Deep Learning Approaches to Multimodal MRI Brain Age Estimation

Andrei-Claudiu Roibu

Kellogg College

University of Oxford

Report submitted for the completion of

Doctor of Philosophy

Michaelmas 2023

This thesis is dedicated to my fiancée,
Oana Pelea,
for being my everyday partner in crime and rock when times are tough.

Acknowledgements

Personal

To all those which have supported me through this long and arduous journey, I give my thanks. In particular, I would like to thank my supervisors, Stephen Smith, Ana Namburete and Frederik (Rick) Lange, for their patience, dedication, moral, personal and academic support, and generally being the best supervisors a student can have. I would like to thank all of them, in particular, for the patience and wisdom they have shown in the face of adversity and having probably the most un-academic (to put it mildly) PhD student to ever grace the University of Oxford with their presence. I would also like to thank Rick in particular for being there for me each day, acting as my first point of contact in any issues that I had, and being a true friend throughout this process, Steve for trusting me with this project and never giving up on me, and Ana for her guidance and advice when projects seemed to be in gridlock.

I would also like to extend my thanks to Torsten Schindler and Stanislaw Adaszewski, my Roche industrial supervisors, for ensuring my projects remained relevant to industry throughout this journey, and for infusing a constant stream of pragmatism and novelty into my work through their advice. I would also like to thank Nicola Dinsdale, Saad Jbabdi, Ludovica Griffanti and Timothy Behrens for their insightful advice at various steps in my PhD journey.

I am also grateful to Ying-Qiu Zheng, Hossein Rafipoor, Miriam Szabo, András Sándor, Paul Lang, Lauren Overend, Charlie Fletcher, Răzvan Apetrei, Alexandra Ștefănescu, Büşra Karaş, as well as all the other friends and colleagues that acted as anchors when I was going downhill, falling into the he depths of despair.

Finally, I would like to extend my warmest thanks to my fiancée Oana Pelea, her parents Cătălin and Diana Pelea, as well as my father, Crinel Roibu, who has generously supported me financially and morally through almost 10 years of superior schooling, his wife Gabriela Roibu, my sister Georgiana Magdan, and both my and Oana's grandparents, for being the best family anybody could ever wish for.

I thank you all, as well all those whom I have not mentioned by name, but who know themselves, for their support through this dark exercise in masochism and self-discovery which is a PhD. You all have my ever lasting gratitude.

Institutional

I am grateful to the various institutions which have supported my research: the University of Oxford and SABS:R^3 CDT through an EPSRC Grant EP/S024093/1, F. Hoffmann-La Roche AG, the Royal Commission for the Exhibition of the 1851 for the 2021 Industrial Fellowship, and the Wellcome Centre for Integrative Neuroimaging (WIN), which is funded by the Wellcome Trust (203139/Z/16/Z). The computational aspects were supported by the Wellcome Trust (203141/Z/16/Z) and the NIHR Oxford BRC.

Abstract

Brain ageing remains an intricate, multifaceted process, marked not just by chronological time but by a myriad of structural, functional, and microstructural changes that often lead to discrepancies between actual age and the age inferred from neuroimaging. Machine learning methods, and especially Convolutional Neural Networks (CNNs), have proven adept in capturing patterns relating to ageing induced changes in the brain. The differences between the predicted and chronological ages, referred to as brain age deltas, have emerged as useful biomarkers for exploring those factors which promote accelerated ageing or resilience, such as pathologies or lifestyle factors. However, previous studies relied overwhelmingly on structural neuroimaging for predictions, overlooking rich details inherent in other MRI modalities, such as potentially informative functional and microstructural changes. This research, utilising the extensive UK Biobank dataset, reveals that 57 different maps spanning structural, susceptibility-weighted, diffusion, and functional MRI modalities can not only predict an individual's chronological age, but also encode unique ageing-related details. Through the use of both 3D CNNs and the novel 3D Shifted Window (SWIN) Transformers, this work uncovered associations between brain age deltas and 191 different non-imaging derived phenotypes (nIDPs), offering a valuable insight into factors influencing brain ageing. Moreover, this work found that ensembling data from multiple maps results in higher prediction accuracies. After a thorough comparison of both linear and non-linear multi-modal ensembling methods, including deep fusion networks, it was found that linear methods, such as ElasticNet, generally outperform their more complex non-linear counterparts. In addition, while ensembling was found to strengthen age prediction accuracies, it was found to weaken nIDP associations in certain circumstances where ensembled maps might have opposing sensitivities to a particular nIDP, thus reinforcing the need for guided selections of the ensemble components. Finally, while both CNNs and SWINs show comparable brain age prediction precision, SWIN networks stand out for their robustness against data corruption, while also proving a degree of inherent explainability. Overall, the results presented herein demonstrate that other 3D maps and modalities, which have not been considered previously for the task of brain age prediction, encode different information about the ageing brain. This research lays the foundation for further explorations into

how different factors, such as off-target drug effects, impact brain ageing. It also ushers in possibilities for enhanced clinical trial design, diagnostic approaches, and therapeutic monitoring grounded in refined brain age prediction models.

Contents

List of Figures

List of Tables

List of Abbreviations

2D . . . . . Two Dimensional

3D . . . . . Three Dimensional

AD . . . . . Axial Diffusivity

ADi . . . . . Alzheimer's Diserase

ADAM . . . . Adaptive Moment Estimation optimiser.

AdamW . . . . Modified version of ADAM optimiser.

ADC . . . . . Apparent Diffusion Coefficient

ANNs . . . . . Artificial Neural Networks

BA-SWIN . . Brain Age-Shifted Window Transformer.

BAG . . . . . Brain Age Gap

BatchNorm . Batch Normalisation

BERT . . . . . Bidirectional Encoder Representations from Transformers

BET . . . . . FSL Brain Extraction Tool

BG . . . . . Basal Ganglia

BIANCA . . . Brain Intensity AbNormality Classification Algorithm

BOLD . . . . . Blood Oxygenation Level Dependent

CBR . . . . . Correlation-Based Regression

CMRR . . . . Centre for Magnetic Resonance Research

CNN . . . . . Convolutional Neural Network

COPE . . . . . Contrast Of Parameter Estimate

CSF . . . . . Cerebrospinal Fluid

CT . . . . . Computed Tomography

DGP . . . . . Deep Gaussian Processes

DL . . . . . Deep Learning

dMRI . . . . . Diffusion Magnetic Resonance Imaging

DOF . . . . . Degrees of Freedom

DTI . . . . . Diffusion Tensor Imaging

DVARS . . . . Differential Variance of the Noise

DWI . . . . . Diffusion Weighted Image

ECoG . . . . . Electrocorticography

EEG . . . . . Electroencephalogram

EPI . . . . . Echo-Planar Imaging

FA . . . . . Fractional Anisotropy

FastICA . . . . A specific algorithm to perform ICA

FCN . . . . . Fully Convolutional Network

FEAT . . . . . FMRI Expert Analysis Tool

FLAIR . . . . Fluid Attenuated Inversion Recovery

FLIRT . . . . . FMRIB Linear Registration Tool

fMRI . . . . . Functional MRI

FNIRT . . . . FMRIB Non-Linear Image Registration Tool

FS T2 . . . . . A confound variable describing if T2 FLAIR was used

GELU . . . . . Gaussian Error Linear Unit

GLM . . . . . Generalised Linear Model

GM . . . . . Grey Matter

GNN . . . . . Graph Neural Network

GPR . . . . . Gaussian Processes Regression

GPT . . . . . Generative Pre-trained Transformer

GPUs . . . . . Graphics Processing Units

HGL . . . . . Happy-Go-Lucky network: my lightweight deep learning architecture.

HMM . . . . . Hidden Markov Model

ICA . . . . . Independent Component Analysis

ICVF . . . . . Intra-Cellular Volume Fraction

IDPs . . . . . Image-Derived Phenotypes

IQ . . . . . Intelligence Quotient

ISOVF . . . . Isotropic of Free Water Colume Fraction

KDEs . . . . . Kernel Density Estimates.

KLD . . . . . Kullback-Leibler Divergence.

LASSO . . . . Least Absolute Shrinkage and Selection Operator

LayerNorm . . Layer Normalisation

LIME . . . . . Local Interpretable Model-agnostic Explanations

LR . . . . . Linear Regression

LRP . . . . . Layer-wise Relevance Propagation

MAE . . . . . Mean Absolute Error

MCI . . . . . Mild Cognitive Impairment

MD . . . . . Mean Diffusivity

MDD . . . . . Major Depressive Disorder

MEG . . . . . Magnetoencephalography

MFN . . . . . Multi-Feature-Based Network

MLP . . . . . Machine Learning

MLP . . . . . Multi-Layer Perceptron

MMORF . . . FSL's MultiMOdal Registration Framework

MNI152 . . . . Montreal Neurological Institute 152 template

MNIST . . . . Modified National Institute of Standards and Technology database

MO . . . . . Mode of Anisotropy

MRI . . . . . Magnetic Resonance Imaging

MS . . . . . Multiple Sclerosis

MSA . . . . . Multi-Head Self-Attention

MSE . . . . . Mean Squared Error

MVNL-R . . . Multivariate Non-Linear Regression

nIDP . . . . . Non-imaging measurement.

