposed Triamese-ViT differs in several important aspects. First, existing multi-view ViT methods typically extract 2D slices or patches and perform feature fusion at the slice level, which may overlook complementary spatial information across orientations. In contrast, Triamese-ViT processes three orthogonal views of the entire 3D brain volume (M_x, M_y, M_z) through independent ViT branches, ensuring that global and orientation-specific features are preserved.

Secondly, unlike most existing architectures, Triamese-ViT incorporates a built-in interpretability mechanism: attention maps generated from each view can be directly integrated into 3D-like maps, providing biologically meaningful explanations. This makes Triamese-ViT not only accurate and efficient but also inherently interpretable, which is critical for brain age estimation and related neuroimaging applications.

4.3 Results

4.3.1 Comparison With State-of-the-Art Algorithms for Brain Age Estimation

The changes of Triamese-ViT's training loss are shown in Figure 4.3.

We employed Triamese-ViT to estimate brain age using MRI scans from a cohort of 1,351 healthy individuals aged 6 to 80 years. The dataset was partitioned into 70% for training, 15% for validation, and 15% for testing, ensuring a rigorous evaluation of the model's performance.

Model performance was assessed using four key metrics: Mean Absolute Error (MAE), the Spearman correlation coefficient between predicted and chronological age (r), the absolute value of the Spearman correlation coefficient between chronological age and the Brain Age Gap (BAG) ($|rp|$), and the coefficient of determination (R^2). The MAE, r, and R^2 evaluate the model's predictive accuracy and the degree of correlation between predicted and

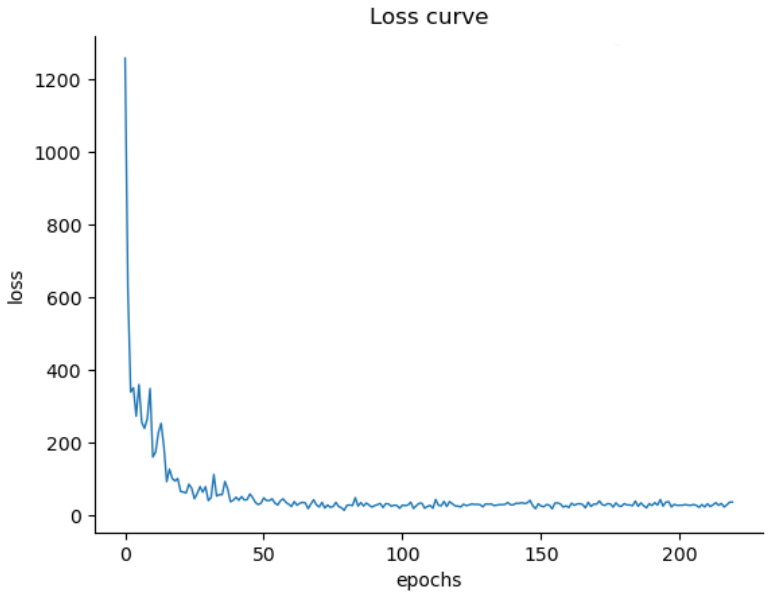


Figure 4.3: Changes of Triamese-ViT's training loss.

chronological ages, whereas $|rp|$ quantifies potential age bias, with a higher $|rp|$ indicating greater bias.

To validate the effectiveness of Triamese-ViT, we compared its performance against state-of-the-art algorithms for brain age estimation. Table 4.3 provides a comprehensive comparison of the Triamese-ViT model with ten other models, encompassing both classical and state-of-the-art (SOTA) approaches in brain age estimation. The comparison includes four established 3D CNN-based models: a 5-layer CNN, ResNet, VGG16, and VGG19. Additionally, our model was benchmarked against six other SOTA methodologies:

- The Two-Stage-Age-Network, which employs a two-stage cascade architecture where the first-stage network estimates a preliminary brain age, and the second-stage network refines this estimate based on the discretized output of the first-stage network.
- The Global-Local Transformer, which utilizes 2D brain slices for age prediction.
- EfficientNet, recognized for its ensemble architecture.
- The Multiple Instance Neuroimage Transformer, a 3D ViT model.
- ITSVR, an improved twin support vector regression method.
- 3D-TDR, a tensor-distribution-regression model built upon 3D convolutional neural networks.

Algorithm	MAE	r	$\|rp\|$	R^2	Memory	Training Time
ResNet (Cole and Franke, 2017)	4.11	0.84	0.33	0.70	958 MB	3978s
VGG19 (T.-W. Huang et al., 2017)	4.09	0.7	0.49	0.68	2.27 GB	6889s
VGG16 (T.-W. Huang et al., 2017)	5.32	0.6	0.41	0.64	2.18 GB	6453s
5-layer CNN (Couvy-Duchesne, Faouzi, et al., 2020)	4.55	0.79	0.47	0.71	2.46 MB	1232s
Global-Local Transformer (He, Grant, and Ou, 2021)	4.68	0.77	0.32	0.73	617 MB	3014s
Two-Stage-Age-Network (Cheng et al., 2021)	3.93	0.91	0.38	**0.81**	1.52 GB	5162s
Efficient Net (Poloni, Ferrari, Alzheimer's Disease Neuroimaging Initiative, et al., 2022)	4.55	0.88	0.4	0.77	72 MB	1296s
Multiple Instance Neuroimage Transformer (Singla et al., 2022)	3.90	0.9	0.36	0.77	4.62 GB	10750s
ITSVR (Ganaie, Tanveer, and Beheshti, 2024)	4.21	0.75	0.35	0.71	3.57 GB	7741s
3D-TDR (L. Chen and Luo, 2023)	3.97	0.85	0.42	0.80	2.15 GB	6026s
Our Triamese-ViT	**3.85**	**0.94**	**0.3**	**0.81**	3.99 GB	8596s

Table 4.3: The details of tested algorithms' performance. Since the input of Global-Local Transformer should be a 2D image, we extract 2D slices around the center of the 3D brain volumes in the axial as input, which is the same process method as (He, Grant, and Ou, 2021). Other algorithms' input are 3D MRIs with dimensions (91,109,91). Our Triamese-ViT has consistently achieved the best among all measures.

As shown in Table 4.3, the Triamese-ViT model achieves the best performance in terms of Mean Absolute Error (MAE), recording a value of 3.85. The Multiple Instance Neuroimage Transformer follows closely with an MAE of 3.90, while the Two-Stage-Age-Network reports an MAE of 3.93. In contrast, the highest MAE, indicating the lowest accuracy, is observed in VGG16 at 5.32.

Regarding the Spearman correlation between predicted and chronological ages (r), the Triamese-ViT model outperforms all competitors, achieving a correlation coefficient of 0.94. The Two-Stage-Age-Network follows with a correlation of 0.91, and the Multiple Instance Neuroimage Transformer reports a correlation of 0.90. VGG16 demonstrates the weakest performance, with a correlation of only 0.60.

In terms of the absolute value of the Spearman correlation between the Brain Age Gap (BAG) and chronological age ($|rp|$), which quantifies the model's fairness, the Triamese-ViT model achieves the most favorable outcome with a correlation of -0.3, indicating reduced age

bias. Conversely, VGG19 exhibits significant age bias with a correlation of 0.49. ResNet and the Global-Local Transformer also demonstrate lower age bias, with correlations of 0.33 and 0.32, respectively.

For the coefficient of determination (R^2) between predicted and chronological ages, both the Triamese-ViT and the Two-Stage-Age-Network exhibit strong performance, achieving $R^2 = 0.81$, followed by 3D-TDR with $R^2 = 0.80$. EfficientNet and the Multiple Instance Neuroimage Transformer also perform well, each achieving $R^2 = 0.77$. In contrast, VGG16 demonstrates the lowest performance with $R^2 = 0.64$.

In terms of memory consumption, the 5-layer CNN exhibits the lowest memory requirement, utilizing only 2.46 MB due to its lightweight architecture. EfficientNet follows, requiring 72 MB, as it operates on only a single MRI slice, significantly reducing computational demands.

In contrast, the 3D ViT has the highest memory requirement, consuming 4.62 GB due to the complexity of processing full 3D volumes. While Triamese-ViT also requires a substantial memory allocation of 3.99 GB, its consumption remains notably lower than that of the 3D ViT, making it a more resource-efficient alternative for large-scale brain age estimation.

To provide a fair comparison of computational efficiency, we report the training times of all baseline models under identical experimental settings. All models were trained for 200 epochs on the same hardware environment consisting of three NVIDIA Tesla V100 GPUs. Under these settings, the lightweight 5-layer CNN required the shortest training time (approximately 1232 seconds), followed by EfficientNet (1296 seconds) due to its single-slice 2D input. The Global-Local Transformer and ResNet exhibited moderate training times of 3014 seconds and 3978 seconds, respectively, reflecting their balance of convolutional and attention-based operations. VGG16 and VGG19 were more computationally demanding, with training times of 6453 seconds and 6889 seconds, respectively. Two-Stage-Age-Network and 3D-TDR also required substantial resources, taking 5162 seconds and 6026 seconds. ITSVR showed relatively higher computational cost (7741 seconds) due to its large memory footprint. The Multiple Instance Neuroimage Transformer was the most computationally expensive baseline, requiring 10750 seconds. By comparison, our proposed Triamese-ViT achieved a measured training time of 8,596 seconds, which, although longer than standard 3D CNNs, is significantly lower than that of a full 3D ViT.

This comparative analysis highlights the superior performance of the Triamese-ViT model in brain age estimation, emphasizing its advantages in both accuracy and fairness over other leading models in the field.

4.3.2 Ablation Study

In this part of our study, we perform ablation experiments to investigate and justify the architectural design choices of Triamese-ViT. Specifically, we examine the impact of the number of layers in the Tri-MLP module. Keeping all other parameters constant, we

systematically vary the number of MLP layers and assess their effect on model performance.

The results, illustrated in Figure 4.4, reveal a distinct trend in Mean Absolute Error (MAE) as a function of the number of MLP layers in Triamese-ViT. Initially, as the number of layers increases from 4 to 6, the MAE rises, followed by a decline after 6 layers, reaching its minimum at 9 layers. Beyond this point, the MAE increases again at 10 layers. This pattern suggests that an optimal number of MLP layers exists to balance model complexity and predictive accuracy. The observed fluctuations in MAE across different layer configurations highlight the intricate interplay between model depth and performance, underscoring the necessity for precise architectural tuning.

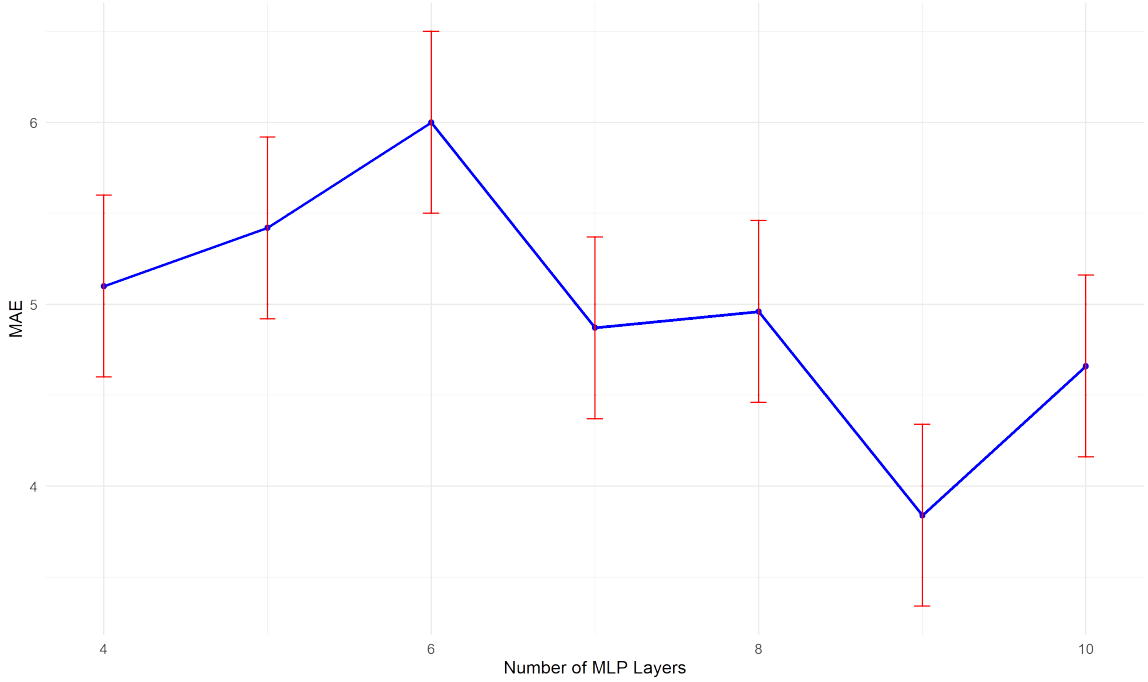


Figure 4.4: The impact of the number of MLP layers in Triamese-ViT.

Next, we examined the backbone architecture of Triamese-ViT. To assess the impact of different backbone models, we replaced the original ViT with alternative architectures, including ResNet, a 5-layer CNN, and VGG19. These models were integrated with the Tri-MLP to evaluate their influence on overall performance. The results of this experiment are presented in Table 4.5.

Our findings indicate that the original ViT backbone is the most effective for the Triamese framework. The 5-layer CNN also demonstrates considerable adaptability, achieving an MAE of 4.00, a Spearman correlation coefficient (r) of 0.85, $\|rp\|$ of 0.45, and an R^2 value of 0.72. In contrast, ResNet and VGG19 exhibit significantly poorer performance within the Triamese structure, both yielding MAEs exceeding 10—suboptimal results for brain age estimation.

These findings underscore the importance of selecting an appropriate backbone model to optimize the performance of the Triamese framework.

We further explored alternative fusion strategies for integrating the outputs from the three ViT branches in our Triamese-ViT model. Specifically, we compared our original MLP-based fusion layer against two alternative fusion mechanisms: convolutional attention (utilizing the Convolutional Block Attention Module—CBAM) and self-attention. In all cases, a four-layer MLP was applied after the fusion step to generate the final predictions.

Our experimental results indicate that the CBAM-based fusion strategy achieved promising performance, yielding a Mean Absolute Error (MAE) of 4.23, a Pearson correlation coefficient (r) of 0.81, a $\|rp\|$ of 0.35, and an R^2 of 0.78, suggesting good accuracy and fairness. In contrast, the self-attention fusion approach demonstrated inferior performance, with an MAE of 6.57, r of 0.52, $\|rp\|$ of 0.41, and an R^2 of 0.64.

Despite the insights gained from these comparative analyses, both alternative fusion methods underperformed relative to our original MLP-based fusion layer in Triamese-ViT. These results substantiate our methodological choice of MLP-based fusion, reinforcing its robustness and effectiveness in multi-view brain age estimation.

We then investigated the contribution of individual components within Triamese-ViT, specifically analyzing the performance of each of the three Vision Transformers (ViTs) aligned along different MRI axes. These are defined as ViT_x (dimensions: $91 \times 109 \times 91$), ViT_y (dimensions: $91 \times 91 \times 109$), and ViT_z (dimensions: $109 \times 91 \times 91$). Evaluating the predictive performance of these orientation-specific ViTs provides insight into the efficacy of the combined Triamese-MLP structure.

Additionally, we introduced and tested a variant model, $Triamese_{map}$, which, like Triamese-ViT, utilizes three ViTs processing different viewpoints. However, unlike the standard Triamese-ViT, each ViT in $Triamese_{map}$ outputs a feature map from the Transformer Encoder rather than directly producing a prediction from the MLP Head. The Triamese MLP in this variant model then concatenates the feature maps from the three ViTs and generates the final prediction.

A comparative performance analysis of these models—including each individual orientation-specific ViT and the $Triamese_{map}$ variant—is detailed in Table 4.5. The results suggest that the integration of the Triamese MLP substantially enhances performance. Among the individual ViTs, ViT_x achieves the second-best MAE of 4.42, while ViT_z exhibits the highest MAE of 5.29, indicating the weakest performance. In terms of Spearman correlation (r), ViT_y attains the highest value of 0.92, closely followed by the full Triamese-ViT model. Notably, $Triamese_{map}$ reports the lowest correlation value of 0.61, suggesting a weaker relationship between predicted and chronological age.

Regarding model fairness, as measured by the absolute value of the Spearman correlation between the Brain Age Gap (BAG) and chronological age ($|rp|$), $Triamese_{map}$ exhibits the lowest correlation, indicating a substantial reduction in age bias. For R^2, ViT_y achieves the highest value of 0.79, whereas $Triamese_{map}$ records the lowest performance with an R^2 of

0.65.

When we divided sMRI into a sequence of flattened 2D squares, we set the side length of the square S as 7 in our model, since after experiments, we found a smaller patch size would increase sensitivity but lead to overly detailed attention maps, and a larger patch size would encompass too many regions within a single patch, reducing the granularity of the attention maps and potentially obscuring meaningful brain structure information.

All ablation experiments were conducted under identical computational settings to ensure fair comparison. Specifically, all models were trained for 200 epochs on three NVIDIA V100 GPUs. The training times of the ablation variants were shown in Table 4.5.

Overall, the results presented in Table 4.5 provide strong evidence supporting the effectiveness of the Triamese-ViT model in improving both accuracy and fairness in brain age estimation, thereby validating its design.

Algorithm	**MAE**	$\boldsymbol{r}$	$\boldsymbol{\lvert rp \rvert}$	$\boldsymbol{R^2}$	Training Time
VGG-Backbone	10.31	0.30	0.31	0.29	6921s
ResNet-Backbone	10.36	0.45	0.25	0.37	5372s
CNN-Backbone	4	0.85	0.45	0.72	4469s
CBAM-fusion layer	4.23	0.81	0.35	0.78	9458s
self-attention-fusion layer	6.57	0.52	0.41	0.64	10564s
ViT_x	4.42	0.78	0.33	0.71	3172s
ViT_y	4.99	0.92	0.29	0.79	**2988s**
ViT_z	5.29	0.73	0.37	0.7	3070s
ViT_{map}	5.04	0.61	0.55	0.65	9153s
Our Triamese-ViT	**3.85**	**0.94**	**0.3**	**0.81**	8596s

Table 4.5: The details of the backbone-changed, fusion-layer changed models and unique structures. ViT_x, ViT_y, and ViT_z are focusing on the individual contributions of the three Vision Transformers (ViTs) oriented along different axes of the MRIs in Triamese-ViT. ViT_{map} also utilizes three ViTs on different viewpoints but each ViT in $Triamese_{map}$ outputs a feature map from the Transformer Encoder, rather than a direct prediction from the MLP Head. Then the MLP in this variant takes as input the concatenated feature maps from the three ViTs to make the final prediction.

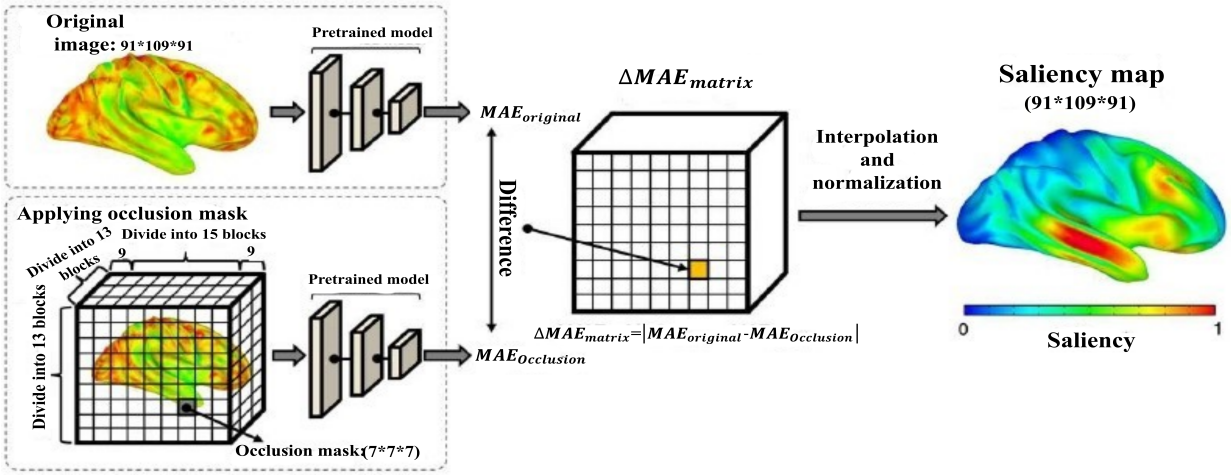


Figure 4.5: Illustration of the framework for occlusion analysis. In this work, occlusion analysis systematically obscures regions in brain MRI images using a 7×7×7 voxel mask to assess their impact on model predictions. By measuring changes in Mean Absolute Error (MAE) as the mask moves across the brain, a saliency map is generated, highlighting critical regions for age estimation. This image is adapted from J. Lee et al. (2022).

4.3.3 Explainable Results for Brain Age Estimation

Deep learning models often function as black boxes, where complex architectures and numerous parameters obscure the decision-making process. In this section, we aim to elucidate the predictive strategy of the Triamese-ViT model and enhance its interpretability using two distinct methods.

The first approach leverages 3D-like attention maps generated by Triamese-ViT, a built-in feature of the model. Given that the input consists of 3D MRI scans processed by three separate ViTs from different viewpoints (as illustrated in Figure 4.2), we obtain three distinct 2D attention maps corresponding to these perspectives. These 2D maps are subsequently expanded into 3D and averaged to generate a composite 3D attention map, providing insights into the spatial distribution of model focus across the brain.

The second method employs a well-established explainable artificial intelligence (XAI) technique known as Occlusion Sensitivity Analysis, as illustrated in Figure 4.5. This technique systematically occludes different regions of the input data to assess their influence on the model's predictions. In our case, specific regions of brain MRI scans are obscured using a cube-shaped occlusion mask of size $7 \times 7 \times 7$ voxels, where the enclosed voxels are set to zero. The mask is systematically moved throughout the entire brain volume, ensuring no overlap between successive positions. As the occlusion mask traverses the brain, variations in the model's predictions are observed. These changes, quantified in terms of Mean Absolute Error (MAE), compare prediction accuracy with and without occlusion. The magnitude

of MAE variation indicates the relative importance of different brain regions, with larger changes signifying greater relevance to the model's decision-making. The aggregation of these variations forms a saliency map, effectively highlighting the regions that the model predominantly relies on for brain age estimation.

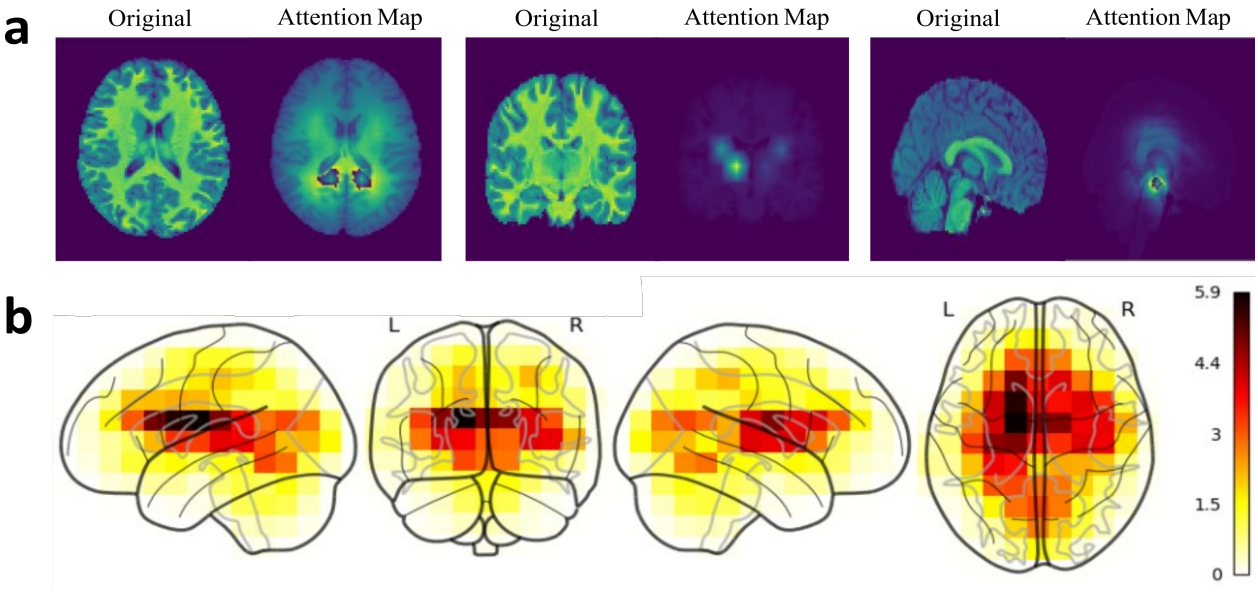


Figure 4.6: Comparison between the Triamese-ViT's attention map and occlusion analysis for healthy people. Figure 4.6.a presents the results from built-in interpretation compared to the original brain, while Figure 4.6.b shows the outcomes of the occlusion analysis. Together, these sections identify the specific brain regions that the Triamese-ViT model finds most crucial for age prediction.

Although attention-based visualization and occlusion-based saliency maps both aim to provide insights into the regions influencing the model's predictions, they differ in nature and methodology.

The attention maps are generated within the Transformer's self-attention mechanism. They represent how the model allocates weights across image patches during prediction, effectively highlighting regions the model attends to. These maps provide an internal view of the decision-making process, but they do not directly quantify the causal effect of each region on the output.

In contrast, occlusion sensitivity analysis is an external perturbation method. By systematically masking regions of the input MRI and observing the resulting change in prediction error, this method estimates the causal importance of each region. Regions where occlusion leads to a larger performance drop are considered more influential.

Together, these two approaches offer complementary perspectives: attention-based maps reveal the model's focus during learning, while occlusion analysis provides causal validation

of regional contributions. Their convergence strengthens the reliability of our interpretability findings.

Figure 4.6 presents the results of our interpretability analyses. Specifically, Figure 4.6.a compares the built-in attention-based interpretation with the original brain structure, while Figure 4.6.b illustrates the outcomes of the Occlusion Sensitivity Analysis. The detailed quantitative results of these two methods are provided in Table A.1.

As observed in Figures 4.6.a and 4.6.b, the most prominent regions, indicated by the brightest areas, are centrally located, suggesting a potential focus on deep brain structures for age estimation. The symmetry of these highlighted regions across both hemispheres aligns with the mirrored organization of many brain processes and structures.

Table A.1 further underscores the consistency between the two explainable AI (XAI) methods. For the attention maps, regions with attention values exceeding 3 are considered significant, while for Occlusion Sensitivity Analysis, regions with values above 4 are deemed critical. Both methods consistently identify the Rolandic Operculum, Cingulum, and Thalamus as key regions for brain age prediction. Additionally, the attention maps emphasize the significance of the Vermis, whereas Occlusion Sensitivity Analysis highlights the Insula, Caudate Nucleus, Putamen, and Heschl's gyrus as important regions in the estimation process.

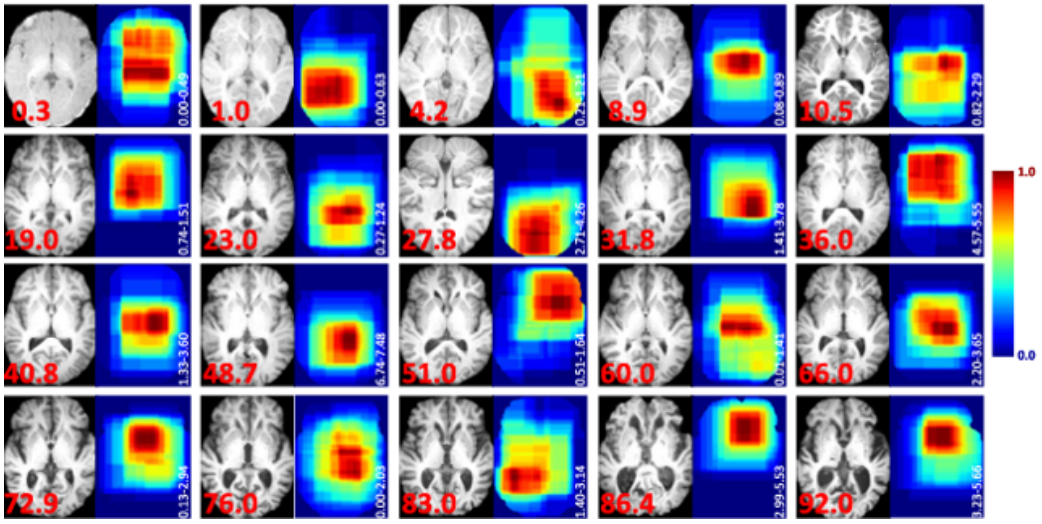


Figure 4.7: This figure is from He, Grant, and Ou, 2021. It shows the interpretability results from the Global-Local Transformer on brain age estimation.

To demonstrate the superior interpretability of Triamese-ViT, we compared its explanation results with those of another inherently interpretable model. Figure 4.7 presents the interpretability outcomes of the Global-Local Transformer (He, Grant, and Ou, 2021) for brain age estimation. A comparison between Figure 4.7 (Global-Local Transformer) and

Figure 4.6 (Triamese-ViT) reveals that the interpretability results of Triamese-ViT provide more detailed and informative insights.

While the Global-Local Transformer provides broad heatmap coverage, this makes it difficult to associate predictive relevance with distinct anatomical structures, as many adjacent regions are highlighted simultaneously. In contrast, Triamese-ViT generates more spatially precise and structurally aligned attention maps, enabling clearer identification of specific brain regions contributing to predictions.

In contrast, the 3D-like attention maps generated by Triamese-ViT effectively associate attention values with distinct brain structures, enabling a more precise identification of regions influencing model predictions. Furthermore, Triamese-ViT provides attention maps from three distinct orientations, offering a comprehensive 3D representation of the brain and enhancing interpretability.

It is important to note that broader coverage is not inherently inferior, nor is narrower focus inherently superior: both can capture valid aspects of the model's decision-making. However, in clinical and neuroscientific contexts, interpretability benefits from clarity and correspondence to known anatomical structures. The advantage of Triamese-ViT lies in its ability to balance coverage and specificity—highlighting critical regions without overextending across unrelated tissue areas.

4.3.4 Gender Differences in Explainable Results During Brain Age Prediction

In this section, we investigate potential gender differences in explainable results using the Triamese-ViT model. Specifically, we aim to identify brain regions that are particularly influential for age prediction in males and females, and to determine whether distinct patterns exist between genders during prediction.

To facilitate this analysis, two separate Triamese-ViT models were trained: one exclusively on healthy male subjects and another exclusively on healthy female subjects. Subsequently, the male-trained model was applied to a test dataset composed solely of male subjects to generate age predictions and associated explainability results. Likewise, the female-trained model was applied exclusively to a test dataset of female subjects to produce corresponding age predictions and explainability analyses. This approach allows us to comparatively assess gender-specific neural contributions to age prediction.

