



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
UNIVERSITY OF PIRAEUS



DEMOKRITOS

Explainable Brain Age Prediction Using Machine Learning: Model Comparison, Longitudinal Evaluation, and Normative Modelling

Nikolas Cheliotis

Submitted

in partial fulfilment of the requirements for the degree of
Master of Artificial Intelligence
at the
UNIVERSITY OF PIRAEUS

Supervisors: Ilias Maglogiannis , Paschalina Lialiou

Dedicated to my family.

And to the memory of Kokos.



ABSTRACT

Brain age prediction is a neuroimaging framework for estimating biological brain aging and identifying deviations from expected aging trajectories. However, the value of brain age models depends not only on predictive accuracy, but also on interpretability, longitudinal reliability, and clinical relevance. This thesis investigates explainable machine learning approaches for brain age prediction using FreeSurfer-derived structural MRI features from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset.

Five regression models were developed and compared: Random Forest, Support Vector Regression, Elastic Net, Ridge Regression, and Gaussian Process Regression. Model performance was evaluated using Mean Absolute Error, Pearson correlation, and Brain Age Gap dispersion. In addition, the study examined hyperparameter sensitivity, explainability using SHAP values and complementary feature-importance methods, short-term and long-term longitudinal stability, and normative modelling in cognitively normal, Mild Cognitive Impairment, and Dementia subjects.

The evaluated models achieved comparable predictive performance, with test-set errors below five years. Random Forest achieved the lowest prediction error, Elastic Net produced the strongest correlation with chronological age, and SVR showed the lowest Brain Age Gap dispersion. Explainability analyses identified biologically plausible predictors, including white matter hypointensities, ventricular measures, hippocampal and thalamic structures, and temporal and parietal cortical regions. SHAP feature rankings were highly stable across random seeds. Longitudinal analyses showed strong short-term reliability, although individual long-term aging trajectories were more difficult to capture. Normative modelling revealed progressively elevated Brain Age Gap values in Mild Cognitive Impairment and Dementia subjects, with group differences remaining significant after adjustment for age and sex.

The findings of this thesis support the importance of evaluating brain age models beyond cross-sectional accuracy alone by incorporating explainability, longitudinal stability, and normative clinical analysis.

Keywords: Brain age; Machine learning; MRI; XAI; Neuroimaging

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor, Professor Paschalina Lialiou, for her guidance, feedback, and support throughout the development of this thesis. Her advice contributed significantly to the completion and improvement of this work.

I would also like to thank Professor Ilias Maglogiannis for his contribution to my academic progress and for his role in the supervision and evaluation of this thesis.

Furthermore, I am grateful to all the professors and teaching staff of the MSc in Artificial Intelligence at NCSR Demokritos for the knowledge, guidance, and academic foundation they provided during the programme. Their teaching and support helped shape my understanding of artificial intelligence and its applications.

Finally, I would like to thank my family for their continuous support, patience, and encouragement throughout my studies.

LIST OF FIGURES

1. Participant and dataset flow diagram for the brain age prediction analyses - 18
2. Age distribution of the baseline-like ADNI cohort and the cognitively normal subset used for brain age model development - 19
3. Chronological age distribution across diagnostic groups in the baseline-like cohort - 20
4. Sex distribution within the cognitively normal baseline-like cohort - 20
5. Cross-validation and independent test-set MAE for the evaluated brain age prediction models - 30
6. Pearson correlation between predicted and chronological age for each brain age prediction model - 31
7. Standard deviation of the Brain Age Gap distribution for each machine learning model - 32
8. Cross-validation MAE of the Random Forest model as a function of min_samples_leaf and max_features - 34
9. Cross-validation MAE of the SVR model for different combinations of the regularization parameter C and gamma - 35
10. Effect of the epsilon parameter on SVR cross-validation performance - 36
11. Cross-validation MAE of the Elastic Net model across different values of alpha and L1 ratio - 36

12. Cross-validation MAE of the Ridge Regression model as a function of the regularization parameter alpha - 37
13. Cross-validation performance of the evaluated Gaussian Process Regression kernel configurations - 38
14. Top 20 Random Forest features ranked by mean absolute SHAP value - 39
15. SHAP summary plot showing the distribution of feature contributions across subjects - 40
16. Top 20 Random Forest features ranked by permutation importance - 41
17. Top positive and negative standardized Elastic Net coefficients - 41
18. Consensus feature ranking derived from SHAP values, permutation importance, and Elastic Net coefficients - 42
19. Pairwise correlation matrix of SHAP feature importance values across five random seeds - 43
20. SHAP waterfall plots for representative subject cases, including two subjects with highly negative BAG, two with near-zero BAG, and two with highly positive BAG - 45
21. Distribution of Brain Age Gap across diagnostic groups - 49

LIST OF TABLES

1. Contextual comparison between the present thesis and selected brain age prediction studies - 33
2. Overlap and Jaccard similarity between SHAP, permutation importance, and Elastic Net feature rankings - 43
3. Short-term test–retest stability results for all brain age prediction models - 46
4. Long-term sensitivity results across all longitudinal subsets - 47
5. Comparison of observed and predicted age progression in the clean longitudinal subset - 48
6. Demographic characteristics of the normative modelling cohort across diagnostic groups - 48
7. Brain Age Gap statistics across diagnostic groups - 49
8. Ordinary least squares regression model examining the association between diagnostic group and Brain Age Gap after adjustment for age and sex - 50

TABLE OF CONTENTS

Introduction

1.1 Motivation	1
1.2 Research Gap	2
1.3 Aim and Objectives	3
1.4 Thesis Contributions	4
1.5 Thesis Structure	5

2. Background and Related Work

2.1 Brain Age Prediction	6
2.2 MRI and FreeSurfer Features	7
2.3 Machine Learning Approaches	8
2.4 Explainability in Brain Age Models	10
2.5 Longitudinal Reliability and Test-Retest Stability	11
2.6 Brain Age Deviation Analysis in Clinical Populations	12
2.7 State-of-the-Art Approaches	13

3. Methodology

3.1 ADNI Dataset	16
3.2 Data Preprocessing	18
3.3 Brain Age Prediction Models	21
3.3.1 Support Vector Regression	22
3.3.2 Ridge Regression	22
3.3.3 Elastic Net Regression	22
3.3.4 Random Forest Regression	23
3.3.5 Gaussian Process Regression	23
3.4 Training and Hyperparameter Optimization	23
3.5 Evaluation Metrics	24
3.6 Explainability Framework	25
3.7 Longitudinal Evaluation	26
3.7.1 Short-Term Stability	27

3.7.2 Long-Term Sensitivity	27
3.8 Normative Modelling on Clinical Subjects	28
3.9 Ethical Considerations	28
4. Results	
4.1 Models Performance.....	30
4.1.1 Comparison with State-of-the-Art Brain Age Prediction Studies	32
4.2 Hyperparameter Optimization Analysis	34
4.3 Explainability Results	38
4.3.1 Global Feature Importance	38
4.3.2 SHAP Stability Analysis	43
4.3.3 Local Explanations.....	44
4.4 Longitudinal Stability Evaluation.....	46
4.4.1 Short-Term Stability	46
4.4.2 Long-Term Sensitivity	47
4.5 Normative Modelling on Clinical Subjects	48
4.6 Summary of Findings.....	50
5 Discussion	
5.1 Models Performance.....	52
5.2 Comparison with State-of-the-Art and Alternative Methodolo-	
gies.....	55
5.3 Hyperparameter Optimization Analysis	57
5.4 Explainability Findings	61
5.5 Longitudinal Stability	64
5.6 Normative Modelling on Clinical Subjects	68
6. Conclusions and Future Work	
6.1 Conclusions	72
6.2 Limitations.....	74
6.3 Future Work.....	76
References	78
APPENDIX A - TRIPOD+AI Checklist.....	82

LIST OF ABBREVIATIONS

AD - Alzheimer's Disease
ADNI - Alzheimer's Disease Neuroimaging Initiative
AI - Artificial Intelligence
BAG - Brain Age Gap
CCC - Concordance Correlation Coefficient
CN - Cognitively Normal
CV - Cross Validation
CI - Confidence Interval
DL - Deep Learning
ENet - Elastic Net
GPR - Gaussian Process Regression
MAE - Mean Absolute Error
MCI - Mild Cognitive Impairment
ML - Machine Learning
MRI - Magnetic Resonance Imaging
OLS - Ordinary Least Squares
PAD - Predicted Age Difference
RF - Random Forest
SD - Standard Deviation
SHAP - SHapley Additive exPlanations
SVR - Support Vector Regression
WM - White Matter

Introduction

This chapter introduces the context, motivation, and scope of the thesis. It first presents the relevance of brain aging and brain age prediction as a framework for studying individual differences in brain health. It then identifies the main research gap addressed in this work, namely the need to evaluate brain age models beyond prediction accuracy alone. The chapter also states the aim and objectives of the thesis, summarizes its main contributions, and outlines the structure of the remaining chapters.

1.1 Motivation

Brain aging is not a uniform process. Individuals of the same chronological age can differ substantially in brain structure, cognitive function, and vulnerability to neurological disease [1]. Structural MRI studies have shown that aging is associated with measurable neuroanatomical changes, including cortical thinning, grey and white matter volume loss, ventricular enlargement, and alterations in brain integrity [2,3]. However, these changes do not occur at the same rate in all individuals and may overlap with structural patterns observed in neurodegenerative disorders. As a result, chronological age alone provides an incomplete description of brain health.

Brain age prediction offers a way to summarize complex neuroimaging information into an individualized estimate of apparent brain aging. Instead of examining each anatomical region separately, a brain age model attempts to learn distributed age-related patterns across the brain and express them as a single predicted age value. This makes brain age attractive as a potential biomarker because it can provide a compact, interpretable measure of deviation from typical aging trajectories [4,5].

MRI-based brain age models have been developed using different types of neuroimaging representations, including morphometric features extracted from structural MRI scans and image-based approaches that use voxel-wise or deep learning representations [6,7]. The present thesis follows a morphometric feature-

based approach, using FreeSurfer-derived anatomical measurements such as cortical thickness, cortical volume, surface area, and subcortical volumes. This representation is particularly suitable for comparing different machine learning models and for examining which neuroanatomical features contribute to predictions.

The clinical motivation for this work is especially relevant in the context of neurodegenerative disease. Previous studies have shown that individuals with MCI, Alzheimer's disease, and other brain disorders may exhibit older-appearing brains compared with cognitively normal individuals [8,9]. Therefore, brain age prediction may provide a useful framework for studying abnormal brain aging, identifying deviations from healthy aging trajectories, and investigating whether cognitively impaired subjects show evidence of accelerated biological brain aging.

1.2 Research Gap

Although brain age prediction has shown considerable promise, its usefulness depends on more than prediction accuracy alone. A model with low prediction error may still be difficult to interpret, unstable across repeated scans, or limited in its ability to capture clinically meaningful differences between healthy and impaired subjects. This is particularly important if brain age is intended to function as a biomarker rather than only as a regression task.

Several methodological issues remain important in brain age research. First, brain age models are affected by prediction bias, where younger subjects may be overestimated and older subjects underestimated because of regression-to-the-mean effects [10]. Second, model generalisation may be affected by differences in datasets, scanner characteristics, acquisition protocols, and preprocessing pipelines [11]. Third, brain age estimates should be reliable across short test-retest intervals while still being sensitive enough to detect meaningful age-related changes over longer periods [12]. Finally, methodological choices such as preprocessing strategy and feature representation may influence model performance and robustness [13].

These challenges motivate the need for brain age studies that evaluate models from multiple complementary perspectives. In addition to cross-sectional accuracy, it is important to assess whether models are interpretable, whether their

predictions are stable over time, and whether their outputs can identify clinically meaningful deviations from normative aging patterns. The present thesis was designed around this broader evaluation perspective.

1.3 Aim and Objectives

The aim of this thesis is to develop and evaluate explainable machine learning models for MRI-based brain age prediction using FreeSurfer-derived structural MRI features from the ADNI dataset. The study focuses not only on predictive performance but also on model interpretability, longitudinal stability, and clinical relevance through normative analysis.

The specific objectives of the thesis are:

1. To compare multiple machine learning regression models for brain age prediction using structural MRI-derived morphometric features.
2. To evaluate model performance using cross-validation and independent test-set metrics, including Mean Absolute Error, Pearson correlation, and Brain Age Gap dispersion.
3. To investigate the effect of hyperparameter optimization on model performance and examine how sensitive each model is to configuration choices.
4. To apply explainability methods, including SHAP values, permutation importance, and Elastic Net coefficients, in order to identify the neuroanatomical features contributing to brain age prediction.
5. To evaluate the stability of SHAP-based explanations across multiple random seeds and examine local explanations for representative individual subjects.

6. To assess short-term test-retest reliability and long-term sensitivity using longitudinal MRI scans from the same subjects.

7. To apply a normative modelling framework in which a model trained on cognitively normal subjects is used to examine Brain Age Gap differences in CN, MCI, and Dementia groups.

1.4 Thesis Contributions

This thesis makes several contributions to the study of explainable brain age prediction.

First, it provides a comparative evaluation of five machine learning models: Support Vector Regression, Random Forest Regression, Elastic Net Regression, Ridge Regression, and Gaussian Process Regression. These models represent different modelling families, including kernel-based learning, ensemble learning, regularized linear regression, and probabilistic regression.

Second, the thesis evaluates brain age prediction beyond conventional accuracy metrics. In addition to MAE and correlation, the analysis considers Brain Age Gap dispersion, age-bias correction, hyperparameter sensitivity, short-term reliability, and long-term longitudinal behaviour.

Third, the study incorporates a multi-method explainability framework. SHAP values, permutation importance, Elastic Net coefficients, consensus feature ranking, SHAP stability analysis, and local subject-level explanations are used to examine whether the models rely on biologically plausible neuroanatomical features.

Fourth, the thesis investigates longitudinal reliability by evaluating whether brain age predictions remain stable over short intervals and whether they capture age-related progression over longer follow-up periods.

Finally, the thesis applies a normative modelling analysis to clinical subjects. By training the model on cognitively normal individuals and applying it to CN, MCI, and Dementia groups, the study examines whether Brain Age Gap values reflect progressively greater deviations from typical aging trajectories in cognitively impaired populations.

1.5 Thesis Structure

The remainder of the thesis is organized as follows. Chapter 2 presents the background and related work, including the brain age prediction framework, MRI and FreeSurfer-derived features, machine learning approaches, explainability methods, longitudinal reliability, clinical brain age deviation analysis, and state-of-the-art approaches. Chapter 3 describes the methodology used in the study, including the ADNI dataset, preprocessing pipeline, machine learning models, hyperparameter optimization, evaluation metrics, explainability framework, longitudinal evaluation, and normative modelling analysis. Chapter 4 presents the experimental results, including model performance, hyperparameter optimization findings, explainability results, longitudinal stability analyses, and normative modelling outcomes. Chapter 5 discusses the findings in relation to existing literature and interprets their methodological and clinical implications. Finally, Chapter 6 summarizes the main conclusions, outlines directions for future work and addresses the limitations of the study.

2. Background and Related Work

This chapter reviews the scientific and methodological background required for the present study. It first introduces the brain age prediction framework and then discusses the structural MRI features and FreeSurfer-derived morphometric measures commonly used in this field. The chapter also reviews traditional machine learning approaches, explainability methods, longitudinal reliability, clinical brain age deviation analysis, and state-of-the-art developments in brain age modelling. Together, these sections provide the theoretical foundation for the methodology presented in Chapter 3.

2.1 Brain Age Prediction

Brain age prediction is a neuroimaging based framework in which machine learning models are trained to estimate chronological age from brain imaging data acquired from healthy individuals. The central idea behind the brain-age paradigm is that deviations between predicted brain age and chronological age may reflect individual differences in brain health and aging trajectories. The difference between predicted and chronological age is commonly referred to as Brain Age Gap (BAG), brain-age delta, BrainAGE, or brain-predicted age difference (brain-PAD).

$$\text{BAG} = \text{Predicted Age} - \text{Chronological Age}$$

Early studies demonstrated that structural MRI data contain age-related neuroanatomical patterns that can be learned by statistical and machine learning models to predict age with relatively high accuracy [5]. The BrainAGE framework was firstly introduced by Franke and colleagues 2010, which used machine learning methods applied to structural MRI data to estimate brain age and investigate deviations from normative aging trajectories [14].

Recent literature has expanded the brain age paradigm beyond traditional morphometric analysis. Brain age prediction has been investigated using multi-

modal neuroimaging, voxel-wise MRI representations, and deep learning approaches operating directly on raw imaging data [6, 7].

Brain age prediction is increasingly viewed as a quantitative neuroimaging framework for studying individual variability in brain aging. However, the literature also emphasizes that prediction accuracy alone is insufficient for establishing clinical utility, motivating ongoing research into model robustness, interpretability, reliability, and generalizability. [6,12]

2.2 MRI and FreeSurfer Features

Structural MRI is widely used in brain age prediction because it provides quantitative information about brain anatomy that changes across the lifespan. Brain age studies commonly use either voxel-based representations of MRI scans or region-based morphometric features extracted from anatomical images. Baecker et al. explicitly compared machine learning models using region-based and voxel-based morphometric data for brain age prediction, showing that both types of structural MRI representations are used in the field [15].

A common feature-based approach is to extract anatomical measurements from T1-weighted MRI scans and use them as input to regression models. Han et al. compared multiple machine learning models for brain age prediction using brain morphological measures derived from structural MRI scans [16], while Cumplido-Mayoral et al. trained brain-age models using imaging-derived measures of cortical thickness, cortical volume, and subcortical volume from the UK Biobank cohort [17].

FreeSurfer is one of the most widely used tools for obtaining such structural MRI-derived features. Fischl describes FreeSurfer as a software for automated analysis of human brain MRI data, including cortical reconstruction, volumetric segmentation, cortical thickness estimation, and anatomical parcellation [18].

The use of FreeSurfer-derived features offers several advantages for brain age prediction. By summarizing MRI scans into regional morphometric measures, such as cortical thickness, cortical volume, and subcortical volume, the dimensionality of the data is substantially reduced compared to voxel-wise representations. This reduction facilitates the application of traditional machine learning algo-

rithms while also improving model interpretability. Furthermore, these measures have direct anatomical meaning, enabling model predictions and feature importance analyses to be linked to specific brain structures. The widespread adoption of FreeSurfer-derived measures in established brain-age frameworks further highlights their utility. For example, the ENIGMA brain-age model uses FreeSurfer-based cortical and subcortical measures as input features for age prediction [12].

At the same time, feature-based MRI approaches differ from image-based or voxel-wise models. Voxel-wise and deep learning methods may preserve more spatial information from the original MRI scan [7,15], but FreeSurfer-based morphometric approaches offer a more compact and interpretable representation of brain anatomy. Recent comparative work has directly contrasted morphometric FreeSurfer-based pipelines with 3D convolutional neural network approaches, emphasizing the trade-off between feature-level interpretability and image-based modeling [19]. This distinction is important for the present thesis, which follows the morphometric feature-based approach. By using FreeSurfer-derived structural MRI features, the analysis remains connected to anatomically interpretable measures while allowing comparison across multiple traditional machine learning models and later explainability analyses.

2.3 Machine Learning Approaches

Traditional machine learning approaches have been widely used in the literature due to their strong predictive performance, ability to handle correlated features, and relatively lower computational requirements compared to deep learning approaches.

Among traditional machine learning methods, Support Vector Regression (SVR) has historically been one of the most commonly used approaches in brain age prediction studies. Franke et al. employed relevance vector regression within the BrainAGE framework, while later studies frequently adopted SVR for brain age prediction [5,14,15]. SVR models can capture non-linear relationships between MRI-derived features and age through the use of kernel functions, making them particularly suitable for structural neuroimaging data. Several comparative studies

have reported competitive predictive performance of SVR-based brain age models across different MRI datasets [15,16].

Ensemble learning methods such as Random Forest Regression have also been investigated in MRI-based brain age prediction. Random Forests construct multiple decision trees and combine their predictions, allowing complex non-linear interactions between neuroanatomical features to be modelled. Comparative studies evaluating multiple regression models for brain age prediction have included Random Forest approaches among commonly used traditional machine learning techniques [16].

Regularized linear models have been widely explored in brain age prediction studies due to the highly correlated nature of MRI-derived morphometric features. Ridge Regression applies L2 regularization in order to reduce overfitting and improve stability in high-dimensional settings, while Elastic Net Regression combines both L1 and L2 regularization penalties, enabling simultaneous regularization and feature selection. These approaches are particularly relevant for neuroimaging applications where large numbers of correlated anatomical measurements are used as predictors. Studies comparing regression methods for structural MRI-based age prediction have included Ridge, Lasso, and Elastic Net models alongside kernel-based approaches such as SVR and GPR [16]

Probabilistic regression approaches such as Gaussian Process Regression (GPR) have also been applied in brain age prediction research. Unlike conventional deterministic regression approaches, GPR incorporates probabilistic modelling principles that enable flexible prediction of continuous outcomes. Cole et al. used Gaussian Processes in early neuroimaging-based brain age studies and demonstrated accurate age prediction from structural MRI data [20].

Although recent years have seen increasing adoption of deep learning and voxel-wise imaging methods, regression-based approaches using morphometric MRI features continue to offer advantages in computational efficiency, interpretability, and compatibility with anatomically meaningful feature representations.

2.4 Explainability in Brain Age Models

Many machine learning models used for brain age estimation can achieve strong predictive performance, however they provide limited insight into the neuroanatomical factors driving their predictions. This lack of transparency is particularly important since brain age estimates are often interpreted as biomarkers of brain health and disease related aging processes. Consequently, recent literature has increasingly emphasized that prediction accuracy alone is insufficient for evaluating brain age models, and that interpretability should also be considered when assessing their scientific and clinical utility [6,19].

Feature-based brain age models provide a relatively interpretable framework because their inputs correspond to anatomically meaningful measurements. In these settings, explainability approaches can be used to identify which neuroanatomical features contribute most strongly to age prediction. Traditional interpretation methods in feature-based brain age models include the examination of regression coefficients in linear models, feature importance measures in ensemble methods, and model-agnostic explainability techniques such as SHAP. These approaches allow researchers to relate prediction behavior to specific brain regions and investigate whether the identified patterns are biologically plausible [19].

More recently, explainable artificial intelligence (XAI) methods have been introduced into brain age prediction research to improve the interpretability of machine learning models. Among these approaches, SHAP has attracted particular attention because it can provide both global and local explanations of model behavior. Global explanations describe the overall contribution of features across a population, whereas local explanations focus on explaining predictions at the individual subject level. SHAP-based analyses have been applied to brain age prediction models to identify important anatomical predictors and improve the interpretation of prediction outcomes. Sun et al., for example, used SHAP to analyze individual brain age predictions and demonstrated how feature attribution methods can provide subject-specific explanations in MRI-based brain age modeling [21].

Explainability has also become increasingly important in deep learning-based brain age prediction. Although deep learning models often achieve strong predictive performance, their internal decision-making processes are typically diffi-

cult to interpret. To address this limitation, researchers have applied a variety of explainability techniques, including SHAP, LIME, saliency maps, layer-wise relevance propagation, and gradient-based visualization methods. Lombardi et al. investigated an explainable deep learning framework for personalized brain age prediction using MRI-derived morphological features and demonstrated that interpretable deep learning approaches can help identify age-related morphological patterns associated with brain age estimation [22].