NIFTI . . . . . Neuroimaging Informatics Technology Initiative (file format)

NIRS . . . . . Near-Infrared Spectroscopy

NMOSD . . . Neuromyelitis Optica Spectrum Disorder

NN . . . . . Neural Network

NODDI . . . . Neurite Orientation Dispersion and Density Imaging

ODI . . . . . Orientation Dispersion Index

OLS . . . . . Ordinary Least Squares

PAD . . . . . Predicted Age Difference

PCA . . . . . Principal Component Analysis

PET . . . . . Positron Emission Tomography

RD . . . . . Radial Diffusivity

ReLU . . . . . Rectified Linear Unit

RENT . . . . . Repeated ElasticNet

RF . . . . . Radio Frequency

rfMRI . . . . . Resting-State Functional MRI

RNNs . . . . . Recurrent Neural Networks

ROI . . . . . Region Of Interest

rsfMRI . . . . Resting State fMRI

RVR . . . . . Relevance Vector Regression

SA . . . . . Self-Attention

SFCN . . . . . Simple Fully Connected Network

SGD . . . . . Stochastic Gradient Descent.

SHAP . . . . . SHapley Additive exPlanations

sMRI . . . . . Structural Magnetic Resonance Imaging

SVD . . . . . Singular Value Decomposition

SVR . . . . . Support Vector Regression

SW-MSA . . . Shifted-Window Multi-Head Self-Attention

SWI . . . . . Susceptibility Weighted Imaging

SWIN . . . . . Shifted Window Transformer

swMRI . . . . Susceptibility-Weighted MRI

T1 . . . . . T1-weighted

T2* . . . . . A specific relaxation time constant in MRI

TBSS . . . . . Tract-Based Spatial Statistics

TE . . . . . Echo Time

tfMRI . . . . . Task fMRI

UK Biobank . A major national and international health resource containing genetic and health information from half a million UK participants

UKB . . . . . UK Biobank

VBM . . . . . Voxel Based Morphometry

VGG-16 . . . . A deep convolutional neural network for object recognition developed and trained by Oxford's renowned Visual Geometry Group (VGG), which achieved very competitive performance in the ImageNet competition.

ViT . . . . . Vision Transformer.

W-MSA . . . . Window Multi-Head Self-Attention

WM . . . . . White Matter

WMH . . . . . White Matter Hyperintensities

*Other **mathematical symbols** not included here and defined locally, as the same symbol can have a context dependent meaning across different chapters.*

1

Introduction

Contents

1.1 The Brain During Ageing

Ageing is a time-dependent process characterised by the accumulation of cellular damage [1], altered cellular communication, and genetic modifications influenced by external factors [2]. A person's apparent age, a metric shaped by individual lifestyle, health, environment, and genetics, provides a more precise understanding of personal health traits and risk patterns than chronological age. It paves the way for interventions tailored to biological benchmarks rather than chronological age [3, 4].

1.1.1 Structural Brain Changes

While biological ageing can easily be observed through cosmetic and behavioural changes, brain ageing is more subtle, manifesting as structural and functional changes throughout one's life [5–7]. Externally, brain ageing shows only as different neuropathologies [8]. Structurally, brain ageing is evidenced predominantly as a nonlinear and population-dependent atrophy process [8, 9], characterised by reductions in brain weight, cortical thickness, grey and white matter (GM and WM) volumes, and an expansion of the cerebral spinal fluid (CSF) spaces [4, 10–13].

1.1.2 Molecular and Cellular Ageing

At the molecular and cellular level, brain ageing unfolds through intricate biochemical and cellular pathways, encompassing oxidative damage, inflammation, protein misfolding, cellular dysfunctions, calcium metabolism imbalances, and genetic alterations [14–18]. This is further compounded by disturbances in neurotransmitter activity [14, 19] and changes in the cerebral vascular system [20]. The ubiquity of brain ageing affects all neurobiological processes, with no age-related changes occurring in isolation. This renders ageing a spatially and temporally diverse phenomenon influenced by a myriad of lifestyle and biological determinants [8, 21].

1.1.3 Consequences of Ageing

Age-related biological changes in the brain accumulate over time, with significant neurological consequences manifesting once specific thresholds are reached [9]. Ageing

is associated with substantial behavioural, psychological, and physiological effects, most notably cognitive decline and heightened vulnerability to neurodegenerative diseases such as Alzheimer's, Parkinson's, dementia, amyotrophic lateral sclerosis, and stroke [8, 22–24]. These disorders, for which age is the primary risk factor, burden individuals, healthcare systems, and societies. Cognitive ageing particularly impacts information processing speed, executive functions, memory, and has cascading repercussions on occupational, social, and mental well-being [6, 7, 25–28].

The rapid growth of the elderly demographic, especially in Western societies, accentuates the healthcare challenges and underscores the need for effective interventions [29]. Distinguishing between normal, healthy ageing (which itself is heterogeneous, displaying a wide range of variations), and early neurodegenerative disease stages is challenging, but crucial for timely interventions [21, 30–34]. This is further complicated by population variation in ageing manifestation, potentially influenced by genetics, lifestyle, and environmental factors [8, 9].

All these factors underscore the need for individualised biomarkers that reflect brain ageing speed, enabling a juxtaposition of one's chronological versus apparent brain ages. Deviations from typical healthy brain trajectories may signal impending age-associated brain disorders [8, 9].

1.2 The Brain Age Gap

Accessing neural tissue in vivo to probe and understand the links between structural and physiological brain changes, and age, is difficult. However, non-invasive neuroimaging techniques offer detailed reconstructions and measurements of brain structure and microstructure, as well as its structural and functional connectivity. The growing availability of vast multimodal neuroimaging datasets ensures reliable and reproducible quantifications of the brain's biological properties [8]. These datasets have paved the way for regression statistical models that connect various brain measures, such as image-derived measures or phenotypes (IDPs), or 2D and 3D brain images, with dependent variables like chronological ages. Once a model is trained, it can produce predictions of age from brain phenotypes or images,

which are then compared to actual chronological ages. Although these models were traditionally trained exclusively on healthy subjects, recent findings suggest similar outcomes even when including subjects with health conditions [21].

These models essentially learn trajectories of healthy ageing. The discrepancy between predicted brain ages and actual ages is termed the brain age gap (BAG), brain age delta, or predicted age difference (PAD). It provides a metric for brain ageing acceleration, showcasing if an individual's brain ageing aligns with or diverges from population norms [35]. Such deviations can signal accelerated ageing influenced by external factors, such as lifestyle, or internal elements like genetics or disease onset. When coupled with metadata or other data sources, such as environmental, biological, and lifestyle measures, or genetic data, these deviations can be assessed for significant associations. The ability to observe and quantify the BAG allows for establishing a healthy ageing baseline, analysing deviations in neurodegeneration, studying longitudinal changes in individuals under the influence of various factors such as medical interventions, and examining the impact of various lifestyle and environmental factors [4].

Given the high-dimensional nature of some datasets, such as UK Biobank [36], machine learning algorithms are preferred for BAG investigations, given their capacity to discern patterns in high-dimensional and highly variable data, outperforming methods like ordinary least squares regression [4, 8, 9].