Figure 4.8 presents the results for male-based Triamese-ViT's interpretability analyses. Specifically, Figure 4.8.a compares the built-in attention-based interpretation with the original brain structure, while Figure 4.8.b illustrates the outcomes of the Occlusion Sensitivity Analysis. The detailed quantitative results of these two methods are provided in Table A.3.

From the explainability analysis of the male-specific Triamese-ViT model, the built-in interpretability results indicate that during male brain age prediction, the model primarily

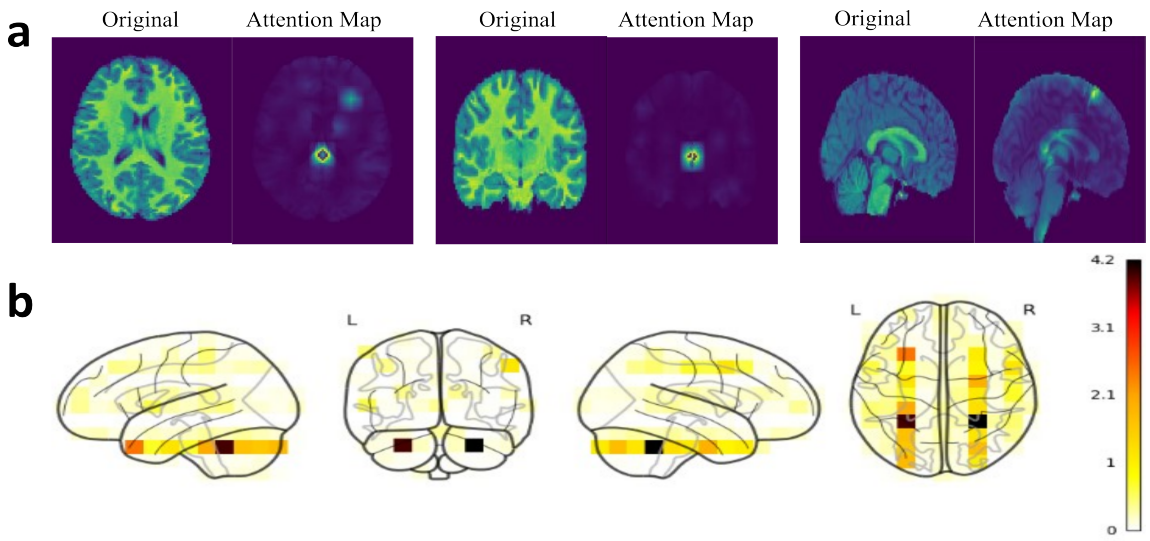


Figure 4.8: Comparison between the male-based Triamese-ViT's attention map and occlusion analysis for male healthy people. Figure 4.8.a presents the results from built-in interpretation compared to the original brain, while Figure 4.8.b shows the outcomes of the occlusion analysis. Together, these sections identify the specific brain regions that the male-based Triamese-ViT model finds most crucial for male individuals during age prediction.

attends to several key regions: the supplementary motor area, the medial portion of the superior frontal gyrus, the cingulum, the cuneus, the thalamus, and the vermis.

When comparing the built-in attention-based interpretation findings to those derived from the all-gender Triamese-ViT model (shown in Figure 4.6), we observe that both models highlight the importance of the cingulum, thalamus, and vermis. However, the male-specific model additionally emphasizes the supplementary motor area, the medial superior frontal gyrus, and the cuneus.

Occlusion analysis provides further insights, identifying the insula, cingulum, amygdala, calcarine fissure, cuneus, middle occipital gyrus, caudate nucleus, and thalamus as critical regions for male brain age prediction. These findings reinforce the significance of the cingulum, cuneus, and thalamus, while also indicating methodological differences in interpretability: the built-in interpretability approach more prominently highlights the superior frontal gyrus and vermis, whereas occlusion analysis places greater emphasis on the insula, amygdala, calcarine fissure, middle occipital gyrus, and caudate nucleus.

In a direct comparison of occlusion analysis between the male-specific and all-gender Triamese-ViT models (Figure 4.6), both models consistently identify the insula, cingulum, caudate nucleus, and thalamus as crucial regions. However, the male-specific model uniquely emphasizes the amygdala, calcarine fissure, cuneus, and middle occipital gyrus.

In summary, the cingulum, thalamus, vermis, insula, and caudate nucleus consistently

appear as essential regions in both male-specific and all-gender models. The male-specific Triamese-ViT additionally prioritizes areas such as the supplementary motor area, medial superior frontal gyrus, cuneus, amygdala, calcarine fissure, and middle occipital gyrus. This implies that these regions exhibit significant inter-individual variability among males and may serve as critical biomarkers for assessing male brain health and age-related differences.

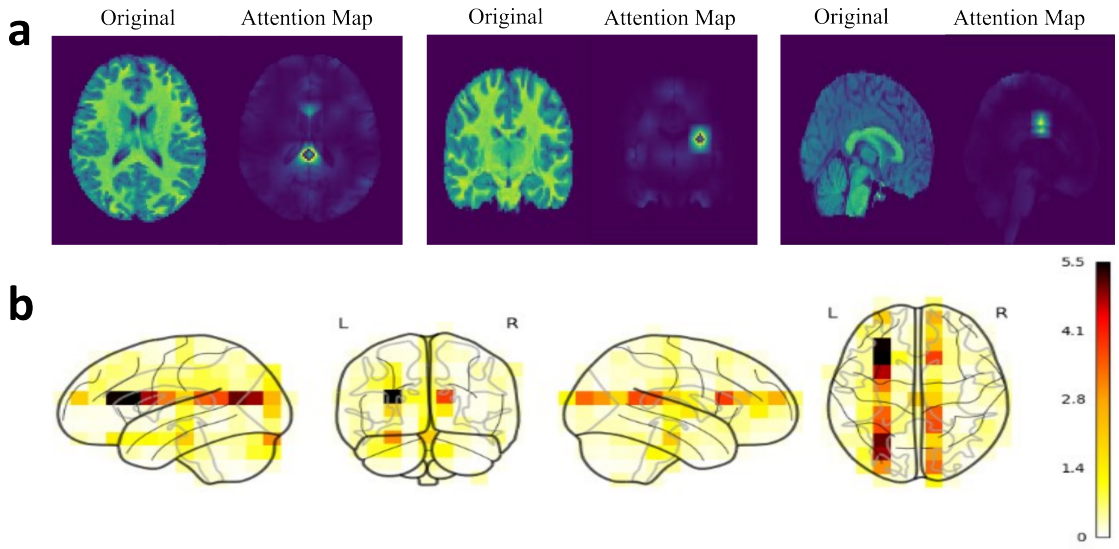


Figure 4.9: Comparison between the female-based Triamese-ViT's attention map and occlusion analysis for female healthy people. Figure 4.9.a presents the results from built-in interpretation compared to the original brain, while Figure 4.9.b shows the outcomes of the occlusion analysis. Together, these sections identify the specific brain regions that the female-based Triamese-ViT model finds most crucial for female individuals during age prediction.

Next, we analyze the results from the female-specific Triamese-ViT model. Figure 4.9 presents the interpretability analyses for this model. Specifically, Figure 4.9.a compares the built-in attention-based interpretability maps against the corresponding anatomical brain structures, while Figure 4.9.b depicts the outcomes from the occlusion sensitivity analysis. Detailed quantitative comparisons between these two interpretability approaches are further illustrated in Table A.3.

From the explainability analysis of the female-specific Triamese-ViT model, the built-in interpretability results demonstrate that the model primarily emphasizes several critical brain regions during female brain age prediction, namely the Rolandic operculum, insula, cingulum, cuneus, caudate nucleus, thalamus, and vermis.

Comparing these built-in attention-based findings with the interpretability outcomes from the all-gender Triamese-ViT model (Figure 4.6), both models consistently highlight the Rolandic operculum, cingulum, thalamus, and vermis. However, the female-specific model further emphasizes the insula, cuneus, and caudate nucleus.

Occlusion sensitivity analysis further complements these findings by identifying the insula, cingulum, amygdala, calcarine fissure, cuneus, middle occipital gyrus, caudate nucleus, and thalamus as particularly influential regions for female brain age prediction. These results reinforce the critical roles of the insula, cingulum, cuneus, caudate nucleus, and thalamus. Notably, the built-in interpretation prominently highlights the Rolandic operculum and vermis, whereas occlusion analysis places greater emphasis on the amygdala, calcarine fissure, and middle occipital gyrus.

When directly comparing the occlusion analysis between the female-specific and all-gender Triamese-ViT models (Figure 4.6), both consistently underline the importance of the insula, cingulum, caudate nucleus, and thalamus. Nevertheless, the female-specific model uniquely emphasizes the amygdala, calcarine fissure, cuneus, and middle occipital gyrus.

In summary, the Rolandic operculum, cingulum, thalamus, vermis, insula, and caudate nucleus emerge as consistently critical regions across both the female-specific and all-gender models. The female-specific Triamese-ViT additionally places greater emphasis on the cuneus, amygdala, calcarine fissure, and middle occipital gyrus. These regions exhibit pronounced inter-individual variability among females, indicating their potential as significant biomarkers for assessing female brain health and age-related changes.

Comparing explainability results directly between male- and female-specific models, both genders share critical regions such as the thalamus, cingulum, vermis, insula, caudate nucleus, cuneus, amygdala, calcarine fissure, and middle occipital gyrus. Most of these areas are also emphasized in the all-gender model, signifying their broad importance in brain age estimation. However, male predictions uniquely prioritize the supplementary motor area and medial superior frontal gyrus, while female predictions uniquely emphasize the Rolandic operculum. This divergence suggests gender-specific differences in brain structures following developmental processes.

Interestingly, our findings align with prior connectivity research, C. C. Yang et al. (2022) found that males exhibit higher regional network efficiency in the right Rolandic operculum, a region linked to motor control and response planning. Additionally, Callaghan et al. (2014) have reported sex-differential involvement of the supplementary motor area and middle frontal gyrus, with suppressed activation observed in males and slight enhancement in females, which corroborating the gender-specific emphasis observed in our attention maps. Together, these findings suggest that our model's interpretability results may reflect meaningful sex-dependent neurobiological differences in brain aging.

Additionally, all interpretability results consistently indicate higher importance for left-hemisphere regions compared to those in the right hemisphere. This lateralization might reflect the predominance of right-handedness among subjects, as greater utilization of the left hemisphere may result in more pronounced structural variability, thereby facilitating easier identification of age-related differences.

4.3.5 Normal Aging Analysis

The experimental results presented above demonstrate that the Triamese-ViT model achieves superior performance in brain age prediction for healthy individuals compared to both classical and state-of-the-art (SOTA) algorithms and may also provide enhanced interpretability through its attention maps, surpassing traditional explainable AI (XAI) methods.

Building on these findings, this section applies the Triamese-ViT model to analyze the normal aging process in the human brain.

Figure 4.10 presents the attention maps generated by the Triamese-ViT model across three different axes when predicting brain age for healthy individuals aged 0 to 80 years. Specifically, Figures 4.10.a, 4.10.b, and 4.10.c correspond to the attention maps along the x-axis, y-axis, and z-axis, respectively. Each attention map was computed by averaging attention values across each decade. Analyzing these maps provides valuable insights into the natural aging process from a machine learning perspective, helping to identify brain regions that play a significant role in age estimation.

In Figure 4.10, bright areas indicate regions of high relevance to the model's age predictions. Notably, the most prominent attention regions consistently appear near the center of the images, likely corresponding to deep brain structures. The symmetrical distribution of attention observed in Figure 4.10.a aligns with the mirrored organization of many brain structures and processes, which supports the reliability of the model's focus. Additionally, attention intensity tends to decrease toward the brain's periphery, suggesting that central structures may carry more informative features for age estimation than cortical areas. In several age groups, particularly in the lateral views, increased attention is directed toward the occipital lobe regions. The occipital lobes are associated with visuospatial processing, distance and depth perception, color discrimination, object and face recognition, and memory formation—functions that are known to change with age (Tohid, M. Faizan, and U. Faizan, 2015).

In younger age groups (0–10 years), the attention maps display a broader distribution of highlighted regions, potentially reflecting the rapid neurodevelopmental and maturation processes occurring during early life. As individuals progress into their teens, twenties, and thirties, the attention maps exhibit a more localized pattern, which may correspond to the stabilization of brain structure following developmental changes and the onset of subtle age-related modifications. From the thirties onward, there is a consistent emphasis on midline structures, potentially indicating alterations in white matter tracts, which are among the first to exhibit age-related changes. In the forties and fifties, deep brain structures are more frequently highlighted, reinforcing their relevance in the aging process. In the oldest age groups (sixties and seventies), attention becomes more diffusely distributed across the brain, suggesting that a broader range of structural changes becomes increasingly informative for age estimation.

In conclusion, these patterns align with established principles of brain development

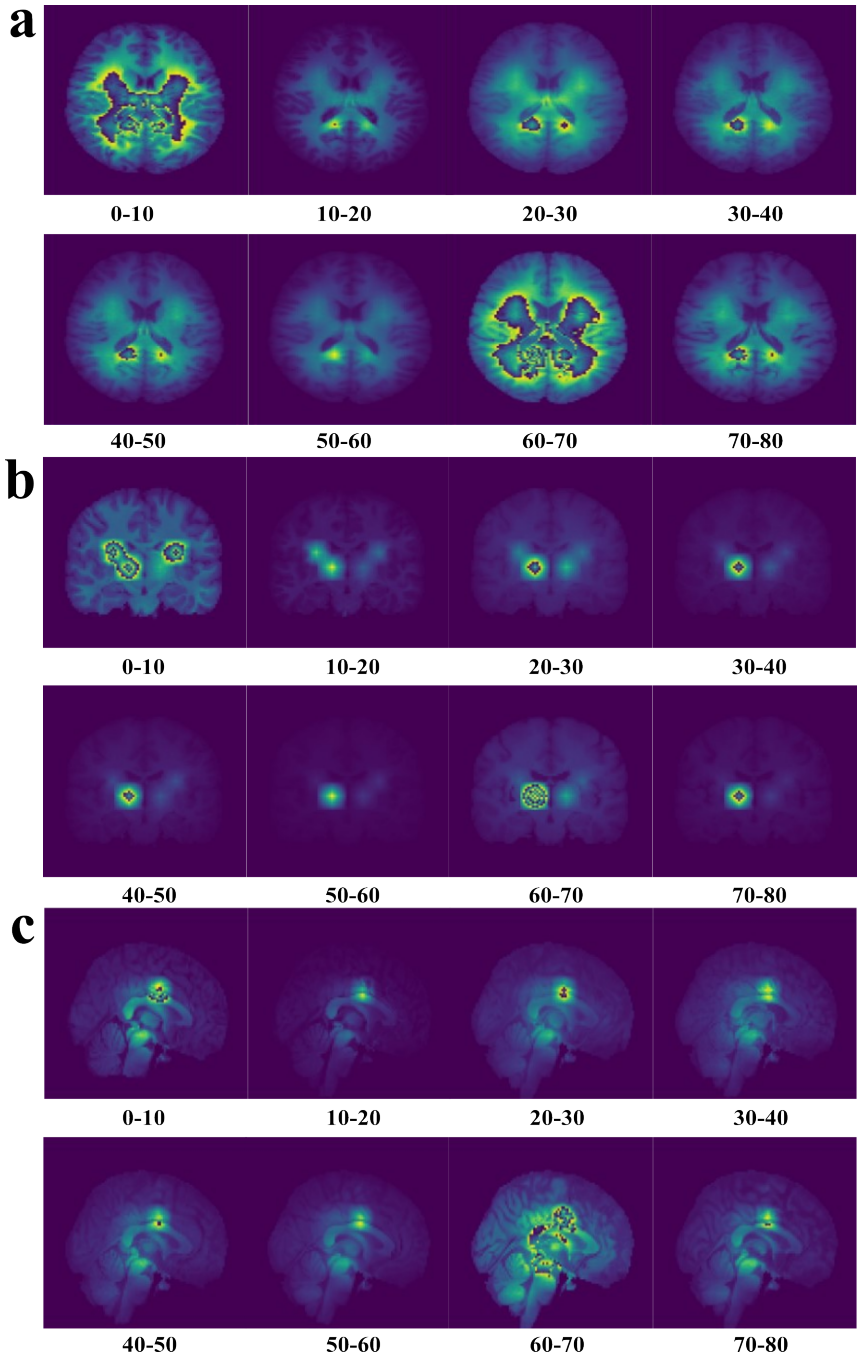


Figure 4.10: This figure represents the Triamese-ViT's attention maps from different axes of the MRIs during natural aging from 0 to 80 years old. **a** shows x-axis attention maps, **b** shows y-axis attention maps, and **c** shows z-axis attention maps. Each attention map was calculated by averaging the attention values over each decade.

and aging. The early years are characterized by dynamic neural changes, followed by a period of relative stability in early adulthood. Middle age marks the emergence of more localized age-related structural changes, particularly in deep brain regions, while older adulthood is associated with widespread structural alterations that become more prominent and informative for age estimation.

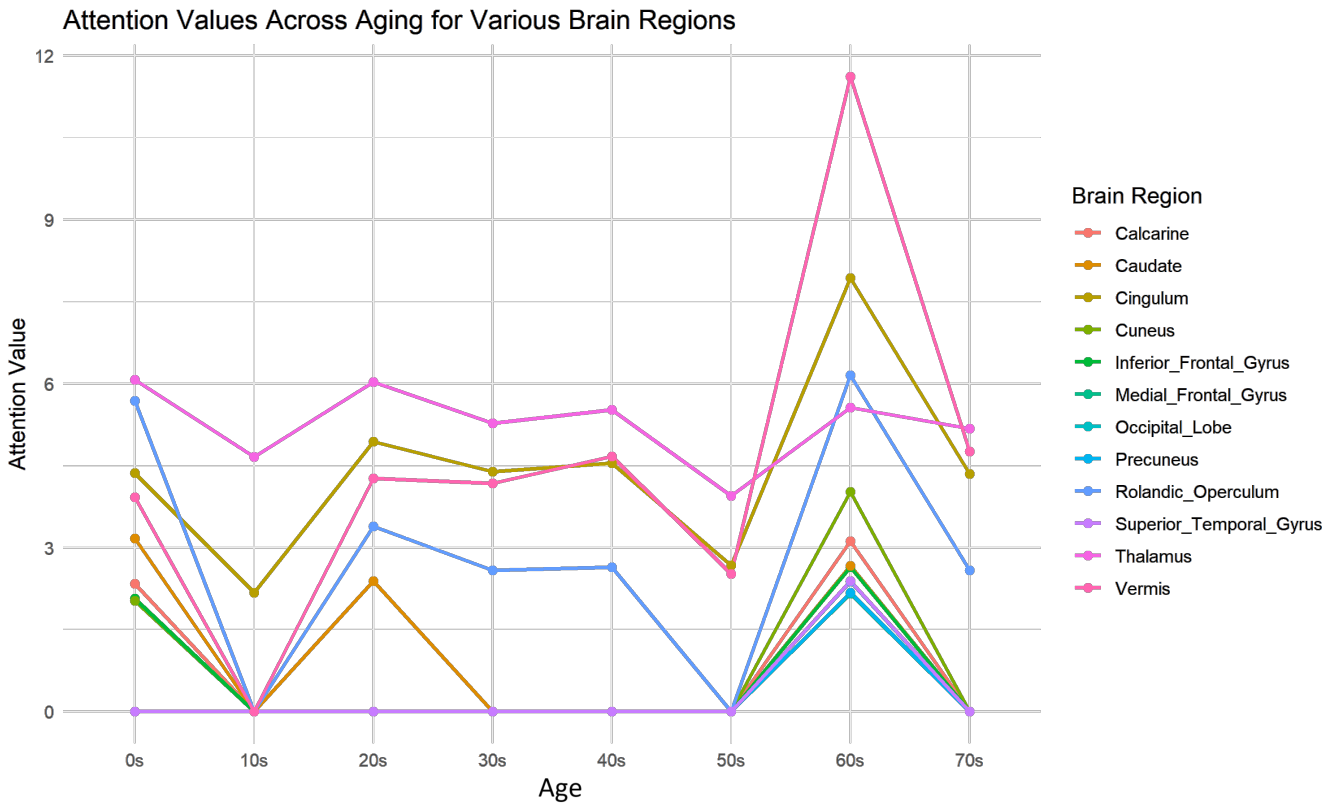


Figure 4.11: This figure presents the attention trend lines for the most important regions throughout natural aging based on the Triamese-ViT built-in interpretation.

Since the MRI scans used in this study are aligned to the standard MNI space, we can map the highlighted regions in Figure 4.10 to specific anatomical structures in the brain. Figure 4.11 provides a detailed analysis of these regions, illustrating the attention trend lines of highlighted areas during natural aging based on the attention maps generated by the Triamese-ViT model. Additionally, Table A.2 presents the corresponding attention values for each identified brain region.

To compute these attention values, we first extracted attention maps from three different views, as shown in Figure 4.10, and then expanded each into a 3D map with dimensions of $91 \times 109 \times 91$. The final 3D attention values for each brain region were obtained by averaging these 3D attention maps.

Figure 4.11 reveals distinct machine-learning-driven patterns in how different brain regions are highlighted throughout the aging process.

Early childhood (0–10 years): Regions such as the Inferior Frontal Gyrus, Rolandic Operculum, Cingulum, Calcarine, Caudate Nucleus, Thalamus, and Vermis exhibit significant attention, with the Thalamus and Rolandic Operculum receiving the highest values. These findings highlight their critical roles in early brain development and neural function.

Adolescence (10–20 years): Attention decreases across most regions, except for the Cingulum and Thalamus, which retain relatively high attention values. This trend likely reflects the maturation and stabilization of neural networks during this developmental stage.

Young adulthood (20–30 years): Moderate attention is observed in the Rolandic Operculum, Cingulum, Caudate Nucleus, Thalamus, and Vermis, potentially linked to ongoing cognitive and emotional development.

Middle age (40–50 years): Attention stabilizes across most brain regions, with slight increases in the Thalamus and Vermis, possibly reflecting their roles in maintaining cognitive function during this period.

Older adulthood (60–70 years): Attention resurges in several regions, including the Inferior Frontal Gyrus, Rolandic Operculum, Medial Frontal Gyrus, Cingulum, Calcarine, Cuneus, Occipital Lobe, Precuneus, and Vermis. The Vermis, in particular, shows a significant increase, which may be associated with age-related changes in coordination and balance.

Notably, the Thalamus and Cingulum demonstrate consistent significance across all stages, underscoring their crucial roles in neural and cognitive processes throughout life. These findings align with prior research, such as (Cera et al., 2019), which highlights the Cingulum's vulnerability to pathological aging, and (Fama and Sullivan, 2015), which emphasizes the Thalamus's role in cognitive networks and its structural and functional changes across the lifespan.

4.3.6 Artifacts Analysis

In Figure 4.12, the attention maps exhibit peculiar patterns, including ringing and wrapping artifacts in panels (a) and (c). In contrast, the target-like focus observed in the 10–20 age group (panel b) does not exhibit such artifacts. To further investigate these discrepancies, we conducted an in-depth analysis of the origin of these artifacts.

To systematically examine their source, we performed a series of controlled experiments using the attention map from the 40–50 age group (panel b) as a representative case, given its clear and analyzable structure.

We progressively subtracted a constant value from all attention values in the original attention map to observe how the pattern evolved (the maximum attention value in the selected attention map is 254). The results are visualized in Figure 4.12:

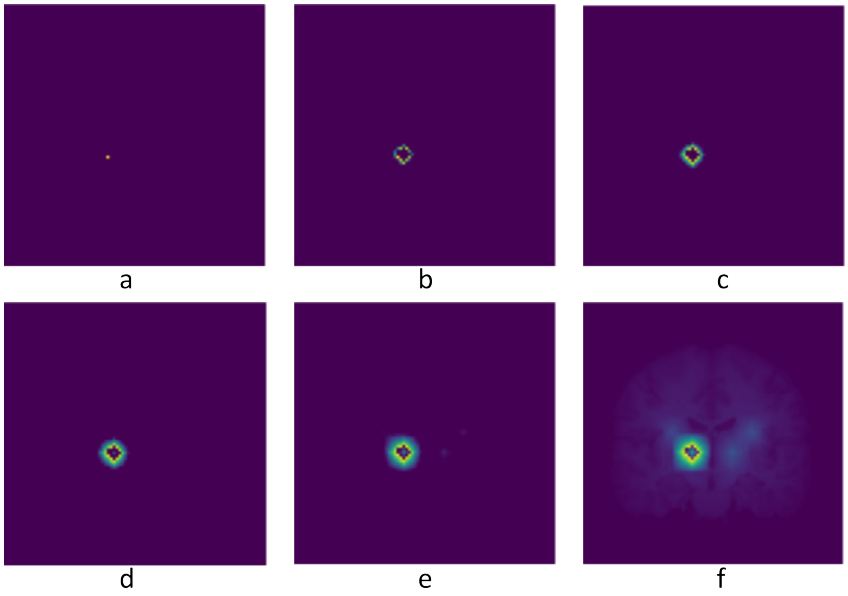


Figure 4.12: We progressively subtracted a constant value from all attention values in the original attention map from the 40–50 age group (panel b), to observe the pattern of artifacts.

- Figure (a): Subtracting 250—A single small bright spot remained, with no observable artifacts, suggesting no structural anomalies at high attention thresholds.
- Figure (b): Subtracting 200—A square-shaped region emerged, where the edges retained high attention values while the interior remained zero, forming an initial boundary.
- Figures (c)–(e): Subtracting 150, 100, and 50—Attention progressively diffused outward from the initial boundary, with decreasing brightness as distance increased. Meanwhile, previously zero-valued areas within the central region gradually gained nonzero values, leaving behind a narrow zero-value boundary, which visually resembles a ring-like artifact.
- Figure (f): The original attention map—After full intensity diffusion, an approximately square-shaped region with a dark boundary surrounding a bright inner region became evident, clearly illustrating the origin of these artifacts.

Our analysis confirms that these visual artifacts primarily result from a sharply delineated boundary within high-attention-value regions, where a thin border of zero intensity contrasts sharply with adjacent high-intensity areas. Re-examination of attention values confirmed that no computational errors or abnormalities were present.

Importantly, our interpretability approach averages attention values across all patches within specific brain regions, effectively reducing the influence of artifacts while ensuring

robust biological interpretability. Furthermore, the reliability of our interpretative results has been independently validated through:

- Occlusion analysis, confirming the biological relevance of high-attention regions.
- Alignment with established medical research findings, reinforcing the credibility of our conclusions.

We acknowledge that high-frequency intensity variations may impact interpretability clarity. These fluctuations likely reflect the model's sensitivity to fine-grained, biologically meaningful patterns inherent to high-resolution MRI inputs ($91 \times 109 \times 91$). To enhance interpretability in future studies, we propose exploring:

- Spatial smoothing techniques to mitigate high-frequency variations and enhance attention coherence.
- Attention regularization strategies to suppress excessive fluctuations while preserving critical information.

By addressing these challenges, we aim to further refine the interpretability and robustness of attention maps in Triamese-ViT, improving its applicability in brain age estimation and related neuroimaging tasks.

4.3.7 Contribution to ASD Diagnosis

To evaluate the impact of Triamese-ViT in disease diagnosis, we applied it to datasets of individuals with Autism Spectrum Disorder (ASD) to identify brain regions most associated with ASD.

Built-in Attention Mechanism in ASD Data: Since Triamese-ViT was trained exclusively on healthy samples, its attention mechanism was learned during training and remains fixed during inference. However, attention weights are influenced not only by the learned mechanism but also by the input features. If the ASD dataset differs from the healthy training data—whether in brain structure or function—the self-attention mechanism generates distinct attention maps due to altered relationships between input patches.

Applying the Triamese-ViT model, trained on healthy samples, to ASD data resulted in attention maps that deviated from those observed in healthy individuals. Using attention maps from healthy samples as a baseline, we analyzed these discrepancies to identify how the highlighted regions in ASD patients diverge from normal brain aging patterns. This comparison allows us to pinpoint critical brain regions associated with ASD, providing valuable insights into the neuroanatomical characteristics of ASD brains.

Occlusion Sensitivity Analysis: To further validate the findings from the attention maps, we conducted Occlusion Sensitivity Analysis as a benchmarking method. This analysis

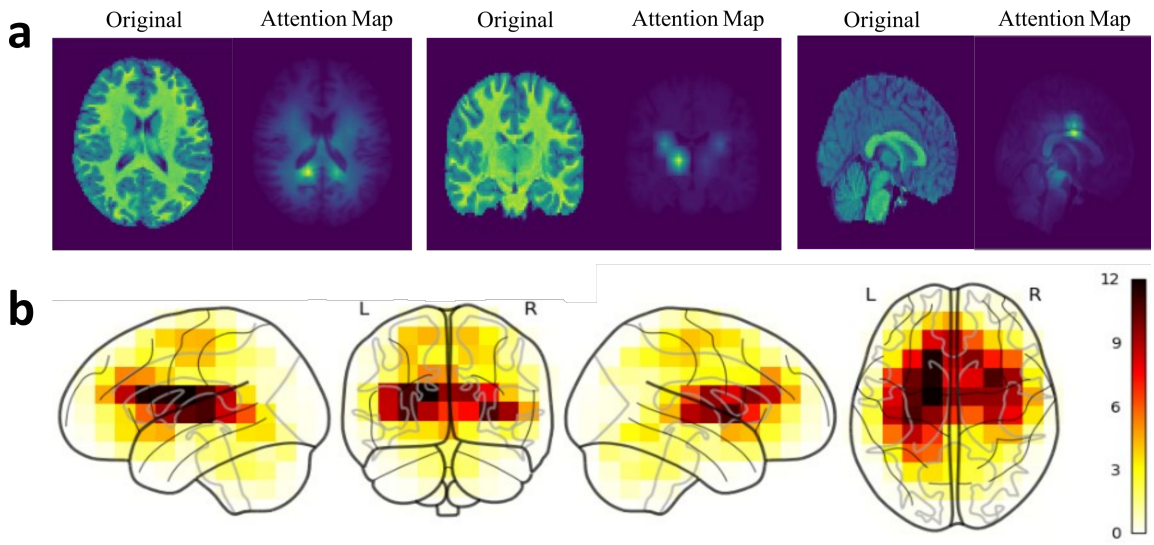


Figure 4.13: Comparison between the Triamese-ViT's attention map and occlusion analysis for ASD patients. Figure 4.13.a presents the attention map results compared to the original brain, while Figure 4.13.b shows the outcomes of the occlusion analysis. Together, these sections identify the specific brain regions that the Triamese-ViT model finds most crucial for ASD diagnosis.

systematically moves a mask across the entire brain volume without overlap to determine the relative importance of each region. The significance of a region is quantified by calculating the difference in Brain Age Gap (BAG) before and after occlusion ($BAG_{original} - BAG_{occlusion}$). A larger difference indicates a higher regional importance. Notably, positive differences highlight crucial areas, while negative values suggest regions of lesser significance.