Despite substantial progress in explainable artificial intelligence, interpretability in brain age prediction remains an active research challenge. Different explanation methods may yield different feature attribution patterns, while interpretation can be affected by factors such as model architecture, preprocessing strategies, feature dependencies, and dataset characteristics. Consequently, recent studies emphasize that explainability techniques should not be regarded solely as visualization tools, but also as complementary evaluation mechanisms that can provide insights into model robustness, biological plausibility, and clinical interpretability in brain age prediction research [19, 22, 23].

2.5 Longitudinal Reliability and Test-Retest Stability

Test-retest reliability refers to the consistency of brain age predictions across repeated MRI scans acquired over relatively short periods of time where substantial biological aging is not expected to occur. If a model produces large fluctuations in predicted brain age between scans obtained within days or months, the resulting Brain Age Gap estimates may reflect methodological instability rather than meaningful neurobiological variation. Kaufmann et al. investigated brain age reliability across multiple cohorts and highlighted the importance of robust prediction consistency when interpreting brain age as a biomarker [24]. Similarly, Hahn et al. highlighted the importance of robust and reliable brain age estimation, emphasizing that uncertainty quantification and prediction stability are critical for the clinical applicability and interpretability of brain age biomarkers [25].

In addition to short-term stability, longitudinal sensitivity is also considered important in brain age modelling. A clinically meaningful brain age model should remain relatively stable across short intervals while still being capable of detecting

gradual aging related changes over longer periods. Dörfel et al. systematically evaluated publicly available brain age prediction tools and demonstrated that predictive accuracy, test-retest reliability, and the ability to track longitudinal age progression may differ substantially across models. Their work further emphasized that cross-sectional prediction performance alone may provide an incomplete assessment of model quality [12].

Several factors such as scanner differences ,acquisition protocols ,preprocessing pipelines and model architecture may influence longitudinal stability in brain age prediction [11, 12, 26]. Bashyam et al., in the large-scale ENIGMA brain age study, highlighted the challenges associated with multi-site neuroimaging data and demonstrated the need for careful harmonization and validation when applying brain age models across diverse cohorts [26]. These issues become particularly relevant in longitudinal analyses, where small methodological inconsistencies may obscure subtle aging-related changes.

Recent literature therefore increasingly argues that robustness and longitudinal consistency should be considered core evaluation criteria in brain age prediction research. Rather than treating brain age solely as a cross-sectional prediction task, longitudinal frameworks allow researchers to examine whether brain age models capture stable and biologically plausible aging trajectories over time.

2.6 Brain Age Deviation Analysis in Clinical Populations

A common application of brain age prediction is the analysis of deviations from expected healthy aging trajectories in clinical populations. In most brain age studies, machine learning models are trained using MRI data acquired from cognitively healthy individuals in order to learn normative relationships between brain structure and chronological age. The trained models can then be applied to subjects with neurological or psychiatric disorders to investigate whether their brains appear older or younger than expected relative to healthy aging patterns [4,5].

Within this framework, the BAG is interpreted as a measure of deviation from typical aging trajectories. Positive BAG values indicate that the brain appears older than expected for a subject's chronological age, whereas negative values indicate a younger appearing brain. Several studies have reported elevated BAG

values in clinical populations including Alzheimer’s disease, Mild Cognitive Impairment (MCI), schizophrenia, bipolar disorder and major depressive disorder, suggesting that brain age prediction may capture disease related alterations in brain structure [5, 9, 24].

This approach has become particularly relevant in neurodegenerative research, where deviations from healthy aging trajectories may emerge before severe clinical symptoms become apparent. Gaser et al. demonstrated that increased BrainAGE values were associated with higher risk of conversion from MCI to Alzheimer’s disease, supporting the idea that brain age measures may reflect abnormal neuroanatomical aging processes [8].

By comparing predicted brain age between healthy controls and clinical populations, researchers can investigate how different disorders relate to deviations from expected neuroanatomical aging trajectories. This perspective forms the basis for many normative brain age analyses in contemporary neuroimaging research.

2.7 State-of-the-Art Approaches

In recent years, there has been a significant shift in brain age prediction research toward deep learning and large scale image based approaches. Early brain age studies primarily relied on handcrafted morphometric features extracted from structural MRI scans, however recent state-of-the-art methods increasingly operate directly on voxel-wise neuroimaging data using deep neural networks. These approaches aim to automatically learn hierarchical imaging representations associated with brain aging without requiring predefined anatomical feature extraction. Cole et al. demonstrated that convolutional neural networks (CNNs) trained directly on raw T1-weighted MRI scans could achieve highly accurate brain age prediction [7].

CNNs have become one of the dominant deep learning architectures in brain age prediction because they can capture spatial patterns directly from volumetric MRI data. Peng et al. proposed a lightweight deep neural network for accurate brain age prediction from MRI scans and reported state-of-the-art predictive performance using image-based learning approaches [27]. Similarly, Bashyam et

al. developed a large scale deep learning brain age framework within the ENIGMA consortium using MRI data collected from multiple cohorts worldwide [26]. Together, these studies demonstrated that deep learning approaches can achieve highly competitive, and often lower, prediction errors than traditional feature-based methods, particularly when trained on large neuroimaging datasets.

Voxel-wise approaches have also become increasingly common in modern brain age modelling. Instead of reducing MRI scans to regional anatomical features, voxel-wise methods preserve much of the original spatial information contained within structural MRI data. Franke et al. used voxel-based morphometry in early BrainAGE studies, while more recent deep learning approaches have extended voxel-wise analysis through three-dimensional CNN architectures operating directly on volumetric brain images [14, 7, 27]. These methods may capture subtle spatial aging patterns that could be partially lost in feature-based morphometric pipelines.

More recently, multimodal and transformer-based approaches have begun to emerge in brain age prediction research. Multimodal methods combine multiple imaging modalities or integrate neuroimaging with non-imaging information such as genetics, cognitive scores, or clinical variables in order to improve prediction performance and capture complementary aspects of brain aging [28]. Transformer architectures have also attracted growing interest in neuroimaging because of their ability to model long-range relationships within imaging data. However, compared with CNN-based approaches, transformer-based brain age models remain relatively less explored and currently represent an emerging direction of research

Despite their strong predictive performance, state-of-the-art deep learning approaches also introduce important challenges. Deep neural networks generally require substantially larger datasets and higher computational resources compared to traditional machine learning models trained on morphometric features. In addition, image based deep learning methods are often more difficult to interpret, raising concerns regarding biological plausibility and clinical transparency. Several studies have therefore emphasised the trade off between predictive performance and interpretability in brain age prediction [19, 22]. Feature-based approaches using anatomical MRI measurements may achieve slightly lower prediction accuracy

in some settings, but they often provide more interpretable relationships between neuroanatomical structures and prediction outcomes.

Current brain age research increasingly explores the balance between predictive accuracy, robustness, interpretability, computational complexity, and clinical applicability. While state-of-the-art deep learning models continue to advance prediction performance, traditional morphometric approaches remain highly relevant, particularly when anatomical interpretability, explainability, and robustness are considered important objectives.

3. Methodology

This chapter describes the methodological framework used to develop and evaluate the brain age prediction models in this thesis. It presents the ADNI dataset, the preprocessing pipeline, the selected machine learning models, the hyperparameter optimization procedure, and the evaluation metrics used to assess predictive performance. The chapter then describes the explainability framework, the longitudinal evaluation strategy, and the normative modelling analysis applied to clinical subjects. These methods define the complete experimental pipeline used to generate the results presented in Chapter 4.

The methodological design and reporting of this study were guided by the TRIPOD+AI (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis using Artificial Intelligence) Statement, an international reporting guideline for studies developing and evaluating prediction models based on regression or machine learning methods. The study was structured to transparently report the dataset, participant selection, preprocessing pipeline, predictor variables, model development, hyperparameter optimization, validation strategy, performance evaluation, explainability analyses, longitudinal assessment, and normative modelling framework in accordance with the recommendations of TRIPOD+AI. The completed TRIPOD+AI checklist is provided in Appendix A as supplementary material. [29]

3.1 ADNI Dataset

The data used in this study were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database, a large-scale longitudinal neuroimaging study designed to investigate biomarkers associated with aging, Mild Cognitive Impairment (MCI), and Alzheimer’s disease (AD). ADNI includes structural MRI scans, demographic information, clinical assessments, and additional biomarker data collected from multiple research sites over several longitudinal follow up periods [30].

A custom dataset was constructed using the ADNI cohort builder provided through the ADNI platform. The dataset included demographic variables such as subject age, sex, and diagnostic labels, together with structural MRI derived FreeSurfer measurements and scanner related information. The FreeSurfer features were selected as the neuroimaging representation for the brain age prediction framework used in this study.

The dataset included subjects from cognitively normal (CN), Mild Cognitive Impairment (MCI) and Dementia groups. ADNI is a longitudinal study so multiple MRI acquisitions from different time points were available for several subjects. This longitudinal structure enabled both cross sectional brain age prediction experiments and longitudinal stability analyses using repeated scans acquired over time.

Scanner related metadata were also included because MRI data in ADNI originate from multiple acquisition sites and scanner configurations. Such information is relevant in brain age studies due to the known influence of scanner variability and acquisition protocols on neuroimaging derived measurements [11].

The resulting dataset therefore combined demographic information, diagnostic labels, MRI derived morphometric features, and scanner metadata within a unified framework suitable for MRI based brain age prediction and longitudinal evaluation.

The initial ADNI dataset contained 24,791 structural MRI scan records acquired across multiple longitudinal time points. To construct a subject-level baseline cohort, only the first eligible scan from each participant was retained, resulting in 2,335 unique subjects. Of these, 794 cognitively normal participants met the eligibility criteria for primary brain age model development. This cohort was divided at the subject level into a training subset of 635 participants and an independent test subset of 159 participants using an 80/20 split.

Subjects were eligible for inclusion if they had available chronological age information, diagnostic classification, and structural MRI-derived FreeSurfer morphometric features required for the corresponding analysis. For the primary brain age model development stage, only cognitively normal subjects were used in order to learn normative age-related patterns from individuals without clinical cognitive impairment. Subjects or visits with missing chronological age information were excluded, while missing feature values were handled during the training-only pro-

cessing stage described in Section 3.2. The final sample size was therefore determined by the availability of eligible ADNI participants meeting these criteria. This data-driven sample definition is appropriate for the present retrospective machine learning study because the objective was to evaluate model performance, interpretability, and longitudinal behaviour within the available ADNI neuroimaging cohort.

3.2 Data Preprocessing

Before model development, the ADNI data were organized into separate analytical cohorts according to the objectives of the study. Figure 3.1 provides an overview of this participant allocation and analytical framework. Starting from the ADNI database, eligible subjects were assigned to cohorts used for brain age model development, longitudinal evaluation, and normative modelling. This structure ensured that each analytical objective was addressed using the appropriate subset of participants while maintaining a clear separation between model development and subsequent evaluation procedures.

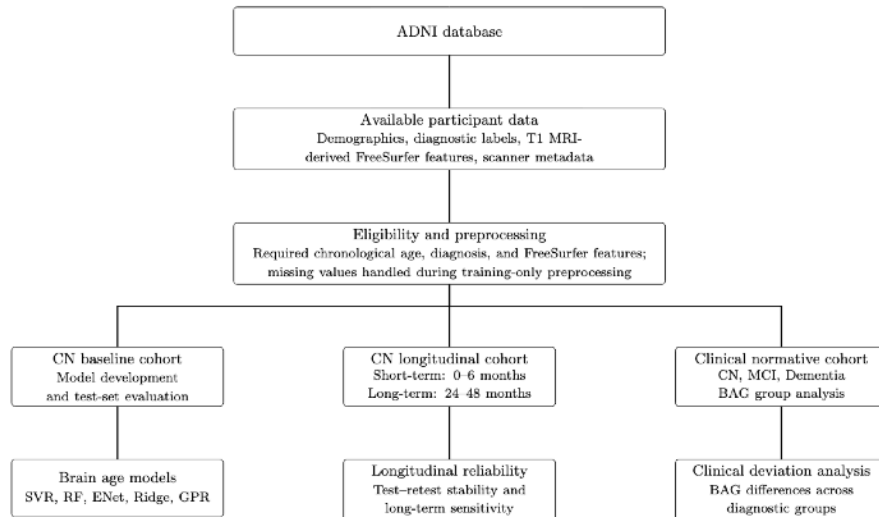


Figure 3.1: Participant and dataset flow diagram for the brain age prediction analyses.

Data preprocessing was performed in multiple stages in order to reduce the risk of data leakage and ensure strict separation between training and testing in-

formation throughout the analysis pipeline. The preprocessing framework consisted of three phases: an initial dataset construction phase, a stratified data splitting phase, and a training-only preprocessing phase applied after dataset partitioning.

During the initial preprocessing stage, the dataset was filtered to include only cognitively normal (CN) subjects for brain age model development. Since the ADNI dataset contains repeated longitudinal visits for many subjects, only baseline visits were retained during construction of the primary training cohort in order to avoid multiple highly correlated observations from the same subject during model training. Rows with missing chronological age values were removed, and demographic, scanner, and FreeSurfer-derived feature tables were merged into a unified dataset.

Descriptive analysis of the resulting cohort was performed to examine the demographic and age distributions of the dataset. Figure 3.2 illustrates the age distribution of the baseline-like cohort together with the cognitively normal subset used for model development.

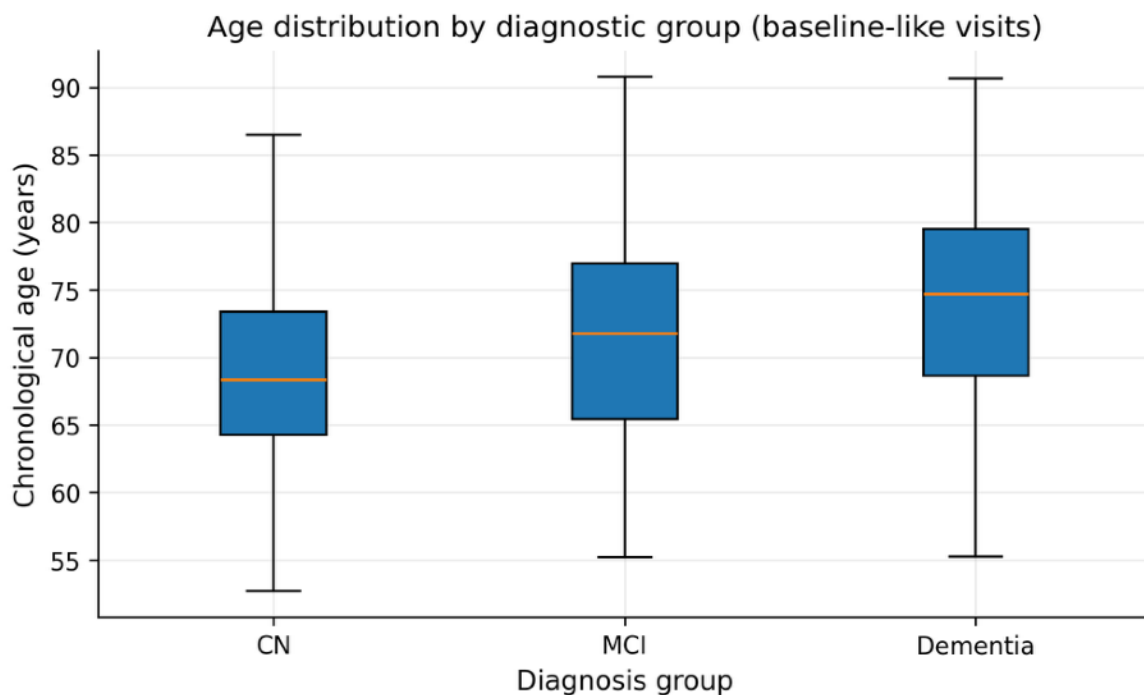


Figure 3.2: Age distribution of the baseline-like ADNI cohort and the cognitively normal subset used for brain age model development.

Additional exploratory analyses were conducted to examine demographic differences across diagnostic groups. Figure 3.3 presents the chronological age

distribution across diagnostic categories, while Figure 3.4 shows the sex distribution within the cognitively normal baseline-like cohort.

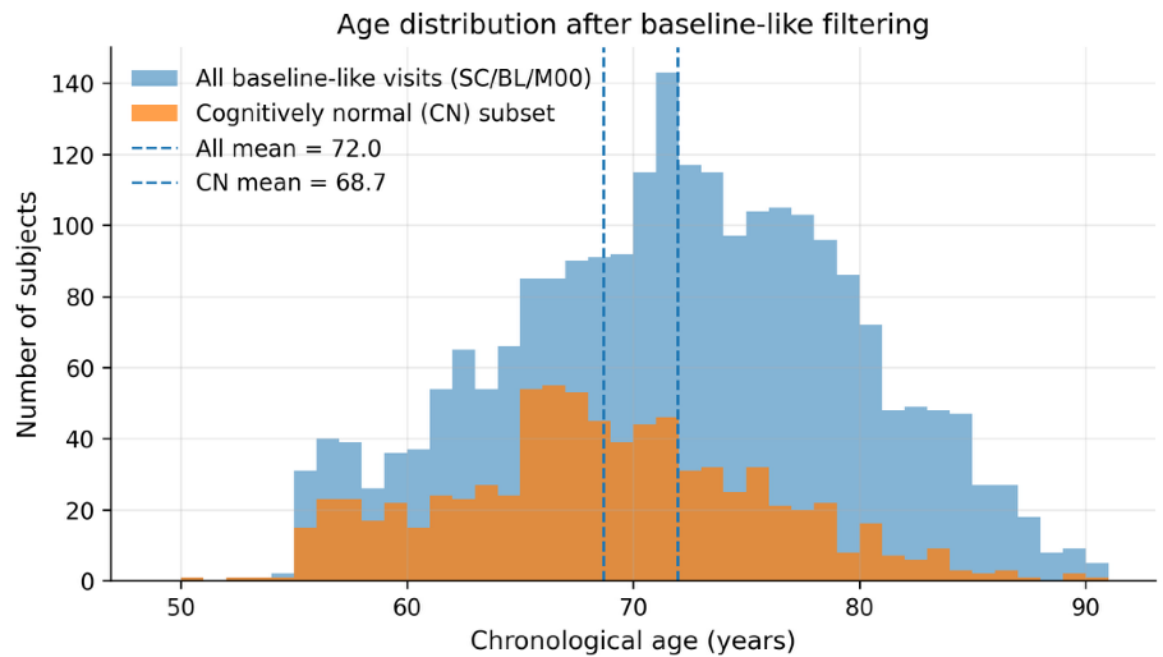


Figure 3.3: Chronological age distribution across diagnostic groups in the baseline-like cohort.

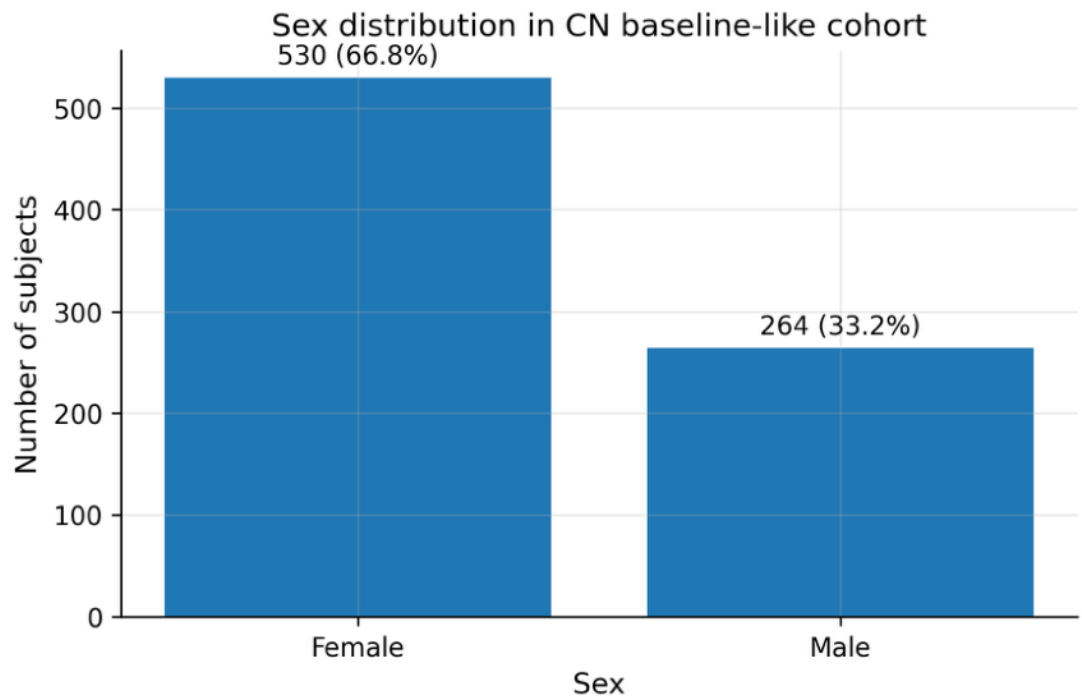


Figure 3.4: Sex distribution within the cognitively normal baseline-like cohort.

After dataset construction, the data were divided into training and testing subsets using an 80/20 split. Stratification was performed using age and sex in order to preserve demographic balance and reduce distributional differences between subsets during model evaluation [16].

A final preprocessing stage was applied exclusively on the training data after dataset splitting in order to prevent information leakage from the test set into the training process. Features with more than 20% missing values were removed from the analysis. Remaining missing values were imputed using median imputation calculated from the training subset, and the same transformations were subsequently applied to the corresponding test data [29].

This multi stage preprocessing framework was designed to maintain separation between training and testing information while ensuring that the resulting dataset was suitable for robust MRI based brain age prediction experiments.

3.3 Brain Age Prediction Models

The predicted outcome in this study was chronological age, estimated from structural MRI-derived FreeSurfer morphometric features. Chronological age was treated as an objective continuous outcome and did not require subjective clinical assessment. The resulting model output was predicted brain age in years, from which Brain Age Gap values were calculated as the difference between predicted brain age and chronological age. Since this was a regression-based prediction task, no classification thresholds or class-imbalance correction procedures were applied.

Candidate predictors consisted of numeric structural MRI-derived FreeSurfer measurements together with sex as a demographic covariate. The original sex variable was binary encoded before model training. Chronological age was used exclusively as the continuous prediction outcome and was not included among the predictors. Participant identifiers, visit information, diagnostic group, ADNI phase, scanner field strength, FreeSurfer version, image identifiers, and examination date were treated as metadata and excluded from the model input.

Five regression models from four different machine learning families were selected to compare alternative approaches for estimating chronological age from-

structural neuroimaging features. The evaluated models included Support Vector Regression (SVR), Ridge Regression, Elastic Net Regression, Random Forest Regression, and Gaussian Process Regression (GPR). These methods have been widely used in brain age prediction studies and represent diverse modelling paradigms, including regularised linear regression, kernel-based learning, ensemble learning, and probabilistic regression [6,16].

3.3.1 Support Vector Regression

SVR was used due to its widespread use in neuroimaging-based age prediction studies. SVR is particularly suitable for high-dimensional neuroimaging data because it can model non-linear relationships between features and chronological age through kernel learning. [5 ,16]

3.3.2 Ridge Regression

Ridge Regression was used as a regularized linear model for brain age prediction. Ridge applies L2 regularization to reduce coefficient magnitude and improve stability when handling large numbers of correlated predictors, which is common in structural MRI feature spaces. Because many FreeSurfer-derived measurements exhibit substantial inter-feature correlation, Ridge Regression provided a stable linear framework for age estimation [16].