1.3 Neuroimaging and Brain Ageing

While neuroimaging encompasses several techniques, such as PET, EEG, MEG, CT, NIRS, Ultrasound, and ECoG, Magnetic Resonance Imaging (MRI) stands out as a versatile neuroimaging tool that offers a plethora of imaging modalities that shed light on different aspects of neurobiology [8]. For this reason, MRI has been widely used in age-related brain studies [31]. Varying the MRI acquisition parameters produces images (modalities) that are sensitive to different structural and functional properties of the brain. While a more detailed exploration of MRI can be found in Chapter 2 Section 2.1, here, the 5 primary, or core, MRI modalities are outlined:

- **Structural MRI (sMRI)**: Sensitive to differences in tissue properties, primarily differentiating between gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). This allows for detailed visualisations of neuroanatomy, including cortical thickness alterations and white matter hyperintensities [37–39];

- **Susceptibility-weighted MRI (swMRI)**: Enhances sensitivity to elements distorting the local magnetic field, offering insights into the brain's vascular system, minor haemorrhages, calcium and iron depositions [40, 41];

- **Diffusion MRI (dMRI)**: Monitors water molecule diffusion, allowing white matter pathway mapping and presenting insights into white matter tissue microstructure [41–44];

- **Functional MRI (fMRI)**: Utilises blood flow and oxygenation levels to estimate neural metabolic demand, which is a proxy for neural activity [41, 45]. **Resting states fMRI (rsfMRI)** reflects spontaneous neural activity at rest [46–48], while **Task fMRI (tfMRI)** captures activity during specific tasks;

1.3.1 T1 as the Dominant Map

Of all the modalities presented above, sMRI, and in particular the T1-weighted (T1) map, has been the central focus of most past brain age studies. This is due to both its widespread availability in clinical and research databases, but also the fact that it provides basic brain structure and tissue information, and how these change with ageing, making them more easily interpretable [21, 41]. Consequently, brain age prediction using the T1-weighted map has amassed a vast body of literature, which is summarised in Tables A.1-A.2 in Appendix A.

However, when attempting to compare across the studies utilising the T1-weighted map, several difficulties emerge. Firstly, given the large number of neuroimaging databases and processing techniques, there is no established reference dataset, such as MNIST [49], which raises harmonisation challenges [50, 51]. This

is compounded by the fact that different studies report brain age accuracy using different metrics, and the reporting can be carried out using either in-sample cross-validation [52] or left-out test sets [21, 53]. This is why, across all tables shown in this thesis, a weighted mean absolute error (MAE), as described by Cole et al., is used. This facilitates a more balanced comparison between different age cohorts by dividing the reported MAE by the age range of the utilised subject population [8].

When observing historical trends in brain age research with T1-weighted maps, it can be seen that earlier research largely relied on either random or guided voxel sampling [52, 54, 55], or image-derived phenotypes (IDPs), such as measures of cortical area, thickness, and regional tissue volume [35, 56], for training models. Such early work relied primarily on linear techniques for estimating brain age, using methods such as Linear Regression, LASSO [57, 58], Ridge [59] and ElasticNet [35, 60], but also relevance vector machines [61], support vector machines [62] and artificial neural networks [63–65]. These were sometimes coupled with additional processing techniques for data reduction and feature selection, using methods such as Principal Component Analysis (PCA) [35, 52]. While many of these techniques possess inherent feature selection or regularisation capabilities, thus reducing the influence of potentially subjective feature choices, the primary limitation was that the initially derived IDPs might miss subtle changes associated with the ageing brain, especially those related to spatial interactions.

To overcome these issues, more recent work has focused on utilising raw or minimally processed 2D and 3D MRI data [9, 66, 67], which has also been facilitated by advances in available computational hardware. The algorithm of choice for working with these data types has been the Convolutional Neural Network (CNN), given its ability to automatically learn relevant filters and extract patterns from large amounts of data [21], thus avoiding the need for hand-crafted features. CNNs have seen widespread use in brain age prediction studies over the past five years [21, 53, 67], producing increasingly accurate predictions. However, these architectures have their own set of limitations, such as the need for large volumes of training data and computational resources. Additionally, they often lack straightforward

interpretability and may struggle to capture patterns that span large distances or complex interrelationships within the data. While some of these downsides can be mitigated by using large-scale medical databases, such as UK Biobank [36, 41] and large computer clusters, others require radical architectural changes. This is why new studies have started exploring other types of deep learning algorithms, such as graph neural networks [65] and CNN-Vision Transformer hybrids [68].

In the quest for increased prediction accuracy, a number of studies have followed a different route. They discovered that combining multiple complementary orthogonal features derived from the same base modality or map, either during training [63, 69–71] or post-training [67] can lead to model accuracy improvements.

1.3.2 Other Modalities and Maps

Despite the existence of numerous other modalities and maps, of 146 previous brain-age studies, 75% utilised sMRI inputs, and in particular T1-weighted maps. For CNN-based methods, this increases to 95% of the past 39 studies [72, 73]. Given that brain ageing is highly heterogeneous and nonlinear, impacting many structural and functional brain aspects, this over-utilisation of sMRI maps can lead to missed opportunities. Certain age-related changes are not captured by sMRI maps, but can be seen in those derived from other modalities, such as increased mineral and iron depositions, captured by susceptibility-weighted MRI (swMRI) maps, microstructural changes in the white matter, to which dMRI is sensitive, or the impact of prolonged alcohol consumption on cerebral blood flow, reflected by fMRI.

Tables A.3-A.6 in Appendix A, summarise the brain-age prediction work carried out with modalities other than sMRI. Of these, most work has focused on dMRI and rsfMRI, and employed different IDPs as inputs. The only exceptions are the work done by Wood [74] for dMRI, and Liem et al. [75] and Li et al. [76] for rsfMRI, which used volumetric data. When evaluating brain age prediction accuracy, sMRI-based models can be observed to perform better than those using other modalities [8, 75]. A recent development which should be mentioned is the work carried out by Gao et al. [77], who are expanding brain age prediction work

to also incorporate time series functional data, enabling the establishment of links between brain ageing and brain functional network dynamics over time.

1.3.3 Multi-Modal Investigations

Another area of research which is currently under-explored, but which could offer a richer understanding of the brain ageing process, consists of the integration of multiple imaging modalities or maps simultaneously for a single prediction. A summary of past multi-modal studies can be found in Table A.7 in Appendix A. Several studies have underscored the potential of combining modalities, particularly as ageing manifests in ways that a single modality like sMRI might miss [75]. The resulting multi-modality models often outperform their single-modality counterparts, highlighting the complementary insights provided by different imaging techniques in decoding brain ageing [8, 75, 78–80].

Of particular note in this area are the studies conducted by Cole [81] and Smith et al. [56]. Both employed a very large number of IDPs from UK Biobank derived from all 5 core modalities. While Cole discovered that multimodality can substantially improve brain age predictions, Smith determined that numerous modes of population ageing can be found, each showing distinct patterns of structural and functional change.

While most multimodal studies have relied on IDPs, recent work by Mouches et al. [82], Wood et al. [74] and Hofmann [83] has expanded these studies to volumetric data. In addition, by combining 5 blood biochemical indicators and 9 dementia-associated biomarkers with sMRI volumetric data to construct a bilinear fusion model, Ren et al. [84] have shown that it is possible to fuse non-imaging and imaging data to increase prediction accuracy.

1.3.4 Explainability

Several recent efforts in brain age prediction studies have focused on understanding which brain regions significantly contribute to age prediction. Earlier studies qualitatively assessed outputs of neural network layers representing specific aspects

of the input data, also known as feature maps, which displayed varying spatial patterns depending on the composition of the training set [8, 54, 55, 58, 85]. Later works applied attention gates [21, 86] and saliency maps [87] to emphasise influential regions in 3D MRI images. Although some insights matched known linear variations with age, the interpretability of attention and saliency maps was sometimes limited. This aligns with the non-linear modes of ageing suggested by Smith et al., indicating a multi-faceted ageing process [56].

Subsequent research explored various methodologies for explainability. Taylor et al. [88] utilised Layer-wise Relevance Propagation (LRP) and DeepLIFT saliency mapping, observing differential regional relevances as ageing progressed. Popescu [89] proposed localised training, with Sanford corroborating the validity of the approach by identifying sex-specific differences in ageing [90]. Hepp's study [91] employed Grad-CAM, highlighting the importance of certain brain sub-regions. Lombardi et al. [92] integrated local explainable DL techniques like SHAP and LIME with MRI scans to discern age-related morphological patterns, finding SHAP as a more reliable method for discerning ageing mechanisms. Finally, He et al. [68] proposed using a Vision Transformer model, taking advantage of its inherent explainability to identify which features the network utilises when making predictions.

Yet, although interesting, explainability is not a solved problem, and is not the focus of this thesis.

1.4 Translational Studies and Clinical Applications

Once a precise statistical model is established and brain age deltas are calculated for test subject cohorts, numerous studies have sought to correlate deviations from standard brain ageing with factors such as brain diseases, psychological disorders, other pathologies, genetics, and environmental factors. These correlations suggest that the brain age deltas, obtained through various imaging modalities, may offer clinical insights [9, 21, 53]. Such studies generally fall into two categories: those

centred exclusively on populations with specific pathologies, and those exploring broader populations encompassing both healthy subjects and those with pathologies. Typically, the former involves smaller study sizes, with tens to hundreds of subjects [8], whereas the latter includes subjects in the several hundreds to thousands.