The results are presented in Figure 4.13. Specifically, Figure 4.13.a displays the built-in interpretation results in comparison to the original brain, while Figure 4.13.b illustrates the outcomes of the occlusion analysis. Additionally, Table A.2 provides detailed importance scores for different brain regions, corresponding to the highlighted areas in Figure 4.13.a and Figure 4.13.b.

Both the built-in attention analysis and occlusion sensitivity analysis indicate that the Thalamus plays a significant role in ASD. Additionally, occlusion analysis highlights the Caudate Nucleus as another crucial region for ASD diagnosis. These findings align with existing medical research. For instance, (Schuetze et al., 2016) reported that individuals with ASD exhibit an expanded surface area in the right posterior thalamus, particularly in the pulvinar nucleus. They also observed a steeper increase in concavity of the caudal putamen with age in ASD individuals. Similarly, (Fu et al., 2019) analyzed dynamic functional network connectivity (dFNC) between 51 intrinsic connectivity networks in 170 individuals with ASD and 195 age-matched typically developing (TD) controls using independent component

analysis and a sliding window approach. Their study found that ASD is characterized by atypical large-scale subcortical-cortical connectivity, including disrupted resting-state functional connectivity between the thalamus and sensory regions. Furthermore, (Voelbel et al., 2006) compared neuropsychological test scores and caudate volumes in children with ASD, bipolar disorder (BD), and TD children. Their findings concluded that children with ASD exhibit larger bilateral caudate volumes and modest executive function deficits compared to TD controls.

4.3.8 Gender Differences in Explainable Results During ASD Diagnosis

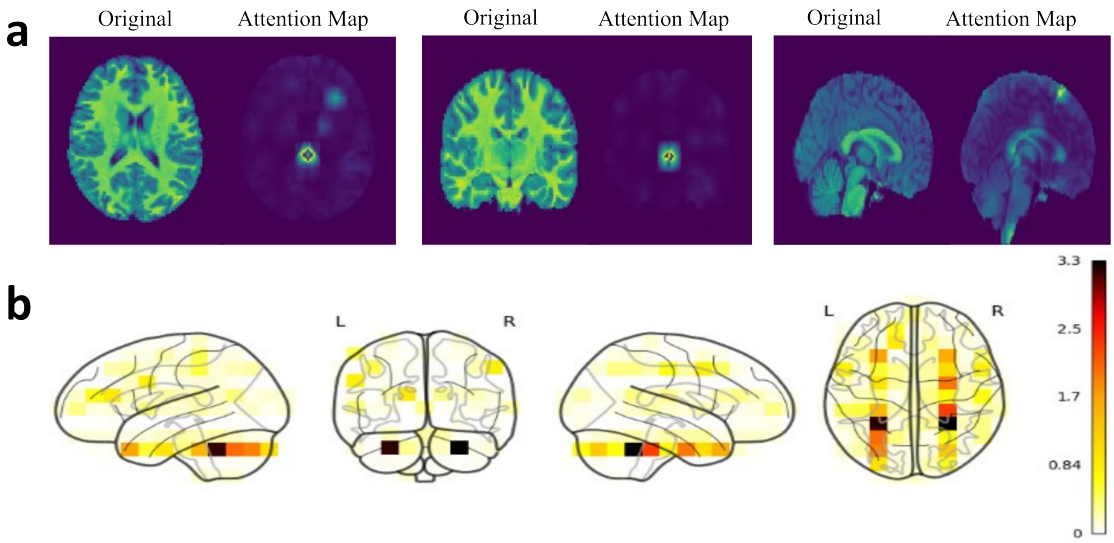


Figure 4.14: Comparison between the male-based Triamese-ViT's attention map and occlusion analysis for male ASD patients. Figure 4.14.a presents the results from built-in interpretation compared to the original brain, while Figure 4.14.b shows the outcomes of the occlusion analysis. Together, these sections identify the specific brain regions that the male-based Triamese-ViT model finds most crucial for male ASD individuals' diagnosis.

In this section, we explore potential gender differences in the interpretability results obtained from the Triamese-ViT model, specifically focusing on identifying critical brain regions for diagnosing ASD. Our goal is to highlight brain areas that significantly influence ASD diagnosis in males and females, and to investigate whether distinct regional patterns emerge between genders.

In the preceding analyses, two separate Triamese-ViT models were trained: one exclusively using male data and the other exclusively using female data. Here, we apply male-based models to male ASD patients and female-based models to female ASD patients, to

obtain interpretability results, allowing for direct comparison between the male- and female-trained models.

Figure 4.14 illustrates the interpretability analyses obtained from the male-trained Triamese-ViT model when diagnosing male ASD patients. Specifically, Figure 4.14(a) provides a comparison between the built-in attention-based interpretability results and the original anatomical brain structures. Figure 4.14(b) demonstrates the corresponding results obtained through occlusion sensitivity analysis. Detailed quantitative comparisons of these two methods are further presented in Table A.4.

From the explainability analysis of the male-specific Triamese-ViT model, the built-in interpretability results indicate that, during ASD diagnosis, the model primarily attends to several critical brain regions, including the supplementary motor area, the medial portion of the superior frontal gyrus, rectus gyrus, cingulum, cuneus, thalamus, and vermis.

Comparing these findings with the built-in attention-based interpretability results derived from the all-gender Triamese-ViT model (Figure 4.13), both models consistently highlight the thalamus as important. However, the male-specific model uniquely emphasizes additional regions, including the supplementary motor area, medial superior frontal gyrus, rectus gyrus, cingulum, cuneus, and vermis.

Further insights from occlusion sensitivity analysis highlight the cerebellum as another crucial region specifically for ASD diagnosis in males. These findings illustrate methodological differences between interpretability approaches: built-in interpretability prominently identifies the supplementary motor area, medial superior frontal gyrus, rectus gyrus, cingulum, cuneus, thalamus, and vermis, whereas occlusion analysis specifically emphasizes the cerebellum.

Directly comparing occlusion analysis outcomes between male-specific and all-gender Triamese-ViT models (Figure 4.13), the male-specific model distinctly prioritizes the cerebellum, while the all-gender model primarily emphasizes the caudate nucleus and thalamus.

In summary, the thalamus consistently emerges as a critical region for ASD diagnosis in both male-specific and all-gender models. Additionally, the male-specific Triamese-ViT model highlights the supplementary motor area, medial superior frontal gyrus, rectus gyrus, cingulum, cuneus, vermis, and cerebellum as particularly influential. These findings suggest that, among males, these areas exhibit substantial differences between ASD patients and healthy individuals, potentially serving as essential biomarkers for male ASD diagnosis.

We now analyze the interpretability results obtained from the female-specific Triamese-ViT model. Figure 4.15 presents these findings. Specifically, Figure 4.15(a) provides a comparison between the built-in attention-based interpretability maps and the corresponding anatomical brain structures, whereas Figure 4.15(b) illustrates results from the occlusion sensitivity analysis. Detailed quantitative comparisons between these two interpretability methods are further depicted in Table A.4.

The built-in interpretability analysis of the female-specific Triamese-ViT highlights the

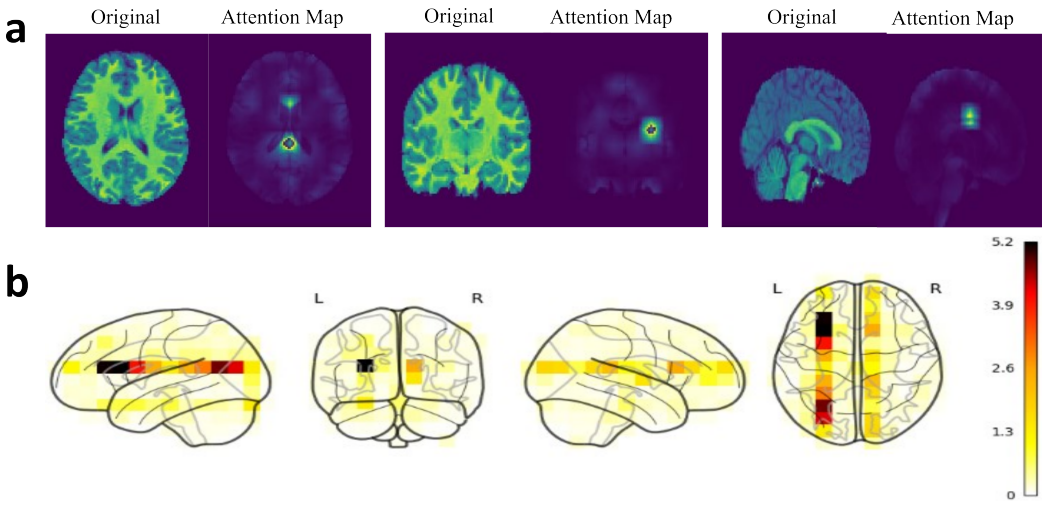


Figure 4.15: Comparison between the female-based Triamese-ViT's attention map and occlusion analysis for female ASD patients. Figure 4.15.a presents the results from built-in interpretation compared to the original brain, while Figure 4.15.b shows the outcomes of the occlusion analysis. Together, these sections identify the specific brain regions that the female-based Triamese-ViT model finds most crucial for female ASD individuals' diagnosis.

cingulum and thalamus as critical brain regions involved in ASD diagnosis for female patients.

When comparing these built-in attention-based results with those from the all-gender Triamese-ViT model (Figure 4.13), both models consistently emphasize the importance of the thalamus. However, the female-specific model additionally prioritizes the cingulum region.

Further insights provided by occlusion sensitivity analysis identify the cingulum, middle occipital gyrus, thalamus, insula, cuneus, and caudate nucleus as particularly influential areas for ASD diagnosis in females. These findings reinforce the critical importance of the cingulum and thalamus while also highlighting additional significant regions such as the middle occipital gyrus, insula, cuneus, and caudate nucleus.

Directly comparing occlusion sensitivity analyses between the female-specific and all-gender Triamese-ViT models (Figure 4.13), both models consistently underline the importance of the caudate nucleus and thalamus. However, the female-specific model uniquely emphasizes the cingulum, middle occipital gyrus, insula, and cuneus.

In summary, the thalamus and caudate nucleus consistently emerge as critical regions across both female-specific and all-gender models. Additionally, the female-specific Triamese-ViT model places greater emphasis on the cingulum, middle occipital gyrus, insula, and cuneus, suggesting that these regions exhibit significant variability among females and serve as valuable biomarkers for ASD diagnosis in female individuals.

When directly comparing the interpretability results between male- and female-specific models, both genders share key regions such as the cingulum, cuneus, and thalamus, which

are similarly emphasized by the all-gender model, indicating their general relevance to ASD diagnosis. Nevertheless, male-specific models uniquely prioritize the supplementary motor area, medial superior frontal gyrus, rectus gyrus, vermis, and cerebellum. In contrast, female-specific models specifically highlight the middle occipital gyrus, insula, and caudate nucleus. These divergences underscore gender-specific differences in brain regions critical for ASD diagnosis.

4.3.9 Improvements for Occlusion Analysis

In this section, we modify the occlusion methodology by masking entire anatomical brain regions rather than uniformly sized voxel cubes. This allows us to clearly identify the influence of each distinct brain structure on predictions of brain age and ASD diagnosis.

Firstly, we investigate the influence of this new occlusion approach on brain age estimation. Figure 4.16 visualizes the outcomes, while detailed quantitative results are presented in Table A.5. Comparing these new findings with the previous attention map and voxel-based occlusion results, we again observe the prominence of the thalamus and cingulum—regions consistently highlighted as critical across different analytical approaches. This consistency underscores the central role these structures play in various manifestations of brain aging. Notably, the updated occlusion method uniquely emphasizes the lingual gyrus, which was not identified as significant in previous analyses. The lingual gyrus, located in the medial occipital lobe, is associated with visual processing and memory. Supporting this finding, recent research by (Duan et al., 2024) revealed distinct patterns of brain aging, notably accelerated gray matter volume loss within medial occipital areas, including the lingual gyrus. This accelerated regional atrophy correlates strongly with biological aging and cognitive decline, thereby validating our new occlusion-based results and highlighting the lingual gyrus as an important area for understanding age-related brain changes.

Next, we applied the region-based occlusion analysis approach to the ASD patient dataset to identify crucial brain regions for ASD diagnosis. Figure 4.17 visualizes the results, while detailed quantitative findings are presented in Table A.5. Comparing these new outcomes to our previous attention map and voxel-based occlusion analyses, the thalamus emerges again as a consistently critical region across all interpretability methods, underscoring its significant role in ASD diagnosis. However, the region-based occlusion method additionally highlights the lingual gyrus and the cerebellum, regions that were not emphasized in prior voxel-based analyses.

The significance of these findings is supported by existing literature. For instance, Habata et al. (2021) reported increased cortical thickness within the lingual gyrus in adults with ASD, correlating with heightened visual sensory sensitivity and social impairments, suggesting a critical role of this region in the sensory processing anomalies characteristic of ASD. Similarly, Chandran et al. (2021) observed that increased gyrification and regional gray matter volume in the right lingual gyrus were associated with higher autistic traits, indicating morphological

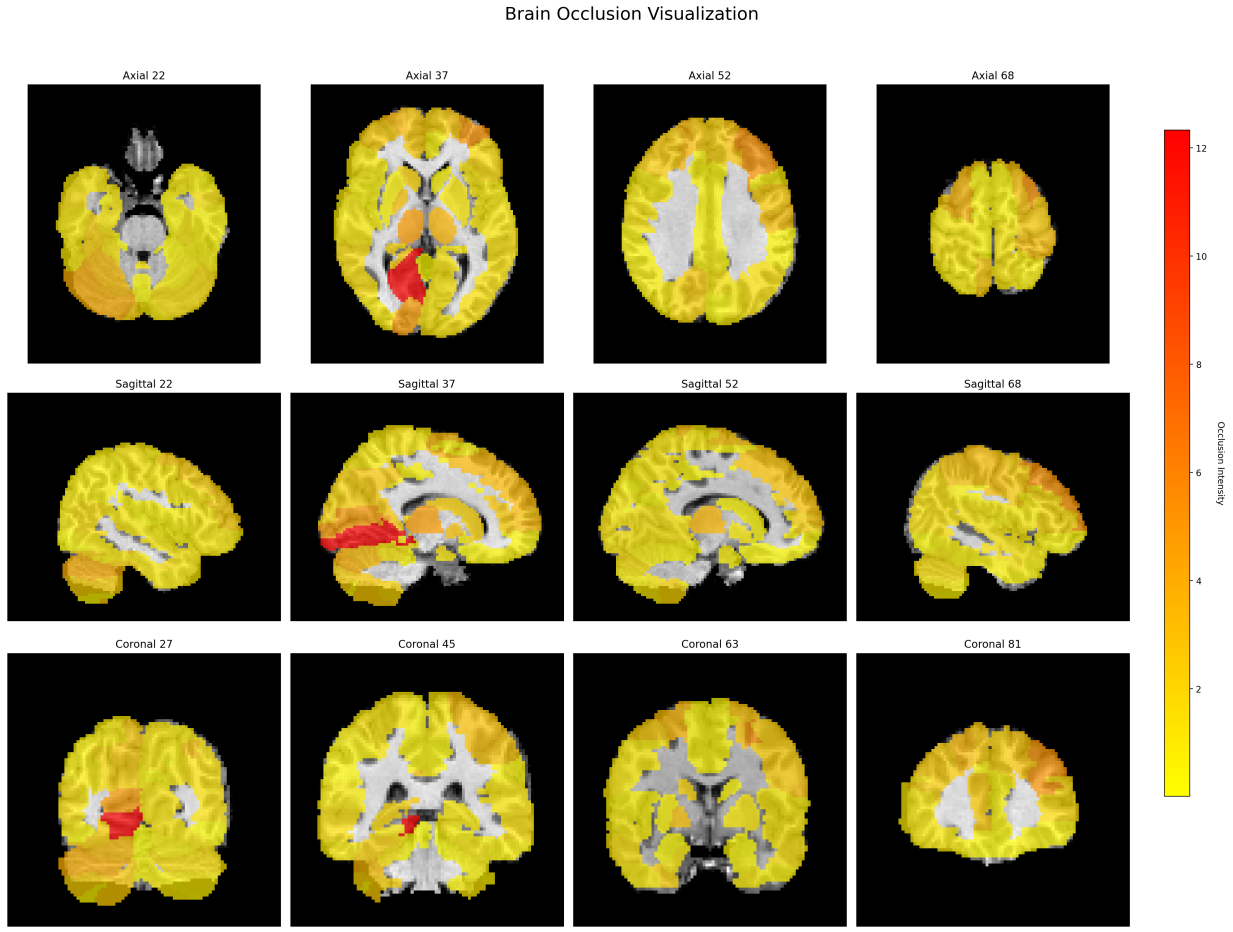


Figure 4.16: Results of region-based occlusion sensitivity analysis for brain age estimation. Each anatomical brain region was systematically masked in turn, enabling evaluation of its individual contribution to the model's predictions.

alterations in this region may reflect autism severity.

Regarding the cerebellum, prior research by D'Mello and Stoodley (2015) revealed atypical functional connections between the cerebellum and both motor and non-motor cortical regions in ASD patients. These atypical connections were hypothesized to contribute directly to core ASD symptoms. Additionally, their functional connectivity analyses suggested that disruptions in cerebellar coordination of cognitive and motor processes could underpin aspects of the disorder's symptomatology.

Thus, our region-based occlusion analysis further reinforces these findings, highlighting the lingual gyrus and cerebellum as significant biomarkers for understanding the neural underpinnings of ASD.

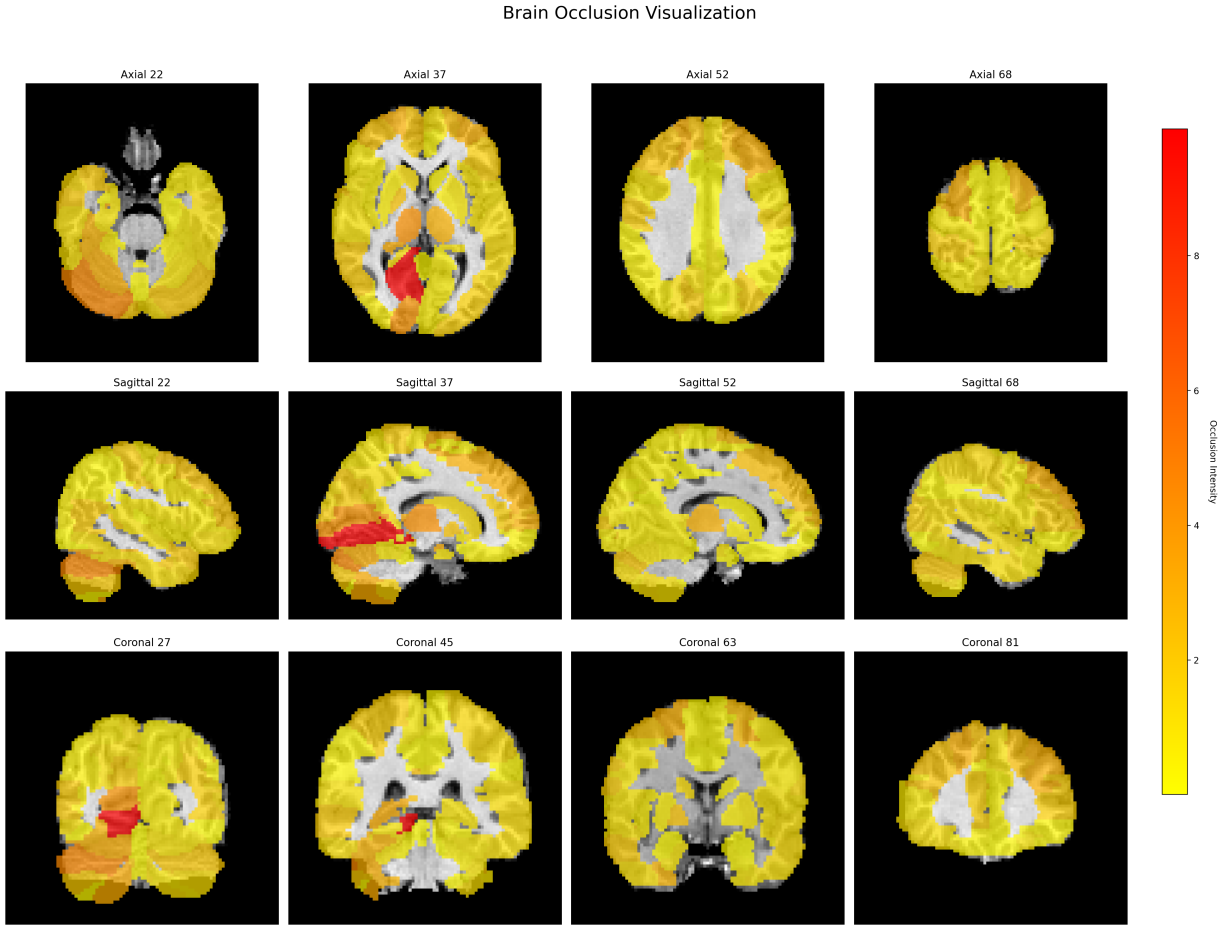


Figure 4.17: Results of region-based occlusion sensitivity analysis for ASD diagnosis. The analysis systematically obscured entire anatomical brain regions individually, allowing identification of key regions influencing the diagnostic predictions for ASD patients.

4.4 Discussion

Our research introduces Triamese-ViT, a deep-learning model specifically designed for brain age estimation. This model has been evaluated against other state-of-the-art (SOTA) models in the field, demonstrating superior performance. The most notable innovation of Triamese-ViT is its unique Tri-architecture, which integrates comprehensive contextual understanding with detailed image analysis. By exploring complex relationships between image patches, it achieves more precise, accurate, and interpretable predictions. This advancement holds significant potential for clinical applications, particularly in the early detection of neurodegenerative diseases and the development of personalized interventions based on individual brain health assessments.

In experiments conducted on a public dataset, Triamese-ViT achieved remarkable results, with a Mean Absolute Error (MAE) of 3.85, a Spearman correlation of 0.94 with chronological age, and a Spearman correlation of -0.3 between the Brain Age Gap (BAG) and chronological age. These results highlight both high predictive accuracy and a significant reduction in age bias, marking a substantial advancement in brain age estimation. Such accuracy is critical in clinical settings, where reliable brain age assessments can help detect deviations from typical aging, potentially signaling early neurodegenerative changes.

Beyond its predictive accuracy, Triamese-ViT offers substantial interpretability. We applied it to the analysis of natural brain aging and examined the attention values across different brain regions over time. The results reveal distinct patterns that align with known neurodevelopmental and aging processes. In early childhood, key regions such as the Inferior Frontal Gyrus, Rolandic Operculum, and Thalamus exhibit high attention values, underscoring their crucial roles in neural development. During adolescence, attention decreases across most regions, except for the Cingulum and Thalamus, reflecting the maturation of neural pathways. In young adulthood, increased activity is observed in regions such as the Rolandic Operculum and Thalamus, potentially associated with cognitive and emotional development. Middle age is characterized by stable attention values, with slight increases in the Thalamus and Vermis, suggesting their involvement in maintaining cognitive function. In older adults, attention resurges in multiple regions, particularly in the Vermis, likely due to age-related changes affecting coordination and balance. These findings are consistent with existing neuroscience research (Sutoko et al., 2020; Humbert et al., 2010; Karaman et al., 2021; Y. Zhang et al., 2007; Wiltshire et al., 2010; Xuereb et al., 1990; Maiti et al., 2021; Ho Park et al., 2001), demonstrating the potential of Triamese-ViT as a valuable tool for brain research. Clinically, this interpretability could assist healthcare professionals in understanding the neural mechanisms underlying aging and identifying deviations indicative of disease onset.

We also examined gender-specific differences during brain age prediction through interpretability analyses of the Triamese-ViT model. To achieve this, we trained separate Triamese-ViT models using exclusively male or female datasets. By comparing their built-in

interpretability results and occlusion sensitivity analyses, we identified several brain regions consistently important for both genders during brain age estimation, including the thalamus, cingulum, vermis, insula, caudate nucleus, cuneus, amygdala, calcarine fissure, and middle occipital gyrus. The significance of these regions indicates substantial structural variability associated with the aging process across both genders.

Additionally, distinct gender-specific biomarkers for brain vitality were observed. Specifically, male brain age prediction models uniquely prioritized the supplementary motor area and medial superior frontal gyrus, indicating these regions exhibit pronounced age-related differences among males. Conversely, female brain age predictions uniquely emphasized the Rolandic operculum, suggesting this region undergoes significant structural changes specific to females as they age. These divergent findings highlight gender-specific patterns in brain structural development and aging processes.

The interpretability of Triamese-ViT also has significant implications for disease diagnosis. Traditional diagnostic methods for brain disorders often require extensive time and rely on subjective clinical judgment, increasing the workload of medical professionals and potentially affecting diagnostic accuracy. Machine learning-based interpretability provides a complementary and objective perspective, facilitating faster and more reliable diagnoses. For instance, Triamese-ViT's attention maps highlight brain regions most relevant to a given diagnosis, aiding clinicians in making informed decisions and reducing the risk of diagnostic oversight.

We tested Triamese-ViT on a dataset of individuals with ASD and identified the Thalamus and Caudate Nucleus as key regions associated with ASD. These findings align with existing medical studies (Schuetze et al., 2016; Fu et al., 2019; Voelbel et al., 2006), further demonstrating the utility of Triamese-ViT in brain disease research. By identifying critical brain regions implicated in ASD, Triamese-ViT has the potential to support early diagnosis and facilitate the development of targeted interventions, ultimately improving patient outcomes.

We also explored gender differences in brain region importance for Autism Spectrum Disorder (ASD) diagnosis using the Triamese-ViT model. Our analysis indicated that the cingulum, cuneus, and thalamus emerged as consistently crucial biomarkers for ASD diagnosis across both genders, suggesting these areas are broadly involved in the pathology of ASD. Additionally, distinct, gender-specific biomarkers were identified: for males, the supplementary motor area, medial superior frontal gyrus, rectus gyrus, vermis, and cerebellum were uniquely highlighted, whereas for females, the middle occipital gyrus, insula, and caudate nucleus were particularly emphasized. These differences underscore important gender-specific variations in the neural structures implicated in ASD diagnosis.

Limitations and Future Work: While Triamese-ViT has demonstrated strong performance, several limitations warrant discussion. One limitation concerns the high-frequency variations observed in the model's attention maps. These fluctuations can reduce clarity and interpretability, potentially obscuring the biological significance of highlighted

regions. To address this, future work will focus on enhancing the stability of attention maps by exploring:

- **Spatial Smoothing:** Reducing noise to improve spatial coherence and emphasize biologically meaningful structures.
- **Attention Regularization:** Applying constraints to suppress excessive fluctuations and ensure consistent anatomical focus.

These refinements aim to improve the accuracy and interpretability of attention maps, thereby providing clearer insights into neurobiological processes associated with aging.

A second limitation lies in the normalization strategy used for interpretability. In the current implementation, attention maps are normalized independently within each age group, which enhances local visibility but prevents direct cross-group comparison. To overcome this, future work will explore a global normalization framework that applies a consistent scaling across all age groups. Such an approach would ensure intensity consistency in attention maps and improve comparability across both views and cohorts.

Another limitation relates to data efficiency. The model requires relatively large training datasets to achieve robust generalization, which may restrict applicability in clinical contexts with limited sample sizes. This challenge could be mitigated by leveraging transfer learning from large public datasets or adopting data-efficient approaches such as self-supervised learning.

In addition, Triamese-ViT imposes higher computational and memory demands compared to single-view CNNs or classical models. Training requires multiple GPUs and extended time, potentially limiting deployment in real-time or resource-constrained environments. Future work may employ architectural optimizations and model compression strategies (e.g., pruning or knowledge distillation) to reduce computational cost.

Like many deep learning models, Triamese-ViT is also sensitive to imaging artifacts, in-scanner motion, and site-specific variability. Although ComBat harmonization was applied to reduce scanner effects, other noise sources may still degrade performance. Robustness could be further enhanced through strategies such as adversarial training or noise-robust loss functions.

While our interpretability findings are consistent across methods and supported by prior studies, they have not yet been systematically verified by clinical or neuroscience experts. Future work will involve expert validation to ensure biological plausibility and clinical relevance, particularly in assessing whether the identified gender-specific biomarkers align with established neurobiological knowledge.

Another limitation stems from our approach to ASD analysis. We employed a model trained on healthy participants to indirectly identify atypical regions in ASD, rather than training a model specifically on ASD datasets. This indirect strategy provides useful insights but is not optimized for ASD detection. Due to the limited size of available ASD cohorts,

we prioritized transfer-based interpretability. Future efforts will focus on assembling larger ASD datasets and developing models explicitly trained for ASD, which could yield deeper and more clinically meaningful results.

Although our study highlights critical brain regions implicated in ASD, we did not provide direct quantitative validation of these findings. Future research will incorporate volumetric and morphometric analyses of highlighted regions in both healthy and ASD participants to determine whether attention-identified regions correspond to significant structural differences. Such quantitative validation would strengthen the clinical utility of Triamese-ViT in supporting disease detection and monitoring.

In addition, while the present study focused primarily on demonstrating the benefits of integrating three orthogonal views using a Multi-Layer Perceptron (MLP) fusion mechanism, more advanced strategies could further enhance both performance and interpretability. Future work will explore adaptive weighting mechanisms, attention-based fusion, and graph neural networks (GNNs) to better capture inter-view relationships and improve contextual learning.

Finally, expanding beyond structural MRI represents another promising direction. Multi-modal integration—including T1- and T2-weighted imaging as well as diffusion-weighted imaging (DWI)—could provide complementary information, improving accuracy, robustness, and generalizability in diverse neuroimaging contexts.

4.5 Conclusion

In this chapter, we presented the Triamese-ViT model, a novel transformer-based architecture designed for brain age estimation with built-in interpretability. By leveraging three orthogonal views of sMRI scans, Triamese-ViT integrates complementary structural features through a Tri-MLP fusion strategy, achieving state-of-the-art predictive performance while maintaining fairness across age groups. A key contribution of this work lies in its intrinsic interpretability, which we validated through occlusion sensitivity analysis and by identifying brain regions consistent with established neurobiological findings. Furthermore, the model provided new insights into normal aging trajectories, gender-specific structural differences, and clinically relevant biomarkers for ASD diagnosis.

Despite these strengths, several limitations must be acknowledged. First, the model requires substantial computational resources and training time compared to simpler architectures. Second, its reliance on large-scale datasets may restrict applicability in small-sample clinical settings. Finally, while interpretability results were consistent across methods, they have yet to be systematically validated by domain experts, which will be a crucial step in future work. Potential improvements include integrating multi-modality MRI data, employing advanced fusion strategies, and exploring efficiency-oriented methods such as token pruning, adapter-based fine-tuning, and model compression.