3.3.3 Elastic Net Regression

Elastic Net Regression was included as a regularised linear model combining both L1 and L2 penalties. This approach allows both coefficient shrinkage and feature selection, making it suitable for high dimensional correlated neuroimaging data. Compared to purely linear regression approaches, Elastic Net can produce sparser feature representations while maintaining robustness in the presence of correlated predictors [16].

3.3.4 Random Forest Regression

Random Forest Regression was used as a non-linear ensemble learning approach for brain age prediction. Random Forest models construct multiple decision trees and aggregate their predictions in order to improve generalization performance and reduce variance. This approach is capable of modeling complex non-linear interactions between neuroanatomical features and chronological age without requiring explicit feature engineering. In addition, Random Forests naturally support feature importance estimation, which is useful for subsequent explainability analyses [16].

3.3.5 Gaussian Process Regression

Gaussian Process Regression (GPR) was used as a probabilistic regression approach for brain age prediction. GPR extends classical regression by incorporating Gaussian probabilistic modelling and can capture complex, potentially non-linear relationships between neuroimaging features and age. The method has been successfully applied in previous brain age prediction studies and has demonstrated strong predictive performance in structural MRI-based age estimation [20].

Together, these five models provided a diverse set of regression frameworks for evaluating MRI-based brain age prediction across different machine learning paradigms.

3.4 Training and Hyperparameter Optimization

All brain age prediction models were trained using the training subset generated during the preprocessing stage. Hyper parameter optimization was performed for each regression model using GridSearchCV with 10-fold cross-validation in order to systematically identify parameter configurations associated with improved predictive performance. The hyper parameter configuration producing the lowest average cross-validation Mean Absolute Error (MAE) was selected as the final configuration for each model.

In addition to selecting optimal parameter configurations, hyper parameter sensitivity analysis was also performed for each model. The relationship between individual hyper parameters and cross-validation MAE was examined in order to investigate how model performance changed across different parameter values. These analyses were used to assess the stability and behaviour of the models during optimization and are presented in the Results chapter.

3.5 Evaluation Metrics

Model performance was evaluated using multiple complementary metrics in order to assess both prediction accuracy and the relationship between predicted brain age and chronological age. Evaluation was performed using cross-validation performance during training as well as independent test-set performance after model optimization.

The primary evaluation metric used during hyperparameter optimization was Mean Absolute Error (MAE), calculated as the average absolute difference between predicted age and chronological age. Cross-validation MAE was used during the GridSearchCV optimization procedure in order to compare candidate hyperparameter configurations across all regression models.

Final model evaluation on the independent test set included both uncorrected and age-corrected MAE measurements. Uncorrected MAE was calculated directly from the raw predicted brain age values produced by each model. In addition, age-corrected MAE was computed after applying a post-hoc bias correction procedure designed to reduce regression to the mean effects commonly observed in brain age prediction models.

Regression to the mean bias is a well documented issue in brain age prediction, where models tend to overestimate age in younger subjects and underestimate age in older subjects [5,10]. To reduce this effect, a linear post-hoc age-bias correction approach was applied following model training . A linear regression model was fitted between predicted age and chronological age within the training data, producing correction parameters α and b . Corrected predicted age values were then computed using:

$$\hat{y}_{corrected} = \frac{\hat{y} - a}{b}$$

where $\hat{y}$ represents the raw predicted age and α and b correspond to the intercept and slope estimated from the linear regression model. This correction procedure was used to reduce systematic age related prediction bias prior to final evaluation.

In addition to MAE, Pearson correlation coefficient (r) was used to evaluate the linear association between predicted brain age and chronological age. Pearson correlation was included in order to assess how well the models captured relative age-related variation across subjects independently of absolute prediction error.

Together, these evaluation metrics provided complementary information regarding prediction accuracy, age related bias, and the relationship between predicted and chronological age across the evaluated brain age models.

3.6 Explainability Framework

Explainability analyses were conducted to investigate the neuroanatomical patterns associated with brain age prediction and to evaluate the stability of feature attribution across repeated model training procedures. The detailed explainability analysis focused on Random Forest and Elastic Net because they provided complementary nonlinear and linear perspectives. Random Forest was selected due to its strong and stable predictive performance, whereas Elastic Net was retained as an interpretable linear reference because its standardized coefficients provide direct information regarding the direction and magnitude of feature associations.

Global feature attribution was examined using SHAP values and permutation importance. SHAP analysis was used to quantify the contribution of individual FreeSurfer derived features to model predictions across the dataset, allowing estimation of both feature importance magnitude and directional influence on predicted brain age. Permutation importance was additionally computed by measuring the reduction in predictive performance after random permutation of individual features. Together, these approaches were used to identify anatomical measurements consistently associated with brain age estimation.

To provide a linear interpretability reference, Elastic Net regression coefficients were also analysed. Since Elastic Net applies embedded regularization and feature selection, the resulting coefficients provided an additional feature attribution framework complementary to the non-linear Random Forest analyses.

In addition to global analyses, subject-level explainability was investigated using local SHAP explanations. Three representative subjects were selected based on Brain Age Gap (BAG) values, corresponding to strongly negative, near-zero, and strongly positive BAG cases. Local SHAP analysis was then used to examine the feature level contributions underlying each individual prediction.

Finally, SHAP stability analysis was performed in order to evaluate the reproducibility of feature attribution patterns across repeated Random Forest training procedures. The model was retrained using multiple random seeds, and SHAP values were recomputed independently for each training instance. Stability was evaluated using cross-seed Pearson correlation of feature importance rankings together with overlap analysis of the top-ranked features across model initializations. Specifically, overlap was assessed for the top 10 and top 20 most important features identified by SHAP.

3.7 Longitudinal Evaluation

Longitudinal analyses were conducted to evaluate both the short-term stability and long-term sensitivity of the brain age prediction models. Since brain age is intended to reflect biological aging processes, a reliable model should produce consistent predictions when minimal aging is expected while remaining sensitive to genuine age-related changes over longer time intervals.

Repeated MRI scans from the same CN subjects were identified using the longitudinal structure of the ADNI dataset. All possible within subject scan pairs were initially generated and subsequently filtered according to the desired follow up interval. To avoid over representation of subjects with multiple follow up visits, a single scan pair was kept for each subject in both the short-term and long-term analyses. For the long term analysis the pair closest to a 36-month interval was selected.

3.7.1 Short-Term Stability

Short term stability was evaluated using repeated MRI scans acquired within 0–6 months. This analysis was designed to assess the test-retest reliability of the models because substantial biological aging is not expected over such a short interval,

For each regression model, BAG values obtained from the first and second scans were compared using Pearson correlation coefficient (r), Concordance Correlation Coefficient (CCC), and the Mean Absolute Difference in BAG. Pearson correlation was used to evaluate the consistency of subject ranking across repeated measurements, while CCC quantified overall agreement between time points. The mean absolute BAG difference provided an estimate of prediction variability between repeated scans.

3.7.2 Long-Term Sensitivity

Long-term sensitivity was evaluated using scan pairs separated by 24–48 months. In contrast to the short-term analysis, the objective was to determine whether the models were capable of detecting age related changes occurring over longer follow up periods.

For each subject, chronological age change (Δ Age) and predicted brain age change (Δ Predicted Age) were calculated between the two time points. Model performance was then assessed using Pearson correlation between Δ Age and Δ Predicted Age together with the Mean Absolute Error of the predicted age change. Higher correlation values indicate that subjects exhibiting greater chronological aging also tended to show larger increases in predicted brain age.

To investigate the robustness of the findings, three longitudinal subsets were evaluated. The ‘all’ subset included all eligible longitudinal pairs. The ‘clean’ subset retained only pairs acquired using the same MRI field strength and FreeSurfer processing version at both time points, thereby reducing potential methodological variability. Finally, the ‘tight’ subset restricted the follow-up interval to 30–42 months, producing a more homogeneous longitudinal cohort centered around the target three year follow up period.

3.8 Normative Modelling on Clinical Subjects

To investigate whether the brain age model captured deviations from normative aging trajectories, the Random Forest model trained only on CN subjects was applied to an independent cohort containing CN, MCI and Dementia subjects.

The trained model, together with the preprocessing pipeline developed during training, was applied without further retraining or adaptation. Feature selection, imputation procedures, and bias correction parameters were preserved exactly as defined during model development in order to ensure that predictions reflected the normative aging patterns learned from the healthy population. Brain age predictions were then generated for all diagnostic groups, and BAG values were calculated as the difference between predicted brain age and chronological age.

Group-level analyses were subsequently performed to characterize differences in BAG across diagnostic categories. Summary statistics were computed separately for CN, MCI, and Dementia subjects, and BAG distributions were visualized using box plots in order to examine potential shifts in predicted brain age between groups.

To evaluate whether diagnostic status was associated with BAG independently of demographic factors, an ordinary least squares regression model was fitted using BAG as the dependent variable and diagnostic group, chronological age, and sex as predictors. Diagnostic group was treated as a categorical variable with cognitively normal subjects serving as the reference category. The resulting regression framework allowed assessment of whether MCI and Dementia subjects exhibited significantly different BAG values after accounting for potential age and sex related effects.

3.9 Ethical Considerations

The present study was conducted as a retrospective secondary analysis of data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI data are collected under institutional review board approval at participating centers, and participants provide informed consent through the ADNI study procedures. Since the present thesis used de-identified data obtained through the ADNI

platform, no additional participant recruitment or local clinical intervention was performed.

The analysis used ADNI data available through controlled access upon application to the ADNI platform. The analytical code developed for this thesis is not publicly released at the time of submission but may be made available upon reasonable request. No formal study protocol was prospectively prepared, and the study was not registered in a clinical trial or prediction model registry. This thesis received no specific external funding, and the author declares no conflicts of interest. Patients or members of the public were not directly involved in the design, conduct, analysis, interpretation, or reporting of this MSc thesis.

No dedicated fairness or health-inequality analysis was performed in this thesis. However, age and sex were described during cohort characterization, and age and sex were included as covariates in the normative modelling regression analysis when assessing diagnostic group differences in Brain Age Gap.

4. Results

This chapter presents the experimental results obtained from the brain age prediction framework developed in this thesis. It first reports the predictive performance of the evaluated machine learning models using cross-validation and independent test-set metrics. It then presents the hyperparameter optimization analysis, the global and local explainability results, the SHAP stability analysis, and the longitudinal evaluation findings. Finally, the chapter reports the normative modelling results in clinical populations and summarizes the main findings before their interpretation in Chapter 5.

4.1 Models Performance

This section presents the predictive performance of the five machine learning models evaluated in this study, namely Random Forest, SVR, GPR, Elastic Net, and Ridge Regression. Performance was assessed using MAE, Pearson correlation and BAG dispersion on the independent test set.

Figure 4.1 compares the cross-validation and test-set MAE values obtained by each model. Cross-validation performance was relatively consistent across all five methods, with mean MAE values ranging from 4.12 to 4.22 years.

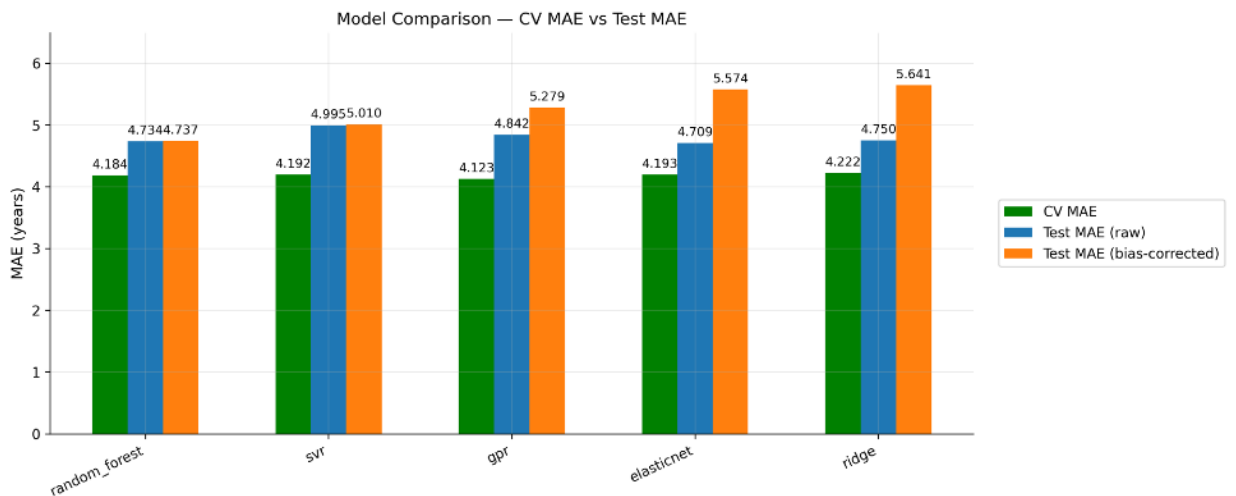


Figure 4.1: Cross-validation and independent test-set MAE for the evaluated brain age prediction models.

The lowest cross-validation MAE was achieved by GPR (4.123 years), followed closely by Random Forest (4.184 years), SVR (4.192 years), Elastic Net (4.193 years), and Ridge Regression (4.222 years). The small differences observed among the cross-validation errors suggest that all models achieved a comparable level of predictive accuracy during model development.

When evaluated on the independent test set, Elastic Net achieved the lowest prediction error, with a raw MAE of 4.709 years, closely followed by Random Forest (4.734 years) and Ridge Regression (4.750 years). SVR and GPR produced slightly higher test-set errors of 4.995 and 4.842 years, respectively. Following bias correction, the MAE remained unchanged for Random Forest but increased for the remaining models, particularly for Elastic Net and Ridge Regression, indicating that bias correction did not improve predictive accuracy in terms of absolute error for these methods.

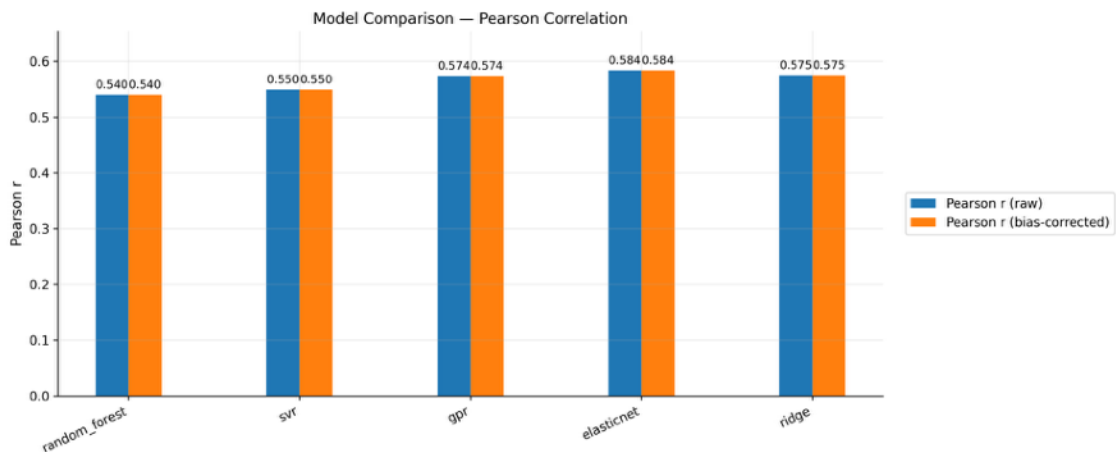


Figure 4.2: Pearson correlation between predicted and chronological age for each brain age prediction model.

Pearson correlation coefficients between predicted and chronological age are shown in Figure 4.2. Correlation values ranged from 0.540 to 0.584 across models, indicating a moderate positive association between predicted and true age. The highest correlation was achieved by Elastic Net ($r = 0.584$), followed by Ridge Regression ($r = 0.575$) and GPR ($r = 0.574$). SVR ($r = 0.550$) and Random Forest ($r = 0.540$) exhibited slightly lower correlations. As expected, bias correction did not alter the correlation coefficients since the procedure applies a linear adjustment to the predictions.

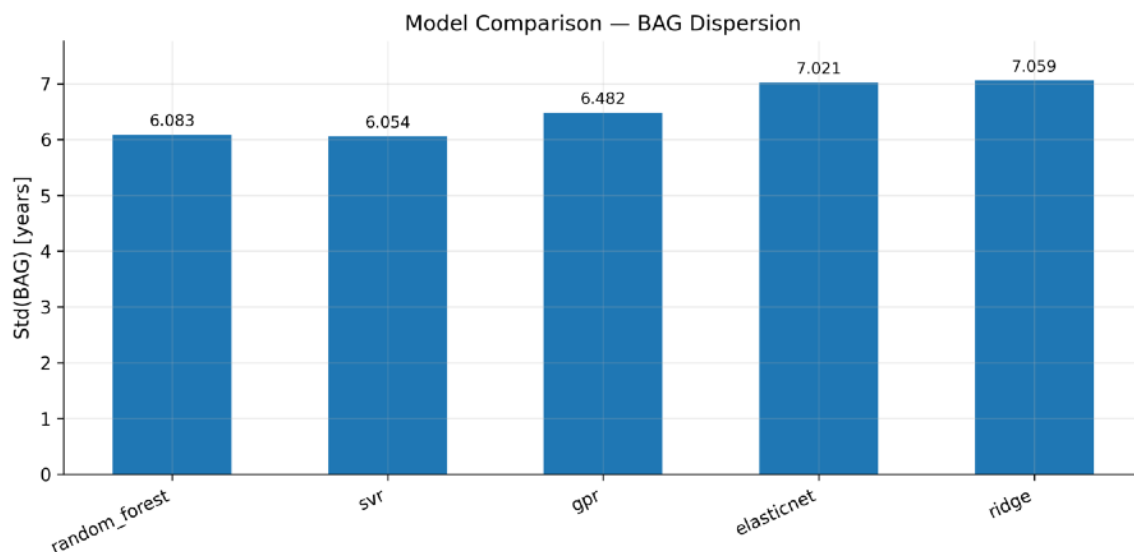


Figure 4.3: Standard deviation of the Brain Age Gap distribution for each machine learning model.

To assess the variability of prediction errors across individuals, the standard deviation of the Brain Age Gap distribution was calculated for each model (Figure 4.3). Lower BAG dispersion indicates more concentrated prediction errors and greater stability of age estimates. SVR achieved the lowest BAG dispersion (6.054 years), followed closely by Random Forest (6.083 years). GPR produced a somewhat wider distribution (6.482 years), while Elastic Net (7.021 years) and Ridge Regression (7.059 years) exhibited the highest variability.

Overall, the results indicate that no single model consistently outperformed all others across every evaluation metric. Random Forest achieved the lowest test-set MAE, Elastic Net produced the highest age correlation, and SVR exhibited the smallest BAG dispersion.

4.1.1 Comparison with State-of-the-Art Brain Age Prediction Studies

To place the present model performance in context, the results were compared with selected state-of-the-art and alternative brain age prediction studies (Table 4.1). Direct numerical comparison should be interpreted cautiously because reported performance depends on several factors, including cohort size, age range, scanner and site variability, preprocessing pipeline, input representation,

model architecture, and validation strategy. Nevertheless, such comparison is useful for evaluating whether the observed results are consistent with the broader literature and for clarifying the methodological contribution of the present framework.

Study	Data / Input Representation	Methodology	Reported Performance	Main Comparison with Present Thesis
Present thesis	ADNI; FreeSurfer-derived regional morphometric features	Random Forest, SVR, Elastic Net, Ridge Regression, GPR	Test-set MAE below 5 years; Pearson r approximately 0.54-0.58	Moderate correlations and non-SOTA accuracy, but interpretable feature-based framework with explainability, longitudinal reliability, and normative clinical analysis
Han et al. [17]	HCP, Cam-CAN, and IXI; FreeSurfer-derived brain morphometric measures	27 traditional machine learning models	HCP: MAE 2.75-3.12, r 0.11-0.42; Cam-CAN: MAE 7.08-10.50, r 0.64-0.85; IXI: MAE 8.04-9.86, r 0.63-0.79	Demonstrates that morphometric ML performance varies strongly across datasets and age ranges; present MAE is lower than Cam-CAN/IXI but Pearson r is more modest
De Bonis et al. [20]	OpenBHB multisite dataset; FreeSurfer morphometric features and minimally pre-processed T1-weighted MRI images	Morphometric DNN pipeline and 3D CNN pipeline	Independent test-set MAE 3.21 with morphometric DNN and 3.08 with DenseNet-121; corrected global r values above 0.90	Shows that both morphometric neural networks and image-based CNNs can outperform the present models in accuracy while also emphasizing the importance of XAI
Peng et al. [28]	UK Biobank and PAC 2019; T1-weighted MRI images	Lightweight 3D CNN / SFCN	MAE 2.14 years in UK Biobank; MAE 2.90 years in PAC 2019	Represents image-based deep learning SOTA accuracy; substantially stronger predictive performance than the present feature-based ML framework

Table 4.1: Contextual comparison between the present thesis and selected brain age prediction studies.

Overall, the comparison shows that the present approach does not achieve state-of-the-art prediction accuracy when evaluated against large-scale image-based deep learning models. The Pearson correlations are also moderate compared with studies using broader age ranges or larger training cohorts. However, the comparison also shows that performance values in brain age prediction vary substantially across datasets and input representations. For example, Han et al. reported low correlations in the HCP dataset despite low MAE values, while higher correlations but larger MAE values were observed in broader age-range datasets [16]. This illustrates that Pearson correlation and MAE are influenced by different dataset properties and should not be interpreted in isolation.

The main contribution of the present thesis therefore lies not in outperforming deep learning models in prediction accuracy, but in providing an interpretable morphometric machine learning framework. By using FreeSurfer-derived regional features, the models produced predictions that could be linked directly to anatomical structures through SHAP values, permutation importance, and coefficient-based analyses. This feature-based design also enabled systematic evaluation of hyperparameter sensitivity, longitudinal reliability, and normative deviations in MCI and Dementia subjects. Thus, the present framework trades state-of-the-art im-

age-based accuracy for greater anatomical interpretability, lower model complexity, and broader methodological evaluation.

4.2 Hyperparameter Optimization Analysis

To investigate the influence of model configuration on predictive performance, the relationship between selected hyperparameters and cross-validation mean absolute error (CV MAE) was examined for each machine learning model. Figures 4.4–4.9 summarize the performance obtained across the evaluated hyperparameter combinations.

For the Random Forest model (Figure 4.4), predictive performance was relatively stable across the explored parameter values. The lowest CV MAE was achieved when using a smaller subset of features at each split (`max_features = 0.3` or `0.5`) combined with a minimum leaf size of one sample. Increasing the minimum leaf size produced only minor changes in performance, while the use of `max_features = sqrt` consistently resulted in slightly higher prediction errors.