The following sections will present a broad overview of the associations observed thus far between brain age deltas and different neurological pathologies, lifestyle factors, genetics, and environmental factors. These lists are by no means exhaustive. In addition, none of these associations imply causality, and many of the presented studies involved small populations. This means that results should not be readily generalised. Furthermore, the overwhelming majority of the studies referenced in the following sections utilised T1-weighted images, but some employed other maps, so readers are advised to delve deeper into these studies for more information. Also, it should be noted that correlations can either be positive (suggesting accelerated ageing), or negative (suggesting resilience to ageing).

1.4.1 Neurological Pathologies

When considering neurological pathologies, brain age deltas have been associated with the following conditions:

- schizophrenia and psychosis [93–105];
- major depressive disorder (MDD) [93, 96, 102, 106–109];
- the use of antidepressants, which was found to mediate the accelerated ageing [108];
- borderline personality disorder [93];
- Alzheimer's disease (ADi) [52, 110–116];
- psychosis [102, 117, 118];
- bipolar spectrum disorder [96, 102];
- multiple sclerosis (MS) [56, 102, 119];

- mild cognitive impairment (MCI) and dementia [75, 102, 110–112, 114, 120–125];
- reduced cognitive performance [81, 124];
- traumatic brain injuries [126–128];
- treatment-resistant refractory focal epilepsy [129];
- eating disorders in male subjects [21];
- obsessive-compulsive symptoms [97];
- temporal lobe epilepsy [130, 131];
- phobias [21];
- autism [132];

Inconsistencies can, however, emerge between studies. For instance, Koutsouleris et al., Hajek et al., and Nenadic et al. [93, 99, 101] found no association between predicted brain age and bipolar spectrum disorder, despite the findings of Kaufmann et al. and Ballester et al. [96, 102]. This should serve as a cautionary tale regarding the generalisation of any findings.

Using transfer learning, Leonardsen et al. [53] showed that models initially trained to predict brain age can also be employed, with a substantial degree of accuracy, in predicting ADi, MS, MCI, Psychotic Disorders, Schizophrenia and Mood Disorders. This reinforces the idea that certain internal model representations underpinning the prediction of brain age are also associated with, or relevant to, these conditions. Bashyam et al. [133] obtained similar results for ADi, Schizophrenia and MCI, and Cheng et al. [134] for ADi and MCI.

1.4.2 Lifestyle and Biomedical Factors

Numerous studies have also found associations to different lifestyle and biomedical measurements, such as:

- medicines (metformin, a diabetes medication [21, 56], cholesterol medication [21], blood pressure medication [21, 53], syndol [21], anti-inflammatory medication such as aspirin [21, 135], gliclazide [21], ramipril [21, 56], amlodipine [35], lithium treatments for subjects with bipolar disorders [136]);
- overall non-fat body size related factors (height [56], strength [56], lung capacity [56], metabolic rate and weight [56]);
- cognition (processing speed [56], reaction time [56]);
- socio-economic status markers (malnourished during fetal development [137], adversity and victimisation [109], relationships with family at home [109], relations with peers [109], academic functioning [109], involvement in general hobbies/interests [109], number of people and vehicles in a household [35], the average household income [35]);
- biological measures (bone density [56], body size [56], fat measures [56], metabolic and cardiovascular functions [56], blood pressure [35, 53, 56, 81, 138, 139], hypertension [21, 137], heart rate [56], vascular and heart problems [56], haemoglobin [56], red blood cell count [35, 56], cardiac output [56], blood markers of liver and kidney functions [139], weaker grip [140], reduced lung function [140], slower walking speed [140], age at menopause [56], menstrual cycle hormonal variations [4, 141], reductions in estradiol and progesterone post birth [142]);
- life factors (alcohol consumption [35, 53, 56, 59, 81, 137, 138, 143–145], smoking [35, 53, 56, 81, 143, 146, 147], maternal smoking [56], physical activity [56, 147], number of siblings [56], sleep duration [56], preterm delivery [21, 148], gestational diabetes [21], delivery complications in male subjects [21], cereal intake [21, 35, 53]);

- cognitive test scores (processing speed [56], IQ [56], verbal fluency [113, 143], lower fluid intelligence [140]);
- mental health (anxiety [56], depression [56]);
- disease and chronic conditions (diabetes [21, 35, 53, 56, 81, 143], diabetes-related eye problems [53], stroke [149], obesity [21, 35, 100, 150], presence of Tumour Necrosis Factor alpha (TNF alpha) [143] and IGF-1 [35, 53, 151], joint pain [21], stroke [81, 138]);

It should be noted that contradictions exist between the findings of various brain age prediction studies. For instance, Franke et al. state that subjects born pre-term displayed younger appearing brains, suggesting a delayed structural brain maturation [148]. This is disputed by Hedderich et al. [152] who found elevated brain age deltas in premature-born adults. These discrepancies might arise from the small sample sizes used in certain studies. Larger cohorts can provide greater statistical power and confidence. In addition, numerous studies have found different associations between male and female subject cohorts [21, 139].

When considering all of the above, certain underlying mechanisms shared across multiple pathologies emerge, such as neuroinflammation, which might influence long-term brain health in those with chronic conditions [8]. Diabetes and obesity, for instance, are known to induce chronic low-level systemic inflammation [153–155]. For women, the anti-inflammatory properties of oestrogen are associated with reduced BAGs [141]. Moreover, the use of anti-inflammatory medications has shown protective effects [135]. Inflammation and oxidative stresses also have an impact on telomeres, with telomere length being strongly associating with older appearing brains [121].

Yet, despite these observations, it is difficult to establish causal links between the different factors and deviations from normal brain ageing trajectories. This is because it is unclear what the compounding effect of numerous factors is. Thus, for each identified association, longitudinal studies are required. An example of

this is the work by Elliott et al. [123], which established a link between brain ageing and cognitive decline.

1.4.3 Environmental Factors

Studies also revealed that certain environmental factors can also be associated with brain age deviations, acting as protective factors. These include:

- meditation [156];
- playing musical instruments [157];
- education [158];
- physical activity [158];
- participation in social activities [53];
- time spent outdoors in summer [35];

1.4.4 Genetics

Extensive work by Lowe et al. [112], Cole et al. [9], Scheller et al. [159], Kaufmann et al. [102], Smith et al. [56], and Millar et al. [115] also investigated the associations between various genetic factors and brain ageing, with potential implications for understanding and predicting neurodegenerative diseases and cognitive decline.

Studies like those involving the Apolipoprotein E (APOE-ε4) genotype reveal correlations with accelerated progression of Alzheimer's and Mild Cognitive Impairment (MCI), reflected in older apparent brain ages and distinct neural activation patterns, hinting at neuronal compensation mechanisms [112, 159]. However, more work in this space is required, as the effect of APOE-ε4 was only marginally significant in the full model control- ling for all biomarkers and covariates [115]. Studies have also found that predicted brain age is a heritable biomarker across both small [9] and large subject cohorts [56, 102] demonstrating that genetic factors can influence brain ageing, beyond mere noise or disease-related atrophy.

Recent genome-wide association studies (GWAS) have identified specific genetic variants associated with brain age deltas. For instance, variants like rs2435204-G and rs1452628-T have been linked to structural brain changes and are implicated in broader neuropsychiatric and neurodevelopmental disorders, including dementia, neuroticism, and cerebral ischemia [160–169]. Comprehensive GWAS further highlight the association of single nucleotide polymorphisms (SNPs) with brain age gaps and a spectrum of neurological disorders, implicating genes involved in biological processes such as blood pressure regulation, neuron migration, Alzheimer's disease, neurogenesis, and inflammation [56, 102].

These findings not establish a genetic basis of brain ageing but also connect these genetic markers with a wide array of health outcomes, emphasising the importance of genetics in the predictive and diagnostic realms of neurodegenerative diseases. However, as no genetic associations were performed in this thesis, this topic is not further discussed or investigated.

1.5 Clinical Transition of Brain Age Studies

Long term, the transition of BAG work from academia to clinical practice is desirable. A major limitation, however, is the reliance of many studies on high-quality research data, making clinical adoption challenging [74]. However, Wood et al. [74] demonstrated the feasibility of using clinical-grade scans with a CNN for accurate brain-age predictions across multiple scanner vendors. Also, Leonardsen et al. [53] showed that brain age prediction models can be used to predict different neural pathologies.

Transitioning to clinical applications can lead to early diagnosis of neurodegeneration, leverage large hospital databases for model refinement, drive research into the causality of brain ageing, and validate other medical image analysis tools. Cole et al. [170] showcased the viability of brain age predictions for survival analysis, utilising real-world hospital datasets, and the potential of point-of-care MRI scanners. Baecker et al. [171] highlighted five clinical applications of the BAG: as a marker for brain health, early disorder detection, prognosis tracking, differential diagnosis, and

treatment outcomes. They emphasised challenges like scanner variability, granularity in brain age interpretation, and the dynamic nature of brain age across a lifespan.