Overall, this chapter demonstrated that Triamese-ViT balances accuracy, fairness, and

interpretability, establishing a strong foundation for explainable and clinically meaningful neuroimaging AI. In the next chapter, we introduce the User-Centric Democratic AI Framework. This framework addresses fairness and bias more explicitly at the service and interaction level, aiming to democratize AI applications and enhance their accessibility and trustworthiness in real-world contexts.

Chapter 5

User Centric Democratic AI Framework

5.1 Introduction

The primary objective of artificial intelligence (AI) is to develop technologies that enhance human well-being and address critical societal challenges (Hodges, 2006; Taddeo and Floridi, 2018). Its origins can be traced back to efforts to simulate human intelligence (Hodges, 2006). Over the past five years, AI has been increasingly applied across various domains, including drug and vaccine discovery (Jumper et al., 2021), environmental problem-solving (Gomes et al., 2019), humanitarian crisis prediction (Tomašev et al., 2019), and policymaking (M. K. Lee et al., 2019). As AI continues to gain prominence, ethical considerations have become a central concern (Conitzer et al., 2017), with the public demanding greater transparency and autonomy in AI-driven decision-making processes (Montes and Goertzel, 2019), in other words, a more democratic AI.

Democratic systems rely fundamentally upon citizens' equitable rights of participation and fair representation (Diakopoulos, 2019). Although this principle remains incompletely achieved and frequently contested in practical contexts (Phillips, 2021), democracies persistently strive to broaden rights and include groups historically marginalized or excluded. Artificial intelligence, by contrast, inherently depends upon datasets reflective of past conditions and therefore poses a danger to these democratic aspirations by potentially perpetuating historical biases and inequalities into future contexts, thereby undermining democratic progress. Specifically, by forecasting human behaviors across varying scenarios using historical observations, AI distinguishes between individuals according to characteristics embedded within data. Consequently, this approach risks solidifying existing societal biases and reviving discriminatory patterns that society has sought—legally, socially, and politically—to abandon (Eubanks, 2018; Mayson, 2018; Mehrabi et al., 2021). It is therefore imperative to consistently monitor and rigorously audit AI systems and their practical

deployments.

Moreover, an individual's visibility and representation within AI systems are contingent upon their historical presence within the data. Artificial intelligence struggles to accurately identify and categorize individuals belonging to groups that have been inadequately represented historically in training datasets. For instance, minorities historically absent or minimally represented in visual datasets remain largely unrecognized by computer vision algorithms (Buolamwini and Gebru, 2018), and historically marginalized populations may be systematically excluded from associations with particular occupations, risking increased bias within employment procedures mediated by AI technologies (Caliskan, Bryson, and Narayanan, 2017). This overarching phenomenon holds profound implications for democratic governance: persistent invisibility of certain groups in datasets implies their reduced presence within AI-generated portrayals of the citizenry and diminished influence over predictive models concerning political attitudes, behaviors, interests, and grievances. Consequently, populations already disadvantaged or disenfranchised risk experiencing further exclusion and discrimination, particularly in the contexts of government service provision, policy formation informed by digitally captured preferences, or exposure to intensified state surveillance and persecution.

AI systems can also amplify the visibility of particular groups in ways that exacerbate inequalities. For instance, historically marginalized communities frequently appear disproportionately in criminal justice data, leading AI-driven policing and sentencing algorithms to disproportionately target individuals from these groups (Chouldechova, 2017; M. Matthews, S. Matthews, and Kelemen, 2022). Particularly in jurisdictions such as the United States, where convicted felons face varying restrictions on voting rights depending upon state regulations, systematic biases embedded within AI-based criminal justice approaches might cumulatively skew voter demographics, thereby further disenfranchising historically marginalized populations (Aviram, Bragg, and Lewis, 2017). Additionally, AI-driven methodologies have significant potential to influence electoral boundary delineation, raising the possibility of perpetuating or even deepening existing inequities (Cho and Cain, 2020). AI thus risks perpetuating structural discrimination by embedding historically biased patterns into present and future decision-making processes, even amidst societal efforts to move towards greater equality and fairness.

In other words, AI-generated depictions of public sentiment, collective political identity, and electoral district boundaries may consistently disadvantage groups that were historically marginalized. Unequal visibility to AI systems could correspondingly alter democratic influence, benefiting some groups while disadvantaging others. For example, AI technology could enhance the representation and policy influence of already advantaged populations by making their preferences, priorities, grievances, and opinions more conspicuous and accessible to policymakers. Meanwhile, groups less represented in data sets might have their interests and views systematically overlooked or undervalued within AI-driven political forecasting and policy impact analyses.

Moreover, AI applications can substantially reshape labor markets, often adversely. Although companies could theoretically harness automation technologies to complement human workers—allowing them to engage in more productive tasks and thus enhance the overall value of labor—in practice, enterprises tend predominantly to pursue automation as a strategy for labor cost reduction, substituting human tasks with AI systems (Acemoglu and Restrepo, 2018). This substitution diminishes employees' negotiating strength and income, shifting the balance toward capital at labor's expense, thus potentially exacerbating economic disparities and undermining collective bargaining capacities. Consequently, this economic weakening of labor could translate into diminished political representation and influence for workers (Gallego and Kurer, 2022).

It remains uncertain precisely which types of labor will be most affected by AI-driven technological advancement. Automation traditionally replaces routine-based human tasks, primarily impacting lower-skilled occupations (Acemoglu and Restrepo, 2022b). Nevertheless, successive waves of AI innovation have demonstrated that even tasks within white-collar and knowledge-intensive professions—previously considered resistant to automation—contain substantial routine components. Consequently, the political and economic repercussions of AI-induced transformations may extend far more broadly than earlier automation waves. Recent debates surrounding the effects of large language models and generative AI technologies on creative industries and software development exemplify this broader vulnerability. Illustratively, the 2023 Hollywood writers' strike revealed emergent tensions, as screenwriters sought contractual safeguards against studios utilizing AI for scriptwriting and related tasks (Wilkinson, 2023).

Simultaneously, AI holds significant promise for addressing labor shortages in aging populations by automating substitutable tasks, thus allowing shrinking workforces to concentrate on non-automatable roles. Such a redistribution of labor could sustain productivity amid demographic pressures confronting developed economies (Acemoglu and Restrepo, 2022a). However, fully leveraging AI's economic potential requires deliberate measures to ensure broad-based benefit distribution, rather than concentration within elite groups alone. Particularly with wealth gains arising from digital technologies, the traditional connection between innovation and widely shared prosperity appears increasingly fragile or disrupted. This provokes serious concern about whether elite groups disproportionately capture AI-generated economic benefits, while the general population disproportionately absorbs the risks associated with automation. If such imbalances persist, they could intensify societal inequalities and further erode democratic norms.

To date, AI technology has been predominantly developed by a limited number of major technology corporations, such as Microsoft, Google, and Facebook. Shen (2017) reported that approximately 10,000 individuals across just seven countries were responsible for coding the vast majority of AI systems worldwide. This concentration of control poses significant challenges, as it restricts the broader potential and accessibility of AI applications. In extreme cases, AI systems may exhibit biases and lack generalizability to diverse populations, thereby

undermining their intended societal benefits.

The potentially destabilizing nature of these developments underscores the critical importance of precisely how AI technologies are deployed and regulated. Clearly, AI intersects directly with democratic ideals of equality. Inequities might emerge in the distribution of public goods and government services mediated by AI systems, in citizens' visibility and representation within AI-derived analyses, and in economic prospects for individuals whose occupational responsibilities become replaceable through AI innovations.

In response to these concerns, Democratic AI (DemAI) has emerged as a framework aimed at ensuring public participation in AI-related decision-making. By aligning AI development with democratic principles, this approach seeks to promote Social Good beyond conventional professional ethical codes and legislative restrictions on technology and industry.

The democratization of AI refers to the principle of providing equitable access to AI-related resources, opportunities, and benefits (Strouse et al., 2021). However, the concept of Democratic AI remains loosely defined (Garvey, 2018). Broadly speaking, it can be interpreted as a system in which individuals, regardless of their technical expertise, are empowered to contribute to AI development and decision-making processes.

The democratization of artificial intelligence presents multiple benefits, particularly through its potential to diminish monopolistic control of AI technologies (Ahmed, Mula, and Dhavala, 2020). By lowering entry barriers, democratization allows individuals lacking specialized AI expertise to utilize and benefit from these technologies. The open dissemination of data, algorithms, and cloud computing resources further enables widespread access to AI, independent of users' financial resources or institutional affiliations. Publicly accessible datasets and algorithms facilitate more efficient and cost-effective resolution of complex AI challenges. Additionally, democratized AI holds considerable promise in reducing biases and promoting fairness within AI systems, mitigating disparities arising from gender, ethnicity, socioeconomic status, and related factors. Moreover, open-source platforms such as PyTorch and TensorFlow have significantly propelled advancements in deep learning research, encouraged talent cultivation, and accelerated the overall growth of AI knowledge. Collectively, these elements foster greater inclusivity, stimulate innovation, and amplify the societal benefits derived from AI.

Despite its potential benefits, Democratic AI faces several challenges. A key concern is the involvement of non-technical participants, which may compromise the rigorously designed ethical values embedded in AI systems by experts (Rao, 2020). Bias remains a significant risk in AI, even within systems developed by highly skilled engineers (Zou and Schiebinger, 2018). Allowing inexperienced contributors to influence AI development may lead to biased outcomes, erroneous conclusions, and unintended consequences. Moreover, identifying the sources of bias is inherently difficult, and remediation efforts can be costly and uncertain in their effectiveness. Additionally, the absence of a formalized definition or mathematical framework for Democratic AI leaves it largely conceptual, lacking a widely accepted implementation model. These unresolved issues have hindered its practical

development and widespread adoption.

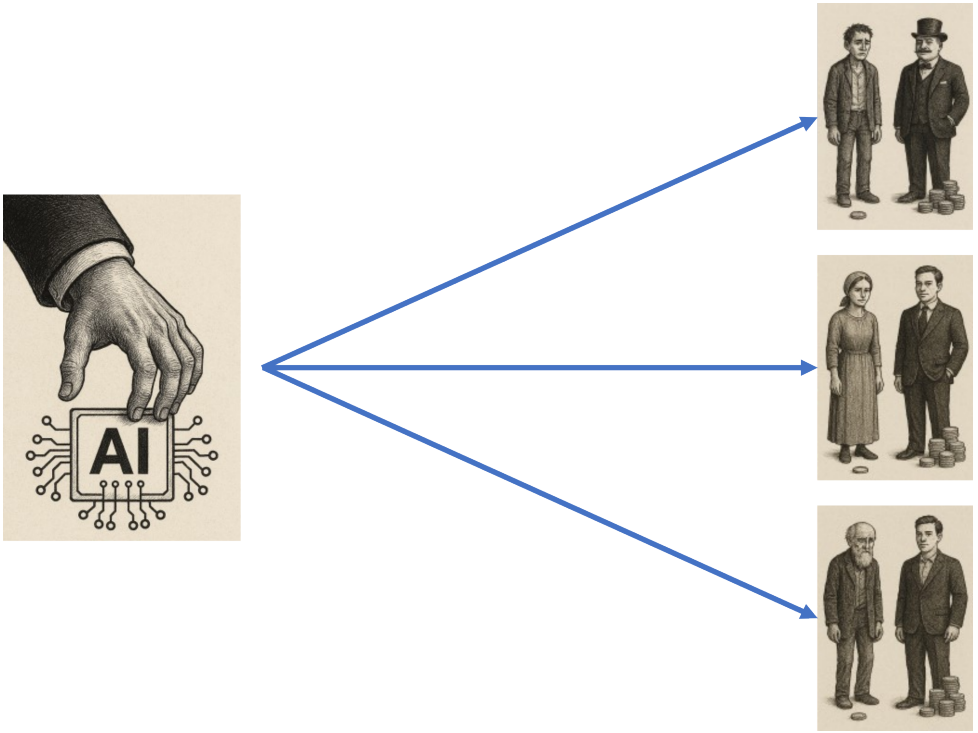


Figure 5.1: Despite advances in accessibility, AI development and deployment remain largely dominated by major corporate entities, which continues to influence the technology's effectiveness and fairness. As a result, systemic biases—such as those related to socioeconomic status, gender, and age—persist within AI systems and may be inadvertently reinforced. The figure was generated using ChatGPT (OpenAI, 2025).

In this work, we seek to address the challenges associated with Democratic AI by investigating its potential to tackle both societal and technological issues. To this end, we propose a user-centric Democratic AI (u-DemAI) framework, which optimizes individual benefits through an iterative user-in-the-loop process.

To validate the effectiveness of our framework, we conducted a case study on brain age estimation in medical services. Our experimental results demonstrate the substantial advantages of the u-DemAI framework, which actively involves users in the decision-making process and fosters autonomy within a community-driven AI ecosystem. Notably, the success of our experiments serves as proof of concept for the practical implementation of democratic AI, underscoring its benefits over traditional expert-driven AI systems by facilitating the participation of non-experts.

5.2 Preliminary

5.2.1 What is Democratic AI

In political theory, democracy is a system in which the people possess the power and liberty to determine their governing structure. With the advancement of Artificial Intelligence (AI), new discussions have emerged within the social sciences, particularly in areas such as AI governance and the implications of integrating AI into democratic systems. Conversely, when democratic principles are embedded into AI systems, a new research domain, referred to as Democratic AI (DemAI), arises. This concept, also known as AI democratization, aims to bring democratic processes into AI development and deployment. In AI-driven services, this implies that users should have greater authority in AI governance. To gain a deeper understanding of AI democratization, it is essential to first examine some fundamental concepts that serve as the basis for constructing a theoretical framework for DemAI.

As the demand for democratic AI increases, researchers have proposed various conceptualizations of ideal Democratic AI systems:

- (Nguyen et al., 2022) describe Democratic AI as a machine learning system that operates based on hierarchical self-organization within a distributed environment. Their framework consists of well-connected, decentralized learning agents that possess limited yet highly personalized data and dynamically adjust themselves through an interplay of specialized and generalized processes.
- (Shashi et al., 2022) argue that Federated Analytics (FA) provides a suitable foundation for Democratic AI. However, they note that the current implementation of Federated Learning (FL) follows a single-server, multiple-client architecture, which limits the generalization capacity of AI models. To address this, they propose a Democratic AI framework based on Federated Learning, aimed at enhancing the generalization capabilities of AI models across cloud-based systems.
- (Montes and Goertzel, 2019) highlight that AI development is currently dominated by a small number of centralized mega-corporations, whose priorities often align with their stakeholders' interests rather than public welfare. To counteract this oligopolistic control, they propose a Democratic AI framework as a decentralized, distributed market for AI services, leveraging distributed ledger technology to ensure transparency and equity in AI deployment.

By synthesizing these perspectives, we propose the following generalized definition of Democratic AI:

> *Democratic AI refers to an AI implementation that involves relevant stakeholders in the optimization of AI services, ensuring that these services promote social values and benefit the broader community.*

A key tenet of Democratic AI is the inclusion of people in the AI optimization loop, ensuring that AI services align with social values and remain accessible to all individuals. This approach fulfills (Montes and Goertzel, 2019) vision of breaking corporate monopolies in AI governance, while also addressing (Nguyen et al., 2022) expectation of enhancing generalization and personalization in AI services.

To formalize this concept, we propose a mathematical framework for Democratic AI, termed user-centric Democratic AI (u-DemAI), which will be introduced in the following sections and evaluated through a case study.

5.2.2 Overview of Our u-DemAI Framework

The u-DemAI framework proposed in this study is a comprehensive, multi-functional, and publicly accessible system that provides users with various optimization strategies, models, and datasets. It fosters collaborative AI development, allowing individuals to share trained models based on local datasets, while enabling cloud-based AI services that are openly available to users.

Users can upload trained models to the u-DemAI system, where the framework evaluates their performance and records their predictive capabilities. These performance records are compiled into a ranked list, allowing users to select models based on empirical evaluation metrics.

When users request a prediction, the u-DemAI framework leverages multiple cloud-based AI services and actively involves users in the optimization loop. Users are given the flexibility to:

- Select preferred models from the uploaded list, using recorded performance metrics as a reference.
- Specify target preferences, which may include fairness criteria across demographic attributes such as gender, age, and ethnicity.
- Choose optimization objectives, including:
  - Maximizing prediction accuracy.
  - Ensuring fairness across demographic groups.
  - Balancing accuracy and fairness for socially responsible AI deployment.

The u-DemAI framework integrates multiple cloud-based AI services to enhance predictive accuracy and fairness. Predictions are aggregated from multiple AI models, and evolutionary optimization techniques are employed to determine the optimal weighting for each service. These weights are dynamically adjusted based on user feedback collected from different demographic communities. This feedback-driven approach enables the system to optimize social values, such as ensuring equitable AI predictions across user groups.

In summary, u-DemAI establishes an open and democratized AI service ecosystem by integrating users into the AI optimization process. Through transparent model selection, customizable preferences, and continuous feedback loops, the framework empowers users to shape AI services according to their values and needs.

Figure 5.2 illustrates the architecture of the proposed u-DemAI framework, highlighting its key components and the interaction between users and cloud-based AI services.

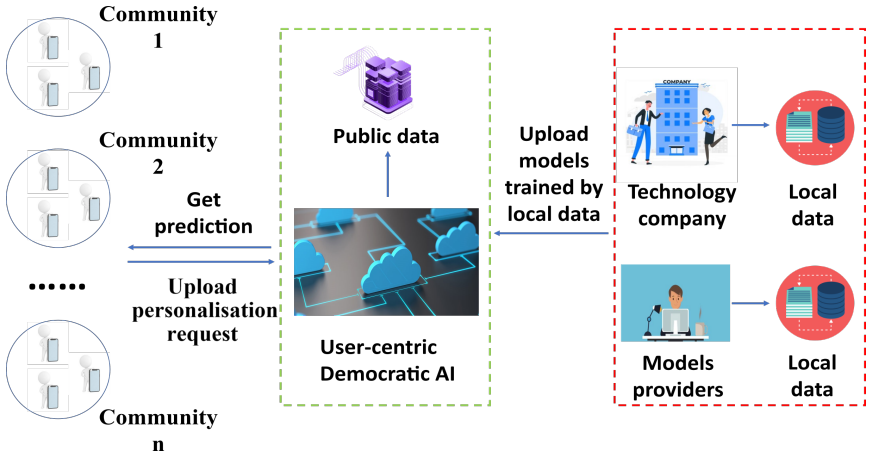


Figure 5.2: The community-adaptive democratic process of the proposed u-DemAI framework over cloud-based AI services.

5.2.3 Case Study: Medical Brain Age Estimation

In this study, we focus on a specific medical task as our case study due to its relevance to Democratic AI and its potential benefits to patients. We select brain age estimation, a contemporary challenge in medical AI, which requires both high predictive accuracy and age-wise fairness (Cole and Franke, 2017; Luders, Cherbuin, and Gaser, 2016).

The human brain can be conceptualized as a Turing's Type-B machine, characterized by randomly interconnected neurons (R. Jiang and Crookes, 2019). Brain function is closely linked to mental health, making brain age estimation an essential indicator in neurological and psychiatric research (R. Jiang, P. Chazot, et al., 2022).

Brain age estimation is commonly framed as a classification or regression task, serving as a key biomarker of brain health (Peng et al., 2021). Previous studies have successfully employed AI-driven neuroimaging analysis to predict individual brain age, achieving state-of-the-art performance (Cole and Franke, 2017; Luders, Cherbuin, and Gaser, 2016). By training deep learning models on neuroimaging data from healthy individuals, it is possible to develop AI models capable of estimating biological brain age (Cole and Franke, 2017).

5.3 Modelling Democratic AI

5.3.1 Democratic AI beyond Clouds

Despite its literal interpretation, Democratic AI is inherently interconnected with cloud computing technology. The rationale behind this is that Democratic AI aims to make AI services widely accessible to the general public, necessitating internet-based access from any location. Consequently, cloud computing serves as an indispensable infrastructure for enabling scalable, distributed AI services (Nguyen et al., 2022; Shashi et al., 2022).

Leading AI companies, including Microsoft, Google, and IBM, have expressed their commitment to democratizing AI, with cloud computing playing a pivotal role in this initiative. These prominent AI providers offer cloud-based AI services to a diverse range of users, such as edge devices, smart home systems, healthcare institutions, and industrial applications.

A fundamental challenge in this paradigm is selecting the most suitable AI service for a given user or application. This challenge is precisely addressed by our user-centric Democratic AI (u-DemAI) framework. In this paper, we introduce u-DemAI as a novel algorithm designed to actively involve users in AI service selection, thereby maximizing both individual benefits and broader societal values derived from AI-driven solutions.

5.3.2 Fairness in Brain Age Prediction

In machine learning, fairness refers to the absence of systematic bias in model predictions with respect to sensitive attributes such as age, gender, or ethnicity. A fair model ensures that its predictive performance is consistent across different sub-populations, preventing unequal treatment of particular groups.

In the context of brain age prediction, fairness is most directly related to mitigating ageism, which means avoiding systematic overestimation or underestimation of brain age for specific chronological age groups. If prediction errors vary consistently with age, the model is considered biased. Such age-related bias not only undermines the scientific validity of brain age as a biomarker but also risks misinforming clinical interpretations.

To quantify fairness in this study, we employed three criteria:

- Pearson correlation coefficient between the brain age gap and true age: This measures whether the error is linearly associated with chronological age. A value close to zero indicates fair predictions that are not systematically biased by age.

- Slope of fitted regression lines for the brain age gap: This criterion evaluates whether errors increase or decrease systematically with age. A slope close to zero reflects age-invariant error distribution.

- Standard deviation of absolute errors: This captures the variability of prediction errors across individuals. A lower standard deviation indicates that errors are more evenly distributed across the population, supporting fairness.

Ensuring fairness in brain age prediction is critical because brain age has been proposed as a biomarker of neurological health. If the model's predictions were systematically biased toward specific age ranges, this could lead to misleading conclusions in both clinical and research contexts. By explicitly incorporating fairness into our training objective, we ensure that brain age predictions are reliable and equitable across different age groups, enhancing both the scientific validity and clinical applicability of our results.

5.3.3 Community Based User-Centric DemAI

From the user's perspective, individuals within an AI service community prefer to evaluate AI services based on their own experiences rather than relying solely on predefined specifications provided by the service provider. To formalize this concept, we define a set of evaluation measures denoted as $\{m_k\}$, reflecting user preferences. To simplify the decision-making process and enable more intuitive choices, we introduce a combined loss function L over k measures:

$$L = \sum_k \alpha_k m_k, \tag{5.1}$$

where α_k represents the preference weights assigned by users. For example, in brain age estimation, we want to keep accuracy and fairness balance, so we set m_1 as accuracy (MAE), m_2 as fairness (the Pearson correlation coefficient between the brain age gap and chronological age). If accuracy is a higher priority, then α_1 corresponding to accuracy will be assigned a larger value. Conversely, if fairness is prioritized, a higher α_2 is assigned to fairness. Consequently, the overall measure L enables users to make informed choices when selecting an AI service i.

In this study, we define:

- m_1 as MAE.
- m_2 as the Pearson correlation coefficient between the brain age gap and true age (PC), which serves as an indicator of ageism in the predicted results.

For our case study, we assume that all users assign equal importance to accuracy and fairness, setting $\alpha_1 = \alpha_2 = 0.5$ to balance both factors. In this way, fairness is integrated directly into the optimization objective. This ensures that the model is penalized not only for inaccurate predictions but also for systematic age-related bias.

Given that users have access to multiple AI services, these services can be combined using a weighted approach to optimize the overall loss function L:

$$c_D^j = \sum_i w_i c_i^j, \tag{5.2}$$

where:

- c_i^j represents the prediction result for the j-th user from AI service i.
- w_i is the user-defined weight for AI service i.
- c_D^j denotes the final combined prediction, which is determined through the democratic AI process.

It is important to note that Equation (5.2) is user-driven, as the weight parameters $\{w_i\}$ are determined by individual user preferences. Since c_D^j is derived through user feedback, the entire optimization process in our user-centric Democratic AI (DemAI) framework actively includes users in the decision-making loop. Users can directly influence the selection of AI services and set their preferences for social values through the assignment of $\{\alpha_k\}$. The democratic selection process can then be mathematically formulated as an optimization problem:

$$w_i = \arg\min L(w_i, \alpha_k, c_i^j) \tag{5.3}$$

This formulation ensures that users retain full control over AI service utilization, as:

- Users select the AI services they prefer.
- Users define the weights assigned to each AI service.
- The final optimization outcome of the Democratic AI framework depends entirely on user-defined preferences.

The implementation of a user-centric Democratic AI system raises two fundamental mathematical challenges:

1. **User Integration:** How can users be effectively incorporated into the democratic optimization loop?
2. **Convergence to Optimality:** How can we ensure that the democratic process leads to the best possible optimization outcome?

To address these challenges, we introduce a novel natural democratic computing process tailored for our u-DemAI framework, which will be further elaborated in the following sections.

5.3.4 Evolutionary Democratic Process

In principle, Democratic AI can be regarded as a natural computing process driven by human behavior within a community of users (Koster et al., 2022). To effectively integrate AI services over the cloud, incorporating user feedback into the optimization loop for achieving socially beneficial AI outcomes, we draw inspiration from Particle Swarm Optimization (PSO) to emulate an evolutionary democratic model. In this framework, PSO serves as the optimization core that aggregates and fine-tunes diverse user-contributed models. Its primary objective is to determine the optimal weighted ensemble of these models that best satisfies the collective goals defined by user preferences—such as balancing predictive accuracy and fairness. A comparable approach was introduced in Chapter 3, where the nl-AAE model addressed variability across age-specific subgroups to enhance predictive accuracy. Here, that proven modeling strategy is extended and embedded within the broader framework of u-DemAI. While the ensemble principles from Chapter 3 are retained to ensure robust accuracy, the key innovation of Chapter 5 lies in the democratic integration of user-defined objectives. In this setting, fairness is explicitly incorporated as an optimization criterion, particularly targeting the mitigation of ageism in brain age predictions. As a result, u-DemAI advances beyond accuracy alone to achieve socially aligned, user-driven, and equitable model outcomes.

We initialize n particles, each assigned a random initial position with a random initial velocity. Each particle represents an n-dimensional weight vector corresponding to the AI services:

$$(\beta_1, \beta_2, ..., \beta_n)$$

where β_i denotes the weight assigned to the i-th AI service.
For the i-th particle, we define:

- $x_i = (\beta_{i1}, \beta_{i2}, ..., \beta_{in}) \rightarrow$ Current position of the particle.
- $v_i \rightarrow$ Current velocity of the particle.
- $y_i \rightarrow$ Best position of the particle based on previous iterations.
- $t \rightarrow$ Number of iterations.
- $g \rightarrow$ Loss function used to evaluate each particle's performance.

The best position for each particle is updated iteratively using user feedback from the AI service community. The update rule follows:

$$y_i(t+1) = \begin{cases} y_i(t), & g(x_i(t+1)) \geq g(y_i(t)) \\ x_i(t+1), & g(x_i(t+1)) < g(y_i(t)) \end{cases} \tag{5.4}$$

This mechanism ensures that each particle retains the best-performing configuration encountered so far.

The global best position among all particles, denoted as $\hat{y}$, is defined as:

$$\begin{aligned} &\hat{y}(t) \in \{y_0(t), y_1(t), \ldots, y_s(t)\} \\ &\quad \text{such that } g(\hat{y}(t)) = \min\{g(y_0(t)), g(y_1(t)), \ldots, g(y_s(t))\} \end{aligned} \tag{5.5}$$

The velocity update equation for each particle in dimension j is:

$$\begin{aligned} V_{ij}(t+1) = \omega V_{ij}(t) + \gamma_1 \text{rand}(0,1)(y_{ij}(t) - x_{ij}(t)) \\ + \gamma_2 \text{rand}(0,1)(\hat{y}_j(t) - x_{ij}(t)) \end{aligned} \tag{5.6}$$

where:

- $\omega \rightarrow$ Inertia weight, controlling the contribution of the previous velocity.
- $\gamma_1, \gamma_2 \rightarrow$ Acceleration coefficients, determining the influence of personal and global best positions.
- rand(0,1) $\rightarrow$ A randomly sampled value between 0 and 1.

The particle's position is then updated using:

$$x_i(t+1) = x_i(t) + V_i(t+1) \tag{5.7}$$

The trajectory of a particle is analyzed in discrete time steps, with x_t representing the particle's position at time t. This system can be described by the following non-homogeneous recurrence relation:

$$x_{t+1} = (1 + \omega - \rho_1 - \rho_2)x_t - \omega x_{t-1} + \rho_1 y + \rho_2 \hat{y}, \tag{5.8}$$

where $\rho_1 = C_1 \cdot random(0,1)$, $\rho_2 = C_2 \cdot random(0,1)$. Given the initial conditions $x(0) = x_0$, $x(1) = x_1$, and assuming that y and $\hat{y}$ remain constant over time, we obtain the closed-form solution to the recurrence relation as:

$$x_t = k_1 + k_2\alpha^t + k_3\theta^t, \tag{5.9}$$

with coefficients defined as follows:

$$k_1 = \frac{\rho_1 y + \rho_2 \hat{y}}{\rho_1 + \rho_2} \tag{5.10}$$

$$\mu = \sqrt{(1 + \omega - \rho_1 - \rho_2)^2 - 4\omega} \tag{5.11}$$

$$\alpha = \frac{1 + \omega - \rho_1 - \rho_2 + \mu}{2} \tag{5.12}$$

$$\theta = \frac{1 + \omega - \rho_1 - \rho_2 - \mu}{2} \tag{5.13}$$

$$x_2 = (1 + \omega - \rho_1 - \rho_2)x_1 - \omega x_0 + \rho_1 y + \rho_2 \hat{y} \tag{5.14}$$

$$k_2 = \frac{\theta(x_0 - x_1) - x_1 + x_2}{\mu(\alpha - 1)} \tag{5.15}$$

$$k_3 = \frac{\alpha(x_1 - x_0) + x_1 - x_2}{\mu(\theta - 1)} \tag{5.16}$$

Equation (5.9) can thus be utilized to calculate the trajectory of the particle. According to equations (5.1) and (5.13), the convergence behavior of the sequence $\{x_t\}_{t=0}^{+\infty}$ is determined by the parameters α and θ.