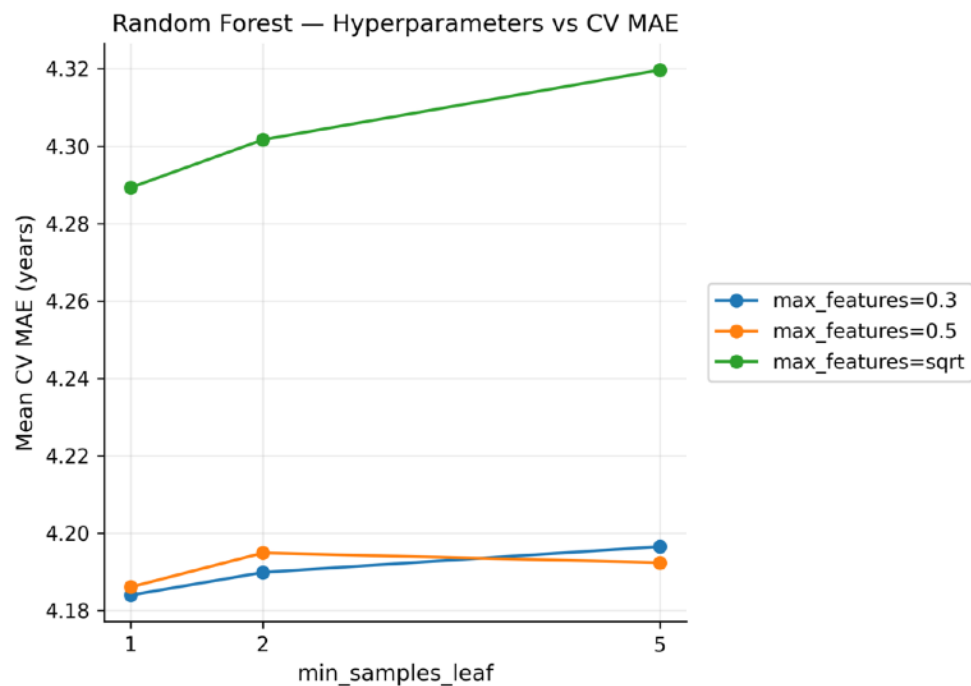


Figure 4.4: Cross-validation MAE of the Random Forest model as a function of `min_samples_leaf` and `max_features`

The SVR model exhibited greater sensitivity to hyperparameter selection. As shown in Figure 4.5, increasing the regularisation parameter C substantially improved performance when the kernel parameter was set to $\gamma = \text{scale}$. The CV MAE decreased from approximately 4.65 years at $C=1$ to around 4.20 years for larger values of C , after which performance plateaued. In contrast, alternative γ values (0.01, 0.03, and 0.1) resulted in substantially higher prediction errors across all tested values of C .

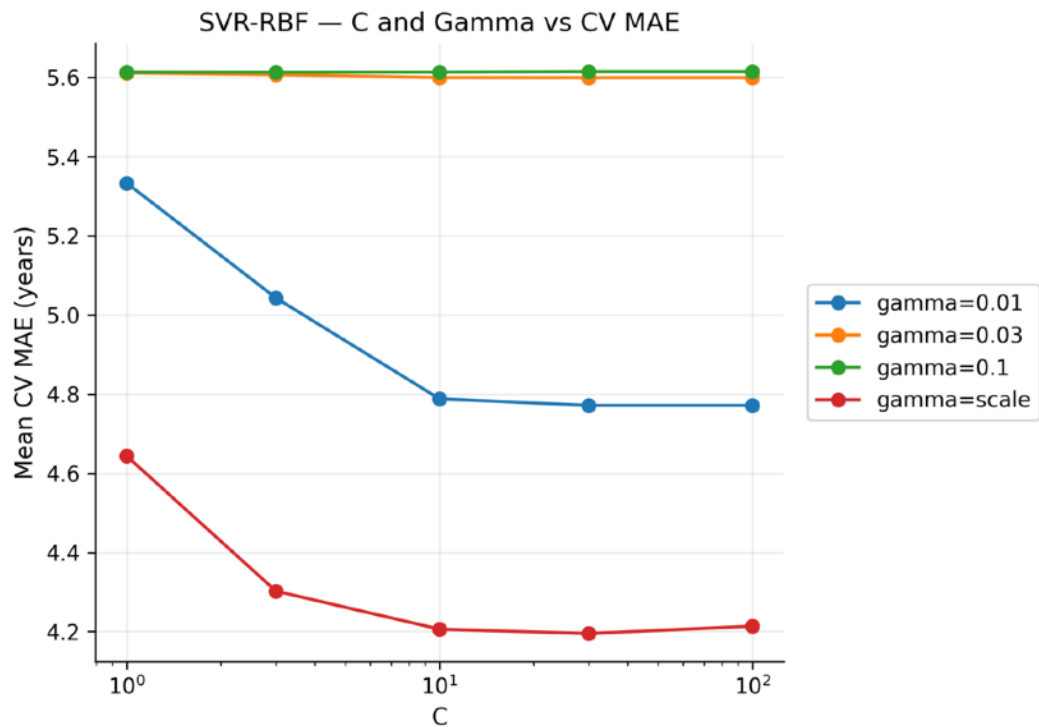


Figure 4.5: Cross-validation MAE of the SVR model for different combinations of the regularization parameter C and γ .

The analysis of the ϵ parameter (Figure 4.6) demonstrated only negligible variation in CV MAE, indicating that model performance was largely insensitive to the explored epsilon values.

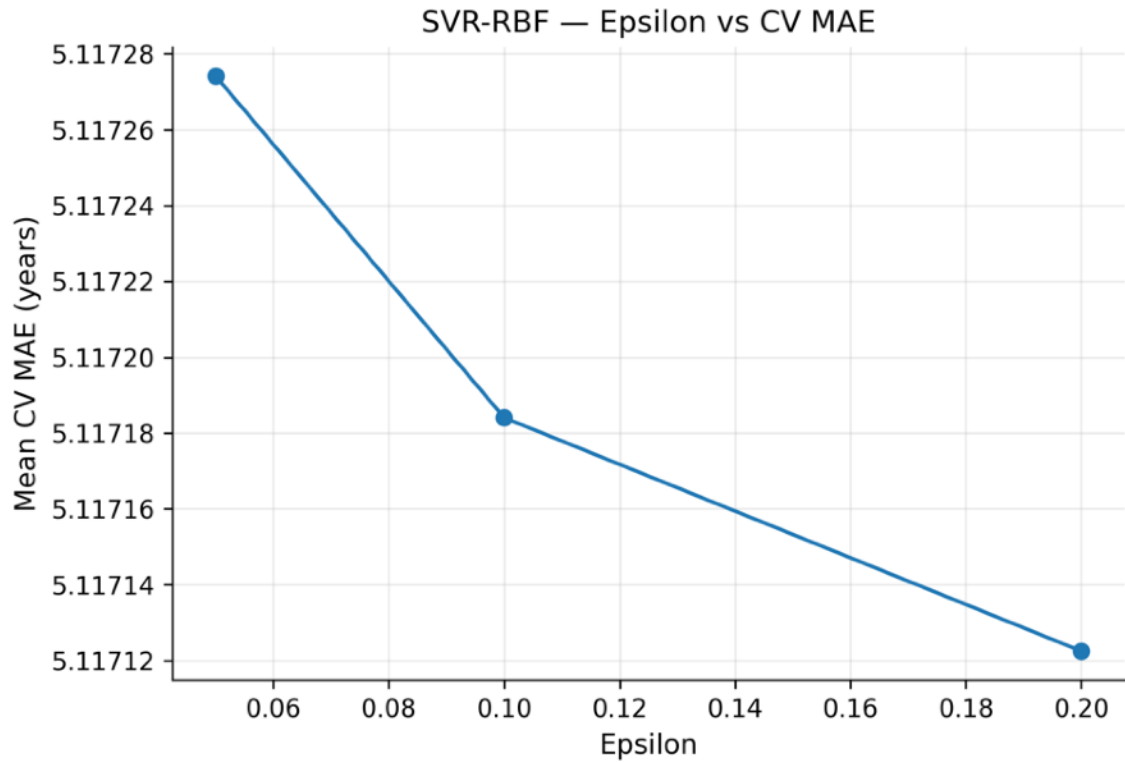


Figure 4.6: Effect of the epsilon parameter on SVR cross-validation performance.

For Elastic Net regression (Figure 4.7), both the regularization strength (Alpha) and the L1 ratio influenced predictive accuracy. Across all tested L1 ratios, performance improved as α increased from 10^{-2} to intermediate values.

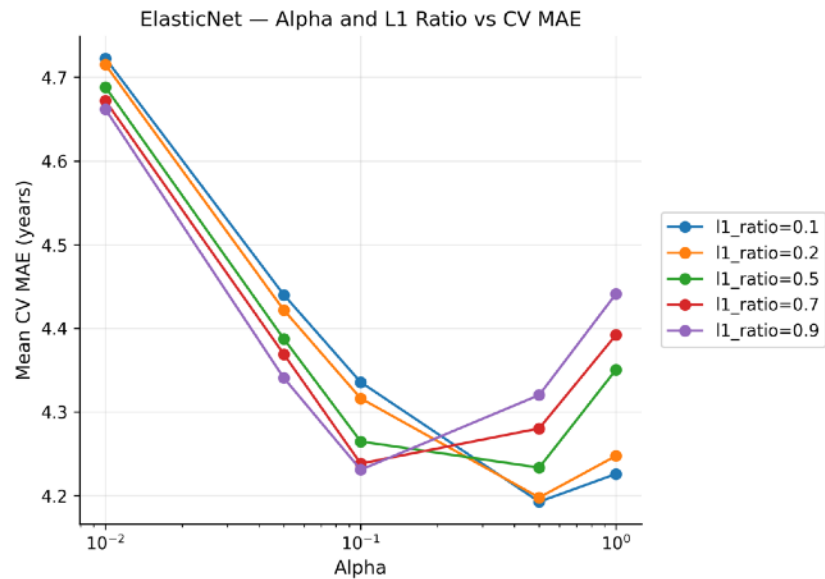


Figure 4.7: Cross-validation MAE of the Elastic Net model across different values of Alpha and L1 ratio.

The lowest CV MAE was observed for moderate regularization strengths, while larger values of α led to a gradual deterioration in performance, particularly for higher L1 ratios. Overall, the results suggest that a balance between L1 and L2 regularization produced the most favorable predictive performance.

The Ridge Regression model demonstrated a clear dependence on the regularization parameter Alpha (Figure 4.8). Increasing α from very small values substantially reduced CV MAE, with performance improving continuously until approximately $\alpha=500$. Beyond this point, only minimal changes were observed, suggesting that the model reached a stable operating region for larger regularization strengths.

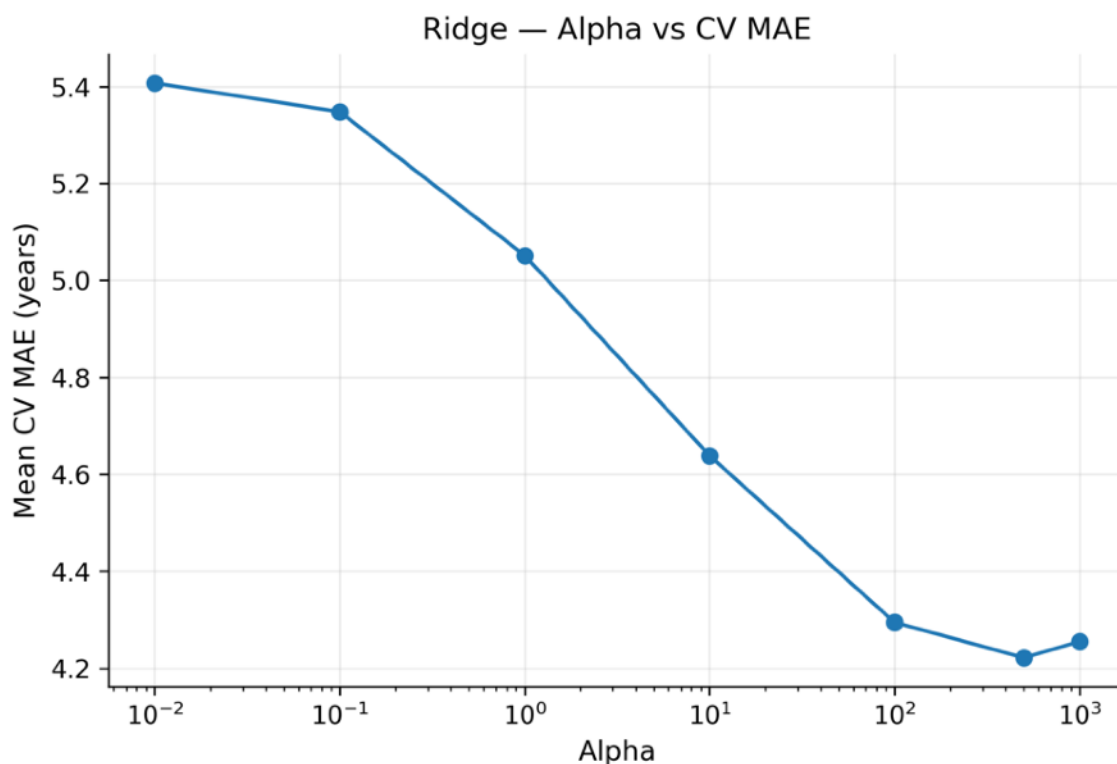


Figure 4.8: Cross-validation MAE of the Ridge Regression model as a function of the regularization parameter Alpha.

Finally, the GPR model showed limited sensitivity to most kernel configurations evaluated in this study (Figure 4.9). The majority of tested kernel parameter combinations achieved identical cross-validation performance (CV MAE = 4.123 years), indicating that the predictive accuracy of the model was largely robust to

the selected kernel settings within the explored search space. Only a small subset of configurations yielded noticeably higher errors.

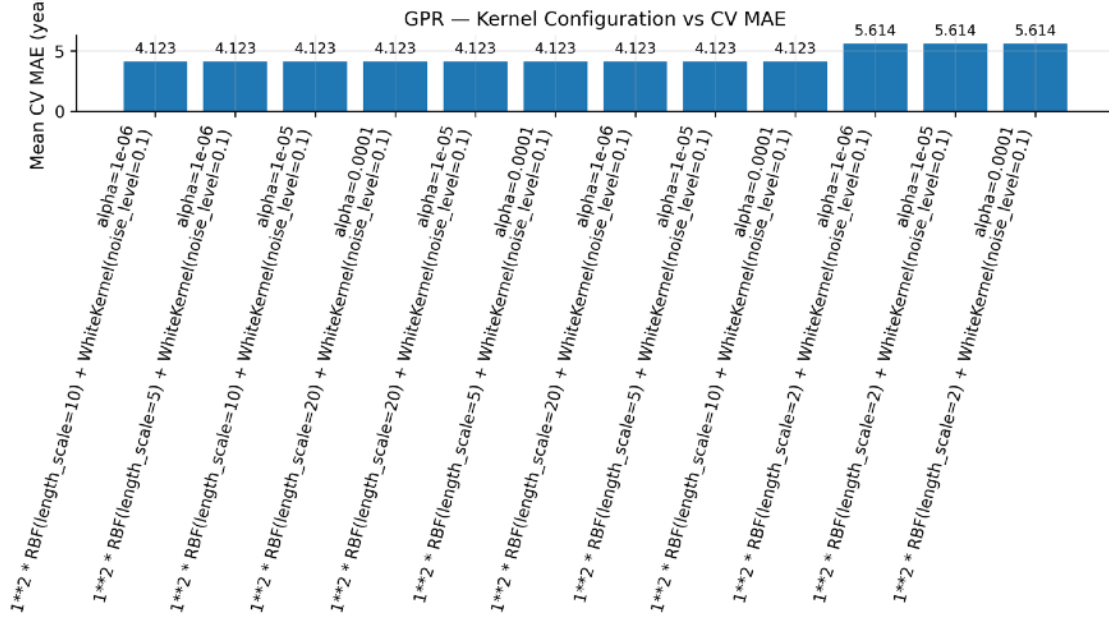


Figure 4.9: Cross-validation performance of the evaluated Gaussian Process Regression kernel configurations.

4.3 Explainability Results

To investigate the neuroimaging features contributing to brain age prediction, multiple explainability approaches were applied to the trained models. Global feature importance was assessed using SHAP values, permutation importance, and Elastic Net coefficients, while the robustness of SHAP-based rankings was evaluated across multiple random seeds. In addition, local explanations were examined for representative subjects with highly negative, near-zero, and highly positive BAG values. The results of these analyses are presented in the following sections.

4.3.1 Global Feature Importance

Global feature importance was investigated using SHAP values, permutation importance, and Elastic Net coefficients. Figures 4.10–4.13 summarize the most influential features identified by each approach.

The Random Forest SHAP analysis identified White Matter Hypointensities (SV) as the most influential feature, followed by ventricular measures including the right inferior lateral ventricle, third ventricle, and left inferior lateral ventricle (Figure 4.10). Additional highly ranked features included the left accumbens area, bilateral lateral ventricles, superior temporal regions, thalamus, precuneus, choroid plexus, and hippocampus.

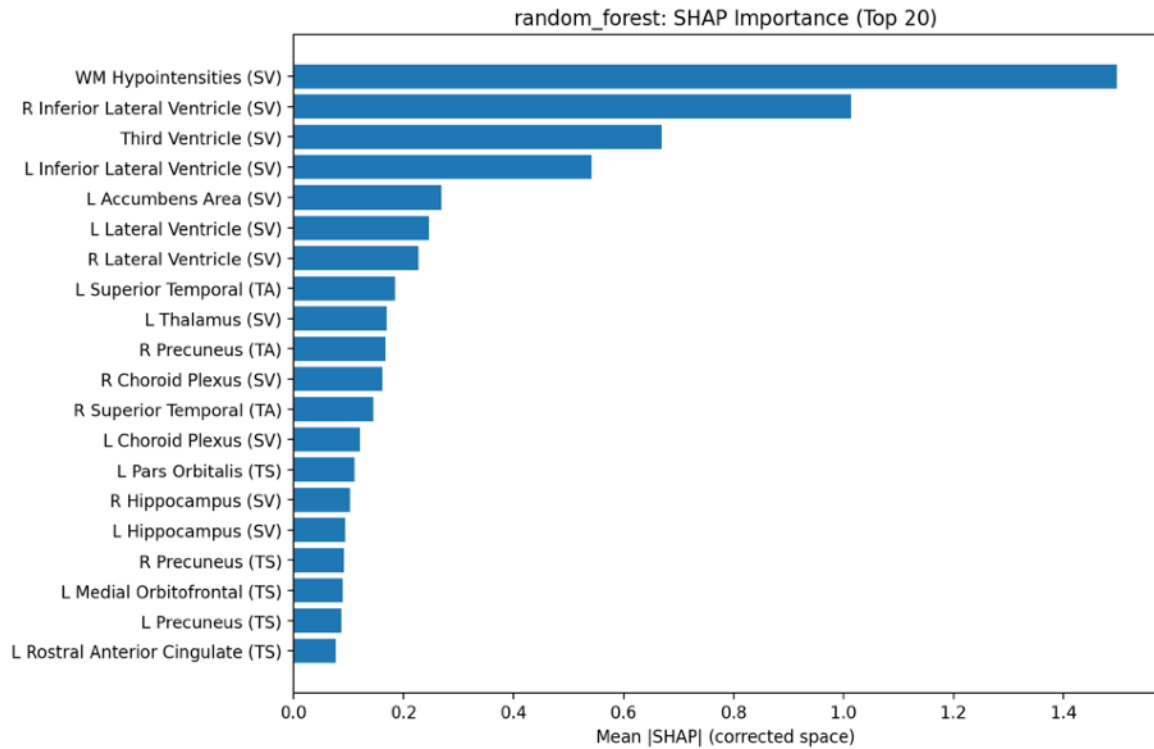


Figure 4.10: Top 20 Random Forest features ranked by mean absolute SHAP value.

The SHAP summary (beeswarm) plot (Figure 4.11) further illustrated the contribution of individual feature values to model predictions. Features identified as highly important in the SHAP ranking also exhibited the largest spread of SHAP values, indicating a stronger influence on predicted brain age across subjects.

Permutation importance analysis produced a similar ranking (Figure 4.12). White Matter Hypointensities (SV) remained the most influential feature, followed by ventricular measures and subcortical structures. Several cortical regions, including temporal, orbitofrontal, and precuneus regions, were also consistently represented among the highest-ranked features.

For the Elastic Net model, feature importance was assessed using standardized model coefficients (Figure 4.13). The largest positive coefficients were

associated with the third ventricle, White Matter Hypointensities (SV), brainstem volume, and orbitofrontal regions, whereas the strongest negative coefficients were observed for the accumbens area, thalamus, hippocampus, precuneus, and superior temporal regions.

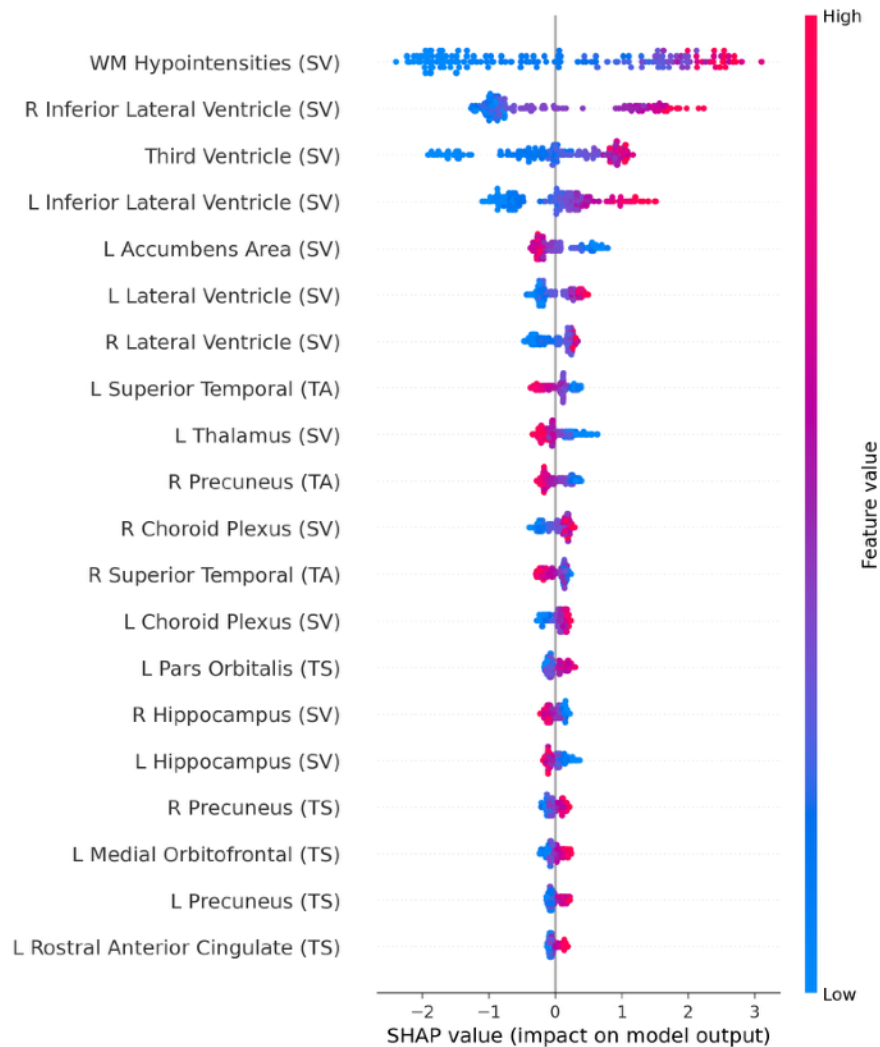


Figure 4.11: SHAP summary (beeswarm) plot showing the distribution of feature contributions across subjects.