1.6 Literature Gaps and Thesis Contributions

Despite the large amount of work already carried out in the brain age prediction field, numerous gaps still exist. Some of these are addressed in this thesis, including:

- Providing a the selection criteria, and associated code, allowing the reproduction of the findings described in this thesis when utilising data from UK Biobank, as well as facilitating comparisons between current and future brain age prediction models;
- Expanding brain age prediction work to include volumetric data from all 5 core MRI modalities: sMRI, dMRI, swMRI, rsfMRI, and tfMRI;
- Addressing the question of what is an appropriate deep learning architecture to explore the brain age paradigm, and what metrics should be utilised to judge what constitutes "good" predictions;
- Introducing a novel, Vision-Transformer based brain age prediction network, which addresses several of the issues common in modern CNNs;
- Determining if different modalities and maps encode different information relating to the ageing brain;
- Exploring the integration of multimodal data to enhance the accuracy and robustness of brain age prediction, addressing the challenges associated with combining multiple maps;
- Investigating various techniques for the optimal fusion of multimodal brain imaging data, including linear and non-linear approaches, as well as deep learning ensembling methods, to determine the most effective strategy for multimodal brain age prediction;

- Determining if multiple independent modes of ageing can be identified within a population of subjects based solely on map predicted brain ages;

1.7 Thesis Structure

This thesis is organised as follows:

- **Chapter 2** provides a general background to MRI, the various modalities and maps employed, and a discussion of the machine learning techniques utilised throughout this thesis;
- **Chapters 3-5** present the contributions discussed above. Each chapter is composed of an overview, general methods, and individual sections presenting and discussing the results. Each chapter ends with a discussion of the most important findings;
- **Chapter 6** is the thesis conclusion, summarising the main findings, discussing limitations, and proposing a set of future directions;

1.8 Statement of Originality

I declare that I am the sole author of this thesis, and I produced all of the tables and figures included herein. Comments were provided by Prof. Stephen M. Smith, Prof. Ana I.L. Namburete, Dr. Frederik J. Lange, Dr. Torsten Schindler, and Dr. Stanislaw Adaszewski. Unless otherwise specified, the pre-processing, code development, and analysis are my own individual work.

- **Chapter 3**: The code for the Simple Fully Connected Network (SFCN) was provided by Dr. Han Peng. The code for determining the correlations between non-image derived phenotypes (nIDPs) and brain age deltas was written by Dr. Emma Bluemke, modified by Dr. Nicola K. Dinsdale, and then adapted by me. Dr. Dinsdale also provided general advice on the implementation of the chapter. UK Biobank data was processed according to UK Biobank pipeline by Dr. Fidel A. Almagro. For multi-modal ensembling, the Repeated

ElasticNet (RENT) code was provided by Dr. Anna Jenul. The PCA-ICA method was based on Prof. Smith's code, modified by me.

- **Chapters 4-5**: UK Biobank data and the nIDP associations code were also used here. In Chapter 5, the BA-SWIN code was adapted by me from the SWIN-UNETR code developed by Dr. Ali Hatamizadeh, included in the Project MONAI codebase.

1.9 List of Publications

Peer Reviewed Conference Proceedings:

- **Roibu, Andrei-Claudiu**, Stanislaw Adaszewski, Torsten Schindler, Stephen M. Smith, Ana IL Namburete, and Frederik J. Lange. "Brain Ages Derived from Different MRI Modalities are Associated with Distinct Biological Phenotypes." In 2023 10th IEEE Swiss Conference on Data Science (SDS), pp. 17-25. IEEE, 2023.

1.10 Code Repositories

The following git repositories contain the codes utilised in this thesis:

- **Chapter 3**:
  - Network codes: `https://github.com/AndreiRoibu/AgeMapper`
  - Analysis codes: `https://github.com/AndreiRoibu/AgeMapper-Analysis`
- **Chapter 4**:
  - Network codes: `https://github.com/AndreiRoibu/MultiAgeMapper`
  - Analysis codes: `https://github.com/AndreiRoibu/MultiAgeMapper-Analysis`
- **Chapter 5**:
  - Network codes: `https://github.com/AndreiRoibu/SwinAgeMapper`

  - Analysis codes: `https://github.com/AndreiRoibu/SwinAgeMapper-Analysis`

- Overall thesis LaTeX code: `https://github.com/AndreiRoibu/Thesis-Draft`

2

General Materials and Methods

Contents

This chapter serves as a foundational introduction to the different data types and methodologies employed within this thesis. While the aim is to offer readers a broad understanding and contextualise the techniques, modalities, maps, and networks incorporated, it is important to emphasise that this is a generalised overview designed to facilitate comprehension. Each subsequent chapter will delve deeper into the specific methods pertinent to its content, ensuring that each segment is self-contained and comprehensive in its exposition.

2.1 Magnetic Resonance Imaging

In this section, a brief overview of the data utilised in this thesis is given. First, the concept of magnetic resonance imaging (MRI) is presented, followed by a

description of the various MRI modalities and maps, and concluding with general considerations surrounding the data preprocessing utilised throughout this work. UK Biobank [36], a large-scale biomedical database, is the sole source for all the data, and thus any data processing was carried out using the FMRIB Software Library (FSL) toolkit [172] following the UK Biobank Protocol [41] and the established UK Biobank Pipeline [173].

2.1.1 Introduction to MRI

MRI is one of several non-invasive medical imaging techniques, which help medical personnel in the diagnosis and monitoring of diseases by providing insight into the internal structures of the body. It uses radio waves and strong magnetic fields to generate very detailed scans of various soft tissues. MRI has become one of the imaging modalities of choice for the investigation of the nervous system, as it provides information about the structures of the brain, its structural and functional connectivity, neural activity, and the deposition of certain elements or compounds, such as calcium and iron, in the brain.

When a person is placed inside an MRI scanner, the magnetic field temporarily aligns the hydrogen atoms present in the water contained in their tissues. Radio-frequency (RF) pulses, tuned to the resonance frequency of the hydrogen atoms, are then used to disturb this alignment. When the RF pulses are turned off, the hydrogen atoms emit signals as they undergo relaxation processes, returning to their equilibrium state and realigning with the magnetic field. These relaxation processes are characterised by two main constants: $T1$ (longitudinal relaxation) and $T2$ (transverse relaxation), with an extended version, $T2^*$, accounting for field inhomogeneities. The emitted signals, which are influenced by these relaxation properties, are detected and converted into images by the computer system.

The main strength of the magnetic field is denoted as B_0, and typically varies from 0.2 to 7 Tesla, influencing the quality of the images obtained. The strength of this magnetic field influences the resolution and clarity of MRI images. Additionally, the MRI system utilises gradients in the magnetic field, superimposed on the B_0

field, to determine the exact location of the hydrogen atoms. This enables the creation of detailed spatial images.

Different settings in the MRI protocol, referred to as imaging parameters, can highlight various tissue properties. This enables the obtaining of information about different aspects of the body's internal structures. These settings are adjusted to exploit differences in the rate at which hydrogen atoms realign with the magnetic field, providing contrast between different tissue types.

2.1.2 MRI Modalities

Varying the MRI acquisition parameters produces images (modalities) that are sensitive to different structural and functional properties of the brain. The MRI data in UK Biobank corresponds to 5 core modalities: structural MRI (sMRI), susceptibility-weighted MRI (swMRI), diffusion MRI (dMRI), resting-state functional MRI (rsfMRI), and task functional MRI (tfMRI). Multivariate modalities, such as dMRI and fMRI, can be further processed to produce sets of images (maps) that represent summary measures of interest. In UK Biobank, these maps include:

- 6 sMRI maps, reflecting the gross anatomy of the brain;
- 1 swMRI map ($T2^*$), revealing information about compounds distorting local magnetic fields, such as cerebral microbleeds, or iron, calcium, or myelin concentration;
- 19 diffusion MRI maps, providing information about the brain white matter and structural connectivity, including:
  - 9 quantifying aspects of water diffusion and microstructural properties;
  - 9 representing the same features but projected onto a white matter skeleton using tract-based spatial statistics (TBSS);
  - 1 summation of 27 major axonal tracts obtained with probabilistic tractography [41], which at the voxel level show the probability of a streamline from a seed region traversing that particular voxel;