By considering expectations, we have (the detailed proof can be found in reference (Van Den Bergh, 2001)):

$$E[\rho_1] = C_1 \int_0^1 \frac{x}{1-0}\, dx = \frac{C_1}{2} \tag{5.17}$$

$$E[\rho_2] = C_2 \int_0^1 \frac{x}{1-0}\, dx = \frac{C_2}{2} \tag{5.18}$$

Assuming that parameters ρ_1, ρ_2, and ω are chosen such that the sequence $\{x_t\}_{t=0}^{+\infty}$ converges, from equation (5.10), we obtain the limit:

$$\lim_{t \to +\infty} x_t = k_1 = \frac{\rho_1 y + \rho_2 \hat{y}}{\rho_1 + \rho_2} \tag{5.19}$$

Using the expected values derived from equations (5.17) and (5.18), we further simplify the expression to obtain:

$$
\begin{aligned}
\lim_{t\to+\infty} x_t &= \frac{\frac{C_1}{2}y + \frac{C_2}{2}\hat{y}}{\frac{C_1}{2} + \frac{C_2}{2}} \\
&= \frac{C_1 y + C_2 \hat{y}}{C_1 + C_2} \\
&= \frac{C_1}{C_1 + C_2} y + \frac{C_2}{C_1 + C_2}\hat{y} \\
&= \left(1 - \frac{C_2}{C_1 + C_2}\right) y + \left(1 - \frac{C_1}{C_1 + C_2}\right)\hat{y} \\
&= (1-\alpha)y + \alpha\hat{y}
\end{aligned}
\tag{5.20}
$$

Here, $x(t)$ denotes the position of the particle at time t.

Equation (5.20) indicates that the particle converges to a point determined by the line connecting the current best position y and the global best position $\hat{y}$.

For the experiments conducted in our case study, we set the following parameters:

- $\omega = 0.2$
- $C_1 = C_2 = 0.2$
- Number of iterations: 100
- Number of initial particles: 300

The mathematical formulation of this evolutionary democratic model provides a rigorous proof of convergence, demonstrating that Democratic AI optimizes AI services toward socially desirable values.

In our framework, user feedback operates at two complementary levels. First, users can contribute locally trained models to u-DemAI. Each contributed model is systematically evaluated on an internal benchmark dataset maintained by the system. Only models that achieve satisfactory performance are retained as candidate base models, ensuring that the pool of available models remains robust and reliable regardless of the quality of individual submissions. Second, users are empowered to express their preferences through the configuration of the loss function. For instance, they may adjust the relative weights assigned to accuracy and fairness in the optimization objective, thereby tailoring the system to their specific priorities. In this way, user feedback is not merely qualitative but is quantitatively embedded into the optimization process, enabling u-DemAI to produce models that are both high-performing and aligned with diverse user-defined values.

It is important to clarify that Particle Swarm Optimization (PSO) in u-DemAI does not aim to directly simulate human feedback. Instead, PSO serves as the optimization algorithm for determining the optimal weighting of user-contributed models within an ensemble. Users provide feedback implicitly by specifying their objectives through preference weights in the

loss function. PSO then optimizes the ensemble weights of the available models to minimize this user-defined loss. In this way, user input defines the optimization target, while PSO operates as the computational tool to achieve the optimal ensemble configuration. This ensures that u-DemAI effectively integrates diverse user preferences while maintaining stable convergence properties.

5.3.5 AI Services for Brain Age Estimation

As demonstrated in the previous chapters, several independent deep learning models have already shown promising performance on brain age estimation tasks, establishing this domain as a reliable benchmark for evaluating model accuracy and fairness. Building on these results, we now employ brain age estimation as the case study for the proposed u-DemAI framework. This choice not only allows us to validate the framework in a well-studied application but also highlights how user-defined preferences (e.g., accuracy–fairness trade-offs) can be systematically integrated into model selection and optimization.

In this context, we assume the availability of four cloud-based AI services, each providing brain age estimation through state-of-the-art models. These models have demonstrated strong performance in predicting brain age, and their details are as follows:

1. CNN developed in this study was implemented using the Keras framework with TensorFlow as the backend. The model leverages deep 3D convolutional architectures to extract structural patterns associated with brain aging from volumetric MRI data. The network architecture comprises seven sequential blocks. The first five blocks each include a 3D convolutional layer with a kernel size of $3 \times 3 \times 3$, followed by batch normalization, an Exponential Linear Unit (ELU) activation function, and a max pooling layer. The sixth block incorporates a dropout layer to prevent overfitting, while the seventh block includes a fully connected (dense) layer for regression output. The input to the network is a 3D brain volume of size $121 \times 145 \times 121$ voxels. Through successive convolutional and pooling operations, the model reduces the spatial dimensions and outputs 128 feature maps of size $4 \times 5 \times 4$. These feature maps are subsequently flattened and passed through the final dense layer to produce the estimated brain age. To enrich the input representation, the model is trained on two imaging channels obtained by concatenating gray matter and white matter segmentations. The model is optimized using the Mean Absolute Error (MAE) as the loss function and the Adam optimizer. The learning rate is set to 0.001 with a weight decay of 10^{-4}, and the Adam hyperparameters are $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

2. GoogLeNet (Inception V1) (Couvy-Duchesne, Faouzi, et al., 2020) is a deep convolutional neural network architecture adapted for regression tasks in this study. The model builds upon the original Inception V1 architecture by replacing the final softmax classification layer with a fully connected regression layer, enabling prediction of

continuous brain age values rather than class labels. The architecture consists of a stem network, followed by two Inception modules and a max-pooling layer. This is succeeded by five additional Inception modules, two of which are connected to auxiliary regression branches designed to mitigate the vanishing gradient problem. Following another max-pooling layer, the network includes two more Inception modules, an average pooling layer, a dropout layer, and a final fully connected regression layer. Each convolutional unit within the network comprises a convolutional layer, batch normalization, a ReLU activation function, and an output layer. The stem network consists of an input layer, a convolutional unit, a max-pooling layer, two additional convolutional units, another max-pooling layer, and an output layer. Within each Inception module, the structure includes an input layer, seven parallel convolutional filters, a max-pooling layer, a concatenation layer, and an output layer. Each auxiliary regression branch contains an average pooling layer, a convolutional layer, a fully connected layer followed by a ReLU activation, a dropout layer, and a final regression output layer. The model is trained using 3D gray matter density maps of size $121 \times 145 \times 121$ as input, with the target output being the subject's chronological age. The loss function employed is MAE, and the network is optimized using the Adam optimizer with a learning rate of 0.0001 and a batch size of 8.

3. ResNet (Peng et al., 2021): This model is based on the ResNet architecture and shares similar parameter settings with the custom-built CNNs described earlier. The primary architectural distinction lies in the incorporation of residual blocks, which are absent in our self-designed CNNs. The network comprises five residual blocks, each followed by a 3D max pooling layer with a kernel size of $3\times3\times3$ and a stride of $2\times2\times2$. Each residual block consists of two repetitions of a core sequence: a 3D convolutional layer with a kernel size of $3\times3\times3$ and a stride of $1\times1\times1$, followed by batch renormalization and an ELU activation. To enable residual learning, the input to each residual block is added to the output of the second convolutional layer, facilitating gradient flow and improving training stability. Following the convolutional and residual stages, the output is passed to a fully connected block structured as a multilayer perceptron. The input to the MLP is a flattened feature vector of size $128 \times 4 \times 5 \times 4 = 10,240$. The FC1 contains 256 neurons with ELU activation, followed by a dropout layer with a keep rate of 0.8. The FC2 consists of a single neuron and performs linear regression to predict brain age. The model is trained on 3D gray matter density maps as input data. MAE is used as the loss function. Optimization is performed using the Adam optimizer with a learning rate of 0.001, a weight decay of 10^{-4}, and momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

4. Support Vector Regression (SVR) (Cole and Franke, 2017) is a machine learning approach derived from SVM, originally designed for classification tasks. While traditional SVMs construct optimal hyperplanes to separate classes in a high-dimensional feature

space, SVR extends this concept to regression by finding a function that approximates the target outputs within a specified margin of tolerance. In this study, SVR is employed for brain age estimation using high-dimensional surface-based gray matter features. Specifically, each individual is represented by approximately 650,000 gray matter measurements derived from surface-based preprocessing. A RBF kernel is used to capture the nonlinear relationships between structural brain features and chronological age. The model is implemented using the `scikit-learn` package in Python. To ensure robust performance and generalization, the model is trained for over 300 epochs, with the goal of achieving optimal predictive accuracy across validation trials.

By integrating these cloud-based AI services, we evaluate the effectiveness of our user-centric Democratic AI (u-DemAI) framework in adapting to diverse user communities and automatically optimizing predictions for different population groups.

To assess the adaptability of u-DemAI, we segment users into three distinct age-based communities:

- Young group: $16 \sim 30$ years old.
- Middle-aged group: $31 \sim 60$ years old.
- Elderly group: $61 \sim 100$ years old.

Each community is evaluated separately to determine whether u-DemAI can dynamically optimize AI services, ensuring personalized performance improvements for different age groups. The upcoming experiments aim to demonstrate the framework's ability to enhance accuracy and fairness across user populations.

5.3.6 Datasets

The dataset used in this study is derived from (Cole and Franke, 2017) and comprises 2,641 healthy individuals' brain structural MRI (sMRI) scans, along with additional demographic information such as age and gender. The age range of participants spans from 16 to 90 years, with an average age of 35.8 years and a standard deviation of 16.2 years. Among the participants, 53% are female and 47% are male. Further details regarding the dataset can be found in (Cole and Franke, 2017).

It is worth noting that this same dataset was also employed in Chapter 3 (Nonlinear Age-Adaptive Ensemble Learning), where it was used to evaluate ensemble-based strategies for brain age estimation. By reusing this dataset here, we ensure consistency across chapters and enable a more direct comparison between the previously proposed ensemble learning approaches and the user-centric Democratic AI (u-DemAI) framework introduced in this chapter.

In this study, we utilize two different types of neuroimaging data as input for the AI services:

- Gray Matter and White Matter Maps: These maps were provided by the PAC organization and serve as input for the self-defined CNN, ResNet, and GoogLeNet models used in this study.
- Surface-Based Processing of Gray Matter: This dataset is derived from vertex-wise cortical thickness and surface area measurements extracted from sMRI scans using FreeSurfer 6.0 (Fischl, 2012). These processed features are used as input for the Support Vector Regression (SVR) model.

By incorporating both volumetric and surface-based neuroimaging features, our approach ensures a comprehensive representation of brain structure, enhancing the accuracy and interpretability of brain age estimation.

5.4 Experimental Results

5.4.1 Experimental Setup

The dataset used in this study comprises 2,641 healthy individuals' brain structural MRI (sMRI) scans, along with demographic attributes such as age and gender. For model training and evaluation, we allocate 75% of the dataset for training and 25% for testing to assess the performance of the AI services.

To quantitatively evaluate the estimation accuracy of AI services, we use the Mean Absolute Error (MAE), which measures the discrepancy between the chronological age and the predicted age. MAE is a widely adopted metric in brain age estimation research (Couvy-Duchesne, Faouzi, et al., 2020; Peng et al., 2021; Cole and Franke, 2017), where a lower MAE indicates higher predictive accuracy.

To assess the fairness of AI services and their susceptibility to ageism, we introduce three evaluation criteria:

The first criterion is the Pearson correlation coefficient between the brain age gap (chronological age minus predicted age) and chronological age. A lower correlation suggests that the model's performance is less influenced by true age, indicating higher fairness and resistance to age-related bias. The Pearson correlation coefficient is computed as:

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{5.21}$$

where:

- X represents the brain age gap.
- Y represents the true age.
- $\text{cov}(X, Y)$ is the covariance between X and Y.

- σ_X and σ_Y denote the standard deviations of X and Y, respectively.

The second fairness criterion is the slope rate of the brain age gap with increasing chronological age (Couvy-Duchesne, Faouzi, et al., 2020). A lower slope indicates that the predicted age remains stable across different age groups, suggesting reduced age bias and higher fairness. The slope rate is computed as:

$$\lim_{\Delta a \to 0} \left| \frac{G(a + \Delta a) - G(a)}{\Delta a} \right| \tag{5.22}$$

where:

- a represents chronological age.
- $G(a)$ represents the brain age gap as a function of age.

This metric is analyzed by examining the slope of the brain age gap vs. chronological age regression line.

The third criterion is the standard deviation of absolute error between chronological age and predicted age, which reflects the degree of variability in the model's predictions. A lower standard deviation indicates a higher ability to mitigate age-related bias, ensuring greater fairness in predictions. The standard deviation is computed as:

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}} \tag{5.23}$$

where:

- n represents the total number of samples.
- x_i is the absolute brain age gap of the i-th sample.
- $\bar{x}$ is the mean absolute brain age gap across all samples.

By incorporating these three fairness criteria, we ensure a comprehensive evaluation of age bias in AI-based brain age estimation. These metrics provide insights into how well AI services generalize across different age groups and whether they unintentionally favor or disadvantage specific populations.

5.4.2 Ageism in Single Models

To examine the presence of ageism in AI-based brain age estimation, we evaluate four widely used AI models: CNN (Couvy-Duchesne, Faouzi, et al., 2020), ResNet (Peng et al., 2021), GoogLeNet (Inception V1) (Couvy-Duchesne, Faouzi, et al., 2020), and Support Vector

Regression (SVR) (Cole and Franke, 2017). This analysis aims to determine whether age bias is method-dependent and to assess the degree of fairness across different models.

Figure 5.3 illustrates GoogLeNet's brain age gap as a function of chronological age across different age groups. The results indicate that ageism is present in GoogLeNet's brain age estimation, as evidenced by the varying slope of the brain age gap-chronological age regression line in different age ranges. Specifically:

- Young Age Group (17-30 years old): The red regression line predominantly lies below the black horizontal line, suggesting that the predicted brain age is consistently overestimated for younger individuals.
- Middle Age Group (30-60 years old): The red regression line intersects the black horizontal line, indicating a transition where the brain age gap shifts from negative to positive.
- Elderly Group (60-90 years old): The red regression line remains above the black horizontal line, implying that the predicted brain age is consistently underestimated for older individuals.

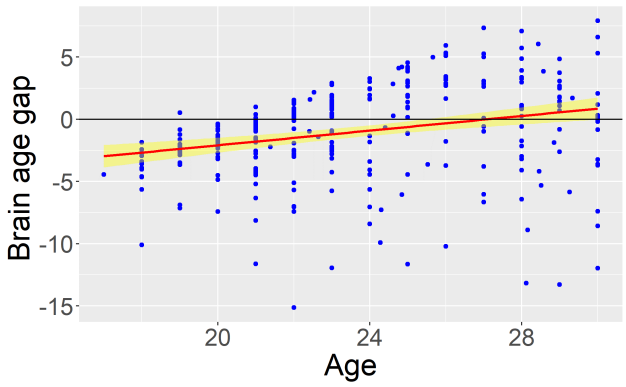
These findings demonstrate that GoogLeNet exhibits systematic age bias: the brain ages of younger individuals are more likely to be overestimated, whereas the brain ages of older individuals are more likely to be underestimated. This pattern of bias highlights the importance of evaluating and mitigating ageism in AI-driven brain age estimation models.

Table 5.1 presents the slopes of the fitted regression lines for the brain age gap across different age groups for the four single AI services. The results indicate that each model exhibits varying degrees of age bias, and the fairest algorithm differs across age groups. In other words, certain models are more suitable for predicting brain age in younger individuals, while others perform better for elderly populations.

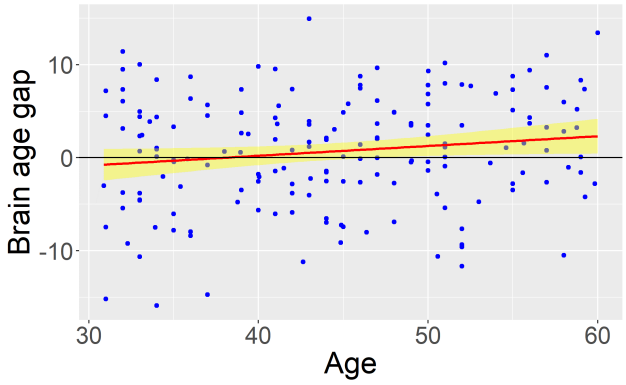
Key observations from Table 5.1 include:

- GoogLeNet demonstrates the highest fairness for young individuals (17–30 years old), as its slope is the least affected by age.
- CNN exhibits the most stable predictions for middle-aged individuals (30–60 years old), suggesting it is well-suited for this demographic.
- ResNet outperforms other models in terms of fairness for elderly individuals (60–90 years old), as its predictions show the least bias in this age group.
- SVR consistently exhibits the highest degree of age bias across all age groups, making it the least fair model among the four services.

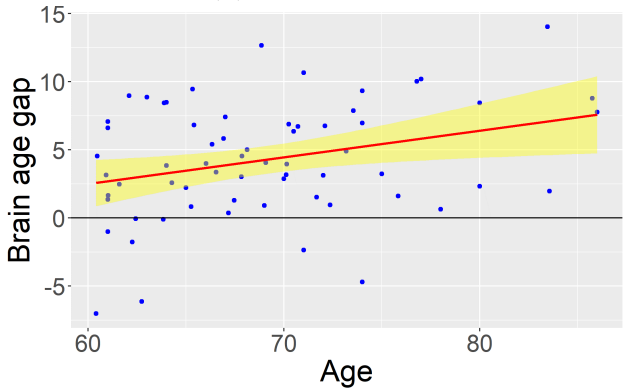
These findings highlight the method-dependent nature of age bias in brain age estimation and emphasize the importance of selecting appropriate AI models based on the target age group.



(a) Young



(b) Middle Aged



(c) Elder

Figure 5.3: Unfairness and ageism in brain age estimation for single service. Here, age ranges for Young, Middle and Elder groups are: 16-30, 31-60 and 61-100. It shows the brain age gap as a function of the chronological age using the GoogLeNet model in different groups. The best fit of line regression (red) in each plot with the 95% prediction interval (yellow area) denotes the degree of bias.

Services/ Indicators	Young	Middle	Elder
Single service 1: GoogLeNet	0.29	0.10	0.20
Single service 2: ResNet	0.39	-0.08	0.16
Single service 3: CNN	0.30	0.01	0.19
Single service 4: SVM	0.42	0.21	0.20

Table 5.1: The details of single services' slopes of fitted lines for age gap in different age groups.

5.4.3 Evaluation of the Democratic Process

In this section, we evaluate the performance of our proposed user-centric Democratic AI (u-DemAI) framework in brain age estimation, specifically assessing its ability to mitigate ageism. The prediction outcomes of u-DemAI are derived from the individual AI services, including CNN, ResNet, GoogLeNet, and SVR.

The changes of u-DemAI's training loss are shown in Figure 5.4. Here, training specifically denotes the optimization process performed by Particle Swarm Optimization (PSO). At this stage, no additional parameters within the base models are updated.

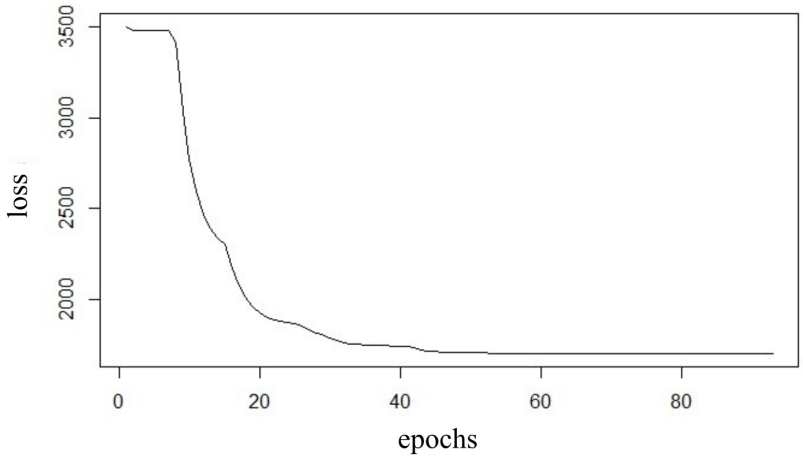


Figure 5.4: Changes of u-DemAI's training loss.

To address age-related biases in brain age estimation, we define three distinct age groups—young, middle-aged, and elderly—as user-specific preferences within the u-DemAI framework. For each age group, a separate u-DemAI model is constructed by integrating the predictions of the four single-model services. During inference, the u-DemAI model corresponding to a user's age group generates predictions by utilizing optimized weight allocations that evolve through user-inclusive learning loops.

By dynamically adjusting the weights of individual AI services, the u-DemAI framework effectively accounts for age-specific brain characteristics, thereby enhancing prediction fairness and reducing ageism.

We can see that this technique was also used in Chapter 3, as nl-AAE focuses on improving predictive accuracy across age-specific subgroups. In this chapter, we extend that line of work by embedding this proven modelling technique within the broader framework of u-DemAI. While the modelling methods from Chapter 3 are reused here to ensure reliable prediction accuracy, the novelty of Chapter 5 lies in introducing the concept of democratic, where user-defined objectives explicitly balance accuracy and fairness. In particular, fairness is operationalized through the mitigation of ageism in predicted outcomes, making brain age estimation not only accurate but also socially aligned.

Figure 5.5 illustrates the evolution of individual service weights and the optimization objective within the u-DemAI models across different age groups during training.

- The blue, brown, green, and purple lines represent the evolving weights of GoogLeNet, ResNet, self-defined CNN, and SVR, respectively, within the u-DemAI models for each age group.
- The red line denotes the variation of the loss function across iterations within the u-DemAI framework.

As observed in Figure 5.5:

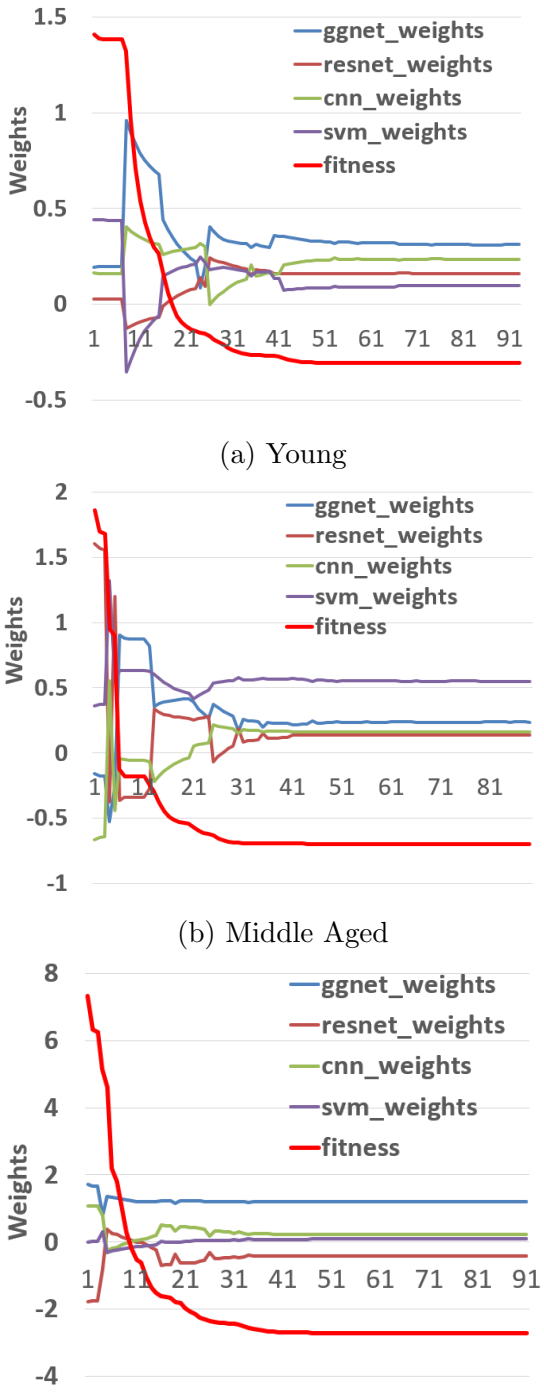
- For the young group, both individual service weights and the loss function converge to a stable value after approximately 45 iterations.
- For the middle-aged group, convergence is achieved in approximately 36 iterations, a pattern that is also consistent with the elderly group.

These results suggest that u-DemAI effectively optimizes service weights through iterative training, achieving a stable and adaptive model tailored to different age groups. This demonstrates that the democratic learning process successfully fine-tunes model predictions by incorporating user preferences and mitigating age bias in brain age estimation.

Figure 5.6 illustrates the brain age gap as a function of chronological age for five different AI services. The results demonstrate that u-DemAI outperforms all other methods in mitigating ageism, as its brain age gap remains the most stable across different age groups.

Key observations from Figure 5.6 include:

- u-DemAI exhibits the least variation in brain age gap across aging, indicating its superior ability to address age-related biases in brain age estimation.
- SVR demonstrates the highest degree of ageism, displaying the largest deviations in predicted age across different age groups.



(a) Young

(b) Middle Aged

(c) Elder

Figure 5.5: The iterative process in the u-DemAI framework for different age communities. We can see both the service weights and the cost function converge successively.

- Among the individual AI models, the self-designed CNN emerges as the fairest single service, exhibiting the lowest degree of bias in brain age predictions.
- All single services tend to underestimate brain age in elderly individuals while overestimating brain age in younger individuals.
- In contrast, u-DemAI consistently predicts a slightly higher brain age than the chronological age, effectively reducing systematic bias.

These findings further confirm that u-DemAI successfully mitigates ageism in brain age estimation, providing more reliable and equitable predictions across different age groups compared to conventional single-model approaches.

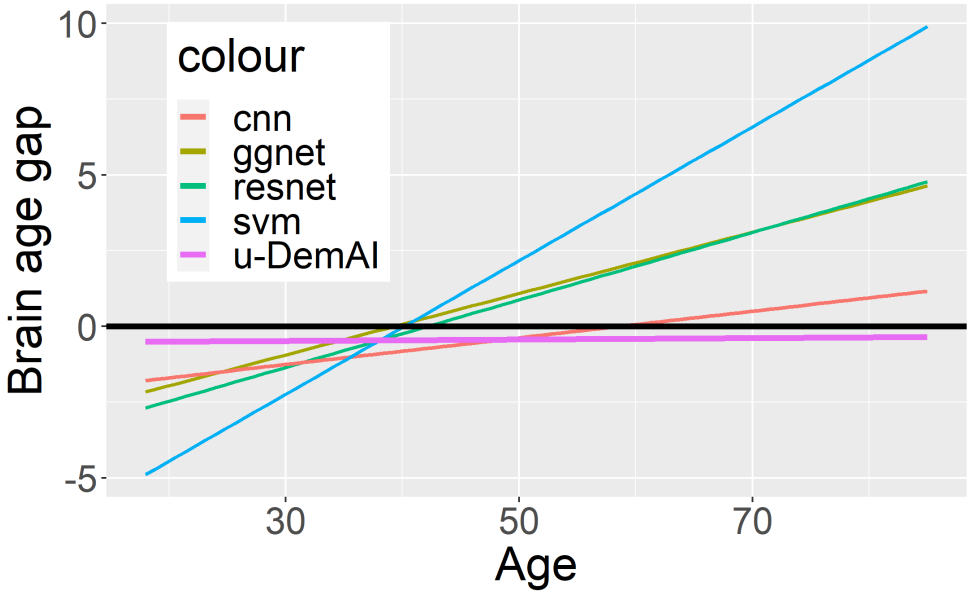


Figure 5.6: Test performance of AI models/services. It represents the brain age gap in 5 different services as function of the whole chronological age. Here, ggnet means GoogLeNet, resnet means ResNet, cnn means a self-defined CNN, and svm refers to SVM. We can see that our u-DemAI reduced the bias.

The test results are detailed in Table 5.2, where we evaluate fairness and ageism in predicted brain ages using three key criteria. The results indicate that:

- Slopes of the fitted age gap regression lines: u-DemAI demonstrates the highest fairness, followed by the self-designed CNN. In contrast, ResNet and GoogLeNet show similar levels of fairness degradation, while SVR exhibits the highest degree of age bias.

Services	Slopes	SDAE	PC	MAE
Single service 1: GoogLeNet	1.13	3.09	0.34	3.70
Single service 2: ResNet	0.96	3.87	0.33	3.92
Single service 3: CNN	0.38	3.49	0.13	4.34
Single service 4: SVM	1.39	3.86	0.55	5.19
Ensemble Service A (Couvy-Duchesne, Faouzi, et al., 2020)	0.51	1.97	0.21	3.33
Ensemble Service B (Da Costa, Dafflon, and Pinaya, 2020)	0.78	2.83	0.27	3.76
Ensemble Service C, developed in Chapter 3 (Z. Zhang, R. Jiang, et al., 2022)	0.27	**0.12**	0.06	3.19
Our u-DemAI	**0.04**	2.67	**0.01**	**2.99**

Table 5.2: The details of tested services' performance. Here, the 1st column is the slopes of fitted lines for the age gap, the 2nd column is the standard deviation of absolute error (SDAE), the 3rd column is the Pearson coefficient (PC) between the brain age gap and true age, and the last column is MAE. Our u-DemAI has consistently achieved the best among all measures.

- Standard deviation of absolute error: u-DemAI achieves the lowest standard deviation (2.67), indicating the highest robustness against age-related bias.
- Pearson correlation coefficient between the brain age gap and chronological age: u-DemAI achieves the best fairness, with an exceptionally low Pearson coefficient of 0.01, whereas SVR performs the worst, with a coefficient of 0.55.

In terms of estimation accuracy, u-DemAI achieves the highest prediction accuracy, with a Mean Absolute Error (MAE) of 2.99. Among the single AI models:

- GoogLeNet demonstrates the best performance with an MAE of 3.7.
- SVR remains the least accurate, with an MAE of 5.19.

We further compare u-DemAI against expert-designed ensemble models evaluated on the same PAC 2019 dataset:

- Couvy-Duchesne, Faouzi, et al. (2020) developed an ensemble model integrating seven different algorithms, achieving Slopes of 0.51, SDAE of 1.97, Pearson coefficient of 0.21 and MAE of 3.33.
- Da Costa, Dafflon, and Pinaya (2020) implemented a shallow machine learning ensemble model, resulting in Slopes of 0.78, SDAE of 2.83, Pearson coefficient of 0.27 and MAE of 3.76.

- Z. Zhang, R. Jiang, et al. (2022), developed in Chapter 3, proposed a nonlinear age-adaptive ensemble learning approach, yielding Slopes of 0.27, SDAE of 0.12, Pearson coefficient of 0.06 and MAE of 3.19.

In contrast, u-DemAI—developed through user-driven model combination, without expert intervention, almost outperforms all the expert-designed ensemble methods in both accuracy and fairness (except the standard deviation of absolute error).

Overall, u-DemAI consistently outperforms existing models across all fairness and accuracy metrics while integrating a PSO process involving non-expert users. These results highlight the tremendous potential of Democratic AI in promoting fairness, accuracy, and social value for individuals and communities.

5.5 Discussion

AI has become a driving force in technological transformation, yet concerns regarding its fairness and inclusivity have intensified in recent years (Posner, Fei-Fei, et al., 2020). The discourse surrounding AI fairness primarily revolves around two critical aspects: the diversity of AI users and the presence of biases in AI predictions. As noted by (Posner, Fei-Fei, et al., 2020), the AI field is currently experiencing a diversity crisis, making this a pivotal moment for addressing inclusivity in AI development.