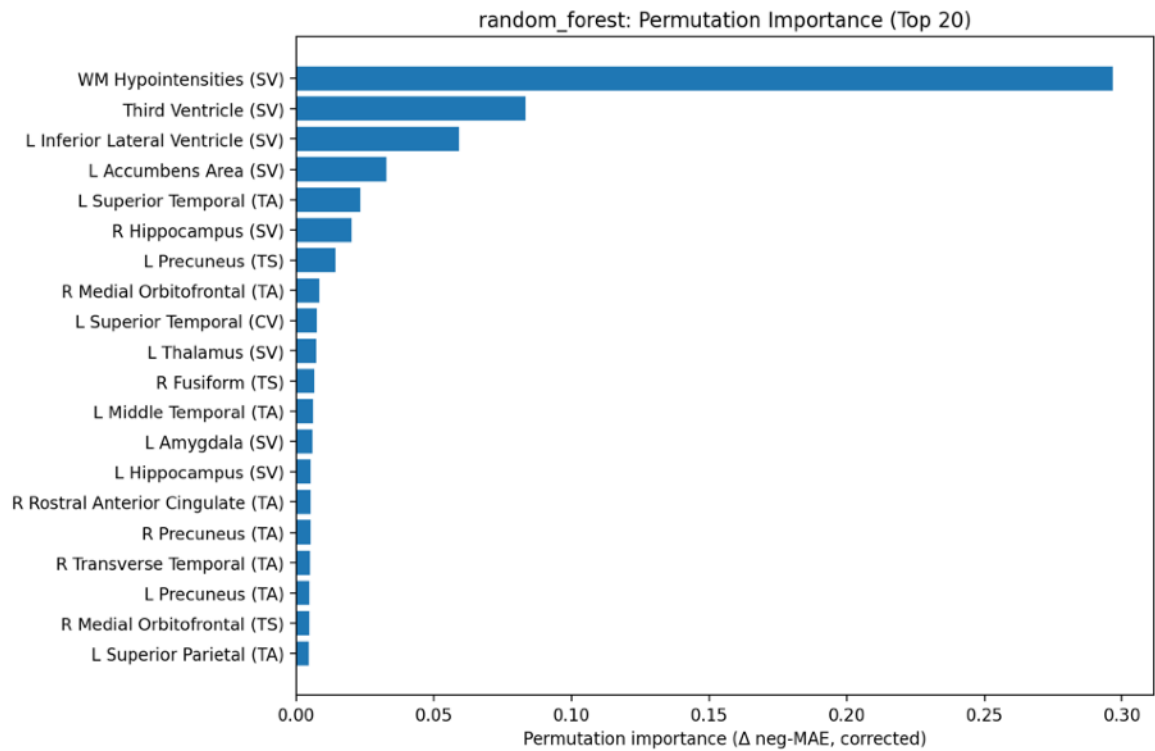


Figure 4.12: Top 20 Random Forest features ranked by permutation importance.

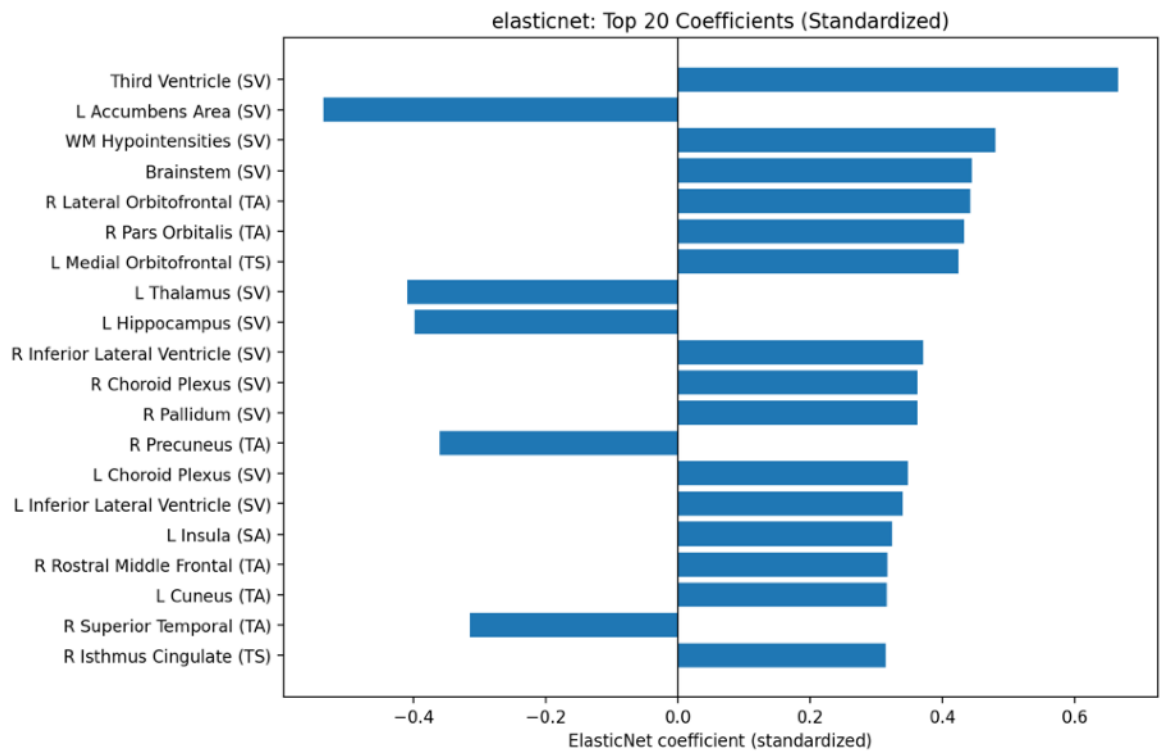


Figure 4.13: Top positive and negative standardized Elastic Net coefficients.

To identify features that were consistently important across multiple explainability methods, a consensus ranking was computed (Figure 4.14). White Matter Hypointensities (SV), ventricular measures, the accumbens area, thalamus, hippocampus, precuneus, choroid plexus, and superior temporal regions appeared among the highest-ranked features in the combined analysis.

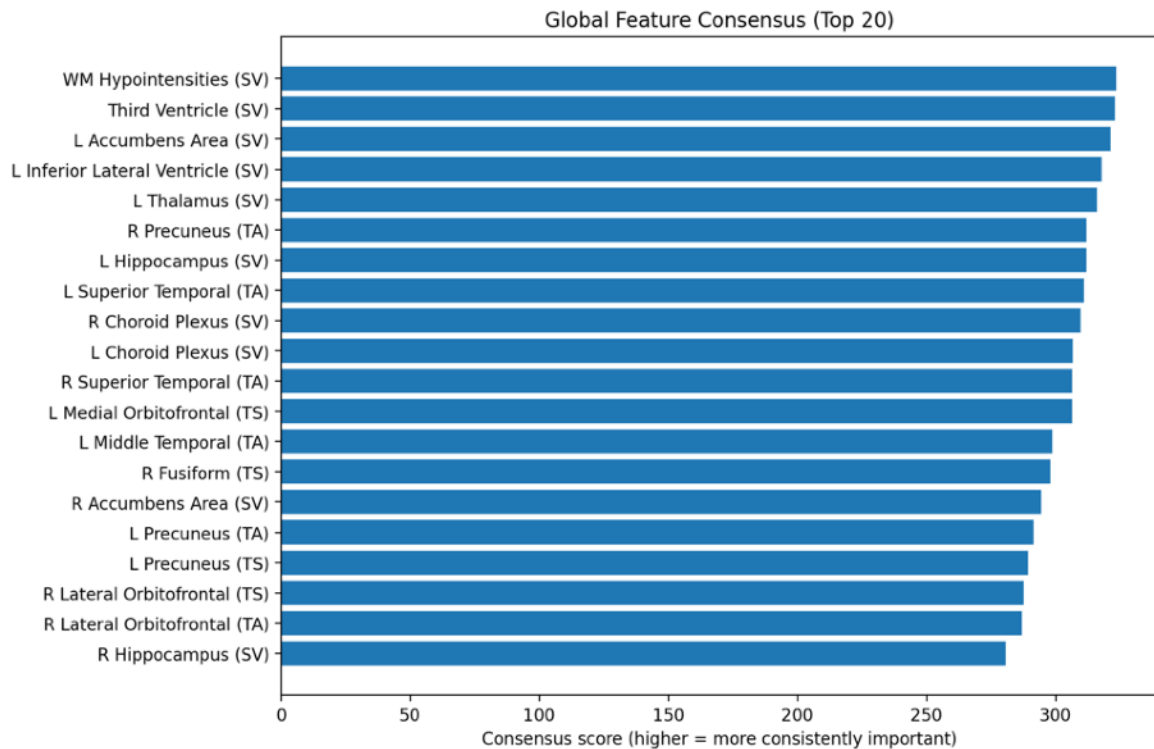


Figure 4.14: Consensus feature ranking derived from SHAP values, permutation importance, and Elastic Net coefficients.

To quantify the agreement between the different explainability approaches, the overlap between the top 20 features identified by SHAP, permutation importance, and Elastic Net coefficients was evaluated (Table 4.2). SHAP and Elastic Net exhibited the greatest agreement, sharing 12 of their top 20 features (Jaccard similarity = 0.429), followed by SHAP and permutation importance, which shared 10 features (Jaccard similarity = 0.333). Permutation importance and Elastic Net showed lower agreement, with 7 common features (Jaccard similarity = 0.212). Overall, seven features were identified among the top 20 features by all three methods.

Comparison	Shared Features	Jaccard Similarity
SHAP vs Permutation Importance	10	0.333
SHAP vs Elastic Net	12	0.429
Permutation Importance vs Elastic Net	7	0.212
Common to all three methods	7	—

Table 4.2: Overlap and Jaccard similarity between SHAP, permutation importance, and Elastic Net feature rankings

The seven features identified among the top 20 features by all three explainability methods were White Matter Hypointensities (SV), Third Ventricle (SV), Left Accumbens Area (SV), Left Hippocampus (SV), Left Thalamus (SV), Right Precuneus (TA), and Right Superior Temporal (TA)

4.3.2 SHAP Stability Analysis

To assess the robustness of the SHAP-based feature rankings, the Random Forest model was retrained using five different random seeds and feature importance rankings were compared.

Figure 4.15 presents the pairwise similarity matrix of SHAP importance values across seeds. Similarity values remained consistently high across all seed combinations, ranging from approximately 0.995 to 0.998. The mean pairwise correlation of SHAP importance values across seeds was 0.997, indicating a high degree of consistency in the resulting feature rankings.

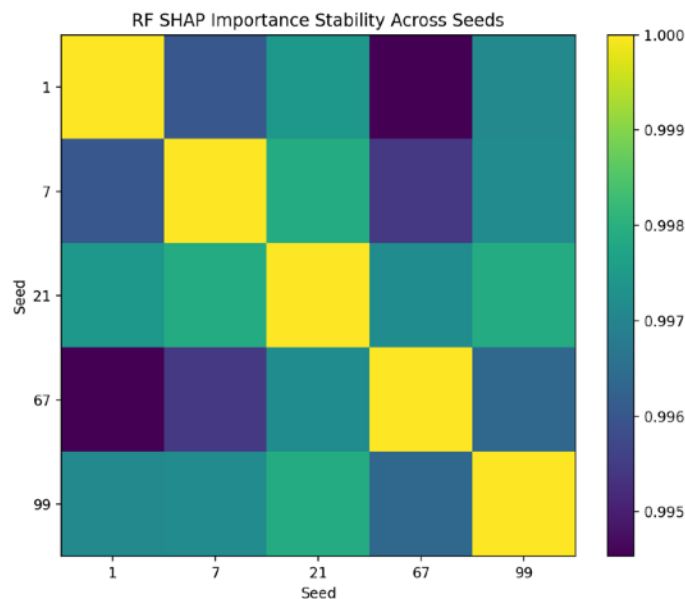


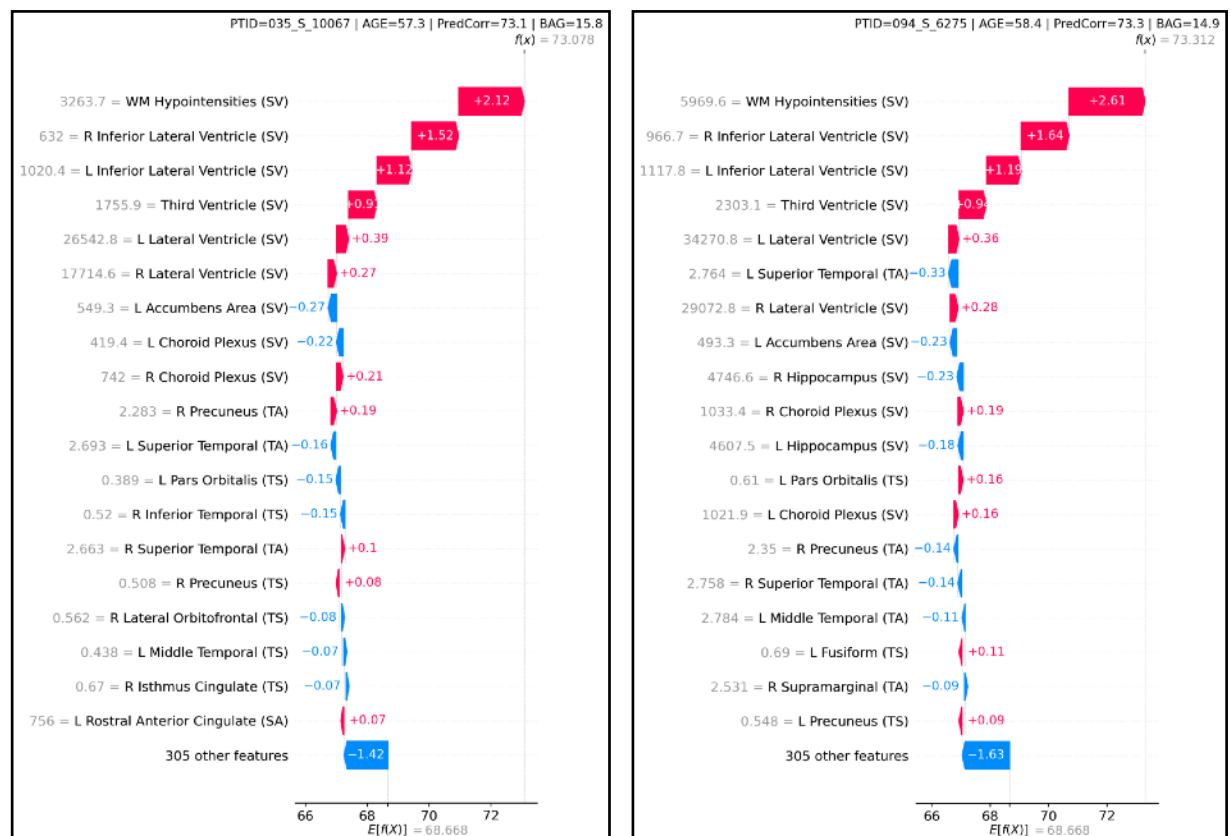
Figure 4.15: Pairwise correlation matrix of SHAP feature importance values across five random seeds.

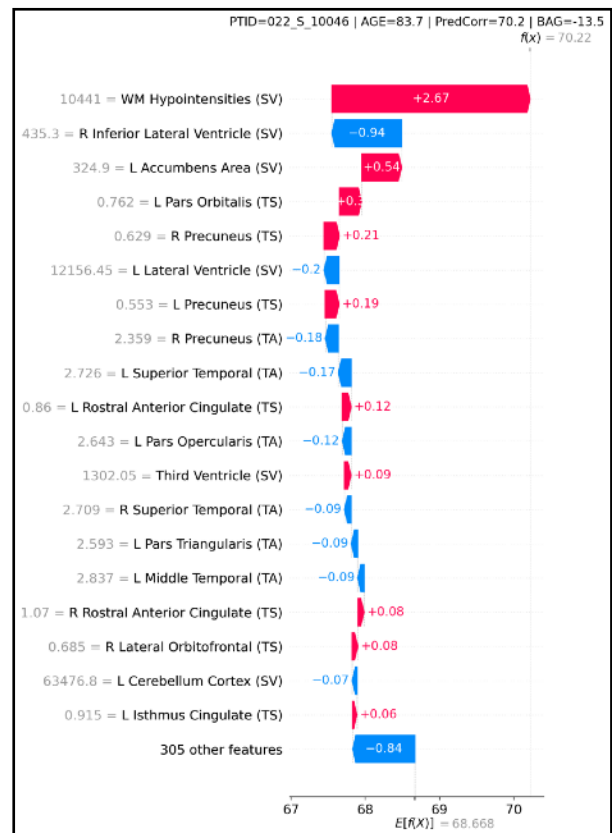
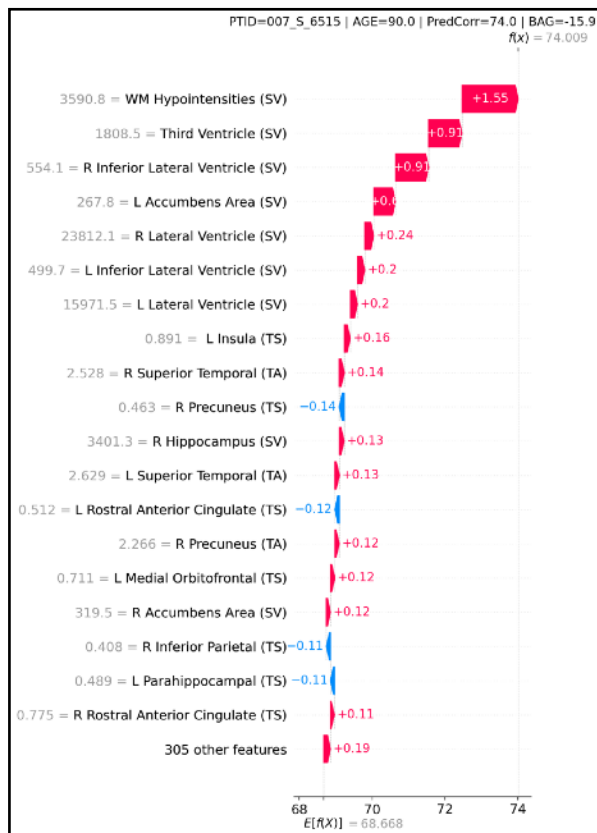
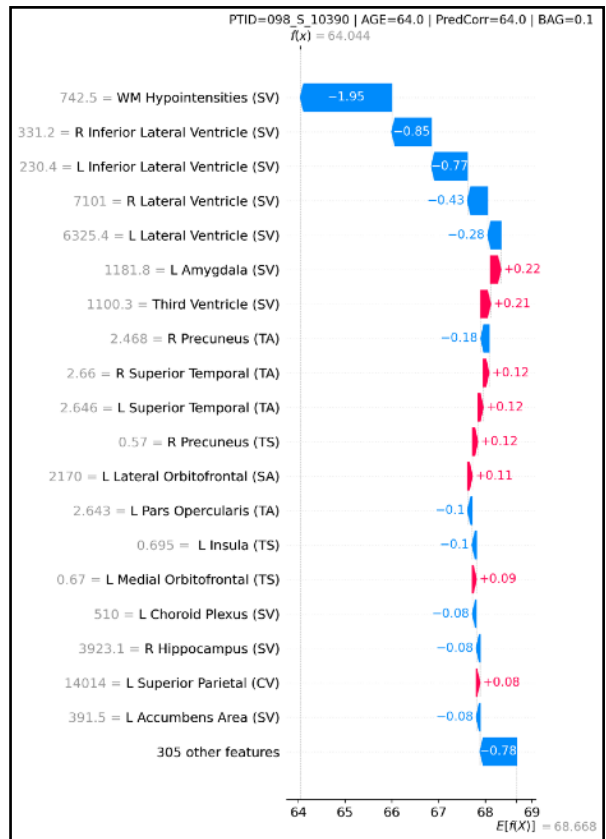
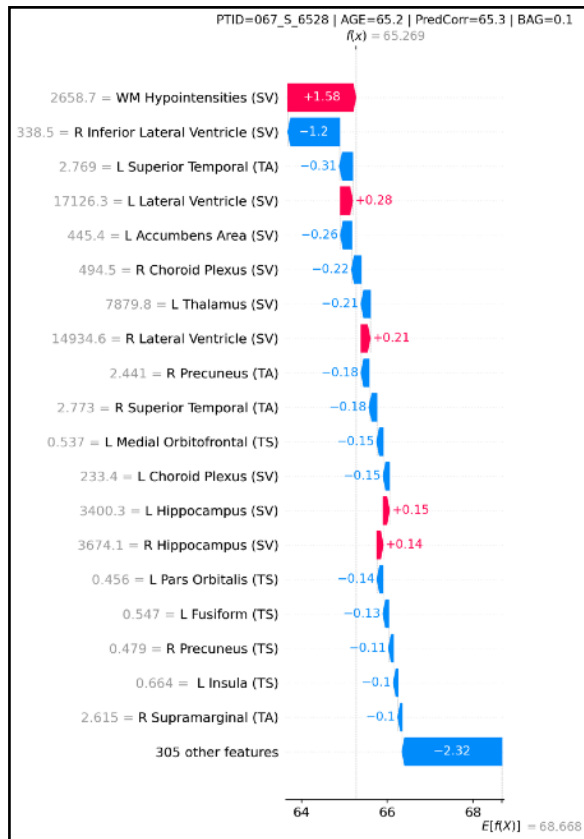
Stability was further evaluated by examining the overlap of the highest-ranked features across runs. The mean overlap among the top 10 features was 0.900, while the mean overlap among the top 20 features was 0.945. These results demonstrate that the most influential features identified by the Random Forest model remained highly consistent across different random initializations.

4.3.3 Local Explanations

Local explainability was examined using representative subjects selected from three Brain Age Gap (BAG) categories: highly negative BAG, near-zero BAG, and highly positive BAG.

Figures 4.16 – 4.21 present the local SHAP explanations for the selected subjects. For each case, the contribution of individual neuroimaging features to the final brain age prediction was visualized. Subjects with highly positive BAG values exhibited a greater contribution from features that increased predicted brain age, whereas subjects with highly negative BAG values showed a larger contribution from features that reduced predicted brain age. Cases with BAG values close to zero displayed a more balanced distribution of positive and negative feature contributions.





Figures 4.16 - 4.21: SHAP waterfall plots for representative subject cases: 2 with highly negative BAG, 2 with near zero BAG and 2 with highly positive BAG.

4.4 Longitudinal Stability Evaluation

The longitudinal evaluation was conducted to assess both the short-term stability and long-term sensitivity of the investigated brain age prediction models. Short-term analyses were based on repeated scans acquired within six months, whereas long-term analyses utilized scan pairs separated by approximately three years.

4.4.1 Short-Term Stability

Short-term test–retest results are summarized in Table 4.3. All models demonstrated strong agreement between repeated measurements, with Pearson correlation coefficients exceeding 0.88 and Concordance Correlation Coefficients exceeding 0.88 across both longitudinal subsets.

The Random Forest model exhibited the highest short-term stability, achieving Pearson correlations of 0.951 and 0.954 in the raw and clean subsets, respectively. It also produced the smallest mean absolute BAG differences, below one year in both subsets. Similarly, the SVR model demonstrated strong repeatability, with correlations above 0.93 and mean absolute BAG differences close to one year.

Linear models showed slightly lower stability. Elastic Net and Ridge achieved correlations ranging from 0.90 to 0.92, with mean absolute BAG differences between 2.3 and 2.5 years. GPR produced the lowest repeatability among the evaluated models, although correlation and agreement metrics remained high overall.

Across all models, results obtained from the clean subset were marginally improved relative to the raw subset.