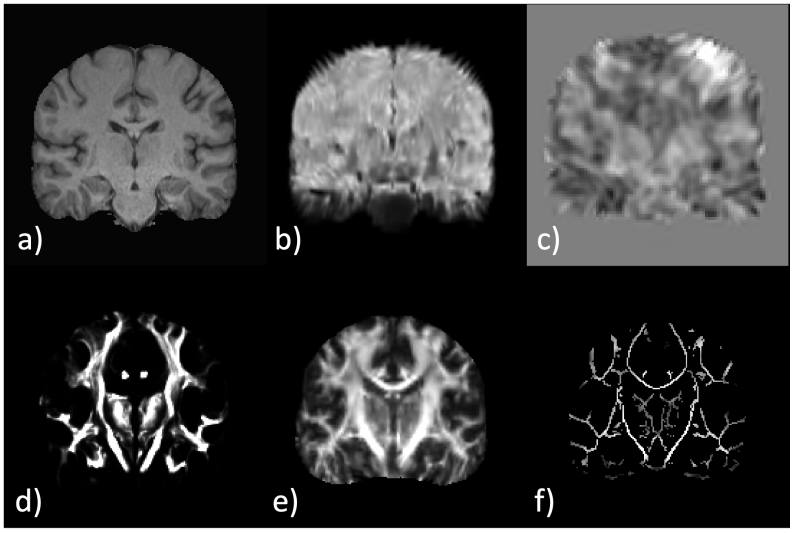


Figure 2.1: MRI modality and map examples, viewed in the coronal plane: **(a)** sMRI T1 Nonlinear, **(b)** swMRI, **(c)** tfMRI, **(d)** dMRI summed probabilistic tracts, **(e)** dMRI fractional anisotropy (FA), **(f)** dMRI TBSS FA

- 25 resting-state fMRI (rsfMRI) ICA dual-regressed z-score normalised maps, representing neuronal activity at rest;

- 6 task fMRI (tfMRI) maps, representing brain activations in the presence of a stimulus, 3 being z-statistic maps and 3 contrasts of parameter estimate maps (COPE);

The full list of 57 maps utilised in this thesis is presented in Table 2.1 and examples in Figure 2.1. The remainder of this section provides intuition regarding what the signal intensity means for each map.

Neuroimaging data from UK Biobank was processed with the standard UK Biobank Pipeline [173]. This includes, among other things, performing brain extractions via the FSL Brain Extraction Tool (BET) [174] and linearly registering the images to the MNI152 template [175–177] via FMRIB Linear Registration Tool (FLIRT) [178]. All imaging data comes in the NIFTI file format [179]. A standardised coordinate frame is utilised: the sagittal plane in the y-z axis, the coronal in the x-z, and the transverse in the x-y.

Table 2.1: UK Biobank MRI core modalities and associated maps utilised throughout this thesis. Where applicable, maps are further grouped by overall type or processing method, to facilitate comprehension.

Modality	Maps
Structural MRI (sMRI)	T1 Nonlinear, T1 Linear, Jacobian, VBM, T2 FLAIR Nonlinear, T2 Lesions
Susceptibility-Weighted MRI (swMRI)	swMRI ($T2^*$)
Diffusion MRI (dMRI)	Fractional Anisotropy (FA), Mean Diffusivity (MD), Mode of Anisotropy (MO), Eigenvalues in three primary directions (L1, L2, L3), Isotropic or free water volume fraction (ISOVF), Intra-cellular volume fraction (ICVF), Orientation dispersion index (OD); TBSS FA, TBSS MD, TBSS MO, TBSS L1, TBSS L2, TBSS L3, TBSS ISOVF, TBSS ICVF, TBSS OD; Summed Probabilistic Tracts;
Resting State Functional MRI (rsfMRI)	rsfMRI-0, rsfMRI-1, rsfMRI-2, rsfMRI-3, rsfMRI-4, rsfMRI-5, rsfMRI-6, rsfMRI-7, rsfMRI-8, rsfMRI-9, rsfMRI-10, rsfMRI-11, rsfMRI-12, rsfMRI-13, rsfMRI-14, rsfMRI-15, rsfMRI-16, rsfMRI-17, rsfMRI-18, rsfMRI-19, rsfMRI-20, rsfMRI-21, rsfMRI-22, rsfMRI-23, rsfMRI-24, rsfMRI-25;
Task Functional MRI (tfMRI)	tfMRI-1, tfMRI-2, tfMRI-5; tfMRI-COPE-1, tfMRI-COPE-2, tfMRI-COPE-5;

2.1.2.1 Structural MRI

sMRI produces visualisations showing the gross anatomy of the brain. Depending on the acquisition parameters, sMRI can yield either T1-weighted images or T2-weighted images, where differences in intensities between tissues across the image correspond to differences in either the $T1$ or $T2$ values, respectively.

T1-weighted images [180] display the brain's tissue volumes and morphology: white matter (WM), grey matter (GM), and cerebrospinal fluid (CSF). Besides detailing individual anatomy, T1 serves as a reference for other modalities, especially during the registration process. Typically captured at $1mm$ isotropic resolution in research settings, T1 preprocessing entails, among other steps, either linear or non-linear registration to the MNI152 space, leading to the existence of two maps used in this work: T1 Nonlinear and T1 Linear. In clinical settings, anisotropic voxels might be present due to limited acquisition times.

An output of the non-linear registration process, Jacobian maps provide a visual representation of the Jacobian determinant for every voxel. They represent the voxelwise change in volume due to the deformation. Consequently, Jacobian

voxel intensities greater than 1 represent expansion, and less than 1 (but greater than 0) represent contraction.

Voxel Based Morphometry (VBM) enables voxel-wise comparisons of local GM density across populations. UK Biobank VBM images are derived from T1 images using the FSL-VBM protocol [181]. In VBM images, the signal intensity in each voxel corresponds to the local concentration of GM, and thus, unlike T1 images, VBM images provide a quantitative assessment of GM density. In addition, these images are typically processed to a $2mm$ isotropic resolution, contrasting with the $1mm$ resolution of other structural MRI (sMRI) maps.

T2 FLAIR (Fluid Attenuated Inversion Recovery) images, also known as T2 FLAIR Nonlinear, are sensitive to the $T2$-relaxation times and provide valuable information for identifying tissue alterations associated with various pathologies, including WM lesions, tissue changes near the hemisphere periphery, or the periventricular region [182].

Using the T2 FLAIR Nonlinear images and the Brain Intensity AbNormality Classification Algorithm (BIANCA) [183], 3D binary maps representing segmentations of the white matter hyperintensities (WMH) can be obtained. These T2 Lesions are significant in the study, diagnosis and treatment of cardiovascular disease [183–185], as well as rare conditions such as NMOSD - Neuromyelitis Optica Spectrum Disorder [186].

2.1.2.2 Susceptibility-Weighted MRI

swMRI is sensitive to tissue compounds exhibiting paramagnetic (like deoxyhaemoglobin, ferritin, and haemosiderin) [187], diamagnetic (such as bone minerals and dystrophic calcifications) [187], and ferromagnetic attributes (for instance, iron or gadolinium depositions) [188, 189]. Collectively, these compounds interact with the localised magnetic field, inducing RF signal distortions [40, 190, 191]. The distortions, or field inhomogeneities, lead to differences in intensities between different parts of the image, which correspond to differences in the $T2^*$ values. Clinically, this enables swMRI images to provide insight into venous brain vasculature and

cerebral microbleeds resulting from deposits like haemosiderin [192], as well as iron, calcium, or myelin concentrations in the brain.

2.1.2.3 Diffusion MRI

dMRI captures the Brownian motion of water molecules and their diffusion in the brain, offering insights into the structure and microstructure of tissues. In the WM, anisotropic diffusion prevails, with water molecules moving primarily parallel to neuronal axons, whereas the GM and regions occupied by the CSF exhibit isotropic diffusion [193], indicating a less structured movement. To generate the dMRI images, the scanner gradient coils apply a magnetic field in specific directions, allowing the detection of water diffusion along those axes. The dMRI signal is sensitised to the diffusion by these gradients, being attenuated when diffusion is aligned with the gradient direction, and stronger when diffusion is perpendicular to the gradient direction.

As dMRI is susceptible to image distortions from various sources, it requires post-processing corrections [194–196]. Following this, the dMRI images are further processed to extract summary measures of interest. Two models are fit to UK Biobank data: Diffusion Tensor Imaging (DTI) [197] and Neurite Orientation Dispersion and Density Imaging (NODDI) [198].