To mitigate these challenges, Democratic AI (DemAI) has emerged as a potential solution, aiming to lower barriers to AI accessibility and empower a broader user base. However, several challenges hinder its widespread adoption:

- Lack of a precise definition and mathematical formalization of the democratic AI process.
- Increased risk of bias in AI predictions due to uncontrolled variability in user contributions.
- Potential risks posed by inexperienced AI contributors, which may lead to inaccurate or misleading outcomes, with significant consequences.

Our study clarifies and addresses these concerns surrounding Democratic AI through the following key contributions:

- Formal Definition of Democratic AI: We provide a comprehensive definition of Democratic AI, emphasizing its role in engaging users in the optimization process toward achieving socially beneficial outcomes.
- Mathematical Framework: We establish a rigorous mathematical formalization of the democratic AI process, demonstrating that DemAI can be interpreted as a natural computing process inspired by human behavior in diverse communities.

- Algorithmic Optimization via Evolutionary Methods: We introduce an evolutionary algorithm-based optimization process within DemAI, leading to the development of our u-DemAI framework, which is proven to exhibit guaranteed convergence in its iterative optimization loop.

- Empirical Validation Through a Medical AI Case Study: We validate the effectiveness of u-DemAI through an AI-driven brain age estimation task, demonstrating that by incorporating non-expert users into the optimization loop, Democratic AI can surpass expert-designed single models and ensemble methods in both accuracy and fairness.

It is important to recognize that integrating democracy into AI is inherently interdisciplinary, spanning social sciences, mathematics, and computer science. However, its practical implementation may inevitably rely on cloud-based AI services, where users interact with AI models via cloud platforms. As a result, the study of Democratic AI is deeply intertwined with cloud computing technologies, and future research must address the security implications of deploying AI services in a decentralized, user-driven manner.

Our proposed u-DemAI framework introduces several novel aspects to the study of Democratic AI:

- A well-defined conceptual framework that underscores the social values of AI democratization.

- A mathematical interpretation of Democratic AI as a natural computing process, capturing the role of human behavior in optimizing AI services.

- Formal proof of convergence in the democratic optimization process, ensuring that u-DemAI refinement leads to stable and beneficial outcomes.

- Demonstration of u-DemAI in a real-world medical AI application, providing empirical evidence that Democratic AI can surpass expert-guided AI systems in performance and fairness.

Users in u-DemAI serve as both model contributors and model consumers, extending beyond clinicians to include any interested individual or institution. In practice, users can (1) train a local model on their own data and upload the trained model to the u-DemAI platform, and (2) request a task-specific model by specifying preferences such as accuracy–fairness trade-offs. All submitted models are evaluated centrally on a held-out public reference dataset, where their performance is benchmarked and stratified according to fairness metrics. For each request, u-DemAI then selects the top-performing models and constructs a weighted ensemble optimized under the user-specified objective, ultimately providing a downloadable model tailored to the user's needs. Importantly, because the framework only integrates models that meet baseline performance, overall system reliability is preserved. The upload of

weaker models does not degrade the system, ensuring that users always benefit from robust and well-performing solutions.

It is also important to clarify how the proposed u-DemAI differs from other established machine learning paradigms, such as federated learning and active learning. While all three approaches aim to improve AI performance and inclusivity, they differ fundamentally in objectives and mechanisms.

Federated learning (FL) is designed to train models across decentralized datasets without requiring raw data exchange. Its primary goal is privacy preservation and efficient distributed training (McMahan et al., 2017). In FL, users contribute their local model updates to a central server, which aggregates these updates into a global model. However, individual users typically do not directly control the optimization objectives (such as fairness vs accuracy trade-offs).

Active learning (AL) seeks to reduce labeling costs by iteratively selecting the most informative samples to be labeled by an oracle (Settles, 2009). Its focus is on improving model performance under limited annotation budgets. The user interacts with the training loop by providing new labels, but users generally do not influence how multiple models are combined or how fairness is incorporated.

In contrast, the u-DemAI framework is not primarily concerned with data privacy (as in FL) or label efficiency (as in AL). Instead, it introduces a democratic optimization process where users directly participate in shaping the AI service output. Users specify their preferences (for example, weighting accuracy and fairness equally or prioritizing one over the other), and the framework integrates multiple pre-trained AI services by dynamically adjusting their weights through an evolutionary democratic process. Thus, u-DemAI is centered on *empowering users to control AI outcomes*, rather than solely on optimizing training efficiency or privacy.

Limitations and Future Work: Although u-DemAI highlights the value of user-in-the-loop optimization, applying such mechanisms in clinical practice faces practical limitations. First, clinical users vary in expertise: while domain experts may provide highly informed feedback, non-specialist clinicians or patients may introduce subjective or inconsistent inputs. Second, the integration of feedback processes into busy clinical workflows may impose significant time burdens, limiting adoption. For future work, efficient interfaces that minimize clinical workload are required. And it should also investigate hybrid approaches that combine automated evaluation with targeted expert feedback to ensure both robustness and usability in healthcare environments.

While the present study demonstrates the feasibility of u-DemAI through simulations, this approach inevitably has limitations. Simulated user preferences and interactions, although informative for proof-of-concept, cannot fully capture the diversity, inconsistency, or unpredictability of real human behavior. In practice, users may have heterogeneous objectives, provide noisy or conflicting feedback, or change their preferences over time—all factors that are difficult to replicate in controlled simulations. As a result, the current

evaluation may overestimate system stability and underestimate challenges associated with large-scale deployment. Future work should therefore incorporate human-in-the-loop evaluation, where researchers and lay users directly interact with the system to train local models, specify fairness–accuracy trade-offs, and provide iterative feedback. Such user studies would not only validate the robustness of u-DemAI in real-world settings but also reveal new insights into how democratic participation can shape AI optimization in practice.

Our case study only focuses on brain age estimation, but the u-DemAI framework is inherently general and applicable across domains. The core mechanism—evaluating contributed models on a common benchmark dataset, ranking them by performance and fairness metrics, and constructing user-specific ensembles guided by preference-weighted loss functions—does not depend on the medical imaging setting. For example, in healthcare, u-DemAI could be adapted to disease risk prediction or treatment outcome forecasting; in non-medical domains, it could support recommender systems, financial forecasting, or policy simulations. What makes u-DemAI generalizable is its user-in-the-loop design, which allows task-specific objectives (e.g., accuracy vs. fairness vs. robustness) to be flexibly defined. Future research should explore cross-domain case studies to demonstrate its robustness and scalability beyond neuroimaging.

5.6 Conclusion

In this chapter, we have introduced the concept of Democratic AI (DemAI) and proposed a user-centric Democratic AI (u-DemAI) framework that integrates user preferences into the optimization loop of AI services. We formalized DemAI mathematically as a natural computing process, and demonstrated how user-defined objectives (such as accuracy vs fairness trade-offs) can be operationalized within an evolutionary optimization scheme using Particle Swarm Optimization (PSO). Importantly, our formulation ensures both convergence guarantees and adaptability to diverse user-defined priorities.

Through a case study on brain age estimation, we validated the effectiveness of u-DemAI in addressing fairness, specifically with respect to age-related bias in predictive outcomes. By incorporating fairness-related metrics into the optimization objective, u-DemAI was shown to outperform both individual AI services and expert-designed ensembles in terms of accuracy and fairness. This highlights the potential of Democratic AI to democratize not only access to AI resources, but also their governance and optimization.

However, several limitations remain. First, our current evaluation was conducted under simulation using a held-out dataset, rather than incorporating real-time human-in-the-loop interactions. This limits the ecological validity of our findings, particularly in clinical contexts where user feedback may be heterogeneous, inconsistent, or resource-constrained. Second, while we demonstrated fairness in relation to age, the extension of u-DemAI to other socially salient attributes (such as gender, ethnicity, socioeconomic background) remains to be fully explored. Finally, the reliance on centralized benchmarking datasets may

constrain generalizability, as real-world deployments will inevitably face distribution shifts and heterogeneous data quality.

Future work will aim to extend u-DemAI in several directions. First, we plan to incorporate real human-in-the-loop evaluation to test the robustness of the system when exposed to diverse user communities. Second, we will investigate fairness across multiple demographic dimensions beyond age, and explore mechanisms for balancing potentially competing fairness objectives. Third, we intend to generalize the framework to other domains, including healthcare decision support, education, and policy forecasting, to further assess its scalability and impact. Collectively, these efforts will strengthen the case for Democratic AI as a practical paradigm for inclusive, transparent, and socially beneficial AI systems.

Chapter 6

Conclusions

This chapter concludes the dissertation by summarizing the key contributions, reflecting on their significance for both neuroscience and artificial intelligence, discussing the limitations of the proposed methods, and outlining future research directions. The work presented here followed a trajectory of increasing scope—from methodological innovation in ensemble learning, to explainable architectures in deep learning, and finally to embedding democratic principles into AI frameworks. Together, these contributions demonstrate interpretability, fairness, and inclusivity, offering a holistic approach to the challenge of brain age estimation and beyond.

6.1 Contributions

The first contribution of this dissertation was nl-AAE, a nonlinear age-adaptive ensemble model designed to address the imbalance inherent in brain age datasets. By dividing data into age-specific subgroups and training dedicated ensemble models, nl-AAE demonstrated that nonlinear weighting across independent learners (GoogLeNet, ResNet, CNN, and SVR) could improve prediction accuracy while reducing systematic bias. Importantly, this framework showed that age-sensitive modeling can mitigate the skewed influence of demographic imbalance, achieving both higher accuracy and greater fairness compared to single-model approaches.

The second contribution advanced brain age estimation by Triamese-ViT, a multi-view transformer-based model with built-in interpretability. Unlike traditional CNNs or ensemble methods, Triamese-ViT processed MRI volumes along three orthogonal views (axial, coronal, sagittal), combining their outputs through a fusion mechanism to produce final predictions. Beyond achieving state-of-the-art predictive accuracy, Triamese-ViT has built-in interpretation, which generated attention maps, and occlusion sensitivity analysis became a baseline for it to prove its credibility. The explainable results from Triamese-ViT highlighted crucial regions for brain age estimation. These explainable outcomes revealed the significant

regions for normal aging, gender-specific, and ASD diagnosis.

The third contribution expanded the methodological scope to the societal level by embedding brain age estimation within a user-centric Democratic AI framework. Unlike nl-AAE and Triamese-ViT, which were primarily expert-driven, u-DemAI explicitly incorporated fairness as part of its optimization objective and placed users at the center of the AI. By allowing users to specify preferences—such as accuracy vs fairness trade-offs—and by combining user-contributed models through optimization, the framework democratized access to AI services. The case study on brain age estimation demonstrated that u-DemAI not only reduced ageism in predictions but also surpassed expert-designed ensemble models, underscoring the potential of democratized AI.

6.2 Broader Significance

The findings of this thesis have significance in three dimensions: neuroscience, machine learning, and the societal governance of AI.

From a neuroscience perspective, the proposed models reinforce the utility of brain age as a biomarker, offering reliable tools for assessing deviations in normal aging trajectories. Triamese-ViT, in particular, identified brain regions whose age-related changes align with prior literature, thereby validating the biological plausibility of deep learning explanations. These insights may guide early diagnosis of neurodegenerative disorders, inform treatment monitoring, and support personalized medicine.

From a machine learning perspective, the work advances ensemble learning, transformer architectures, and fairness-aware optimization. nl-AAE demonstrated the power of nonlinear adaptive ensembles in imbalance datasets; Triamese-ViT contributed to the growing body of interpretable vision transformers; and u-DemAI introduced a novel paradigm in which fairness is mathematically embedded into optimization objectives. Collectively, these innovations enrich the methodological toolkit of AI beyond the specific application of brain age prediction.

From a societal perspective, u-DemAI highlighted how democratic participation can be operationalized in AI. By treating users not only as consumers but also as contributors and decision-makers, the framework offers a vision of AI that is transparent, accountable, and aligned with diverse social values. In this sense, the thesis contributes to democratic AI, demonstrating its feasibility through a concrete medical case study.

6.3 Limitations

Despite donating these contributions, several limitations must be acknowledged.

First, the reliance on publicly available datasets such as PAC 2019 constrains the diversity of populations represented in both training and testing. Most participants originate from Western cohorts, which limits generalizability across different demographic, cultural, and

clinical populations. Future studies should incorporate multi-site, cross-cultural data to evaluate model robustness under broader population variability.

Second, the practical deployment of the proposed models—nl-AAE, Triamese-ViT, and u-DemAI—faces notable challenges in clinical settings. The nl-AAE model, while improving predictive accuracy across age groups, depends on large and well-balanced datasets that are rarely available in hospital environments. Additionally, its ensemble structure, which integrates multiple base learners, increases computational demands and memory consumption, complicating integration with clinical data systems. Triamese-ViT, although achieving state-of-the-art accuracy and interpretability, also requires extensive computational resources and high-quality preprocessed MRI data. These requirements make it less suitable for real-time clinical deployment or for institutions with limited computational infrastructure. Moreover, the interpretability maps—although biologically plausible—have yet to be systematically validated by clinical experts, leaving their diagnostic utility unconfirmed.

Similarly, while the u-DemAI framework represents a step toward democratized and fairness-aware AI, it was evaluated primarily through simulation rather than real-world user interaction. The use of Particle PSO served as a proxy for modeling collective user behavior and preference-driven optimization, but it cannot fully replicate the complexity, variability, and contextual decision-making processes of actual human feedback. Implementing such a human-in-the-loop framework in practice will require well-designed user interfaces, ethical oversight, and regulatory compliance to ensure safety and accountability in clinical use.

Finally, fairness in this study was operationalized primarily in relation to ageism—mitigating systematic overestimation or underestimation of brain age across chronological age groups. Although this is a central fairness concern in brain age prediction, other sensitive attributes such as gender, ethnicity, and socioeconomic status were not explicitly addressed. Extending fairness auditing across multiple demographic dimensions, and exploring bias mitigation strategies that generalize across diverse populations, remains an essential future direction.

In summary, the deployment of the proposed frameworks in clinical practice requires further validation across multi-site datasets, model compression for resource-limited environments, expert collaboration for interpretability assessment, and expansion of fairness considerations beyond age-related bias. Addressing these limitations will be critical to achieving reliable, equitable, and clinically applicable brain age estimation systems.

6.4 Future Work

Several future research directions are presented here.

In our study, model performance was primarily evaluated using widely adopted metrics in the brain age estimation literature, including mean absolute error (MAE) and Pearson correlation. These measures allow direct comparison with prior work, ensuring consistency and reproducibility across studies. However, we acknowledge that statistical significance

testing of performance differences (e.g., paired hypothesis testing between models) would provide a stronger validation of the observed improvements. Incorporating such analyses represents an important direction for future work, as it would allow us to confirm whether the differences between models are not only numerically but also statistically meaningful.

Future studies should explore multi-modality MRI integration, incorporating modalities such as T2-weighted imaging, diffusion-weighted imaging (DWI), and quantitative susceptibility mapping (QSM) to capture complementary structural and microstructural features. This could enhance both predictive performance and robustness to noise.

On the methodological side, data-efficient learning strategies—including self-supervised pretraining, transfer learning from large neuroimaging datasets, and federated learning—offer pathways to reduce dependence on large labeled datasets. Furthermore, model compression techniques, such as pruning, quantization, or knowledge distillation, may mitigate the computational burden, making models more suitable for clinical deployment.

For interpretability, stabilizing attention maps and implementing global normalization strategies will be crucial for ensuring cross-cohort comparability. Incorporating expert-in-the-loop validation will also enhance the credibility and clinical utility of model explanations.

For the democratic AI framework, the next step is to conduct real human-in-the-loop evaluations. This includes involving clinicians, patients, and domain experts in the preference-setting process and testing usability in real clinical workflows. Additionally, fairness optimization should extend beyond age to encompass gender, ethnicity, and other demographic factors, ensuring that the benefits of AI are equitably distributed. Beyond healthcare, the u-DemAI framework could be generalized to domains such as education, environmental monitoring, and policy-making, further demonstrating its versatility.

6.5 Conclusion

In conclusion, this dissertation demonstrates that advancing brain age estimation requires more than building accurate predictive models. It requires interpretability to ensure trust, fairness to ensure inclusivity, and democratization to ensure that AI reflects the values of the communities it serves. By integrating methodological rigor with neuroscientific insight and democratic principles, the research presented here contributes to a vision of AI that is technically advanced, clinically relevant, and socially responsible.

The trajectory traced in this work—from nl-AAE to Triamese-ViT to u-DemAI—illustrates a gradual expansion of focus: from accuracy, to interpretability, to fairness and democratization. Together, these contributions form a forward-looking research direction, bridging the gap between innovation and societal impact. While limitations remain, the insights and methods developed here provide a strong foundation for future work at the intersection of machine learning, neuroscience, and democratic AI.

Appendix A

Appendices

Table A.1: The sensitivity of various brain regions in healthy individuals and ASD patients during brain age estimation. BMI (Built-in Model Interpretation) reflects attention values, while OSA (Occlusion Sensitivity Analysis) shows impact when occluding regions.

Brain Region	**Healthy People**		**ASD Patients**	
	BMI	**OSA**	**BMI**	**OSA**
Precentral_L	0.30	1.47	0.08	0.42
Precentral_R	0.26	1.15	0.07	0.48
Frontal_Sup_L	0.40	1.43	0.02	0.45
Frontal_Sup_R	0.43	0.97	0.03	0.34
Frontal_Sup_Orb_L	0	0.56	0	0.24
Frontal_Sup_Orb_R	0	0.52	0	0.23
Frontal_Mid_L	0.69	1.06	0.05	0.36
Frontal_Mid_R	0.54	0.75	0.03	0.40
Frontal_Mid_Orb_L	0	0.83	0	0.32
Frontal_Mid_Orb_R	0	0.39	0	0.19
Frontal_Inf_Oper_L	2.15	1.93	0.22	0.60
Frontal_Inf_Oper_R	0.82	1.83	0.07	0.69
Frontal_Inf_Tri_L	1.15	1.07	0.09	0.32
Frontal_Inf_Tri_R	0.75	1.19	0.05	0.42
Frontal_Inf_Orb_L	0	1.26	0	0.43
Frontal_Inf_Orb_R	0	1.03	0	0.38

Continued on next page...

Brain Region	**Healthy People**		**ASD Patients**	
	BMI	**OSA**	**BMI**	**OSA**
Rolandic_Oper_L	4.31	5.09	0.63	1.79
Rolandic_Oper_R	3.94	2.57	0.55	1.39
Supp_Motor_Area_L	0.25	2.28	0.11	0.68
Supp_Motor_Area_R	0.05	2.18	0.05	0.51
Olfactory_L	0.29	2.93	0.06	1.02
Olfactory_R	0	2.62	0	0.98
Frontal_Sup_Medial_L	0.89	1.69	0.10	0.53
Frontal_Sup_Medial_R	0.69	0.78	0.05	0.28
Frontal_Med_Orb_L	0.20	0.85	0.08	0.27
Frontal_Med_Orb_R	0.04	0.94	0.02	0.32
Rectus_L	0.22	1.15	0.07	0.39
Rectus_R	0.01	1.09	0	0.42
Insula_L	1.19	5.24	0.26	1.85
Insula_R	0.97	5.02	0.26	2.13
Cingulum_Ant_L	1.81	4.47	0.29	1.57
Cingulum_Ant_R	2.02	4.20	0.20	1.54
Cingulum_Mid_L	1.47	2.30	0.53	1.41
Cingulum_Mid_R	0.20	2.40	0.09	1.09
Cingulum_Post_L	3.79	1.86	0.99	1.65
Cingulum_Post_R	2.36	1.52	1.01	1.61
Hippocampus_L	0.07	1.96	0.07	1.19
Hippocampus_R	0.07	1.22	0.07	0.81
ParaHippocampal_L	0.05	0.63	0.05	0.51
ParaHippocampal_R	0.06	0.82	0.06	0.45
Amygdala_L	0	0.52	0	0.41
Amygdala_R	0	0.83	0	0.49
Calcarine_L	1.31	0.85	0.24	1.35
Calcarine_R	2.29	0.46	0.36	1.44
Cuneus_L	2.16	0.30	0.30	1.00
Cuneus_R	1.40	0.42	0.16	1.01

Continued on next page...

Brain Region	**Healthy People**		**ASD Patients**	
	BMI	**OSA**	**BMI**	**OSA**
Lingual_L	0.06	1.87	0.04	1.54
Lingual_R	0	0.99	0	1.26
Occipital_Sup_L	1.00	0.17	0.08	0.42
Occipital_Sup_R	1.95	0.41	0.22	0.58
Occipital_Mid_L	1.16	0.15	0.11	0.28
Occipital_Mid_R	1.22	0.20	0.10	0.37
Occipital_Inf_L	0	0.19	0	0.20
Occipital_Inf_R	0	0.08	0	0.16
Fusiform_L	0.02	0.78	0.02	0.59
Fusiform_R	0.02	0.54	0.02	0.50
Postcentral_L	0.79	0.81	0.13	0.37
Postcentral_R	0.33	1.05	0.05	0.51
Parietal_Sup_L	0	0.31	0	0.27
Parietal_Sup_R	0	0.37	0	0.22
Parietal_Inf_L	0	0.40	0	0.37
Parietal_Inf_R	0	0.74	0	0.66
SupraMarginal_L	0.58	0.68	0.05	0.37
SupraMarginal_R	0.43	0.59	0.05	0.47
Angular_L	0	0.29	0	0.26
Angular_R	0	0.37	0	0.39
Precuneus_L	0.87	0.97	0.21	0.81
Precuneus_R	0.72	0.98	0.29	0.96
Paracentral_Lobule_L	0.11	1.78	0.07	0.51
Paracentral_Lobule_R	0.03	1.77	0.01	0.41
Caudate_L	2.66	7.27	0.41	3.19
Caudate_R	2.96	6.60	0.50	2.83
Putamen_L	0.05	6.42	0.05	2.34
Putamen_R	0.04	6.69	0.04	2.78
Pallidum_L	0	5.40	0	1.87
Pallidum_R	0	5.88	0	2.42

Continued on next page...

Brain Region	**Healthy People**		**ASD Patients**	
	BMI	**OSA**	**BMI**	**OSA**
Thalamus_L	2.60	8.09	0.76	3.60
Thalamus_R	3.50	6.18	1.85	3.09
Heschl_L	0.08	7.43	0.08	2.57
Heschl_R	0.09	5.52	0.09	2.97
Temporal_Sup_L	1.74	2.23	0.21	0.81
Temporal_Sup_R	1.18	1.63	0.15	0.90
Temporal_Pole_Sup_L	0	0.43	0	0.20
Temporal_Pole_Sup_R	0	0.74	0	0.30
Temporal_Mid_L	0.82	0.50	0.10	0.31
Temporal_Mid_R	0.93	0.26	0.11	0.25
Temporal_Pole_Mid_L	0	0.14	0	0.07
Temporal_Pole_Mid_R	0	0.08	0	0.05
Temporal_Inf_L	0.03	0.20	0.03	0.17
Temporal_Inf_R	0.02	0.12	0.02	0.10
Cerebelum_Crus1_L	0	0.31	0	0.20
Cerebelum_Crus1_R	0	0.34	0	0.21
Cerebelum_Crus2_L	0.03	0.21	0.01	0.16
Cerebelum_Crus2_R	0	0.18	0	0.10
Cerebelum_3_L	0	1.64	0	1.18
Cerebelum_3_R	0	1.32	0	0.89
Cerebelum_4_5_L	0	1.88	0	1.49
Cerebelum_4_5_R	0	1.36	0	1.33
Cerebelum_6_L	0	0.94	0	0.76
Cerebelum_6_R	0	0.89	0	0.83
Cerebelum_7b_L	0	0.15	0	0.13
Cerebelum_7b_R	0	0.13	0	0.08
Cerebelum_8_L	0	0.32	0	0.25
Cerebelum_8_R	0	0.33	0	0.20
Cerebelum_9_L	0.02	0.85	0.01	0.56
Cerebelum_9_R	0	0.84	0	0.55

Continued on next page...

Brain Region	**Healthy People**		**ASD Patients**	
	BMI	**OSA**	**BMI**	**OSA**
Cerebelum_10_L	0	0.48	0	0.31
Cerebelum_10_R	0	0.38	0	0.21
Vermis_1_2	2.04	1.74	0.53	1.01
Vermis_3	3.22	1.83	0.70	1.52
Vermis_4_5	1.56	2.11	0.45	1.89
Vermis_6	0.61	1.54	0.27	1.51
Vermis_7	0.43	1.24	0.26	0.93
Vermis_8	0.58	1.24	0.38	1.00
Vermis_9	0.80	1.74	0.56	1.04
Vermis_10	0.74	1.83	0.39	1.05

Table A.2: The sensitivity of various brain regions in normal aging during brain age estimation. The values are derived from Built-in Model Interpretation (BMI).

Brain Region	**Normal Aging**							
	0s	**10s**	**20s**	**30s**	**40s**	**50s**	**60s**	**70s**
Precentral_L	0.86	0.19	0.42	0.25	0.25	0.14	0.62	0.24
Precentral_R	1.03	0.18	0.43	0.24	0.25	0.12	0.67	0.25
Frontal_Sup_L	0.25	0.06	0.17	0.10	0.13	0.07	0.35	0.14
Frontal_Sup_R	0.27	0.05	0.19	0.13	0.14	0.08	0.39	0.17
Frontal_Sup_Orb_L	0	0	0	0	0	0	0	0
Frontal_Sup_Orb_R	0	0	0	0	0	0	0	0
Frontal_Mid_L	0.47	0.08	0.28	0.19	0.22	0.12	0.63	0.23
Frontal_Mid_R	0.33	0.07	0.22	0.15	0.16	0.10	0.49	0.18
Frontal_Mid_Orb_L	0	0	0	0	0	0	0	0
Frontal_Mid_Orb_R	0	0	0	0	0	0	0	0
Frontal_Inf_Oper_L	2.07	0.42	1.28	0.91	0.98	0.61	2.64	1.02
Frontal_Inf_Oper_R	0.53	0.14	0.44	0.28	0.30	0.19	0.92	0.36
Frontal_Inf_Tri_L	0.92	0.15	0.55	0.38	0.42	0.24	1.19	0.47
Frontal_Inf_Tri_R	0.46	0.11	0.34	0.24	0.27	0.15	0.77	0.29
Frontal_Inf_Orb_L	0	0	0	0	0	0	0	0
Frontal_Inf_Orb_R	0	0	0	0	0	0	0	0
Rolandic_Oper_L	5.69	1.49	3.39	2.59	2.64	1.66	5.98	2.59
Rolandic_Oper_R	5.30	1.34	3.18	2.33	2.49	1.44	6.15	2.54
Supp_Motor_Area_L	0.86	0.21	0.87	0.77	0.87	0.42	1.74	0.83
Supp_Motor_Area_R	0.59	0.12	0.27	0.13	0.15	0.06	0.32	0.13
Olfactory_L	0.44	0.16	0.60	0.60	0.57	0.33	1.27	0.38
Olfactory_R	0	0	0	0	0	0	0	0
Frontal_Sup_Medial_L	1.00	0.25	1.08	0.84	0.92	0.52	2.16	0.91
Frontal_Sup_Medial_R	0.48	0.10	0.31	0.22	0.23	0.14	0.71	0.27
Frontal_Med_Orb_L	0.74	0.19	0.91	0.76	0.85	0.40	1.89	0.79
Frontal_Med_Orb_R	0.15	0.04	0.16	0.16	0.16	0.09	0.53	0.16
Rectus_L	0.63	0.18	0.64	0.58	0.61	0.31	1.59	0.64
Rectus_R	0.02	0	0.02	0.01	0.01	0.01	0.05	0.02

Continued on next page...

Brain Region	**Normal Aging**							
	0s	**10s**	**20s**	**30s**	**40s**	**50s**	**60s**	**70s**
Insula_L	1.81	0.62	1.26	0.90	0.96	0.56	1.23	0.90
Insula_R	1.84	0.69	1.04	0.68	0.71	0.40	1.34	0.64
Cingulum_Ant_L	2.58	0.81	2.63	2.22	2.24	1.21	5.61	2.02
Cingulum_Ant_R	1.76	0.47	1.19	0.88	1.01	0.57	2.77	1.05
Cingulum_Mid_L	3.62	1.14	3.13	2.88	3.04	1.68	5.14	2.79
Cingulum_Mid_R	0.70	0.21	0.44	0.33	0.32	0.16	0.58	0.28
Cingulum_Post_L	4.37	2.18	4.94	4.39	4.55	2.68	7.93	4.35
Cingulum_Post_R	2.47	2.16	2.66	2.45	2.31	2.31	2.22	2.17
Hippocampus_L	0.84	0.21	0.36	0.19	0.20	0.09	0.40	0.15
Hippocampus_R	0.90	0.21	0.37	0.19	0.22	0.09	0.40	0.16
ParaHippocampal_L	0.58	0.14	0.21	0.11	0.12	0.05	0.23	0.09
ParaHippocampal_R	0.76	0.17	0.26	0.13	0.15	0.06	0.30	0.11
Amygdala_L	0	0	0	0	0	0	0	0
Amygdala_R	0	0	0	0	0	0	0	0
Calcarine_L	1.62	0.48	1.51	1.23	1.30	0.77	3.12	1.20
Calcarine_R	2.34	0.69	1.74	1.46	1.55	0.95	1.96	1.51
Cuneus_L	2.03	0.63	1.77	1.43	1.48	0.88	4.02	1.51
Cuneus_R	1.25	0.35	0.80	0.61	0.71	0.41	1.71	0.67
Lingual_L	0.17	0.06	0.20	0.20	0.21	0.12	0.50	0.20
Lingual_R	0	0	0	0	0	0	0	0
Occipital_Sup_L	0.68	0.18	0.48	0.33	0.36	0.23	1.01	0.35
Occipital_Sup_R	1.82	0.46	1.09	0.86	0.91	0.55	2.38	0.90
Occipital_Mid_L	1.01	0.23	0.60	0.42	0.47	0.30	1.31	0.47
Occipital_Mid_R	0.88	0.23	0.63	0.46	0.51	0.31	1.40	0.51
Occipital_Inf_L	0	0	0	0	0	0	0	0
Occipital_Inf_R	0	0	0	0	0	0	0	0
Fusiform_L	0.28	0.06	0.09	0.04	0.05	0.02	0.10	0.04
Fusiform_R	0.29	0.06	0.08	0.04	0.05	0.02	0.09	0.03
Postcentral_L	1.55	0.33	0.78	0.46	0.49	0.27	1.34	0.50
Postcentral_R	0.69	0.13	0.36	0.20	0.21	0.11	0.58	0.22

Continued on next page...