Model	Raw Subset				Clean Subset			
	N	r	CCC	MAE Diff	N	r	CCC	MAE Diff
Elastic Net	206	0.902	0.900	2.485	183	0.910	0.910	2.396
Ridge	206	0.906	0.904	2.410	183	0.915	0.914	2.330
Random Forest	206	0.951	0.947	0.899	183	0.954	0.950	0.893
SVR	206	0.934	0.933	1.180	183	0.956	0.953	1.038
GPR	206	0.881	0.880	1.816	183	0.897	0.896	1.726

Table 4.3: Short-term test–retest stability results for all brain age prediction models.

4.4.2 Long-Term Sensitivity

Long-term sensitivity results are presented in Table 4.4. Correlations between chronological age change and predicted brain age change were generally modest across all models and longitudinal subsets.

Random Forest achieved the highest correlations, reaching 0.125 in the full longitudinal cohort and 0.158 in the clean subset. It also produced the lowest prediction error for age change, with MAE values close to two years. Ridge and Elastic Net achieved similar performance, with correlations around 0.08–0.11 and MAE values between 2.9 and 3.2 years.

SVR and GPR demonstrated weaker associations between chronological age change and predicted age change. Correlation coefficients remained close to zero across all subsets, despite MAE values comparable to those observed in the linear models.

Restricting the analysis to the clean subset generally resulted in slightly higher correlations and lower prediction errors, whereas the tight subset produced similar overall conclusions.

Model	All			Clean			Tight		
	N	r	MAE Change	N	r	MAE Change	N	r	MAE Change
Elastic Net	514	0.080	3.054	384	0.106	2.976	329	0.066	3.167
Ridge	514	0.080	3.002	384	0.108	2.922	329	0.071	3.083
Random Forest	514	0.125	2.061	384	0.158	1.982	329	0.033	2.033
SVR	514	0.000	2.339	384	-0.036	2.313	329	0.008	2.506
GPR	514	0.025	2.412	384	0.021	2.351	329	0.034	2.508

Table 4.4: Long-term sensitivity results across all longitudinal subsets.

Although correlations between chronological age change and predicted age change were generally low across all models, positive mean predicted age changes were observed in every case (Table 4.5). The average chronological age increase in the clean longitudinal subset was 2.72 years. Ridge and Elastic Net produced mean predicted age changes of 2.59 and 2.53 years, respectively, whereas Random Forest, SVR, and GPR produced smaller average increases of 1.06, 0.76, and 1.57 years. These findings indicate that all models detected age-related progression at the cohort level, although substantial differences were observed in the magnitude of the predicted change.

Model	Pearson r	MAE Change (years)	Mean Δ Age (years)	Mean Δ Predicted Age (years)
Ridge	0.108	2.92	2.72	2.59
Elastic Net	0.106	2.97	2.72	2.53
Random Forest	0.158	1.98	2.72	1.06
SVR	-0.036	2.31	2.72	0.76
GPR	0.021	2.35	2.72	1.57

Table 4.5: Comparison of observed and predicted age progression in the clean longitudinal subset.

4.5 Normative Modelling on Clinical Subjects

To evaluate whether the trained brain age model captured deviations from normative aging trajectories, the Random Forest normative model was applied to an independent cohort containing cognitively normal (CN), mild cognitive impairment (MCI), and Dementia subjects. Brain Age Gap (BAG) values were subsequently analysed across diagnostic groups in order to determine whether subjects with cognitive impairment exhibited evidence of accelerated brain aging relative to the healthy reference population.

The normative modeling cohort consisted of the 159 CN test subjects, 550 MCI subjects, and 166 Dementia subjects. Mean chronological age increased across diagnostic groups, with average ages of 68.78 years for CN subjects, 71.39 years for MCI subjects, and 73.60 years for Dementia subjects (Table 4.6).

Group	N	Mean Age	SD	Median Age
CN	159	68.78	7.16	68.34
MCI	550	71.39	7.75	71.85
Dementia	166	73.60	8.11	74.69

Table 4.6: Demographic characteristics of the normative modeling cohort across diagnostic groups.

Group-level BAG statistics revealed a progressive increase in predicted brain age across diagnostic categories (Table 4.7 and Figure 4.22). CN subjects exhibited a mean BAG close to zero (-0.20 years), indicating that predictions were well calibrated for the healthy reference population. In contrast, MCI subjects showed a positive mean BAG of 0.95 years, while Dementia subjects exhibited the largest mean BAG of 2.34 years. The boxplot distributions further demonstrated an upward shift in BAG values from CN to MCI and Dementia groups.

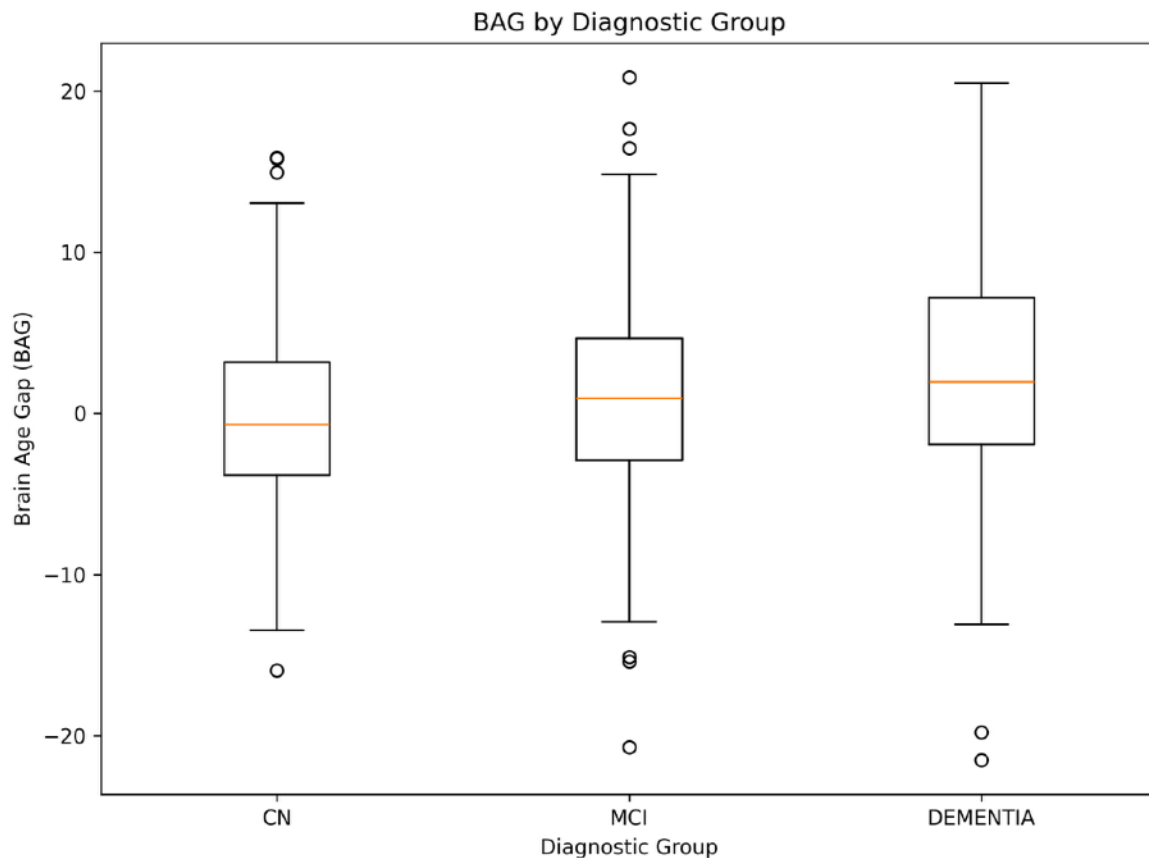


Figure 4.22: Distribution of Brain Age Gap (BAG) across diagnostic groups. Positive BAG values indicate brains predicted to be older than chronological age.

Group	N	Mean BAG	SD	Median BAG
CN	159	-0.20	6.08	-0.69
MCI	550	0.95	5.71	0.92
Dementia	166	2.34	7.11	1.93

Table 4.7: Brain Age Gap (BAG) statistics across diagnostic groups. Positive BAG values indicate brains predicted to be older than chronological age.

The association between diagnostic status and BAG was further evaluated using an ordinary least squares regression model adjusting for chronological age and sex. Relative to CN subjects, MCI subjects exhibited significantly elevated BAG (Table 4.8) values ($\beta = 2.54$), while Dementia subjects showed an even larger increase ($\beta = 5.29$). These findings indicate that the relationship between diagnostic status and BAG remained significant after accounting for demographic co-variables.

Predictor	β	95% CI	p-value
MCI	2.54	[1.85, 3.23]	< 0.001
Dementia	5.29	[4.43, 6.14]	< 0.001
Age	-0.62	[-0.65, -0.58]	< 0.001
Sex	-1.19	[-1.71, -0.67]	< 0.001

Table 4.8: Results of the ordinary least squares regression model examining the association between diagnostic group and Brain Age Gap (BAG) after adjustment for age and sex. Cognitively normal (CN) subjects served as the reference category.

Overall, normative model predictions revealed progressively larger BAG values in clinically impaired groups. Both the descriptive analyses and the adjusted regression model demonstrated higher BAG values in MCI and Dementia subjects compared with cognitively normal individuals.

4.6 Summary of Findings

This chapter presented the results obtained from the machine learning models developed for brain age prediction. Among the evaluated models, Random Forest achieved the strongest overall predictive performance, obtaining the lowest prediction error after age correction. Hyperparameter optimization analysis demonstrated that model performance was influenced differently by individual hyperparameters, with some algorithms exhibiting greater sensitivity to parameter selection than others.

Explainability analyses revealed consistent identification of biologically plausible brain regions and structures associated with aging. The stability analysis demonstrated high reproducibility of the SHAP-based feature importance rankings across different random seeds, supporting the robustness of the interpretability results.

Longitudinal evaluation showed strong short-term test–retest reliability for all models, particularly for Random Forest and SVR. In the long-term analysis, all models exhibited positive mean increases in predicted brain age over time, although associations between chronological age change and predicted age change remained relatively weak.

Finally, normative modeling demonstrated progressively higher Brain Age Gap values across diagnostic groups. Both MCI and Dementia subjects exhibited significantly elevated BAG compared with cognitively normal individuals after ad-

justment for age and sex, suggesting that accelerated brain aging is associated with cognitive impairment and dementia.

The implications of these findings are discussed in the following chapter.

5 Discussion

This chapter interprets the findings presented in Chapter 4 and discusses their methodological, biological, and clinical implications. It examines the performance of the evaluated models, the effect of hyperparameter optimization, the neuroanatomical relevance of the explainability findings, the longitudinal stability of the predictions, and the normative modelling results in clinical subjects. The chapter also relates the findings to previous brain age prediction literature and highlights the strengths, limitations, and broader significance of the study.

5.1 Models Performance

The results demonstrate that all five machine learning models achieved remarkably similar levels of predictive accuracy during model development, as reflected by the narrow range of cross-validation MAE values. This finding suggests that the FreeSurfer-derived structural MRI features contained a robust age-related signal that could be exploited effectively by multiple learning algorithms. In brain age prediction studies, feature representation is often a major determinant of performance, and when informative neuroanatomical features are available, differences between model architectures may become less pronounced [6, 15].

Although cross-validation performance was highly comparable across models, differences emerged when evaluation was performed on the independent test set. Random Forest achieved the lowest test-set MAE, whereas Elastic Net and Ridge Regression produced the strongest correlations with chronological age. The fact that different models performed best according to different metrics highlights an important consideration in brain age research: no single metric fully characterizes model quality. Mean Absolute Error measures the average magnitude of prediction errors, while Pearson correlation evaluates the degree to which predicted ages preserve age-related ordering across individuals. Consequently, models with similar prediction errors may exhibit substantially different correlation structures, and vice versa.

The strong test-set performance of Random Forest (MAE after bias correction: 4.737) may reflect its ability to model nonlinear relationships between structural brain features and chronological age. Neurobiological aging is characterized by heterogeneous trajectories across brain regions, with different cortical and sub-cortical structures exhibiting distinct rates and patterns of age-related change [5, 6]. Consequently, methods capable of capturing nonlinear feature interactions may offer advantages when modeling complex age-related variation in neuroanatomical features. Previous comparative work using regional structural MRI features showed that Random Forest achieved test-set MAEs of 2.83 years in HCP, 8.99 years in Cam-CAN, and 8.99 years in IXI, compared with the best-performing values of 2.76, 7.08, and 8.05 years, respectively [16].

In contrast, Elastic Net and Ridge Regression achieved the highest correlations with chronological age (Ridge r : 0.575; Elastic Net r : 0.584) despite exhibiting larger Brain Age Gap (BAG) variability. One possible interpretation is that regularized linear models captured the overall age-related structure of the morphometric data effectively. However, a high correlation between predicted and chronological age does not necessarily imply low individual prediction error. Accordingly, predictions may remain strongly associated with chronological age while still showing substantial variability in subject-level BAG values. The present findings are consistent with this interpretation, as both linear models demonstrated relatively high correlations but also the largest BAG dispersion among the evaluated methods.

The analysis of BAG dispersion provides additional information beyond conventional prediction metrics. In the brain age literature, BAG is frequently interpreted as a potential marker of accelerated or delayed brain aging and is often used in downstream clinical analyses [4, 5]. Consequently, the stability and variability of BAG estimates may be as important as prediction accuracy itself. The relatively low BAG dispersion observed for SVR and Random Forest suggests that these models produced more concentrated prediction errors across individuals, potentially yielding more stable estimates of age-related deviation. Although lower BAG variability does not necessarily imply superior predictive performance, it may improve the reliability of subsequent analyses that rely on BAG as a biological marker.

An interesting observation was that bias correction increased MAE for most models, from 4.709–4.995 years before correction to 4.737–5.641 years after correction, while leaving correlation coefficients unchanged. This outcome is not unexpected. The primary objective of bias correction is not to improve age prediction accuracy but rather to remove systematic age-related dependencies from BAG estimates [10]. Age-dependent prediction bias is a well-established phenomenon in brain age modelling, typically causing younger individuals to be predicted as older and older individuals to be predicted as younger [5,10]. Correcting this bias improves the interpretability and validity of BAG measurements by reducing their dependence on chronological age, even if it does not always lead to improvements in conventional prediction error metrics. The present findings therefore suggest a trade-off between minimizing prediction error and obtaining statistically valid BAG estimates for downstream analyses.

While the observed MAE values are broadly comparable to those reported in several studies using regional FreeSurfer-derived features and traditional machine learning methods, the correlation coefficients obtained in the present study were relatively modest compared with those achieved by some contemporary brain age models [16,17]. Several factors may contribute to this result. First, Pearson correlation is strongly influenced by the age distribution of the dataset, and reduced age variability limits the maximum attainable correlation even when prediction errors remain relatively low [12]. Second, the present study relied exclusively on summary regional morphometric features rather than higher-dimensional voxel-wise representations or deep learning approaches, which often achieve stronger age correlations at the cost of reduced interpretability and substantially greater computational complexity [7, 19]. Therefore, the moderate correlation values observed here should be interpreted within the methodological context of an explainable and computationally efficient modelling framework.

Overall, the results indicate that no single model consistently dominated across all evaluation criteria. Instead, each algorithm exhibited distinct strengths. Random Forest achieved the lowest prediction error, Elastic Net produced the strongest age correlation, and SVR generated the most concentrated BAG distribution. This pattern suggests that model selection in brain age prediction should be guided not only by predictive accuracy but also by the intended downstream appli-

cation. Models optimized for age prediction, biomarker development, explainability, or clinical interpretation may not necessarily be the same. The findings therefore support the use of a multi-metric evaluation framework when assessing brain age models and reinforce the importance of considering both predictive and biological characteristics of model outputs.

5.2 Comparison with State-of-the-Art and Alternative Methodologies

The performance of the models developed in this thesis should be interpreted in relation to state-of-the-art brain age prediction approaches. The Pearson correlations obtained in the present study were moderate, ranging approximately between 0.54 and 0.58 across the evaluated models. Although these values are lower than those reported in some morphometric brain age studies, they should be interpreted in the context of dataset characteristics. Han et al. reported correlations of 0.63–0.79 in the IXI cohort and 0.64–0.85 in Cam-CAN, but also substantially lower correlations of 0.11–0.42 in the HCP cohort despite low prediction error [16]. This pattern suggests that Pearson correlation can vary markedly across datasets and is strongly influenced by factors such as age range and cohort composition. In the present thesis, the use of an older ADNI sample with a relatively restricted age distribution likely reduced between-subject age variance, thereby limiting the attainable correlation despite reasonable prediction error.

In terms of MAE, the present models achieved test-set errors below five years, which remain above the performance reported by state-of-the-art image-based deep learning approaches but are not outside the broader range observed in morphometric machine learning studies. Peng et al. reported MAE values of 2.14 years in UK Biobank and 2.90 years in the PAC 2019 challenge using a light-weight 3D convolutional neural network trained directly on T1-weighted MRI data [27]. Similarly, De Bonis et al. reported independent test-set MAE values of 3.21 for a morphometric deep neural network pipeline and 3.08 for an image-based DenseNet-121 architecture [19]. However, Han et al. reported substantially higher MAE values in morphometric models applied to the Cambridge Centre for Ageing and Neuroscience (Cam-CAN) dataset (7.08–10.50 years) and the Information eX-

traction from Images (IXI) dataset (8.04–9.86 years), while reporting considerably lower MAE values for the Human Connectome Project (HCP) dataset [16]. Taken together, these findings indicate that the present thesis did not achieve state-of-the-art predictive accuracy, particularly relative to large image-based deep learning models, but nevertheless produced competitive performance within the context of interpretable morphometric machine learning, while additionally incorporating explainability, longitudinal reliability analysis, and normative clinical evaluation.

However, this comparison should not be interpreted as a direct like-for-like benchmark. Brain age prediction performance depends strongly on dataset composition, sample size, age range, scanner and site variability, preprocessing strategy, input representation, model architecture, and validation design. This is evident in the study by Han et al., where the same morphometric machine learning framework produced markedly different performance across datasets: the younger and age-restricted HCP dataset showed low MAE but weak correlations (MAE = 2.75–3.12 years; $r = 0.11$ – 0.42), whereas the broader age-range Cam-CAN dataset showed higher correlations but larger errors (MAE = 7.08–10.50 years; $r = 0.64$ – 0.85), as did the IXI dataset (MAE = 8.04–9.86 years; $r = 0.63$ – 0.79) [17].

The main advantage of the present methodology is not state-of-the-art prediction accuracy, but interpretability and methodological transparency. Image-based deep learning models can learn spatially rich representations directly from MRI volumes and generally achieve stronger predictive performance when trained on large datasets [27]. However, these models are usually more complex and less directly interpretable than feature-based approaches. In contrast, the present thesis used FreeSurfer-derived regional morphometric features, which correspond to named cortical and subcortical structures. This allowed the resulting predictions to be interpreted through SHAP values, permutation importance, and Elastic Net coefficients, linking model behaviour to biologically meaningful anatomical patterns.

This interpretability-oriented design is consistent with recent work emphasizing that brain age models should be evaluated not only by accuracy, but also by the interpretability and clinical usability of their explanations. De Bonis et al. directly compared morphometric and deep learning brain age pipelines and found strong performance for both approaches, while also showing that SHAP provided more consistent and interpretable explanations than some image-based XAI methods

[19]. This supports the broader rationale of the present thesis: although image-based deep learning approaches may provide superior predictive accuracy, morphometric feature-based models remain valuable when the objective includes anatomical interpretability, explanation stability, computational simplicity, and clinical contextualization.

Therefore, the present thesis should be viewed as complementary to state-of-the-art deep learning approaches rather than as a direct competitor to them. Deep learning and image-based models currently represent stronger approaches for minimizing MAE and maximizing correlation with chronological age. The present framework instead prioritizes transparent model comparison, interpretable feature attribution, hyperparameter sensitivity analysis, longitudinal reliability assessment, and normative modelling in clinical groups. This distinction is important because the practical value of a brain age model depends not only on its cross-sectional prediction accuracy, but also on whether its outputs are interpretable, reliable, and clinically meaningful.

5.3 Hyperparameter Optimization Analysis

The hyperparameter optimization results revealed substantial differences in how sensitive the evaluated machine learning models were to parameter selection. While certain algorithms exhibited pronounced performance changes across the explored search space, others remained remarkably stable. This finding is important because hyperparameter sensitivity influences not only predictive performance but also model robustness, reproducibility, and ease of deployment in future brain age prediction applications.

This observation is consistent with previous brain age prediction studies showing that methodological and modelling choices can influence predictive performance. Franke et al. explicitly examined the influence of different parameter settings in MRI-based age estimation using kernel methods, demonstrating that model configuration is an important component of brain age model development [14]. Similarly, comparative studies using structural morphometric data have shown that different machine learning algorithms can achieve different levels of performance depending on the feature representation and modelling strategy used

[15,16]. Within this context, the present hyperparameter analysis provides additional information beyond final model accuracy by showing how stable or sensitive each algorithm was across the explored search space.

Among the evaluated methods, Random Forest demonstrated relatively limited sensitivity to the explored hyperparameter combinations. Only modest differences in cross-validation MAE were observed across variations of `min_samples_leaf` and `max_features`, suggesting that the model was able to extract a stable age-related signal regardless of minor adjustments to tree complexity. The slightly improved performance obtained with smaller feature subsets at each split may reflect one of the core strengths of ensemble tree methods. By restricting the number of candidate features available during tree construction, greater diversity is introduced among individual trees, reducing correlation between ensemble members and potentially improving generalization performance. The overall stability observed for Random Forest is consistent with the robustness observed in the present experiments.

In contrast, SVR exhibited substantially greater sensitivity to hyperparameter selection. The most notable effect was observed for the interaction between the regularization parameter C and the kernel parameter γ . Performance improved considerably as C increased when γ was set to its default scale value, whereas alternative γ settings consistently produced inferior results. This behaviour is expected given the role of these parameters in controlling model flexibility. The parameter C determines the degree to which training errors are penalized, while γ controls the locality of the radial basis function kernel. In high-dimensional neuroimaging feature spaces, inappropriate γ values can lead either to excessive smoothing and underfitting or to highly localized decision functions that generalize poorly. The strong dependence on γ observed in the present study suggests that kernel configuration plays a particularly important role when applying SVR to FreeSurfer-derived morphometric features.

Interestingly, the epsilon parameter exerted almost no influence on predictive performance within the explored range. Since epsilon defines the width of the error-insensitive region used during optimization, this finding suggests that the selected values all fell within a range that adequately accommodated the intrinsic noise present in the data. The negligible performance differences imply that overall

model complexity and kernel configuration were substantially more influential determinants of predictive accuracy than the precise choice of epsilon. This observation is consistent with previous neuroimaging studies reporting that SVR performance is generally more sensitive to overall parameter selection and model configuration than to minor adjustments of individual hyperparameters [14].

The Elastic Net results demonstrated the importance of balancing model complexity and regularization. Performance improved as the regularization strength increased from very small values, indicating that insufficient regularization likely allowed the model to fit noise or unstable feature-specific variation. However, performance deteriorated when regularization became excessively strong, particularly for larger L1 ratios. This pattern is consistent with the fundamental objective of Elastic Net regression, which combines L1 and L2 penalties to simultaneously perform coefficient shrinkage and feature selection. Moderate regularization appears to have achieved the most effective trade-off between retaining informative neuroanatomical features and suppressing noise. The superior performance observed at intermediate parameter values suggests that age-related information was distributed across multiple correlated structural measurements rather than being concentrated within a small subset of highly dominant predictors.