DTI models water diffusion within a voxel using a tensor, which can be visualised as an ellipsoid. The ellipsoid axes represent the principal diffusion directions, the length of each axis indicating the magnitude of diffusion along that direction. The longest axis indicates the direction of maximum diffusion, while the shorter axes, perpendicular to the primary direction, represent the lesser degrees of diffusion freedom. Thus, for each voxel, the tensor eigenvectors and eigenvalues represent the direction and magnitude of diffusion, respectively. From the DTI model, several maps are derived:

- Fractional anisotropy (FA), representing the extent to which diffusion in each voxel is directionally dependent, with values ranging from 0 (isotropic diffusion) to 1 (fully anisotropic diffusion);

- Mean diffusivity (MD), representing the average diffusion magnitude of water molecules irrespective of direction [199];

- Mode of anisotropy (MO), which is a map mathematically orthogonal to FA, quantifying second-order geometric properties [200, 201], its voxel intensities spanning from -1 (indicating oblate diffusion ellipsoid orientation) to 1 (indicating prolate diffusion ellipsoid orientation), with 0 representing isotropic diffusion;

- Three eigenvalues, denoting diffusion in the primary directions (L1, L2, L3), representing diffusion magnitudes. Typically, isotropic diffusion regions exhibit similar values across L1, L2, and L3 and possess low FA, while anisotropic areas present distinct values across these three directions and have elevated FA values. Brain ageing studies have generally employed a map representing either the Axial Diffusivity (AD), which equivalent to L1, or the Radial Diffusivity (RD), which is the average of the L2 and L3. In this work, however, the three components are considered as separate input maps. This is because the work by Gong et al [202] suggest that the individual maps might encode bespoke information relating to the ageing brain;

NODDI, on the other hand, is utilised for extracting voxel-wise microstructural parameters. This produces several more maps, including:

- Isotropic or free water volume fraction (ISOVF), which is an index of extracellular water diffusion, and can be considered a measure of isotropic water diffusion [198, 203];

- Intra-cellular volume fraction (ICVF), which acts as an index of WM neurite density, any changes in its values being indicative of microstructural organisation changes at the voxel level;

- Orientation dispersion index (OD), expressing the extent of directional complexity of diffusion, and being an index of intracellular neurite dispersion;

When processing diffusion data from multiple subjects, standard registration algorithms may produce misregistration or misalignment artefacts. Tract-based spatial statistics (TBSS) addresses this by using nonlinear registration and a skeletonisation process [204, 205]. Initially, a mean FA image is generated and then thinned to form a mean FA skeleton, representing the centres of all tracts common to a reference subject group. The measures derived from both DTI and NODDI are then projected onto this skeleton, by filling the skeleton with values from the closest relevant tract centre by perpendicularly searching the local skeleton structure for the peak FA value, and then projecting the value at that pixel for each measurement. This leads to the creation of 9 additional TBSS maps, which are resistant to alignment discrepancies.

Finally, the structural connectivity of the brain can be investigated by the voxel-wise summation of 27 significant WM tracts. The voxels of each of the 27 tract maps indicate a streamline density map, or the probability of a tract originating from a seed region, as defined by an ROI mask set by AutoPtx [206], passing through each voxel. The fibre orientations are estimated using BEDPOSTX [207] and probabilities are calculated using the PROBTRACKX probabilistic tractography model [207–210]. As there is minimal overlap among the 27 tracts, they were voxel-wise summed to form the Summed Tracts map [204].

2.1.2.4 Functional MRI

Neuronal activity, which occurs instantly either at rest or following a stimulus, is estimated by observing changes in blood flow and oxygenation (hemodynamics) [211, 212], with the MRI signal varying based on the difference in magnetic properties between oxygenated and deoxygenated blood. Since deoxygenated haemoglobin is paramagnetic and oxygenated haemoglobin is diamagnetic, it produces local magnetic field variations, creating the blood oxygenation level dependent (BOLD) effect [212].

Functional scans can be acquired either at rest, during subject relaxation, without a specific task or stimulus [46], or when a task is performed. Resting state fMRI

(rsfMRI) reveals intrinsic brain activity and deduces apparent connectivity between brain regions. It is consistent and organised across individuals [47, 213–218]. In UK Biobank, 25 major functional subdivisions across cortical and sub-cortical GM were discovered using an ICA-based technique [219–221]. Each subdivision consists of voxels that, although appearing non-continuous when thresholded, are actually part of a continuous network, each voxel having a certain degree of association with a specific functional component [41]. After, dual regression was applied [222, 223] to produce subject-specific spatial maps as z-statistic maps. While the 25 resulting maps have no official designation, Lee et al. [224] proposed the following classification based on the network number: visual $(2, 4, 8, 19)$, motor $(10, 11, 12)$, cerebellum (15), auditory (17), subcortical (18), default mode (1), limbic (7), precuneal (20), salience (14), attention (3), right and left fronto-parietal $(5, 6)$, language-related $(9, 13, 21)$, and executive control (16). Maps $22, 23, 24, 25$ are considered group-ICA artefacts [41]. Yet, they are still included in this work as they might still contain useful brain ageing information.

Task fMRI (tfMRI) scans are acquired when subjects are exposed to the Hariri faces/shapes emotion experiment [225, 226]. As fMRI is not quantitative, maps between the stimulated state and a baseline are necessary to determine whether a brain region is associated with a particular task. In UK Biobank, three primary maps are explored: shapes versus rest, faces versus rest, and the differences between faces and shapes. These correspond to maps $1, 2, 5$. These maps are generated using the FMRI Expert Analysis Tool (FEAT) [227], which results in the creation of z-statistic and contrast of parameter estimate (COPE) maps. Though seemingly alike, COPE maps are not normalised by regression and specifically highlight differences between conditions, while z-stat maps are normalised.

2.1.3 IDPs, nIDPs and Confounds

Besides the MRI maps, 3 additional data types were employed in this thesis. Firstly, utilising the automated UK Biobank image processing pipeline [173], 3921 distinct

image-derived phenotypes (IDPs) were derived, which can be further thematically split (Appendix B Section B.1).

Secondly, for exploring the correlations between brain age deltas and diverse biomedical and lifestyle metrics, a collection of 17,526 non-genetic, non-imaging derived phenotypes (nIDPs) from UK Biobank were used. These measurements were also split into 19 thematically defined categories for interpretability [21, 173] (Appendix B Section B.2).

In large-scale health studies, it is crucial to consider how unrelated factors might create false or misleading links between variables, and lead to incorrect conclusions [228]. To prevent this, both IDPs and nIDPs were linearly deconfounded [229] using 613 confounds listed in UK Biobank, split into 6 categories [228] (Appendix B Section B.3).

2.2 Machine Learning

This section provides a brief introduction to the machine learning (ML) methods used throughout this thesis.

Given an input $\boldsymbol{X}$, which can be an individual's IDP or map volumetric data, the aim is to predict a subject's chronological age y by learning a function $f : \mathcal{X} \mapsto \mathcal{Y}$, expressed as an ML algorithm, given a batch of inputs $\mathcal{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_N\}$ and corresponding targets $\mathcal{Y} = \{y_1, y_2, ..., y_N\}$. The complexity of the models underpinning f can vary, and the following sections discuss these in more detail.

2.2.1 Linear Methods

2.2.1.1 Supervised Linear Methods

Supervised linear regression methods aim to model the relationship between a set of independent variables and a continuous dependent variable. Using the learned relationship, predictions can be made for new data. By minimising the differences between the predicted and actual values, the optimal relationship parameters are determined.

Linear Regression (LR - Equation 2.2.1.1), the simplest supervised regression method, finds the line of best fit for a given dataset, characterised by a set of weights, $\mathbf{w}$, and a bias intercept, b. The LR coefficients $\hat{\mathbf{w}} = \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}$ can be calculated either using a closed-form solution (Equation 2.2.1.2) when $\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}$ is invertible, where $\tilde{\mathbf{X}}$ is $\mathbf{X}$ augmented with a column of ones to account for b, or using ML techniques such as gradient descent with an ordinary least squares (OLS) loss function (Equation 2.2.1.3 - loss functions are introduced in Section 2.2.2.6), where $\tilde{\mathbf{x}}_i$ and y_i are the ith rows of $\tilde{\mathbf{X}}$ and $\mathbf{y}$.

$$\mathbf{y} = \mathbf{w} \cdot \mathbf{X} + b \tag{2.2.1.1}$$

$$\hat{\mathbf{w}} = \begin{bmatrix} w \\ b \end{bmatrix} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \mathbf{y} \tag{2.2.1.2}$$

$$L_{OLS}(\hat{\mathbf{w}}) = \sum_{i=1}^{n} (y_i - \hat{\mathbf{w}}^\top \tilde{\mathbf{x}}_i)^2 = \|\mathbf{y} - \tilde{\mathbf{X}}\hat{\mathbf{w}}\|^2 \tag{2.2.1.3}$$

LR can be modified by adding a penalty term either to the sum of squared coefficients (L2 Regularisation or Ridge Regression - Equation 2.2.1.4), or by adding the penalty to the absolute value of coefficients (L1 Regularisation or Lasso Regression - Equation 2.2.1.5). For these, χ is the coefficient controlling the penalty term, m represents the number of predictor variables, or features, in the training dataset, and $\hat{w}_j$ is the jth element of the estimated coefficient vector $\hat{\mathbf{w}}$.