Brain Region	**Normal Aging**							
	0s	**10s**	**20s**	**30s**	**40s**	**50s**	**60s**	**70s**
Parietal_Sup_L	0	0	0	0	0	0	0	0
Parietal_Sup_R	0	0	0	0	0	0	0	0
Parietal_Inf_L	0	0	0	0	0	0	0	0
Parietal_Inf_R	0	0	0	0	0	0	0	0
SupraMarginal_L	0.49	0.12	0.32	0.22	0.26	0.14	0.77	0.26
SupraMarginal_R	0.61	0.15	0.35	0.23	0.24	0.13	0.65	0.21
Angular_L	0	0	0	0	0	0	0	0
Angular_R	0	0	0	0	0	0	0	0
Precuneus_L	1.17	0.45	1.24	1.09	1.14	0.65	2.18	1.17
Precuneus_R	0.78	0.59	0.74	0.75	0.77	0.71	0.71	0.76
Paracentral_Lobule_L	0.67	0.12	0.45	0.34	0.35	0.16	0.79	0.32
Paracentral_Lobule_R	0.10	0.02	0.11	0.07	0.09	0.05	0.22	0.09
Caudate_L	3.11	1.11	2.13	1.51	1.58	1.00	2.61	1.54
Caudate_R	3.17	1.31	2.39	1.79	1.97	1.18	2.67	1.93
Putamen_L	0.57	0.12	0.23	0.11	0.13	0.06	0.21	0.08
Putamen_R	0.30	0.09	0.17	0.08	0.09	0.03	0.15	0.06
Pallidum_L	0	0	0	0	0	0	0	0
Pallidum_R	0	0	0	0	0	0	0	0
Thalamus_L	6.07	2.70	5.02	2.83	2.88	1.69	0	2.21
Thalamus_R	5.69	4.66	6.03	5.28	5.52	3.95	4.13	5.18
Heschl_L	1.17	0.28	0.63	0.33	0.31	0.14	5.56	0.24
Heschl_R	1.32	0.30	0.62	0.30	0.28	0.13	0.63	0.24
Temporal_Sup_L	1.90	0.47	1.19	0.81	0.90	0.52	0.67	0.89
Temporal_Sup_R	1.61	0.34	0.86	0.60	0.65	0.36	2.39	0.69
Temporal_Pole_Sup_L	0	0	0	0	0	0	1.67	0
Temporal_Pole_Sup_R	0	0	0	0	0	0	0	0
Temporal_Mid_L	1.15	0.24	0.58	0.34	0.37	0.20	0	0.40
Temporal_Mid_R	0.98	0.25	0.60	0.45	0.50	0.27	1.08	0.54
Temporal_Pole_Mid_L	0	0	0	0	0	0	1.12	0
Temporal_Pole_Mid_R	0	0	0	0	0	0	0	0

Continued on next page...

Brain Region	**Normal Aging**							
	0s	**10s**	**20s**	**30s**	**40s**	**50s**	**60s**	**70s**
Temporal_Inf_L	0.46	0.08	0.10	0.05	0.05	0.02	0	0.04
Temporal_Inf_R	0.47	0.08	0.11	0.05	0.06	0.02	0.11	0.04
Cerebelum_Crus1_L	0	0	0	0	0	0	0.12	0
Cerebelum_Crus1_R	0	0	0	0	0	0	0	0
Cerebelum_Crus2_L	0.04	0.01	0.09	0.06	0.08	0.04	0	0.07
Cerebelum_Crus2_R	0	0	0	0	0	0	0.19	0
Cerebelum_3_L	0	0	0	0	0	0	0	0
Cerebelum_3_R	0	0	0	0	0	0	0	0
Cerebelum_4_5_L	0	0	0	0	0	0	0	0
Cerebelum_4_5_R	0	0	0	0	0	0	0	0
Cerebelum_6_L	0	0	0	0	0	0	0	0
Cerebelum_6_R	0	0	0	0	0	0	0	0
Cerebelum_7b_L	0	0	0	0	0	0	0	0
Cerebelum_7b_R	0	0	0	0	0	0	0	0
Cerebelum_8_L	0	0	0	0	0	0	0	0
Cerebelum_8_R	0	0	0	0	0	0	0	0
Cerebelum_9_L	0.06	0.02	0.06	0.07	0.08	0.04	0.24	0.10
Cerebelum_9_R	0	0	0	0	0	0	0	0
Cerebelum_10_L	0	0	0	0	0	0	0	0
Cerebelum_10_R	0	0	0	0	0	0	0	0
Vermis_1_2	2.42	0.72	3.04	2.82	2.97	1.72	7.54	3.69
Vermis_3	3.92	1.40	4.27	4.18	4.67	2.52	10.44	4.76
Vermis_4_5	2.14	0.71	2.63	2.59	2.80	1.48	6.71	3.01
Vermis_6	1.44	0.49	1.79	1.67	1.75	0.94	4.55	1.81
Vermis_7	1.28	0.49	1.64	1.49	1.63	0.87	4.50	1.72
Vermis_8	1.97	0.66	2.17	1.96	2.44	1.36	7.16	2.81
Vermis_9	2.56	0.91	3.20	3.08	3.55	2.04	11.62	4.51
Vermis_10	1.88	0.54	2.58	2.38	2.32	1.38	6.96	2.97

Table A.3: The sensitivity of various brain regions in healthy male and female individuals during Triamese-ViT brain age prediction. BMI (Built-in Model Interpretation) reflects attention values, while OSA (Occlusion Sensitivity Analysis) shows impact when occluding regions.

Brain Region	**Male**		**Female**	
	BMI	**OSA**	**BMI**	**OSA**
Precentral_L	0.09	0.05	0.11	0.14
Precentral_R	0.15	0.14	0.12	0.04
Frontal_Sup_L	0.07	0.03	0.09	0.16
Frontal_Sup_R	0.09	0.02	0.10	0.21
Frontal_Sup_Orb_L	0	0.01	0	0.04
Frontal_Sup_Orb_R	0	0.06	0	0.13
Frontal_Mid_L	0.16	0.06	0.13	0.41
Frontal_Mid_R	0.08	0.05	0.10	0.07
Frontal_Mid_Orb_L	0	0.01	0	0.13
Frontal_Mid_Orb_R	0	0.02	0	0.16
Frontal_Inf_Oper_L	0.19	0.10	0.42	0.17
Frontal_Inf_Oper_R	0.09	0.03	0.16	0.10
Frontal_Inf_Tri_L	0.26	0.05	0.29	0.22
Frontal_Inf_Tri_R	0.10	0.05	0.17	0.08
Frontal_Inf_Orb_L	0	0.02	0	0.25
Frontal_Inf_Orb_R	0	0.01	0	0.13
Rolandic_Oper_L	0.35	0.13	0.51	0.13
Rolandic_Oper_R	0.40	0.07	0.37	0.03
Supp_Motor_Area_L	1.60	0.03	0.40	0.17
Supp_Motor_Area_R	0.10	0.03	0.25	0.18
Olfactory_L	0.31	0.02	0.02	0.06
Olfactory_R	0	0.04	0	0.08
Frontal_Sup_Medial_L	1.46	0.05	0.30	0.13
Frontal_Sup_Medial_R	0.14	0.05	0.15	0.47
Frontal_Med_Orb_L	0.88	0.02	0.10	0.04
Frontal_Med_Orb_R	0.16	0.01	0.01	0.12

Continued on next page...

Brain Region	**Male**		**Female**	
	BMI	**OSA**	**BMI**	**OSA**
Rectus_L	1.13	0.08	0.41	0.06
Rectus_R	0.02	0.10	0.01	0.07
Insula_L	0.11	0.04	0.74	0.56
Insula_R	0.07	0.02	0.22	0.08
Cingulum_Ant_L	1.06	0.03	0.72	0.09
Cingulum_Ant_R	0.38	0.08	0.47	0.83
Cingulum_Mid_L	0.96	0.01	1.44	0.18
Cingulum_Mid_R	0.08	0.01	0.19	0.17
Cingulum_Post_L	1.81	0.03	1.76	0.24
Cingulum_Post_R	1.38	0.10	2.09	1.09
Hippocampus_L	0.10	0.04	0.06	0.30
Hippocampus_R	0.10	0.02	0.08	0.20
ParaHippocampal_L	0.10	0.13	0.10	0.28
ParaHippocampal_R	0.12	0.21	0.17	0.14
Amygdala_L	0	0.07	0	0.61
Amygdala_R	0	0.02	0	0.20
Calcarine_L	0.81	0.04	0.25	0.15
Calcarine_R	0.18	0.03	0.31	0.67
Cuneus_L	1.20	0.01	0.54	0.11
Cuneus_R	0.13	0.03	0.19	0.89
Lingual_L	0.15	0.02	0.02	0.12
Lingual_R	0	0.01	0	0.14
Occipital_Sup_L	0.15	0.02	0.12	0.34
Occipital_Sup_R	0.17	0.03	0.19	0.34
Occipital_Mid_L	0.19	0.02	0.30	0.87
Occipital_Mid_R	0.12	0.02	0.18	0.04
Occipital_Inf_L	0	0.01	0	0.23
Occipital_Inf_R	0	0.01	0	0.12
Fusiform_L	0.07	0.16	0.10	0.15
Fusiform_R	0.03	0.17	0.08	0.12

Continued on next page...

Brain Region	**Male**		**Female**	
	BMI	**OSA**	**BMI**	**OSA**
Postcentral_L	0.23	0.08	0.17	0.13
Postcentral_R	0.11	0.09	0.06	0.07
Parietal_Sup_L	0	0.02	0	0.13
Parietal_Sup_R	0	0.04	0	0.05
Parietal_Inf_L	0	0.08	0	0.20
Parietal_Inf_R	0	0.17	0	0.04
SupraMarginal_L	0.08	0.03	0.05	0.12
SupraMarginal_R	0.06	0.10	0.04	0.05
Angular_L	0	0.07	0	0.29
Angular_R	0	0.05	0	0.07
Precuneus_L	0.87	0.02	0.35	0.14
Precuneus_R	0.14	0.04	0.27	0.42
Paracentral_Lobule_L	0.49	0.03	0.15	0.22
Paracentral_Lobule_R	0.27	0.02	0.14	0.28
Caudate_L	0.54	0.04	0.53	0.21
Caudate_R	0.42	0.07	0.34	0.71
Putamen_L	0.03	0.03	0.25	0.37
Putamen_R	0.02	0.02	0.06	0.15
Pallidum_L	0	0.04	0	0.16
Pallidum_R	0	0.02	0	0.35
Thalamus_L	2.35	0.08	1.14	0.30
Thalamus_R	1.64	0.11	0.87	1.07
Heschl_L	0.06	0.09	0.51	0.22
Heschl_R	0.08	0.03	0.67	0.04
Temporal_Sup_L	0.34	0.09	0.22	0.05
Temporal_Sup_R	0.18	0.06	0.22	0.03
Temporal_Pole_Sup_L	0	0.21	0	0.39
Temporal_Pole_Sup_R	0	0.06	0	0.10
Temporal_Mid_L	0.14	0.07	0.21	0.08
Temporal_Mid_R	0.10	0.04	0.10	0.10

Continued on next page...

Brain Region	**Male**		**Female**	
	BMI	**OSA**	**BMI**	**OSA**
Temporal_Pole_Mid_L	0	0.63	0	0.16
Temporal_Pole_Mid_R	0	0.29	0	0.14
Temporal_Inf_L	0.06	0.05	0.03	0.07
Temporal_Inf_R	0.03	0.03	0.04	0.14
Cerebelum_Crus1_L	0	0.19	0	0.28
Cerebelum_Crus1_R	0	0.20	0	0.10
Cerebelum_Crus2_L	0.19	0.37	0.02	0.16
Cerebelum_Crus2_R	0	0.26	0	0.05
Cerebelum_3_L	0	0.06	0	0.11
Cerebelum_3_R	0	0.06	0	0.26
Cerebelum_4_5_L	0	0.15	0	0.10
Cerebelum_4_5_R	0	0.11	0	0.12
Cerebelum_6_L	0	0.34	0	0.17
Cerebelum_6_R	0	0.31	0	0.11
Cerebelum_7b_L	0	0.22	0	0.03
Cerebelum_7b_R	0	0.05	0	0.02
Cerebelum_8_L	0	0.26	0	0.04
Cerebelum_8_R	0	0.16	0	0.05
Cerebelum_9_L	0.12	0.03	0.02	0.06
Cerebelum_9_R	0	0.01	0	0.12
Cerebelum_10_L	0	0.97	0	0.05
Cerebelum_10_R	0	0.89	0	0.19
Vermis_1_2	2.64	0.30	0.51	0.33
Vermis_3	2.36	0.08	0.32	0.20
Vermis_4_5	1.84	0.12	0.08	0.11
Vermis_6	2.10	0.15	0.13	0.04
Vermis_7	2.49	0.24	0.37	0.03
Vermis_8	3.65	0.09	0.68	0.04
Vermis_9	4.02	0.04	0.73	0.11
Vermis_10	2.26	0.09	0.29	0.22

Continued on next page...

Brain Region	**Male**		**Female**	
	BMI	**OSA**	**BMI**	**OSA**

Table A.4: The sensitivity of various brain regions in ASD male and female patients during Triamese-ViT diagnosis. BMI (Built-in Model Interpretation) reflects attention values, while OSA (Occlusion Sensitivity Analysis) shows impact when occluding regions.

Brain Region	**Male**		**Female**	
	BMI	**OSA**	**BMI**	**OSA**
Precentral_L	0.08	0.09	0.11	0.09
Precentral_R	0.14	0.09	0.12	0.04
Frontal_Sup_L	0.06	0.04	0.08	0.13
Frontal_Sup_R	0.09	0.02	0.10	0.16
Frontal_Sup_Orb_L	0	0	0	0.05
Frontal_Sup_Orb_R	0	0.06	0	0.10
Frontal_Mid_L	0.16	0.05	0.14	0.32
Frontal_Mid_R	0.07	0.05	0.11	0.07
Frontal_Mid_Orb_L	0	0.01	0	0.13
Frontal_Mid_Orb_R	0	0.03	0	0.08
Frontal_Inf_Oper_L	0.17	0.09	0.47	0.12
Frontal_Inf_Oper_R	0.07	0.02	0.20	0.08
Frontal_Inf_Tri_L	0.24	0.04	0.33	0.19
Frontal_Inf_Tri_R	0.09	0.03	0.20	0.08
Frontal_Inf_Orb_L	0	0.02	0	0.19
Frontal_Inf_Orb_R	0	0.01	0	0.10
Rolandic_Oper_L	0.40	0.09	0.54	0.10
Rolandic_Oper_R	0.31	0.03	0.39	0.04
Supp_Motor_Area_L	1.43	0.02	0.46	0.11
Supp_Motor_Area_R	0.10	0.01	0.24	0.11
Olfactory_L	0.35	0.01	0.02	0.07
Olfactory_R	0	0.05	0	0.07
Frontal_Sup_Medial_L	1.29	0.06	0.32	0.13

Continued on next page...

Brain Region	**Male**		**Female**	
	BMI	**OSA**	**BMI**	**OSA**
Frontal_Sup_Medial_R	0.15	0.04	0.15	0.34
Frontal_Med_Orb_L	0.92	0.02	0.13	0.07
Frontal_Med_Orb_R	0.17	0.01	0.01	0.13
Rectus_L	1.46	0.05	0.36	0.08
Rectus_R	0.03	0.07	0.01	0.07
Insula_L	0.11	0.04	0.73	0.48
Insula_R	0.06	0.02	0.21	0.11
Cingulum_Ant_L	1.05	0.08	0.73	0.15
Cingulum_Ant_R	0.37	0.09	0.45	0.60
Cingulum_Mid_L	0.93	0.01	1.47	0.12
Cingulum_Mid_R	0.08	0.01	0.20	0.11
Cingulum_Post_L	1.76	0.04	1.80	0.13
Cingulum_Post_R	1.38	0.09	2.07	0.58
Hippocampus_L	0.17	0.05	0.06	0.21
Hippocampus_R	0.05	0.02	0.08	0.16
ParaHippocampal_L	0.11	0.12	0.09	0.23
ParaHippocampal_R	0.10	0.22	0.12	0.15
Amygdala_L	0	0.05	0	0.38
Amygdala_R	0	0.03	0	0.16
Calcarine_L	0.07	0.05	0.35	0.15
Calcarine_R	0.17	0.03	0.26	0.39
Cuneus_L	1.01	0.04	0.59	0.12
Cuneus_R	0.12	0.04	0.18	0.49
Lingual_L	0.14	0.01	0.02	0.10
Lingual_R	0	0.01	0	0.10
Occipital_Sup_L	0.10	0.04	0.11	0.26
Occipital_Sup_R	0.15	0.03	0.18	0.23
Occipital_Mid_L	0.15	0.02	0.24	0.59
Occipital_Mid_R	0.10	0.01	0.17	0.07
Occipital_Inf_L	0	0.01	0	0.14

Continued on next page...

Brain Region	**Male**		**Female**	
	BMI	**OSA**	**BMI**	**OSA**
Occipital_Inf_R	0	0.02	0	0.07
Fusiform_L	0.07	0.14	0.09	0.14
Fusiform_R	0.02	0.17	0.05	0.08
Postcentral_L	0.24	0.07	0.20	0.06
Postcentral_R	0.09	0.07	0.06	0.04
Parietal_Sup_L	0	0.02	0	0.06
Parietal_Sup_R	0	0.02	0	0.05
Parietal_Inf_L	0	0.09	0	0.11
Parietal_Inf_R	0	0.15	0	0.07
SupraMarginal_L	0.08	0.07	0.06	0.08
SupraMarginal_R	0.04	0.07	0.04	0.04
Angular_L	0	0.06	0	0.19
Angular_R	0	0.04	0	0.08
Precuneus_L	0.84	0.02	0.41	0.09
Precuneus_R	0.13	0.04	0.25	0.21
Paracentral_Lobule_L	0.46	0.01	0.15	0.09
Paracentral_Lobule_R	0.24	0.02	0.10	0.10
Caudate_L	0.58	0.07	0.54	0.22
Caudate_R	0.30	0.06	0.46	0.47
Putamen_L	0.02	0.02	0.28	0.30
Putamen_R	0.01	0.01	0.06	0.11
Pallidum_L	0	0.01	0	0.12
Pallidum_R	0	0.01	0	0.28
Thalamus_L	2.24	0.06	1.15	0.19
Thalamus_R	1.57	0.09	0.81	0.69
Heschl_L	0.07	0.06	0.64	0.16
Heschl_R	0.05	0.03	0.53	0.03
Temporal_Sup_L	0.34	0.06	0.24	0.04
Temporal_Sup_R	0.13	0.05	0.21	0.03
Temporal_Pole_Sup_L	0	0.14	0	0.29

Continued on next page...

Brain Region	**Male**		**Female**	
	BMI	**OSA**	**BMI**	**OSA**
Temporal_Pole_Sup_R	0	0.08	0	0.11
Temporal_Mid_L	0.14	0.04	0.17	0.06
Temporal_Mid_R	0.09	0.03	0.10	0.04
Temporal_Pole_Mid_L	0	0.43	0	0.13
Temporal_Pole_Mid_R	0	0.32	0	0.09
Temporal_Inf_L	0.07	0.03	0.03	0.05
Temporal_Inf_R	0.02	0.03	0.03	0.06
Cerebelum_Crus1_L	0	0.20	0	0.13
Cerebelum_Crus1_R	0	0.14	0	0.07
Cerebelum_Crus2_L	0.16	0.37	0.02	0.06
Cerebelum_Crus2_R	0	0.18	0	0.04
Cerebelum_3_L	0	0.04	0	0.05
Cerebelum_3_R	0	0.04	0	0.10
Cerebelum_4_5_L	0	0.12	0	0.10
Cerebelum_4_5_R	0	0.13	0	0.09
Cerebelum_6_L	0	0.28	0	0.17
Cerebelum_6_R	0	0.28	0	0.10
Cerebelum_7b_L	0	0.20	0	0.02
Cerebelum_7b_R	0	0.03	0	0.02
Cerebelum_8_L	0	0.23	0	0.03
Cerebelum_8_R	0	0.12	0	0.04
Cerebelum_9_L	0.10	0.03	0.01	0.06
Cerebelum_9_R	0	0.01	0	0.07
Cerebelum_10_L	0	0.70	0	0.05
Cerebelum_10_R	0	1.56	0	0.09
Vermis_1_2	2.78	0.16	0.38	0.17
Vermis_3	2.31	0.06	0.36	0.08
Vermis_4_5	1.82	0.09	0.11	0.07
Vermis_6	1.99	0.08	0.13	0.06
Vermis_7	2.49	0.10	0.27	0.04

Continued on next page...

Brain Region	**Male**		**Female**	
	BMI	**OSA**	**BMI**	**OSA**
Vermis_8	3.93	0.06	0.42	0.06
Vermis_9	3.90	0.06	0.48	0.08
Vermis_10	1.79	0.12	0.30	0.12

Table A.5: The sensitivity of various brain regions based on Region-Based Occlusion Sensitivity Analysis comparing healthy individuals and ASD patients.

Brain Region	**Region-Based OSA**	
	Healthy People	**ASD Patients**
Precentral_L	2.37	0.69
Precentral_R	0.78	0.73
Frontal_Sup_L	2.74	2.43
Frontal_Sup_R	3.32	2.95
Frontal_Sup_Orb_L	0.24	0.22
Frontal_Sup_Orb_R	0.41	0.32
Frontal_Mid_L	5.08	2.94
Frontal_Mid_R	2.30	2.18
Frontal_Mid_Orb_L	0.51	0.83
Frontal_Mid_Orb_R	0.78	0.81
Frontal_Inf_Oper_L	0.66	0.85
Frontal_Inf_Oper_R	0.62	0.89
Frontal_Inf_Tri_L	1.68	1.80
Frontal_Inf_Tri_R	1.01	0.67
Frontal_Inf_Orb_L	0.32	0.27
Frontal_Inf_Orb_R	0.92	1.07
Rolandic_Oper_L	0.71	0.86
Rolandic_Oper_R	0.35	0.79
Supp_Motor_Area_L	0.35	0.27
Supp_Motor_Area_R	0.45	0.25
Olfactory_L	0.06	0.04
Olfactory_R	0.05	0.07
Frontal_Sup_Medial_L	2.03	0.99
Frontal_Sup_Medial_R	2.50	2.20
Frontal_Med_Orb_L	0.18	0.15
Frontal_Med_Orb_R	0.53	0.41
Rectus_L	0.08	0.05
Rectus_R	0.15	0.22

Continued on next page...

Brain Region	**Region-Based OSA**	
	Healthy People	**ASD Patients**
Insula_L	0.94	0.91
Insula_R	0.51	0.56
Cingulum_Ant_L	0.44	0.30
Cingulum_Ant_R	2.38	1.72
Cingulum_Mid_L	0.21	0.14
Cingulum_Mid_R	0.42	0.10
Cingulum_Post_L	0.70	0.60
Cingulum_Post_R	2.73	2.33
Hippocampus_L	0.84	0.88
Hippocampus_R	1.81	3.11
ParaHippocampal_L	0.44	0.27
ParaHippocampal_R	1.02	0.82
Amygdala_L	0.20	0.06
Amygdala_R	0.17	0.18
Calcarine_L	1.21	0.57
Calcarine_R	4.71	4.03
Cuneus_L	0.20	0.12
Cuneus_R	2.14	1.85
Lingual_L	1.15	0.57
Lingual_R	12.33	9.88
Occipital_Sup_L	0.80	0.72
Occipital_Sup_R	1.73	1.10
Occipital_Mid_L	1.34	1.72
Occipital_Mid_R	0.87	0.53
Occipital_Inf_L	0.13	0.10
Occipital_Inf_R	0.14	0.12
Fusiform_L	0.97	0.73
Fusiform_R	2.05	2.78
Postcentral_L	2.91	1.38
Postcentral_R	1.10	1.79

Continued on next page...

Brain Region	**Region-Based OSA**	
	Healthy People	**ASD Patients**
Parietal_Sup_L	0.71	0.50
Parietal_Sup_R	0.70	0.69
Parietal_Inf_L	2.64	1.21
Parietal_Inf_R	0.55	0.22
SupraMarginal_L	0.38	0.27
SupraMarginal_R	0.41	0.29
Angular_L	1.29	0.31
Angular_R	0.58	0.43
Precuneus_L	1.10	0.45
Precuneus_R	2.44	1.08
Paracentral_Lobule_L	0.91	0.30
Paracentral_Lobule_R	0.32	0.60
Caudate_L	0.40	0.37
Caudate_R	2.02	1.23
Putamen_L	1.02	0.75
Putamen_R	0.49	0.57
Pallidum_L	0.86	0.57
Pallidum_R	2.85	2.20
Thalamus_L	3.47	2.89
Thalamus_R	4.69	4.19
Heschl_L	0.43	0.61
Heschl_R	0.10	0.12
Temporal_Sup_L	1.06	1.10
Temporal_Sup_R	0.80	0.92
Temporal_Pole_Sup_L	0.73	0.61
Temporal_Pole_Sup_R	1.04	1.96
Temporal_Mid_L	1.26	1.21
Temporal_Mid_R	1.41	1.90
Temporal_Pole_Mid_L	0.33	0.66
Temporal_Pole_Mid_R	0.96	1.34

Continued on next page...

Brain Region	**Region-Based OSA**	
	Healthy People	**ASD Patients**
Temporal_Inf_L	0.50	1.02
Temporal_Inf_R	0.88	1.06
Cerebelum_Crus1_L	1.47	2.02
Cerebelum_Crus1_R	3.72	4.49
Cerebelum_Crus2_L	0.71	1.14
Cerebelum_Crus2_R	2.23	1.92
Cerebelum_3_L	0.04	0.01
Cerebelum_3_R	0.13	0.27
Cerebelum_4_5_L	0.27	0.54
Cerebelum_4_5_R	1.10	0.81
Cerebelum_6_L	0.58	0.87
Cerebelum_6_R	3.06	3.04
Cerebelum_7b_L	0.41	0.71
Cerebelum_7b_R	0.37	0.20
Cerebelum_8_L	0.96	1.15
Cerebelum_8_R	2.46	3.17
Cerebelum_9_L	1.94	0.93
Cerebelum_9_R	1.80	1.66
Cerebelum_10_L	0.38	0.17
Cerebelum_10_R	2.17	2.79
Vermis_1_2	0.02	0.01
Vermis_3	0.06	0.05
Vermis_4_5	0.10	0.20
Vermis_6	0.05	0.04
Vermis_7	0.15	0.16
Vermis_8	0.24	0.21
Vermis_9	0.23	0.51
Vermis_10	0.79	0.36

References

Acemoglu, Daron and Pascual Restrepo (2018). "Artificial intelligence, automation, and work". In: *The economics of artificial intelligence: An agenda*. University of Chicago Press, pp. 197–236.

— (2022a). "Demographics and automation". In: *The Review of Economic Studies* 89.1, pp. 1–44.

— (2022b). "Tasks, automation, and the rise in US wage inequality". In: *Econometrica* 90.5, pp. 1973–2016.

Ahmed, Shakkeel, Ravi S Mula, and Soma S Dhavala (2020). "A framework for democratizing AI". In: *arXiv preprint arXiv:2001.00818*.

Alp, Sait et al. (2024). "Joint transformer architecture in brain 3D MRI classification: its application in Alzheimer's disease classification". In: *Scientific Reports* 14.1, p. 8996.

Amoroso, Nicola et al. (2019). "Deep learning and multiplex networks for accurate modeling of brain age". In: *Frontiers in aging neuroscience* 11, p. 115.

Aviram, Hadar, Allyson Bragg, and Chelsea Lewis (2017). "Felon disenfranchisement". In: *Annual Review of Law and Social Science* 13.1, pp. 295–311.

Ballester, Pedro L et al. (2021). "Predicting brain age at slice level: convolutional neural networks and consequences for interpretability". In: *Frontiers in psychiatry* 12, p. 598518.

Bashyam, Vishnu M et al. (2020). "MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide". In: *Brain* 143.7, pp. 2312–2324.

Beheshti, Iman, MA Ganaie, et al. (2021). "Predicting brain age using machine learning algorithms: A comprehensive evaluation". In: *IEEE Journal of Biomedical and Health Informatics* 26.4, pp. 1432–1440.

Beheshti, Iman, Norihide Maikusa, and Hiroshi Matsuda (2018). "The association between "brain-age score"(BAS) and traditional neuropsychological screening tools in Alzheimer's disease". In: *Brain and Behavior* 8.8, e01020.

Beheshti, Iman, Scott Nugent, et al. (2019). "Bias-adjustment in neuroimaging-based brain age frameworks: A robust scheme". In: *NeuroImage: Clinical* 24, p. 102063.

Beheshti, Iman, Daichi Sone, et al. (2018). "Gray matter and white matter abnormalities in temporal lobe epilepsy patients with and without hippocampal sclerosis". In: *Frontiers in Neurology* 9, p. 107.

Bellantuono, Loredana et al. (2021). "Predicting brain age with complex networks: From adolescence to adulthood". In: *NeuroImage* 225, p. 117458.

Bhalla, Usha, Suraj Srinivas, and Himabindu Lakkaraju (2024). "Discriminative feature attributions: bridging post hoc explainability and inherent interpretability". In: *Advances in Neural Information Processing Systems* 36.

Binder, Lester I et al. (2005). "Tau, tangles, and Alzheimer's disease". In: *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 1739.2-3, pp. 216–223.

Böhle, Moritz, Mario Fritz, and Bernt Schiele (2022). "B-cos networks: Alignment is all we need for interpretability". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10329–10338.

Breiman, Leo (1996a). "Bagging predictors". In: *Machine learning* 24, pp. 123–140.

— (1996b). "Stacked regressions". In: *Machine learning* 24, pp. 49–64.

Buolamwini, Joy and Timnit Gebru (2018). "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Conference on fairness, accountability and transparency*. PMLR, pp. 77–91.

Bzdok, Danilo, Gael Varoquaux, and Ewout W Steyerberg (2021). "Prediction, not association, paves the road to precision medicine". In: *JAMA psychiatry* 78.2, pp. 127–128.

Cai, Hongjie, Yue Gao, and Manhua Liu (2022). "Graph transformer geometric learning of brain networks using multimodal MR images for brain age estimation". In: *IEEE Transactions on Medical Imaging* 42.2, pp. 456–466.

Caliskan, Aylin, Joanna J Bryson, and Arvind Narayanan (2017). "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334, pp. 183–186.

Callaghan, Martina F et al. (2014). "Widespread age-related differences in the human brain microstructure revealed by quantitative magnetic resonance imaging". In: *Neurobiology of aging* 35.8, pp. 1862–1872.

Caro, Josue Ortega et al. (2023). "BrainLM: A foundation model for brain activity recordings". In: *bioRxiv*, pp. 2023–09.

Cera, Nicoletta et al. (2019). "Altered cingulate cortex functional connectivity in normal aging and mild cognitive impairment". In: *Frontiers in Neuroscience* 13, p. 857.