A similar regularization effect was observed for Ridge Regression. Increasing the regularization parameter alpha produced a substantial reduction in prediction error before performance eventually plateaued. This behaviour indicates that the unregularized or weakly regularized models were likely overfitting feature-specific variation, whereas stronger regularization improved generalization by shrinking coefficient magnitudes and reducing variance. The existence of a broad plateau at larger alpha values is noteworthy because it suggests that the model became relatively insensitive to precise parameter selection once an appropriate level of regularization had been reached. Such behaviour is generally desirable from a practical perspective, as it reduces the risk of performance degradation due to minor variations in tuning procedures.

Perhaps the most striking finding emerged from the Gaussian Process Regression analysis. Across the majority of evaluated kernel configurations, nearly identical cross-validation performance was obtained. This result suggests that the predictive signal captured by the model was highly robust to the specific kernel pa-

parameterization explored in this study. One possible interpretation is that the dominant age-related structure present in the FreeSurfer feature space could already be adequately represented by a broad range of kernel configurations, resulting in little opportunity for further optimization. Alternatively, the explored kernel search space may have contained multiple parameter combinations that produced functionally similar covariance structures. Regardless of the precise explanation, the results indicate that model performance was primarily determined by the available data rather than by fine-grained kernel tuning.

The differences observed across models also reinforce the importance of transparent reporting of optimization procedures in brain age prediction. Previous work has emphasized that brain age estimates can be affected by methodological factors such as scanner variability, prediction bias, preprocessing pipelines, and feature representations [11,13,15]. Although the present analysis did not perform external validation across independent cohorts, the hyperparameter results show that algorithmic configuration itself represents another source of variability that can influence model performance. Reporting this behaviour is therefore important because two models with similar final MAE values may differ substantially in their sensitivity to parameter choices and, consequently, in their expected robustness when applied in future studies.

Taken together, the hyperparameter optimization results show that parameter tuning contributed differently to the performance of each algorithm. SVR and Ridge Regression derived substantial benefit from appropriate hyperparameter selection, whereas Random Forest and GPR exhibited greater robustness across the explored search space. This pattern is consistent with previous neuroimaging machine learning studies showing that model performance depends on the interaction between algorithm choice, feature representation, and optimization strategy [14,16]. Moreover, the relatively smooth performance trends observed across most models indicate that the selected FreeSurfer-derived features contained a stable age-related signal that could be recovered under a range of reasonable model configurations. These findings support the overall robustness of the developed brain age prediction framework while also emphasizing the need to report optimization procedures transparently.

5.4 Explainability Findings

A major objective of this thesis was not only to predict chronological age but also to understand which neuroanatomical features contributed to the resulting brain age estimates. This is particularly important in brain age research, where predictive performance alone provides limited insight into the biological processes underlying model predictions. By combining SHAP values, permutation importance, Elastic Net coefficients, consensus ranking, stability analysis, and local explanations, the present study sought to evaluate both the biological plausibility and robustness of the learned age-related patterns.

This approach is consistent with recent work emphasizing the role of explainability in brain age prediction. SHAP has been proposed as a useful framework for interpreting individual model predictions because it assigns feature-level contributions to each prediction [23], and previous brain age studies have used explainability methods to interpret anatomical contributors to predicted brain age [21,22]. More recently, comparative work has highlighted the need to evaluate not only the predictive accuracy of brain age models but also the interpretability of the learned neuroanatomical patterns, particularly when comparing morphometric and deep learning pipelines [19]. Within this context, the present analysis contributes by combining multiple complementary explanation methods rather than relying on a single attribution technique.

One of the most notable findings was the consistent importance of White Matter Hypointensities and ventricular measures across all explainability approaches. White Matter Hypointensities emerged as the highest-ranked feature in both SHAP and permutation importance analyses and also occupied the top position in the consensus ranking. Ventricular structures, including the third ventricle and inferior lateral ventricles, were likewise identified as highly influential predictors. The prominence of these features is biologically plausible and consistent with established neuroimaging studies of aging. Ventricular enlargement is among the most robust structural markers of age-related brain atrophy and has repeatedly been associated with both normal aging and neurodegenerative processes [2, 3]. Similarly, white matter abnormalities have been linked to aging, cerebrovascular burden, cognitive decline, and increased biological brain aging, suggesting that the

model captured meaningful age-related variation rather than relying on arbitrary statistical associations [1, 2].

Beyond ventricular and white matter measures, several subcortical and cortical regions were consistently identified as important. These included the hippocampus, thalamus, accumbens area, precuneus, superior temporal cortex, and orbitofrontal regions. Many of these structures have previously been implicated in age-related morphological change. The hippocampus is particularly well known for exhibiting age-related volume reduction and vulnerability to neurodegenerative processes, especially in Alzheimer's disease [2, 8]. Likewise, structural alterations in the thalamus, temporal cortex, precuneus, and frontal regions have been repeatedly reported in studies of healthy aging and age-related neurodegenerative processes[1, 3]. The convergence of these findings with existing neuroimaging literature provides indirect evidence that the models learned biologically meaningful representations of aging.

An important aspect of the present analysis was the comparison of multiple explainability approaches. Although the rankings were not identical, substantial overlap was observed between SHAP, permutation importance, and Elastic Net coefficients. Seven features were consistently identified among the top twenty features by all three methods, while SHAP and Elastic Net exhibited the greatest overall agreement. This level of convergence is noteworthy because the three approaches quantify feature importance using fundamentally different principles. SHAP evaluates feature contributions to individual predictions, permutation importance assesses performance degradation after feature perturbation, and Elastic Net coefficients quantify linear associations between predictors and age. Therefore, agreement across methods increases confidence that the identified features reflect genuine age-related information rather than artifacts of a particular explainability technique.

The consensus ranking further strengthened this interpretation by demonstrating that several neuroanatomical structures remained highly important regardless of the underlying explanation method. Rather than relying on a single feature or anatomical region, the models appeared to integrate information from multiple complementary markers of brain aging. This finding aligns with current views of

brain aging as a distributed process affecting numerous interconnected structures rather than a localized phenomenon confined to a single region [1, 3].

A particularly important result was the exceptionally high stability of SHAP-based feature importance rankings across random seeds. The mean pairwise correlation of 0.997 and the substantial overlap among the highest-ranked features indicate that the resulting explanations were highly reproducible. This finding has important implications for the trustworthiness of explainable artificial intelligence in neuroimaging. Recent work has emphasized the importance of evaluating the robustness and consistency of explainability methods when interpreting machine learning models [19]. In the present study, the identified feature rankings remained remarkably stable despite repeated model retraining, suggesting that the explanations reflected robust characteristics of the data rather than random fluctuations introduced by model initialization. Such stability is particularly valuable when explainability results are intended to support scientific interpretation rather than merely provide post-hoc visualization.

The local SHAP analyses provided additional support for the global findings. Across subjects with highly positive, near-zero, and highly negative Brain Age Gap values, the same core set of features repeatedly appeared among the strongest contributors to individual predictions. White Matter Hypointensities and ventricular measures were especially prominent in subjects exhibiting markedly elevated predicted brain age, where they consistently increased model output. This observation is consistent with the broader literature linking ventricular enlargement and white matter abnormalities to structural brain aging and neurodegenerative change [1, 2]. Importantly, the local explanations therefore reinforced the conclusions drawn from the global analyses, demonstrating that the same biologically plausible features influenced both population-level patterns and individual predictions.

This subject-level interpretation is consistent with previous work highlighting the value of individual explanations in brain age prediction. Sun et al. applied interpretability analysis to individual brain age predictions based on grey matter volume, showing that feature attribution can relate subject-level predictions to anatomical patterns [21]. In the present thesis, local SHAP explanations served a similar interpretive role by showing how individual BAG values were influenced by

specific FreeSurfer-derived features. However, these explanations should be interpreted as descriptive model-based interpretations rather than diagnostic evidence for individual subjects.

The local analyses also illustrated the heterogeneous nature of brain age estimation. Although common features appeared across subjects, the magnitude and direction of their contributions varied substantially between individuals. Subjects with highly positive Brain Age Gap values tended to exhibit stronger positive contributions from aging-associated structural markers, whereas subjects with highly negative Brain Age Gap values displayed a more balanced combination of positive and negative feature effects. This observation suggests that brain age predictions arise from the cumulative interaction of multiple anatomical characteristics rather than being determined by a single dominant biomarker. Such behaviour is consistent with the concept of biological brain aging as a multidimensional process involving simultaneous changes across several structural systems [1, 3].

Overall, the explainability analyses provide evidence that the developed models relied on neuroanatomical features with established relevance to brain aging and that these findings were reproducible across multiple explanation methods, random seeds, and individual subjects. The convergence of global and local explanations, combined with the exceptionally high stability of the SHAP rankings, supports the interpretability and biological plausibility of the resulting brain age predictions.

5.5 Longitudinal Stability

The short-term test–retest analysis demonstrated high agreement between repeated measurements acquired within six months. Across all models, Pearson correlation coefficients and Concordance Correlation Coefficients exceeded 0.88, indicating that predicted brain age remained largely consistent when minimal biological aging was expected to occur. These findings suggest that the investigated models primarily captured stable anatomical characteristics rather than random measurement noise.

This finding is consistent with the growing emphasis on test–retest reliability in brain age research. Dörfel et al. compared publicly available brain age prediction tools and highlighted that prediction accuracy and test–retest reliability should be evaluated together, since accurate models are not necessarily equally reliable across repeated scans [12]. Similarly, Cole et al. described deep learning-based brain age prediction from structural MRI as producing a reliable and heritable biomarker [7]. In this context, the high short-term agreement observed in the present study supports the reliability of the generated brain age estimates and strengthens their suitability for subsequent longitudinal and normative analyses.

Among the evaluated approaches, Random Forest exhibited the strongest short-term reliability, achieving the highest correlation and agreement values together with the smallest mean absolute BAG differences. The superior performance of Random Forest may be attributable to the ensemble nature of the method, which reduces variance by aggregating predictions across a large number of decision trees. This averaging effect can increase robustness to small fluctuations in feature values that may arise from measurement variability or preprocessing differences.

SVR also demonstrated excellent repeatability, producing correlation coefficients above 0.93 and BAG differences close to one year. This finding is consistent with the widespread use of SVR in brain age prediction literature, where the method has frequently served as a reliable baseline due to its ability to capture nonlinear relationships and achieve strong predictive performance across diverse neuroimaging datasets [5,14].

Although Elastic Net, Ridge, and GPR produced slightly lower stability metrics, their overall agreement remained high. Importantly, even the lowest-performing model achieved correlations above 0.88, indicating that all investigated approaches generated relatively stable predictions over short time intervals. This level of repeatability compares favourably with findings reported in previous longitudinal brain age studies and supports the suitability of these models for applications requiring reliable individual-level measurements [12].

The modest improvement observed in the clean subset further suggests that methodological consistency contributes to longitudinal reliability. Because this subset retained only scan pairs acquired using the same MRI field strength and

processed with the same FreeSurfer version at both time points, potential sources of technical variability were reduced. The resulting improvements in correlation and agreement metrics indicate that differences in acquisition and processing conditions may influence longitudinal brain age estimates. This finding is consistent with previous studies showing that scanner characteristics and image processing pipelines can affect brain age predictions and other MRI-derived biomarkers, even when biological changes are minimal [11, 13].

In contrast to the strong short-term results, long-term sensitivity proved substantially more challenging. Correlations between chronological age change and predicted brain age change remained relatively low across all models, with Random Forest achieving the highest values. Although these correlations indicate that some relationship existed between observed aging and predicted aging, they also suggest that accurately tracking individual aging trajectories over periods of approximately three years remains difficult.

Several factors may contribute to the relatively modest long-term correlations observed across models. First, structural brain changes occurring over a period of approximately three years are often subtle relative to the substantial inter-individual variability present in brain morphology [1, 3]. Brain age models are generally trained to learn cross-sectional age differences spanning several decades rather than longitudinal trajectories within the same individual. Consequently, the signal associated with within-subject aging may be considerably weaker than the cross-sectional age signal used during model development.

Second, although post-hoc bias correction was applied to reduce systematic age-related prediction bias, models optimized for cross-sectional age prediction may not fully capture subtle longitudinal changes in brain structure. As a result, predicted brain age progression may be attenuated relative to true chronological aging, particularly when structural changes remain small compared with the substantial variability present across individuals and brain age estimates [12]. Similar challenges have been reported in previous longitudinal brain age studies, where strong cross-sectional performance did not necessarily translate into high sensitivity to short- or medium-term aging trajectories [31, 12].

The present findings also extend previous longitudinal BrainAGE work by separating short-term stability from longer-term sensitivity. Franke and Gaser ex-

amined longitudinal changes in BrainAGE in healthy aging, MCI, and Alzheimer's disease, demonstrating the relevance of repeated brain age estimation for studying aging and clinical progression [31]. However, the present results show that reliable short-term repeatability does not necessarily imply strong sensitivity to subtle within-subject change over longer intervals. This distinction is also consistent with Dörfel et al., who emphasized that brain age models can differ in accuracy, reliability, and their ability to reflect longitudinal change [12].

Interestingly, Random Forest achieved both the strongest short-term stability and the strongest long-term sensitivity among the evaluated models. This suggests that the model was capable of maintaining stable predictions when biological change was minimal while remaining comparatively responsive to genuine aging-related variation over longer intervals. Such a combination of properties is desirable for longitudinal biomarker applications because it minimizes measurement noise without completely suppressing meaningful biological signals.

The Random Forest model presents a particularly interesting example of the trade-off between population-level accuracy and individual-level longitudinal sensitivity. Although the model underestimated the average magnitude of age progression, predicting a mean age increase of only 1.06 years compared with the observed mean chronological increase of 2.72 years, it consistently achieved the strongest correlations and lowest prediction errors across the longitudinal analyses. This finding suggests that Random Forest may have been more effective at preserving relative differences between individuals, allowing subjects who experienced greater chronological aging to also exhibit larger predicted brain age increases. In contrast, models such as Ridge and Elastic Net more closely reproduced the average cohort-level age progression but showed weaker correspondence at the individual level. From a practical perspective, this distinction may be important because many potential clinical applications of brain age focus on identifying individuals exhibiting accelerated or delayed brain aging trajectories rather than precisely reproducing average chronological change across a population [4, 5].

Despite the modest subject-level correlations, all models produced positive mean predicted age changes across the longitudinal cohort. This finding indicates that the models successfully detected age-related progression at the population

level. Notably, Ridge Regression and Elastic Net generated mean predicted age increases of 2.59 and 2.53 years, respectively, closely matching the observed mean chronological age increase of 2.72 years. These results suggest that while accurately capturing individual differences in aging trajectories remains challenging, certain models were capable of reproducing average cohort-level aging trends with reasonable accuracy. Consequently, the present findings highlight an important distinction between population-level sensitivity and subject-level longitudinal tracking, with the latter representing a substantially more demanding task for brain age prediction models [12].

Taken together, the longitudinal analyses indicate that the investigated models provide highly reliable short-term brain age estimates and are capable of detecting age-related progression at the cohort level. However, accurately quantifying individual aging trajectories over multi-year intervals remains challenging. These findings are consistent with an emerging body of literature suggesting that longitudinal sensitivity may represent one of the most demanding benchmarks for brain age prediction models and should be considered alongside conventional cross-sectional performance metrics when evaluating their practical utility [5, 12].

5.6 Normative Modelling on Clinical Subjects

The primary objective of the normative modeling analysis was to determine whether a brain age model trained exclusively on cognitively normal individuals could identify deviations from expected aging trajectories in subjects with cognitive impairment. The results provide evidence that the Random Forest normative model captured clinically meaningful variation beyond normal aging, as demonstrated by the progressive increase in Brain Age Gap (BAG) values across diagnostic groups.

An important observation is that cognitively normal subjects exhibited a mean BAG close to zero. This finding suggests that the model remained well calibrated when applied to an independent healthy population and supports the validity of the normative modeling framework adopted in this study. Because the model was trained exclusively on cognitively normal individuals, BAG values close to zero in the reference group indicate that the learned aging trajectory generalized

appropriately to unseen healthy subjects rather than systematically overestimating or underestimating brain age.

In contrast, both MCI and Dementia groups exhibited progressively larger positive BAG values. Subjects with MCI demonstrated a mean BAG approaching one year, whereas Dementia subjects exhibited a mean BAG exceeding two years. This pattern suggests that individuals with cognitive impairment possessed brain structural characteristics more commonly associated with older individuals, resulting in brains that appeared older than expected for their chronological age. The gradual increase in BAG from CN to MCI and subsequently to Dementia is consistent with the hypothesis that neurodegenerative processes accelerate biological brain aging and produce structural changes resembling those observed during normal aging, albeit at a faster rate [5, 8].

These findings align closely with a large body of brain age literature reporting elevated BAG values in individuals with neurodegenerative disorders and cognitive impairment [8, 24]. Previous studies have consistently demonstrated that Alzheimer's disease patients exhibit older appearing brains relative to cognitively normal controls, while individuals with MCI typically occupy an intermediate position between healthy aging and dementia [8, 31]. The present results therefore provide additional evidence supporting the biological relevance of BAG as a marker of abnormal brain aging.

The diagnostic ordering observed in the present study is particularly consistent with previous BrainAGE work in cognitive impairment and Alzheimer's disease. Franke and Gaser reported longitudinal BrainAGE changes in healthy aging, MCI, and Alzheimer's disease, supporting the relevance of repeated brain age estimation for monitoring clinical progression [31]. Similarly, Gaser et al. showed that BrainAGE measures in MCI patients were associated with conversion to Alzheimer's disease, indicating that elevated apparent brain age can capture clinically relevant neurodegenerative variation [8]. In this context, the progressive increase in BAG from cognitively normal subjects to MCI and Dementia in the present analysis is consistent with earlier evidence that brain age deviations reflect abnormal aging processes associated with cognitive decline.

Importantly, the association between diagnostic status and BAG remained significant after adjustment for chronological age and sex. This observation

strengthens the interpretation that the observed group differences were not simply a consequence of the Dementia and MCI groups being older on average than the cognitively normal group. Instead, the persistence of significant effects after statistical adjustment suggests that diagnostic status contributed independently to elevated BAG values. In practical terms, the model identified differences in brain aging that could not be fully explained by normal chronological aging alone.

The magnitude of the regression coefficients further supports the notion of progressive deviation from normative aging trajectories. Relative to cognitively normal individuals, MCI subjects exhibited significantly elevated BAG values, while Dementia subjects demonstrated substantially larger increases. This gradient mirrors the expected progression of neurodegenerative disease severity and suggests that BAG may reflect the cumulative burden of pathological brain changes. Although the present study was not designed to investigate disease mechanisms directly, the observed pattern is consistent with established evidence linking cognitive decline to widespread cortical thinning, ventricular enlargement, hippocampal atrophy, and other structural alterations commonly associated with advanced brain aging [2, 3].

The increasing variability of BAG values observed in the clinically impaired groups is also noteworthy. Both MCI and Dementia subjects exhibited substantially larger dispersion compared with cognitively normal individuals. One possible explanation is that neurodegenerative disorders are inherently heterogeneous, with considerable variation in disease stage, progression rate, underlying pathology, and individual resilience. Consequently, some clinically impaired individuals may exhibit relatively modest deviations from normative aging trajectories, whereas others demonstrate substantially accelerated brain aging. Similar increases in BAG variability have been reported in studies of heterogeneous clinical populations, where substantial inter-individual differences in apparent brain aging are commonly observed [24]

From a methodological perspective, these findings support the utility of normative modeling as a framework for detecting abnormal brain aging. By learning expected age-related patterns exclusively from healthy individuals and subsequently evaluating deviations from those patterns in independent populations, normative models provide a biologically interpretable measure of individual brain

health. Unlike traditional case-control classification approaches, which focus primarily on diagnostic discrimination, normative modeling quantifies the degree to which an individual's brain deviates from expected healthy aging trajectories [4, 6].

Nevertheless, the observed effect sizes should be interpreted cautiously. Although the group differences were statistically significant and followed the expected clinical ordering, the average BAG elevations remained relatively modest in absolute terms. This finding suggests that while BAG captures meaningful disease-related variation at the population level, substantial overlap remains between diagnostic groups. Consequently, BAG should not be viewed as a standalone diagnostic marker but rather as a complementary biomarker that may provide additional information alongside clinical, cognitive, and neuroimaging assessments [4, 5].

Overall, the normative modelling analysis demonstrated that the Random Forest brain age model was able to identify progressively greater deviations from normative aging trajectories in MCI and Dementia subjects. The persistence of these effects after adjustment for age and sex, together with their consistency with established neurodegenerative aging patterns, supports the potential utility of brain age as a biologically meaningful marker of abnormal brain aging and cognitive decline [4, 5].

6. Conclusions and Future Work

This chapter summarizes the main conclusions of the thesis, its limitations and outlines directions for future research. It first revisits the central findings of the study in relation to predictive performance, explainability, longitudinal reliability, and normative modelling in clinical populations. It then discusses the main limitations that should be considered when interpreting the results. Finally, the chapter proposes future research directions that could extend and strengthen the present work.

6.1 Conclusions

The prediction of brain age from neuroimaging data has emerged as a promising approach for quantifying biological brain aging and identifying deviations from expected aging trajectories. However, the practical value of brain age models depends not only on predictive accuracy but also on their interpretability, reliability, and clinical relevance. The primary objective of this thesis was therefore to investigate whether explainable machine learning methods could provide accurate and biologically meaningful brain age estimates while also demonstrating robustness across longitudinal and clinical evaluation settings.

To address this objective, five machine learning models were developed and evaluated using FreeSurfer-derived structural MRI features from the ADNI dataset. Model performance was assessed through a comprehensive framework incorporating predictive accuracy, hyperparameter optimization analysis, explainability assessment, longitudinal evaluation, and normative modeling.

The results demonstrated that all evaluated models achieved comparable predictive performance, with test-set mean absolute errors below five years and moderate correlations between predicted and chronological age. Although no single model consistently outperformed all others across every evaluation metric, Random Forest achieved the lowest prediction error and demonstrated particularly strong performance throughout the subsequent analyses. These findings suggest

that multiple machine learning approaches are capable of extracting meaningful age-related information from regional structural MRI measurements.

A major contribution of the present work was the incorporation of explainability analyses at both global and local levels. The explainability results revealed that model predictions were primarily driven by neuroanatomical features with well-established associations with brain aging, including ventricular measures, white matter abnormalities, hippocampal structures, thalamic regions, and cortical temporal and parietal areas. Furthermore, the high agreement observed across SHAP, permutation importance, and Elastic Net coefficient analyses, together with the exceptional stability of SHAP rankings across multiple random seeds, provided evidence that the identified aging patterns were both biologically plausible and highly reproducible. These findings support the use of explainable machine learning as a means of improving transparency and trustworthiness in brain age prediction research.

The longitudinal analyses demonstrated that the investigated models exhibited excellent short-term test–retest reliability, indicating that predicted brain age remained highly stable when minimal biological aging was expected. At the same time, the results highlighted the greater challenge of detecting individual aging trajectories over longer follow-up intervals. Although all models captured positive age-related progression at the cohort level, sensitivity to subject-specific aging rates remained more limited. This finding reinforces the distinction between cross-sectional age prediction and longitudinal aging assessment and suggests that longitudinal sensitivity represents an important area for future methodological development.

The normative modeling analyses further demonstrated the clinical relevance of the proposed framework. A model trained exclusively on cognitively normal individuals identified progressively larger Brain Age Gap values in Mild Cognitive Impairment and Dementia subjects, with these effects remaining significant after adjustment for age and sex. The observed pattern is consistent with the hypothesis that neurodegenerative processes are associated with accelerated biological brain aging and provides additional support for the use of brain age as a marker of abnormal aging trajectories.

Taken together, the findings of this thesis demonstrate that explainable machine learning approaches can provide accurate, interpretable, and clinically meaningful estimates of brain age from structural MRI data. Beyond predictive performance alone, the integration of explainability analyses, longitudinal evaluation, and normative modeling provided a more comprehensive assessment of brain age models than is commonly reported in the literature. The results indicate that brain age estimation has the potential to serve not only as a predictive modeling task but also as a valuable framework for investigating biological aging processes and identifying deviations associated with cognitive impairment and neurodegenerative disease.

In addition, the study was conducted and reported in accordance with the TRIPOD+AI recommendations, which strengthens the transparency, reproducibility, and interpretability of the proposed prediction framework compared with many earlier brain age studies that predate these reporting guidelines

As research in this field continues to evolve, future advances in multimodal neuroimaging, longitudinal modeling, and clinically oriented validation studies may further enhance the utility of brain age as a biomarker of brain health. The present work contributes to this ongoing effort by demonstrating that interpretable machine learning methods can capture meaningful patterns of brain aging while maintaining transparency regarding the neuroanatomical features underlying their predictions.

6.2 Limitations

Several limitations should be considered when interpreting the findings of the present study. Firstly, the developed brain age models were trained and evaluated using data from the ADNI, which primarily consists of older adults. Consequently, the models were optimized to capture age-related neuroanatomical variation within the later stages of adulthood. Brain aging patterns are known to differ across the lifespan, and structural changes observed in older individuals may not fully reflect those occurring during younger or middle adulthood. Therefore, the applicability of the proposed models to substantially younger populations remains uncertain and would require further investigation.

Second, although the investigated models demonstrated strong cross-sectional performance and excellent short-term test–retest reliability, the longitudinal analyses revealed more limited sensitivity to individual aging trajectories over longer follow-up periods. Correlations between chronological age change and predicted brain age change remained relatively modest across all models, despite the observation of positive age-related progression at the cohort level. These findings highlight an important challenge in the brain age literature: accurate prediction of chronological age in cross-sectional settings does not necessarily imply accurate tracking of within-subject aging over time. Consequently, caution is warranted when interpreting brain age as a direct measure of individual longitudinal aging rates.

Third, external validation was not performed using an independent dataset outside ADNI. Although strict training, validation, and testing procedures were employed throughout model development, all analyses were ultimately conducted within a single neuroimaging cohort. Differences in scanner characteristics, acquisition protocols, demographic composition, and preprocessing procedures may influence model behavior when applied to external populations. Therefore, additional validation across independent datasets would be valuable for establishing the robustness and generalizability of the proposed framework.

Finally, the present study focused on interpretable machine learning approaches using FreeSurfer-derived regional morphometric features. These models were selected because they provide a favourable balance between predictive performance, computational simplicity, and anatomical interpretability. However, the study did not include modern image-based deep learning models trained directly on raw or minimally preprocessed MRI data. State-of-the-art deep learning approaches have reported substantially lower prediction errors than those observed in the present thesis, particularly when trained on large datasets with broad age ranges [19, 27]. Therefore, the proposed framework should not be interpreted as outperforming contemporary image-based deep learning methods in terms of prediction accuracy.

Direct comparison with state-of-the-art studies is also limited by differences in cohort composition, scanner protocols, preprocessing pipelines, age distribution,

sample size, and validation strategy. The present analysis was based on an older and clinically focused ADNI cohort, whereas several high-performing deep learning studies used larger population datasets or challenge datasets with different age ranges and imaging characteristics [19, 27]. In addition, no empirical runtime comparison was performed between the present feature-based models and image-based deep learning architectures. Consequently, the advantage of the present approach should be understood primarily in terms of lower model complexity, anatomical interpretability, and suitability for multi-method explainability, longitudinal reliability analysis, and normative clinical evaluation. Future work should directly compare the present framework with deep learning and voxel-wise approaches under the same dataset, preprocessing, and validation conditions.

Despite these limitations, the present study provides a comprehensive evaluation of explainable machine learning methods for brain age prediction. By integrating predictive modeling, explainability analyses, longitudinal evaluation, and normative modeling within a unified framework, the study contributes both methodological and biological insights into the development of interpretable brain age biomarkers.

6.3 Future Work

The findings of the present study suggest several promising directions for future research. An important next step would be the external validation of the proposed framework using independent neuroimaging cohorts. Although the models demonstrated robust performance within the ADNI dataset, evaluation on datasets acquired using different scanners, protocols, and populations would provide stronger evidence regarding their generalizability and clinical applicability.

Future work could also explore the integration of additional data modalities. The present study focused exclusively on structural MRI-derived features, whereas brain aging is influenced by a wide range of biological processes. Combining structural neuroimaging with diffusion imaging, functional imaging, genetic information, fluid biomarkers, or other clinical measures may provide a more comprehensive characterization of biological brain aging and potentially improve predictive performance.

Another important direction involves comparison with contemporary deep learning approaches. Recent advances in convolutional neural networks, vision transformers, and foundation models have produced increasingly accurate brain age estimates from raw neuroimaging data. Future studies could directly compare the predictive performance, interpretability, and robustness of these approaches against the explainable machine learning framework developed in the present work.

The longitudinal findings also highlight the need for models specifically designed to capture within-subject aging trajectories. Although the investigated models demonstrated excellent short-term stability, sensitivity to individual aging progression remained more limited. Future work could therefore focus on developing longitudinally informed brain age models that explicitly incorporate repeated measurements during training and are optimized for detecting subtle age-related changes over time.

Finally, the normative modeling results suggest potential clinical applications of brain age estimation. Future studies could investigate whether elevated Brain Age Gap values predict subsequent cognitive decline, disease progression, or conversion from Mild Cognitive Impairment to dementia. Such work would help determine the practical value of brain age as a prognostic biomarker and further clarify its role in clinical decision-making.

The proposed framework is intended for research use and requires users with experience in neuroimaging preprocessing and machine learning model evaluation. In future applications, scans with poor image quality, failed FreeSurfer processing, or unavailable predictor values should be excluded or handled using preprocessing procedures estimated from the training data.

References

1. Raz, N., & Rodrigue, K. M. (2006). Differential aging of the brain: Patterns, cognitive correlates and modifiers. *Neuroscience & Biobehavioral Reviews*, 30(6), 730–748. <https://doi.org/10.1016/j.neubiorev.2006.07.001>
2. Lockhart, S. N., & DeCarli, C. (2014). Structural imaging measures of brain aging. *Neuropsychology Review*, 24(3), 271–289. <https://doi.org/10.1007/s11065-014-9268-3>
3. Fjell, A. M., & Walhovd, K. B. (2010). Structural brain changes in aging: Courses, causes and cognitive consequences. *Reviews in the Neurosciences*, 21(3), 187–221. <https://doi.org/10.1515/revneuro.2010.21.3.187>
4. Cole, J. H., & Franke, K. (2017). Predicting age using neuroimaging: Innovative brain ageing biomarkers. *Trends in Neurosciences*, 40(12), 681–690. <https://doi.org/10.1016/j.tins.2017.10.001>
5. Franke, K., & Gaser, C. (2019). Ten years of BrainAGE as a neuroimaging biomarker of brain aging: What insights have we gained? *Frontiers in Neurology*, 10, 789. <https://doi.org/10.3389/fneur.2019.00789>
6. Baecker, L., Garcia-Dias, R., Vieira, S., Scarpazza, C., & Mechelli, A. (2021). Machine learning for brain age prediction: Introduction to methods and clinical applications. *EBioMedicine*, 72, 103600. <https://doi.org/10.1016/j.ebiom.2021.103600>
7. Cole, J. H., Poudel, R. P. K., Tsagkrasoulis, D., Caan, M. W. A., Steves, C., Spector, T. D., & Montana, G. (2017). Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, 163, 115–124. <https://doi.org/10.1016/j.neuroimage.2017.07.059>
8. Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer, H., & Alzheimer's Disease Neuroimaging Initiative. (2013). BrainAGE in mild cognitive impaired patients: Predicting the conversion to Alzheimer's disease. *PLOS ONE*, 8(6), e67346. <https://doi.org/10.1371/journal.pone.0067346>
9. Koutsouleris, N., Davatzikos, C., Borgwardt, S., Gaser, C., Bottlender, R., Frodl, T., Falkai, P., Riecher-Rössler, A., Möller, H.-J., Reiser, M., Pantelis, C., & Meisenzahl, E. (2014). Accelerated brain aging in schizophrenia and beyond: A neuroanatomical marker of psychiatric disorders. *Schizophrenia Bulletin*, 40(5), 1140–1153. <https://doi.org/10.1093/schbul/sbt142>
10. Liang, H., Zhang, F., & Niu, X. (2019). Investigating systematic bias in brain age estimation with application to post-traumatic stress disorders. *Human Brain Mapping*, 40(11), 3143–3152. <https://doi.org/10.1002/hbm.24588>

11. Jirsaraie, R. J., Kaufmann, T., Bashyam, V., Erus, G., Luby, J. L., Westlye, L. T., Davatzikos, C., Barch, D. M., & Sotiras, A. (2023). Benchmarking the generalizability of brain age models: Challenges posed by scanner variance and prediction bias. *Human Brain Mapping*, 44(3), 1118–1128. <https://doi.org/10.1002/hbm.26144>
12. Dörfel, R. P., Arenas-Gomez, J. M., Fisher, P. M., Ganz, M., Knudsen, G. M., Svensson, J. E., & Plavén-Sigra, P. (2023). Prediction of brain age using structural magnetic resonance imaging: A comparison of accuracy and test–retest reliability of publicly available software packages. *Human Brain Mapping*, 44(17), 6139–6148. <https://doi.org/10.1002/hbm.26502>
13. Dular, L., Pernuš, F., & Špiclin, Ž. (2024). Extensive T1-weighted MRI preprocessing improves generalizability of deep brain age prediction models. *Computers in Biology and Medicine*, 177, 108320. <https://doi.org/10.1016/j.combiomed.2024.108320>
14. Franke, K., Ziegler, G., Klöppel, S., & Gaser, C. (2010). Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. *NeuroImage*, 50(3), 883–892. <https://doi.org/10.1016/j.neuroimage.2010.01.005>
15. Baecker, L., Dafflon, J., da Costa, P. F., Garcia-Dias, R., Vieira, S., Scarpazza, C., Calhoun, V. D., Sato, J. R., Mechelli, A., & Pinaya, W. H. L. (2021). Brain age prediction: A comparison between machine learning models using region- and voxel-based morphometric data. *Human Brain Mapping*, 42(8), 2332–2346. <https://doi.org/10.1002/hbm.25368>
16. Han, J., Kim, S. Y., Lee, J., & Lee, W. H. (2022). Brain age prediction: A comparison between machine learning models using brain morphometric data. *Sensors*, 22(20), 8077. <https://doi.org/10.3390/s22208077>
17. Cumplido-Mayoral, I., et al. (2023). Biological brain age prediction using machine learning on structural neuroimaging data: Multi-cohort validation against biomarkers of Alzheimer’s disease and neurodegeneration stratified by sex. *eLife*, 12, e81067. <https://doi.org/10.7554/eLife.81067>
18. Fischl, B. (2012). FreeSurfer. *NeuroImage*, 62(2), 774–781. <https://doi.org/10.1016/j.neuroimage.2012.01.021>
19. De Bonis, M. L. N., Fasano, G., Lombardi, A., Ardito, C., Ferrara, A., Di Sciascio, E., & Di Noia, T. (2024). Explainable brain age prediction: A comparative evaluation of morphometric and deep learning pipelines. *Brain Informatics*, 11, 33. <https://doi.org/10.1186/s40708-024-00244-9>
20. Cole, J. H., Leech, R., Sharp, D. J., & Alzheimer’s Disease Neuroimaging Initiative. (2015). Prediction of brain age suggests accelerated atrophy after traumatic brain injury. *Annals of Neurology*, 77(4), 571–581. <https://doi.org/10.1002/ana.24367>

21. Sun, J., Cao, R., Zhou, M., Hussain, W., Wang, B., Xue, J., & Xiang, J. (2022). Interpretation for individual brain age prediction based on gray matter volume. *Frontiers in Aging Neuroscience*, 14, 1028081. <https://doi.org/10.3389/fnagi.2022.1028081>
22. Lombardi, A., Diacono, D., Amoroso, N., Monaco, A., Tavares, J. M. R. S., Bellotti, R., & Tangaro, S. (2021). Explainable deep learning for personalized age prediction with brain morphology. *Frontiers in Neuroscience*, 15, 674055. <https://doi.org/10.3389/fnins.2021.674055>
23. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
24. Kaufmann, T., van der Meer, D., Doan, N. T., Schwarz, E., Lund, M. J., Agartz, I., Alnæs, D., Barch, D. M., Baur-Streubel, R., Bertolino, A., Bettella, F., Beyer, M. K., Bøen, E., Borgwardt, S., Brandt, C. L., Buitelaar, J., Celius, E. G., Cervenka, S., Conzelmann, A., ... Westlye, L. T. (2019). Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nature Neuroscience*, 22(10), 1617–1623. <https://doi.org/10.1038/s41593-019-0471-7>
25. Hahn, T., Ernsting, J., Winter, N. R., Holstein, V., Leenings, R., Beisemann, M., Fisch, L., Sarink, K., Emden, D., & Dannlowski, U. (2022). An uncertainty-aware, shareable, and transparent neural network architecture for brain-age modeling. *Science Advances*, 8(1), eabg9471. <https://doi.org/10.1126/sciadv.abg9471>
26. Bashyam, V. M., Erus, G., Doshi, J., Habes, M., Nasrallah, I. M., Truelove-Hill, M., Srinivasan, D., Mamourian, E., Pomponio, R., Fan, Y., Launer, L. J., Masters, C. L., Maruff, P., Zhuo, C., Völzke, H., Johnson, S. C., Fripp, J., Koutsouleris, N., Satterthwaite, T. D., ... Davatzikos, C. (2020). MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14,468 individuals worldwide. *Brain*, 143(7), 2312–2324. <https://doi.org/10.1093/brain/awaa160>
27. Peng, H., Gong, W., Beckmann, C. F., Vedaldi, A., & Smith, S. M. (2021). Accurate brain age prediction with lightweight deep neural networks. *Medical Image Analysis*, 68, 101871. <https://doi.org/10.1016/j.media.2020.101871>
28. Jirsaraie, R. J., Gorelik, A. J., Camacho, M. C., Ameis, S. H., Voineskos, A. N., Chakravarty, M. M., & Satterthwaite, T. D. (2023). A systematic review of multimodal brain age studies: Uncovering a divergence between model accuracy and utility. *Patterns*, 4(4), 100712. <https://doi.org/10.1016/j.patter.2023.100712>
29. Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., ... Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, e078378. <https://doi.org/10.1136/bmj-2023-078378>

30. Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Liu, E., Morris, J. C., Petersen, R. C., Saykin, A. J., Schmidt, M. E., Shaw, L., Siuciak, J. A., Soares, H., Toga, A. W., Trojanowski, J. Q., & Alzheimer's Disease Neuroimaging Initiative. (2012). The Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception. *Alzheimer's & Dementia*, 8(1 Suppl), S1–S68. <https://doi.org/10.1016/j.jalz.2011.09.172>
31. Franke, K., & Gaser, C. (2012). Longitudinal changes in individual BrainAGE in healthy aging, mild cognitive impairment, and Alzheimer's disease. *GeroPsych*, 25(4), 235–245. <https://doi.org/10.1024/1662-9647/a000074>

APPENDIX A - TRIPOD+AI Checklist

The completed TRIPOD+AI checklist for the present thesis is provided below as supplementary material to document reporting transparency for the development and evaluation of the proposed brain age prediction models.

Section/Topic	Item	Development / evaluation ¹	Checklist item	Reported on page
TITLE				
<i>Title</i>	1	D;E	Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted	P: Title
ABSTRACT				
<i>Abstract</i>	2	D;E	See TRIPOD+AI for Abstracts checklist	II
INTRODUCTION				
<i>Background</i>	3a	D;E	Explain the healthcare context (including whether diagnostic or prognostic) and rationale for developing or evaluating the prediction model, including references to existing models	1-2, 6-15
	3b	D;E	Describe the target population and the intended purpose of the prediction model in the context of the care pathway, including its intended users (e.g., healthcare professionals, patients, public)	P:1-4,77
	3c	D;E	Describe any known health inequalities between sociodemographic groups	NR
<i>Objectives</i>	4	D;E	Specify the study objectives, including whether the study describes the development or validation of a prediction model (or both)	3-4
METHODS				
<i>Data</i>	5a	D;E	Describe the sources of data separately for the development and evaluation datasets (e.g., randomised trial, cohort, routine care or registry data), the rationale for using these data, and representativeness of the data	16-18,74
	5b	D;E	Specify the dates of the collected participant data, including start and end of participant accrual; and, if applicable, end of follow-up	NR
<i>Participants</i>	6a	D;E	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including the number and location of centres	P:16-17
	6b	D;E	Describe the eligibility criteria for study participants	17-18
	6c	D;E	Give details of any treatments received, and how they were handled during model development or evaluation, if relevant	N/A
<i>Data preparation</i>	7	D;E	Describe any data pre-processing and quality checking, including whether this was similar across relevant sociodemographic groups	P:18-21
<i>Outcome</i>	8a	D;E	Clearly define the outcome that is being predicted and the time horizon, including how and when assessed, the rationale for choosing this outcome, and whether the method of outcome assessment is consistent across sociodemographic groups	P:21, 27-28
	8b	D;E	If outcome assessment requires subjective interpretation, describe the qualifications and demographic characteristics of the outcome assessors	N/A (21)
	8c	D;E	Report any actions to blind assessment of the outcome to be predicted	N/A
<i>Predictors</i>	9a	D	Describe the choice of initial predictors (e.g., literature, previous models, all available predictors) and any pre-selection of predictors before model building	17,21
	9b	D;E	Clearly define all predictors, including how and when they were measured (and any actions to blind assessment of predictors for the outcome and other predictors)	P:7-8, 17,21
	9c	D;E	If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of the predictor assessors	N/A
<i>Sample size</i>	10	D;E	Explain how the study size was arrived at (separately for development and evaluation), and justify that the study size was sufficient to answer the research question. Include details of any sample size calculation	P:17-18
<i>Missing data</i>	11	D;E	Describe how missing data were handled. Provide reasons for omitting any data	17,19,21
<i>Analytical methods</i>	12a	D	Describe how the data were used (e.g., for development and evaluation of model performance) in the analysis, including whether the data were partitioned, considering any sample size requirements	17-24,28
	12b	D	Depending on the type of model, describe how predictors were handled in the analyses (functional form, rescaling, transformation, or any standardisation).	P:21
	12c	D	Specify the type of model, rationale ² , all model-building steps, including any hyperparameter tuning, and method for internal validation	21-24
	12d	D;E	Describe if and how any heterogeneity in estimates of model parameter values and model performance was handled and quantified across clusters (e.g., hospitals, countries). See TRIPOD-Cluster for additional considerations ³	NP
	12e	D;E	Specify all measures and plots used (and their rationale) to evaluate model performance (e.g., discrimination, calibration, clinical utility) and, if relevant, to compare multiple models	24-28; 30-32
	12f	E	Describe any model updating (e.g., recalibration) arising from the model evaluation, either overall or for particular sociodemographic groups or settings	24-25; 30-31
	12g	E	For model evaluation, describe how the model predictions were calculated (e.g., formula, code, object, application programming interface)	P:24-25, 28
<i>Class imbalance</i>	13	D;E	If class imbalance methods were used, state why and how this was done, and any subsequent methods to recalibrate the model or the model predictions	21
<i>Fairness</i>	14	D;E	Describe any approaches that were used to address model fairness and their rationale	29 (none)
<i>Model output</i>	15	D	Specify the output of the prediction model (e.g., probabilities, classification). Provide details and rationale for any classification and how the thresholds were identified	21

<i>Training versus evaluation</i>	16	D;E	Identify any differences between the development and evaluation data in healthcare setting, eligibility criteria, outcome, and predictors	P:17-18, 28,48-50
<i>Ethical approval</i>	17	D;E	Name the institutional research board or ethics committee that approved the study and describe the participant-informed consent or the ethics committee waiver of informed consent	P:28-29
OPEN SCIENCE				
<i>Funding</i>	18a	D;E	Give the source of funding and the role of the funders for the present study	29
<i>Conflicts of interest</i>	18b	D;E	Declare any conflicts of interest and financial disclosures for all authors	29
<i>Protocol</i>	18c	D;E	Indicate where the study protocol can be accessed or state that a protocol was not prepared	29
<i>Registration</i>	18d	D;E	Provide registration information for the study, including register name and registration number, or state that the study was not registered	29
<i>Data sharing</i>	18e	D;E	Provide details of the availability of the study data	29
<i>Code sharing</i>	18f	D;E	Provide details of the availability of the analytical code ⁴	29
PATIENT & PUBLIC INVOLVEMENT				
<i>Patient & Public Involvement</i>	19	D;E	Provide details of any patient and public involvement during the design, conduct, reporting, interpretation, or dissemination of the study or state no involvement.	29
RESULTS				
<i>Participants</i>	20a	D;E	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	17-18, 46-48
	20b	D;E	Report the characteristics overall and, where applicable, for each data source or setting, including the key dates, key predictors (including demographics), treatments received, sample size, number of outcome events, follow-up time, and amount of missing data. A table may be helpful. Report any differences across key demographic groups.	P:17-21, 46-49
	20c	E	For model evaluation, show a comparison with the development data of the distribution of important predictors (demographics, predictors, and outcome).	NR
<i>Model development</i>	21	D;E	Specify the number of participants and outcome events in each analysis (e.g., for model development, hyperparameter tuning, model evaluation)	17,46-48
<i>Model specification</i>	22	D	Provide details of the full prediction model (e.g., formula, code, object, application programming interface) to allow predictions in new individuals and to enable third-party evaluation and implementation, including any restrictions to access or re-use (e.g., freely available, proprietary) ⁵	P:21-24, 29
<i>Model performance</i>	23a	D;E	Report model performance estimates with confidence intervals, including for any key subgroups (e.g., sociodemographic). Consider plots to aid presentation.	P:30-32, 46-50
	23b	D;E	If examined, report results of any heterogeneity in model performance across clusters. See TRIPOD Cluster for additional details ³ .	NP
<i>Model updating</i>	24	E	Report the results from any model updating, including the updated model and subsequent performance	30-31
DISCUSSION				
<i>Interpretation</i>	25	D;E	Give an overall interpretation of the main results, including issues of fairness in the context of the objectives and previous studies	P:29, 52-71
<i>Limitations</i>	26	D;E	Discuss any limitations of the study (such as a non-representative sample, sample size, overfitting, missing data) and their effects on any biases, statistical uncertainty, and generalizability	74-76
<i>Usability of the model in the context of current care</i>	27a	D	Describe how poor quality or unavailable input data (e.g., predictor values) should be assessed and handled when implementing the prediction model	77
	27b	D	Specify whether users will be required to interact in the handling of the input data or use of the model, and what level of expertise is required of users	77
	27c	D;E	Discuss any next steps for future research, with a specific view to applicability and generalizability of the model	76-77

P = Partially Reported; NR = Not Reported; NP = Not Performed; N/A = Not Applicable