Chandran, Varun Arunachalam et al. (2021). "Brain structural correlates of autistic traits across the diagnostic divide: A grey matter and white matter microstructure study". In: *NeuroImage: Clinical* 32, p. 102897.

Chen, Chaofan et al. (2019). "This looks like that: deep learning for interpretable image recognition". In: *Advances in neural information processing systems* 32.

Chen, Jianbo et al. (2018). "Learning to explain: An information-theoretic perspective on model interpretation". In: *International conference on machine learning*. PMLR, pp. 883–892.

Chen, Lin and Xin Luo (2023). "Tensor distribution regression based on the 3D conventional neural networks". In: *IEEE/CAA Journal of Automatica Sinica* 10.7, pp. 1628–1630.

Cheng, Jian et al. (2021). "Brain age estimation from MRI using cascade networks with ranking loss". In: *IEEE Transactions on Medical Imaging* 40.12, pp. 3400–3412.

Chiang, Chia-Yen et al. (2020). "Deep learning-based automated forest health diagnosis from aerial images". In: *IEEE Access* 8, pp. 144064–144076.

Cho, Wendy K Tam and Bruce E Cain (2020). "Human-centered redistricting automation in the age of AI". In: *Science* 369.6508, pp. 1179–1181.

Chouldechova, Alexandra (2017). "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: *Big data* 5.2, pp. 153–163.

Chung, Yoonho et al. (2018). "Use of machine learning to determine deviance in neuroanatomical maturity associated with future psychosis in youths at clinically high risk". In: *JAMA psychiatry* 75.9, pp. 960–968.

Cole, James H and Katja Franke (2017). "Predicting age using neuroimaging: innovative brain ageing biomarkers". In: *Trends in neurosciences* 40.12, pp. 681–690.

Cole, James H, Stuart J Ritchie, et al. (2018). "Brain age predicts mortality". In: *Molecular psychiatry* 23.5, pp. 1385–1392.

Conitzer, Vincent et al. (2017). "Moral decision making frameworks for artificial intelligence". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1.

Courchesne, Eric et al. (2000). "Normal brain development and aging: quantitative analysis at in vivo MR imaging in healthy volunteers". In: *Radiology* 216.3, pp. 672–682.

Couvy-Duchesne, Baptiste, Johann Faouzi, et al. (2020). "Ensemble learning of convolutional neural network, support vector machine, and best linear unbiased predictor for brain age prediction: Aramis contribution to the predictive analytics competition 2019 challenge". In: *Frontiers in Psychiatry* 11, p. 593336.

Couvy-Duchesne, Baptiste, Lachlan T Strike, et al. (2020). "A unified framework for association and prediction from vertex-wise grey-matter structure". In: *Human brain mapping* 41.14, pp. 4062–4076.

Cox, Joseph et al. (2024). "BrainSegFounder: Towards 3D foundation models for neuroimage segmentation". In: *Medical Image Analysis* 97, p. 103301.

D'Mello, Anila M and Catherine J Stoodley (2015). "Cerebro-cerebellar circuits in autism spectrum disorder". In: *Frontiers in neuroscience* 9, p. 408.

Da Costa, Pedro F, Jessica Dafflon, and Walter HL Pinaya (2020). "Brain-age prediction using shallow machine learning: predictive analytics competition 2019". In: *Frontiers in psychiatry* 11, p. 604478.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Diakopoulos, Nicholas (2019). *Automating the news: How algorithms are rewriting the media*. Harvard University Press.

Dibaji, Mahsa et al. (2023). "Studying the effects of sex-related differences on brain age prediction using brain mr imaging". In: *Workshop on Clinical Image-Based Procedures*. Springer, pp. 205–214.

Dietterich, Thomas G (2000). "Ensemble methods in machine learning". In: *International workshop on multiple classifier systems*. Springer, pp. 1–15.

Dinsdale, Nicola K et al. (2021). "Learning patterns of the ageing brain in MRI using deep convolutional networks". In: *NeuroImage* 224, p. 117401.

Dosovitskiy, Alexey et al. (2020). "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929*.

Duan, Haojing et al. (2024). "Population clustering of structural brain aging and its association with brain development". In: *Elife* 13, RP94970.

Dular, Lara, Žiga Špiclin, and Alzheimer's Disease Neuroimaging Initiative (2021). "Improving across dataset brain age predictions using transfer learning". In: *Predictive Intelligence in Medicine: 4th International Workshop, PRIME 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 4*. Springer, pp. 243–254.

Ecker, Christine et al. (2013). "Brain surface anatomy in adults with autism: the relationship between surface area, cortical thickness, and autistic symptoms". In: *JAMA psychiatry* 70.1, pp. 59–70.

Eubanks, Virginia (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Fama, Rosemary and Edith V Sullivan (2015). "Thalamic structures and associated cognitive functions: Relations with age and aging". In: *Neuroscience & Biobehavioral Reviews* 54, pp. 29–37.

Farokhian, Farnaz, Iman Beheshti, et al. (2017). "Comparing CAT12 and VBM8 for detecting brain morphological abnormalities in temporal lobe epilepsy". In: *Frontiers in neurology* 8, p. 428.

Farokhian, Farnaz, Chunlan Yang, et al. (2017). "Age-related gray and white matter changes in normal adult brains". In: *Aging and disease* 8.6, p. 899.

Feng, Xinyang et al. (2020). "Estimating brain age based on a uniform healthy population with deep learning and structural magnetic resonance imaging". In: *Neurobiology of aging* 91, pp. 15–25.

Fisch, Lukas et al. (2021). "Predicting brain-age from raw T 1-weighted Magnetic Resonance Imaging data using 3D Convolutional Neural Networks". In: *arXiv preprint arXiv:2103.11695*.

Fischl, Bruce (2012). "FreeSurfer". In: *Neuroimage* 62.2, pp. 774–781.

Fjell, Anders M, Kristine B Walhovd, et al. (2009). "One-year brain atrophy evident in healthy aging". In: *Journal of Neuroscience* 29.48, pp. 15223–15231.

Fjell, Anders M, Lars T Westlye, et al. (2009). "High consistency of regional cortical thinning in aging across multiple samples". In: *Cerebral cortex* 19.9, pp. 2001–2012.

Fraley, Chris et al. (2007). *EnsembleBMA: An R package for probabilistic forecasting using ensembles and Bayesian model averaging*. Tech. rep. Citeseer.

Franke, Katja, Christian Gaser, et al. (2013). "Advanced BrainAGE in older adults with type 2 diabetes mellitus". In: *Frontiers in aging neuroscience* 5, p. 90.

Franke, Katja, Gabriel Ziegler, et al. (2010). "Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters". In: *Neuroimage* 50.3, pp. 883–892.

Freund, Yoav and Robert E Schapire (1997). "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Journal of computer and system sciences* 55.1, pp. 119–139.

Friedman, Nir, Dan Geiger, and Moises Goldszmidt (1997). "Bayesian network classifiers". In: *Machine learning* 29, pp. 131–163.

Frisoni, Giovanni B et al. (2010). "The clinical use of structural MRI in Alzheimer disease". In: *Nature reviews neurology* 6.2, pp. 67–77.

Fu, Zening et al. (2019). "Transient increased thalamic-sensory connectivity and decreased whole-brain dynamism in autism". In: *Neuroimage* 190, pp. 191–204.

Gallego, Aina and Thomas Kurer (2022). "Automation, digitalization, and artificial intelligence in the workplace: implications for political behavior". In: *Annual Review of Political Science* 25.1, pp. 463–484.

Ganaie, MA, M Tanveer, and Iman Beheshti (2024). "Brain age prediction using improved twin SVR". In: *Neural Computing and Applications* 36.1, pp. 53–63.

Ganaie, MA, Muhammad Tanveer, and Iman Beheshti (2022). "Brain age prediction with improved least squares twin SVR". In: *IEEE Journal of Biomedical and Health Informatics* 27.4, pp. 1661–1669.

Gao, Yang et al. (2024). "Using Stacking Machine Learning Models to Predict High-Performance Concrete Compressive Strength". In: *Proceedings of the 2024 4th International Conference on Artificial Intelligence, Big Data and Algorithms*, pp. 75–80.

Garvey, Colin (2018). "A framework for evaluating barriers to the democratization of artificial intelligence". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1.

Gaser, Christian et al. (2013). "BrainAGE in mild cognitive impaired patients: predicting the conversion to Alzheimer's disease". In: *PloS one* 8.6, e67346.

Gomes, Carla et al. (2019). "Computational sustainability: Computing for a better world and a sustainable future". In: *Communications of the ACM* 62.9, pp. 56–65.

Good, Catriona D et al. (2001). "A voxel-based morphometric study of ageing in 465 normal adult human brains". In: *Neuroimage* 14.1, pp. 21–36.

Greicius, Michael D et al. (2004). "Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI". In: *Proceedings of the National Academy of Sciences* 101.13, pp. 4637–4642.

Grodstein, Francine (2012). *How early can cognitive decline be detected?*

Groves, Adrian R et al. (2012). "Benefits of multi-modal fusion analysis on a large-scale dataset: life-span patterns of inter-subject variability in cortical morphometry and white matter microstructure". In: *Neuroimage* 63.1, pp. 365–380.

Guo, JY et al. (2016). "Brain structural changes in women and men during midlife". In: *Neuroscience letters* 615, pp. 107–112.

Gutman, Boris A, Sarah K Madsen, et al. (2013). "A family of fast spherical registration algorithms for cortical shapes". In: *Multimodal Brain Image Analysis: Third International Workshop, MBIA 2013, Held in Conjunction with MICCAI 2013, Nagoya, Japan, September 22, 2013, Proceedings 3*. Springer, pp. 246–257.

Gutman, Boris A, Yalin Wang, et al. (2012). "Shape matching with medial curves and 1-D group-wise registration". In: *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 716–719.

Haar, Shlomi et al. (2016). "Anatomical abnormalities in autism?" In: *Cerebral cortex* 26.4, pp. 1440–1452.

Habata, Kaie et al. (2021). "Relationship between sensory characteristics and cortical thickness/volume in autism spectrum disorders". In: *Translational psychiatry* 11.1, p. 616.

Al-Hammuri, Khalid et al. (2023). "Vision transformer architecture and applications in digital health: a tutorial and survey". In: *Visual Computing for Industry, Biomedicine, and Art* 6.1, p. 14.

Han, Laura KM et al. (2021). "Brain aging in major depressive disorder: results from the ENIGMA major depressive disorder working group". In: *Molecular psychiatry* 26.9, pp. 5124–5139.

Hastie, Trevor J (2017). "Generalized additive models". In: *Statistical models in S*. Routledge, pp. 249–307.

He, Sheng, Yanfang Feng, et al. (2022). "Deep relation learning for regression and its application to brain age estimation". In: *IEEE transactions on medical imaging* 41.9, pp. 2304–2317.

He, Sheng, P Ellen Grant, and Yangming Ou (2021). "Global-local transformer for brain age estimation". In: *IEEE transactions on medical imaging* 41.1, pp. 213–224.

Hedman, Anna M et al. (2012). "Human brain changes across the life span: a review of 56 longitudinal magnetic resonance imaging studies". In: *Human brain mapping* 33.8, pp. 1987–2002.

Ho Park, Seong et al. (2001). "Magnetic resonance reflects the pathological evolution of Wernicke encephalopathy". In: *Journal of Neuroimaging* 11.4, pp. 406–411.

Hodges, Andrew (2006). "B. JACK COPELAND (ed.), The Essential Turing: The Ideas that Gave Birth to the Computer Age. Oxford: Clarendon Press, 2004. Pp. viii+ 613. ISBN 0-19-825079-7.£ 50.00 (hardback). ISBN 0-19-825080-0.£ 14.99 (paperback)." In: *The British Journal for the History of Science* 39.3, pp. 470–471.

Hofmann, Simon M et al. (2022). "Towards the interpretability of deep learning models for multi-modal neuroimaging: Finding structural changes of the ageing brain". In: *NeuroImage* 261, p. 119504.

Hong, Jin et al. (2020). "Brain age prediction of children using routine brain MR images via deep learning". In: *Frontiers in Neurology* 11, p. 584682.

Hu, Lianting et al. (2023). "MRI-based brain age prediction model for children under 3 years old using deep residual network". In: *Brain Structure and Function* 228.7, pp. 1771–1784.

Hu, Yixiao, Haolin Wang, and Baobin Li (2022). "SQET: Squeeze and excitation transformer for high-accuracy brain age estimation". In: *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, pp. 1554–1557.

Huang, Tzu-Wei et al. (2017). "Age estimation from brain MRI images using deep learning". In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, pp. 849–852.

Humbert, Ianessa A et al. (2010). "Early deficits in cortical control of swallowing in Alzheimer's disease". In: *Journal of Alzheimer's Disease* 19.4, pp. 1185–1197.

Hwang, Inpyeong et al. (2021). "Prediction of brain age from routine T2-weighted spin-echo brain magnetic resonance images with a deep convolutional neural network". In: *Neurobiology of Aging* 105, pp. 78–85.

Iandola, Forrest N et al. (2016). "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and¡ 0.5 MB model size". In: *arXiv preprint arXiv:1602.07360*.

Jack Jr, Clifford R et al. (2008). "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods". In: *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* 27.4, pp. 685–691.

Jenkinson, Mark et al. (2012). "Fsl". In: *Neuroimage* 62.2, pp. 782–790.

Jiang, Huiting et al. (2020). "Predicting brain age of healthy adults based on structural MRI parcellation using convolutional neural networks". In: *Frontiers in neurology* 10, p. 1346.

Jiang, Richard, Paul Chazot, et al. (2022). "Private facial prediagnosis as an edge service for parkinson's dbs treatment valuation". In: *IEEE Journal of Biomedical and Health Informatics* 26.6, pp. 2703–2713.

Jiang, Richard and Danny Crookes (2019). "Shallow unorganized neural networks using smart neuron model for visual perception". In: *IEEE Access* 7, pp. 152701–152714.

Jiang, Richard, Danny Crookes, et al. (2022). *Recent Advances in AI-enabled Automated Medical Diagnosis*. CRC Press.

Jiang, Richard, Anthony TS Ho, et al. (2017). "Emotion recognition from scrambled facial images via many graph embedding". In: *Pattern Recognition* 67, pp. 245–251.

Jiang, Ziping, Paul L Chazot, et al. (2019). "Social behavioral phenotyping of Drosophila with a 2D–3D hybrid CNN framework". In: *IEEE Access* 7, pp. 67972–67982.

Jiang, Ziping, Yunpeng Wang, et al. (2022). "Delve Into Neural Activations: Toward Understanding Dying Neurons". In: *IEEE Transactions on Artificial Intelligence* 4.4, pp. 959–971.

Jónsson, Benedikt Atli et al. (2019). "Brain age prediction using deep learning uncovers associated sequence variants". In: *Nature communications* 10.1, p. 5409.

Jumper, John et al. (2021). "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873, pp. 583–589.

Karaman, Onur et al. (2021). "Robust automated Parkinson disease detection based on voice signals with transfer learning". In: *Expert Systems with Applications* 178, p. 115013.

Khan, Salman et al. (2022). "Transformers in vision: A survey". In: *ACM computing surveys (CSUR)* 54.10s, pp. 1–41.

Kim, Jonghun, Mansu Kim, and Hyunjin Park (2024). "Domain Aware Multi-Task Pre-Training of 3D Swin Transformer for Brain MRI". In: *Proceedings of the Asian Conference on Computer Vision*, pp. 2124–2144.

Kolbeinsson, Arinbjörn et al. (2020). "Accelerated MRI-predicted brain ageing and its associations with cardiometabolic and brain disorders". In: *Scientific Reports* 10.1, p. 19940.

Koster, Raphael et al. (2022). "Human-centred mechanism design with Democratic AI". In: *Nature Human Behaviour* 6.10, pp. 1398–1407.

Kramer, Arthur F, Kirk I Erickson, and Stanley J Colcombe (2006). "Exercise, cognition, and the aging brain". In: *Journal of applied physiology* 101.4, pp. 1237–1242.

Krawczyk, Bartosz (2016). "Learning from imbalanced data: open challenges and future directions". In: *Progress in artificial intelligence* 5.4, pp. 221–232.

Kuncheva, Ludmila I and Christopher J Whitaker (2003). "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy". In: *Machine learning* 51, pp. 181–207.

Kuo, Chen-Yuan et al. (2021). "Improving individual brain age prediction using an ensemble deep learning framework". In: *Frontiers in Psychiatry* 12, p. 626677.

Lam, Pradeep K et al. (2020). "Accurate brain age prediction using recurrent slice-based networks". In: *16th international symposium on medical information processing and analysis*. Vol. 11583. SPIE, pp. 11–20.

Larson, Eric B et al. (2006). "Exercise is associated with reduced risk for incident dementia among persons 65 years of age and older". In: *Annals of internal medicine* 144.2, pp. 73–81.

Lee, Jeyeon et al. (2022). "Deep learning-based brain age prediction in normal aging and dementia". In: *Nature aging* 2.5, pp. 412–424.

Lee, Min Kyung et al. (2019). "WeBuildAI: Participatory framework for algorithmic governance". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW, pp. 1–35.

Levakov, Gidon et al. (2020). "From a deep learning model back to the brain—Identifying regional predictors and their relation to aging". In: *Human brain mapping* 41.12, pp. 3235–3252.

Li, Jing, Linda Chiu Wa Lam, and Hanna Lu (2024). "Decoding MRI-informed brain age using mutual information". In: *Insights into Imaging* 15.1, p. 216.

Li, Xinlin et al. (2024). "Brain age prediction via cross-stratified ensemble learning". In: *NeuroImage* 299, p. 120825.

Lin, Lan et al. (2021). "Utilizing transfer learning of pre-trained AlexNet and relevance vector machine for regression for predicting healthy older adult's brain age from structural MRI". In: *Multimedia Tools and Applications* 80.16, pp. 24719–24735.

Lombardi, Angela et al. (2021). "Explainable deep learning for personalized age prediction with brain morphology". In: *Frontiers in neuroscience* 15, p. 674055.

Luders, Eileen, Nicolas Cherbuin, and Christian Gaser (2016). "Estimating brain age using high-resolution pattern recognition: Younger brains in long-term meditation practitioners". In: *Neuroimage* 134, pp. 508–513.

Lundberg, Scott (2017). "A unified approach to interpreting model predictions". In: *arXiv preprint arXiv:1705.07874*.

Maiti, Baijayanta et al. (2021). "Functional connectivity of vermis correlates with future gait impairments in Parkinson's disease". In: *Movement Disorders* 36.11, pp. 2559–2568.

Matthews, Michael, Samuel Matthews, and Thomas Kelemen (2022). "The Alignment Problem: Machine Learning And Human Values." In: *Personnel Psychology* 75.1.

Mayson, Sandra G (2018). "Bias in, bias out". In: *YAle lJ* 128, p. 2218.

McMahan, Brendan et al. (2017). "Communication-efficient learning of deep networks from decentralized data". In: *Artificial intelligence and statistics*. PMLR, pp. 1273–1282.

Mehboob, Fozia et al. (2022). "Towards robust diagnosis of COVID-19 using vision self-attention transformer". In: *Scientific Reports* 12.1, p. 8922.

Mehrabi, Ninareh et al. (2021). "A survey on bias and fairness in machine learning". In: *ACM computing surveys (CSUR)* 54.6, pp. 1–35.

Monteith, Kristine et al. (2011). "Turning Bayesian model averaging into Bayesian model combination". In: *The 2011 international joint conference on neural networks*. IEEE, pp. 2657–2663.

Montes, Gabriel Axel and Ben Goertzel (2019). "Distributed, decentralized, and democratized artificial intelligence". In: *Technological Forecasting and Social Change* 141, pp. 354–358.

Mouches, Pauline, Matthias Wilms, Deepthi Rajashekar, Sonke Langner, et al. (2021). "Unifying brain age prediction and age-conditioned template generation with a deterministic autoencoder". In: *Medical Imaging with Deep Learning*. PMLR, pp. 497–506.

Mouches, Pauline, Matthias Wilms, Deepthi Rajashekar, Sönke Langner, et al. (2022). "Multimodal biological brain age prediction using magnetic resonance imaging and angiography with the identification of predictive regions". In: *Human brain mapping* 43.8, pp. 2554–2566.

Nguyen, Minh NH et al. (2022). "Self-organizing democratized learning: Toward large-scale distributed learning systems". In: *IEEE Transactions on Neural Networks and Learning Systems*.

Ning, Kaida, Ben A Duffy, et al. (2021). "Improving brain age estimates with deep learning leads to identification of novel genetic factors associated with brain aging". In: *Neurobiology of Aging* 105, pp. 199–204.

Ning, Kaida, Lu Zhao, et al. (2020). "Association of relative brain age with tobacco smoking, alcohol consumption, and genetic variants". In: *Scientific reports* 10.1, p. 10.

OpenAI (2025). *ChatGPT*. `https://chat.openai.com`. Large language model used for generating illustrative figures.

Pantelaios, Dimitrios et al. (2024). "Hybrid CNN-ViT models for medical image classification". In: *2024 IEEE international symposium on biomedical imaging (ISBI)*. IEEE, pp. 1–4.

Pardakhti, Nastaran and Hedieh Sajedi (2020). "Brain age estimation based on 3D MRI images using 3D convolutional neural network". In: *Multimedia tools and applications* 79.33, pp. 25051–25065.

Peng, Han et al. (2021). "Accurate brain age prediction with lightweight deep neural networks". In: *Medical image analysis* 68, p. 101871.

Penny, William D et al. (2011). *Statistical parametric mapping: the analysis of functional brain images*. Elsevier.

Phillips, Anne (2021). "Unconditional equals". In.

Pierpaoli, Carlo et al. (1996). "Diffusion tensor MR imaging of the human brain." In: *Radiology* 201.3, pp. 637–648.

Poloni, Katia Maria, Ricardo José Ferrari, Alzheimer's Disease Neuroimaging Initiative, et al. (2022). "A deep ensemble hippocampal CNN model for brain age estimation applied to Alzheimer's diagnosis". In: *Expert Systems with Applications* 195, p. 116622.

Popescu, Sebastian G et al. (2021). "Local brain-age: a U-net model". In: *Frontiers in Aging Neuroscience* 13, p. 761954.

Posner, Tess, Li Fei-Fei, et al. (2020). "AI will change the world, so it's time to change AI". In: *Nature* 588.7837, S118–S118.

Rao, Anand S. (Aug. 2020). "Democratization of AI: A Double-Edged Sword". In: URL: `https://towardsdatascience.com/democratization-of-ai-de155f0616b5`.

Raz, Naftali and Karen M Rodrigue (2006). "Differential aging of the brain: patterns, cognitive correlates and modifiers". In: *Neuroscience & Biobehavioral Reviews* 30.6, pp. 730–748.

Reeve, Amy, Eve Simcox, and Doug Turnbull (2014). "Ageing and Parkinson's disease: why is advancing age the biggest risk factor?" In: *Ageing research reviews* 14, pp. 19–30.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016a). "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.

— (2016b). "Model-agnostic interpretability of machine learning". In: *arXiv preprint arXiv:1606.05386*.

— (2018). "Anchors: High-precision model-agnostic explanations". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1.

Roibu, AC (2023). "Deep learning approaches to multimodal MRI brain age estimation". PhD thesis. University of Oxford.

Sadigh-Eteghad, Saeed et al. (2015). "Amyloid-beta: a crucial factor in Alzheimer's disease". In: *Medical principles and practice* 24.1, pp. 1–10.

Sajedi, Hedieh and Nastaran Pardakhti (2019). "Age prediction based on brain MRI image: a survey". In: *Journal of medical systems* 43.8, p. 279.

Schnack, Hugo G et al. (2016). "Accelerated brain aging in schizophrenia: a longitudinal pattern recognition study". In: *American Journal of Psychiatry* 173.6, pp. 607–616.

Schuetze, Manuela et al. (2016). "Morphological alterations in the thalamus, striatum, and pallidum in autism spectrum disorder". In: *Neuropsychopharmacology* 41.11, pp. 2627–2637.

Selvaraju, Ramprasaath R et al. (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Settles, Burr (2009). "Active learning literature survey". In.

Shashi, Raj Pandey et al. (2022). "Edge-assisted Democratized Learning Towards Federated Analytics". In.

Shen, Lucinda (Nov. 2017). "Former U.S. CTO: The 'Robot Apocalypse' Could Happen. Here's How You Stop It". In: URL: `https://fortune.com/2017/11/14/megan-smith-cto-robot-apocalypse-elon-musk/`.

Shi, Wen et al. (2020). "Fetal brain age estimation and anomaly detection using attention-based deep ensembles with uncertainty". In: *Neuroimage* 223, p. 117316.

Sidey-Gibbons, Jenni AM and Chris J Sidey-Gibbons (2019). "Machine learning in medicine: a practical introduction". In: *BMC medical research methodology* 19, pp. 1–18.

Siegel, Nys Tjade et al. (2025). "Do transformers and CNNs learn different concepts of brain age?" In: *Human Brain Mapping* 46.8, e70243.

Singla, Ayush et al. (2022). "Multiple instance neuroimage transformer". In: *International Workshop on PRedictive Intelligence In MEdicine*. Springer, pp. 36–48.

Smith, Stephen M (2002). "Fast robust automated brain extraction". In: *Human brain mapping* 17.3, pp. 143–155.

Soch, Joram (2020). "Distributional transformation improves decoding accuracy when predicting chronological age from structural MRI". In: *Frontiers in Psychiatry* 11, p. 604268.

Sollich, Peter and Anders Krogh (1995). "Learning with ensembles: How overfitting can be useful". In: *Advances in neural information processing systems* 8.

Srinivas, Suraj and François Fleuret (2019). "Full-gradient representation for neural network visualization". In: *Advances in neural information processing systems* 32.

Steffener, Jason et al. (2016). "Differences between chronological and brain age are related to education and self-reported physical activity". In: *Neurobiology of aging* 40, pp. 138–144.

Storsve, Andreas B et al. (2014). "Differential longitudinal changes in cortical thickness, surface area and volume across the adult life span: regions of accelerating and decelerating change". In: *Journal of Neuroscience* 34.25, pp. 8488–8498.

Strouse, DJ et al. (2021). "Collaborating with humans without human data". In: *Advances in Neural Information Processing Systems* 34, pp. 14502–14515.

Sun, Huili et al. (2023). "Network controllability of structural connectomes in the neonatal brain". In: *Nature communications* 14.1, p. 5820.

Sutoko, Stephanie et al. (2020). "Lesions in the right Rolandic operculum are associated with self-rating affective and apathetic depressive symptoms for post-stroke patients". In: *Scientific reports* 10.1, p. 20264.

Taddeo, Mariarosaria and Luciano Floridi (2018). "How AI can be a force for good". In: *Science* 361.6404, pp. 751–752.

Tak, Divyanshu et al. (2024). "A foundation model for generalized brain MRI analysis". In: *medRxiv*.

Tanveer, M et al. (2023). "Deep learning for brain age estimation: A systematic review". In: *Information Fusion*.

Tohid, Hassaan, Muhammad Faizan, and Uzma Faizan (2015). "Alterations of the occipital lobe in schizophrenia". In: *Neurosciences Journal* 20.3, pp. 213–224.

Tomašev, Nenad et al. (2019). "A clinically applicable approach to continuous prediction of future acute kidney injury". In: *Nature* 572.7767, pp. 116–119.

Van Den Bergh, Frans (2001). *An analysis of particle swarm optimizers*. University of Pretoria (South Africa).

Van Rooij, Daan et al. (2018). "Cortical and subcortical brain morphometry differences between patients with autism spectrum disorder and healthy individuals across the lifespan: results from the ENIGMA ASD Working Group". In: *American Journal of Psychiatry* 175.4, pp. 359–369.

Varatharajah, Yogatheesan et al. (2018). "Predicting brain age using structural neuroimaging and deep learning". In: *BioRxiv*, p. 497925.

Voelbel, Gerald T et al. (2006). "Caudate nucleus volume and cognitive performance: Are they related in childhood psychopathology?" In: *Biological psychiatry* 60.9, pp. 942–950.

Wang, Johnny et al. (2019). "Gray matter age prediction as a biomarker for risk of dementia". In: *Proceedings of the National Academy of Sciences* 116.42, pp. 21213–21218.

Wilkinson, Alissa (2023). "The looming threat of AI to Hollywood, and why it should matter to you". In: *Vox*.

Wiltshire, Katie et al. (2010). "Corpus callosum and cingulum tractography in Parkinson's disease". In: *Canadian Journal of Neurological Sciences* 37.5, pp. 595–600.

Wood, David A, Sina Kafiabadi, et al. (2022). "Accurate brain-age models for routine clinical MRI examinations". In: *Neuroimage* 249, p. 118871.

Wood, David A, Matthew Townend, et al. (2024). "Optimising brain age estimation through transfer learning: A suite of pre-trained foundation models for improved performance and generalisability in a clinical setting". In: *Human brain mapping* 45.4, e26625.

Xuereb, JH et al. (1990). "Parameters of cholinergic neurotransmission in the thalamus in Parkinson's disease and Alzheimer's disease". In: *Journal of the neurological sciences* 99.2-3, pp. 185–197.

Yang, Crystal C et al. (2022). "Normative sex differences in cognition and morphometric brain connectivity: Evidence from 30,000+ UK Biobank participants". In: *bioRxiv*, pp. 2022–10.

Yao, Zhaomin et al. (2023). "Fuzzy-VGG: A fast deep learning method for predicting the staging of Alzheimer's disease based on brain MRI". In: *Information Sciences* 642, p. 119129.

Zeiler, Matthew D and Rob Fergus (2014). "Visualizing and understanding convolutional networks". In: *European conference on computer vision*. Springer, pp. 818–833.

Zeineldin, Ramy A et al. (2024). "Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI". In: *Scientific reports* 14.1, p. 3713.

Zhang, Y et al. (2007). "Diffusion tensor imaging of cingulum fibers in mild cognitive impairment and Alzheimer disease". In: *Neurology* 68.1, pp. 13–19.

Zhang, Zhaonian, Vaneet Aggarwal, et al. (2025). "Modeling Brain Aging with Explainable Triamese ViT: Towards Deeper Insights into Autism Disorder". In: *IEEE Journal of Biomedical and Health Informatics*, pp. 1–14. DOI: 10.1109/JBHI.2025.3574366.

Zhang, Zhaonian and Richard Jiang (2023). "User-Centric Democratization towards Social Value Aligned Medical AI Services." In: *IJCAI*, pp. 6326–6334.

Zhang, Zhaonian, Richard Jiang, et al. (2022). "Robust brain age estimation based on sMRI via nonlinear age-adaptive ensemble learning". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 30, pp. 2146–2156.

Zhao, Haiyan, Hongjie Cai, and Manhua Liu (2024). "Transformer based multi-modal MRI fusion for prediction of post-menstrual age and neonatal brain development analysis". In: *Medical Image Analysis* 94, p. 103140.

Zou, James and Londa Schiebinger (2018). *AI can be sexist and racist—it's time to make it fair*.