$$L_{Ridge}(\hat{\mathbf{w}}) = \|\mathbf{y} - \tilde{\mathbf{X}}\hat{\mathbf{w}}\|^2 + \chi \sum_{j=1}^{m} \hat{w}_j^2 \tag{2.2.1.4}$$

$$L_{Lasso}(\hat{\mathbf{w}}) = \|\mathbf{y} - \tilde{\mathbf{X}}\hat{\mathbf{w}}\|^2 + \chi \sum_{j=1}^{m} |\hat{w}_j| \tag{2.2.1.5}$$

Ridge regression mitigates the issue of multicollinearity, where independent variables are highly correlated, improving the model's stability when dealing with correlated features. However, it does not perform feature selection, leaving potentially irrelevant, or "noisy", features to be included in the final model. Lasso,

on the other hand, encourages the model to be sparse, effectively reducing features to retain only the most important, improving interpretability. Nevertheless, Lasso may not perform well when the number of features exceeds the number of observations, or when predictors are highly correlated, as it might arbitrarily select among them. ElasticNet blends Ridge and Lasso, striking a balance between them. It is guided by two hyperparameters: α, the mixing parameter that controls the trade-off between L1 and L2 regularisation, and λ, the parameter that controls the strength of the regularisation (Equation 2.2.1.6). n represents the training dataset's sample count.

$$L_{EN}(\hat{\mathbf{w}}) = \frac{\sum_{i=1}^{n}(y_i - \hat{\mathbf{w}}^\top \tilde{\mathbf{x}}_i)^2}{2n} + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^{m} \hat{w}_j^2 + \alpha \sum_{j=1}^{m} |\hat{w}_j| \right) \tag{2.2.1.6}$$

Compared to LR, however, these three methods introduce the need for cross-validated hyperparameter tuning experiments for determining the optimal values of χ, λ and α.

2.2.1.2 Unsupervised Linear Methods

Linear methods extend beyond regression tasks. Often, discerning hidden patterns in unlabelled data is useful. Unsupervised linear methods address this, aiding in dimensionality reduction, feature extraction, and visualisation. Dimensionality reduction streamlines complex data to essential features, mitigating overfitting and computational load. Feature extraction isolates important variables, enhancing model performance. And finally, visualisation translates multifaceted data into an interpretable format. This thesis leverages two such techniques: Principal Component Analysis (PCA) and Independent Component Analysis (ICA).

PCA is a data dimensionality reduction method and works by creating a set of orthogonal 'principal components' that capture the maximum data variance. The first component captures the most variance, and each subsequent one captures the next highest variance while remaining orthogonal to the previous ones. PCA can be implemented using Singular Value Decomposition (SVD), which factorises a real matrix $\mathbf{A}$, of dimension $m \times n$, where m and n are the number of observations and

the number of variables respectively, by identifying a new coordinate system of orthogonal vectors, known as singular vectors, that optimally capture the variance in the data (Equation 2.2.1.7). Here, $\mathbf{U}$ is an $m \times r$ orthogonal matrix, where $r = \min(m, n)$, containing the left singular vectors of $\mathbf{A}$. $\mathbf{\Sigma}$ is a diagonal matrix containing the singular values of $\mathbf{A}$ along the diagonal. $\mathbf{V}^T$ is the transpose of an $n \times r$ orthogonal matrix $\mathbf{V}$, containing the right singular vectors of $\mathbf{A}$. By retaining only the top k largest singular values, where $k < r$, and centring $\mathbf{A}$ by subtracting the column-wise feature means, $\overline{\mathbf{A}}$, SVD becomes equivalent to PCA (Equation 2.2.1.8). Transforming a test matrix, $\mathbf{A}_{test}$, is achieved with Equation 2.2.1.9, where $\overline{\mathbf{A}}$ is the mean of $\mathbf{A}$.

$$\underset{(m\times n)}{\mathbf{A}} = \underset{(m\times r)}{\mathbf{U}}\,\underset{(r\times r)}{\mathbf{\Sigma}}\,\underset{(r\times n)}{\mathbf{V}^T}, \qquad r = min\,(m, n) \tag{2.2.1.7}$$

$$\mathbf{A} \approx \mathbf{A}_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T \tag{2.2.1.8}$$

$$\mathbf{A}_{test} \approx \widehat{\mathbf{A}}_{test} = \left[\left(\mathbf{A}_{test} - \overline{\mathbf{A}}\right)\mathbf{V}_k\right]\mathbf{V}_k^T + \overline{\mathbf{A}} \tag{2.2.1.9}$$

ICA is a technique for separating a multivariate signal into additive, independent non-Gaussian signals. It is used in blind source separation, where original signals, or sources, are estimated solely from observed data without additional information about the mixing process. Equation 2.2.1.10 describes an ICA model, where $\mathbf{X}$ is the observed data, $\mathbf{A}$ is the mixing matrix, and $\mathbf{S}$ contains the independent components. FastICA [219] is an efficient ICA algorithm which seeks to find an orthogonal rotation of the pre-whitened data, through negentropy maximisation, to produce the most non-Gaussian signals. Negentropy is a statistical measure indicating how non-Gaussian a signal is, used in ICA to enhance the independence of the separated signals. Equations 2.2.1.11-2.2.1.12 show the simplest form of the update rule in FastICA, where g is a non-quadratic function, g' is its derivative, and n represents the number of observations.

$$\mathbf{X} = \mathbf{AS} \tag{2.2.1.10}$$

$$\mathbf{w} \leftarrow \frac{1}{n}\mathbf{X}^T g(\mathbf{Xw}) - \frac{1}{n}\sum_{i=1}^{n} g'(\mathbf{x}_i\mathbf{w})\mathbf{w} \tag{2.2.1.11}$$

$$\mathbf{w} \leftarrow \frac{\mathbf{w}}{\|\mathbf{w}\|} \tag{2.2.1.12}$$

For a comprehensive understanding of these unsupervised methods and their applications, further reading and exploration are encouraged.

2.2.2 Artificial Neural Networks

Artificial Neural Networks (ANNs) are composed of a series of stacked layers $l \in L$ containing learnable weights $\mathbf{W}_l \in \mathcal{W}$ and bias terms $\mathbf{b}_l$ which transform input data with the aim of extracting useful information from it. These need to be optimised for the given task by minimising a loss function $\mathcal{J}$. The MSE loss (where n is the total number of elements - Equation 2.2.2.1) minimises the reconstruction error between the predictions, $f\left(\boldsymbol{x}^{(i)}\right)$, and targets, $\boldsymbol{y}^{(i)}$, and was used in this thesis.

$$\mathcal{J} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}^{(i)} - f\left(\boldsymbol{x}^{(i)}\right))^2 \tag{2.2.2.1}$$

Three Deep Learning (DL) architectures were utilised in this thesis: fully connected networks, known as Multi-Layer Perceptrons (MLPs), Convolutional Neural Networks (CNNs), and Vision Transformers (ViTs). These are introduced in the following sections and subsections.

2.2.2.1 Fully Connected Layers

Fully connected layers (Figure 2.2) are simple DL operators, sitting at the core of many methods. Each layer l is formed of a series of neurons composed of weights and biases, $\mathcal{W}_l = (\mathbf{W}_l, \mathbf{b}_l)$, which compute an affine transformation (Equation 2.2.2.2) for an input activation $\boldsymbol{a}_{l-1}$, where $\boldsymbol{a}_0 = \boldsymbol{x}$ [230].

In neural networks, an activation, $\boldsymbol{a}_l$, is the output of a neuron after applying element-wise a non-linear activation function, $\phi(\cdot)$, to the weighted sum of its inputs plus a bias term, $\boldsymbol{z}_l$ (Equation 2.2.2.3). $\boldsymbol{a}_0 = \boldsymbol{x}$ indicates that the input to the first layer of a network is the data $\boldsymbol{x}$.

$$\boldsymbol{z}_l = f\left(\boldsymbol{a}_{l-1}, \mathcal{W}_l\right) = \boldsymbol{a}_{l-1} \cdot \mathbf{W}_l + \mathbf{b}_l \tag{2.2.2.2}$$

$$\boldsymbol{a}_l = \phi\left(f\left(\boldsymbol{a}_{l-1}, \mathcal{W}_l\right)\right) = \phi\left(\boldsymbol{a}_{l-1} \cdot \mathbf{W}_l + \mathbf{b}_l\right) = \phi\left(\boldsymbol{z}_l\right) \tag{2.2.2.3}$$

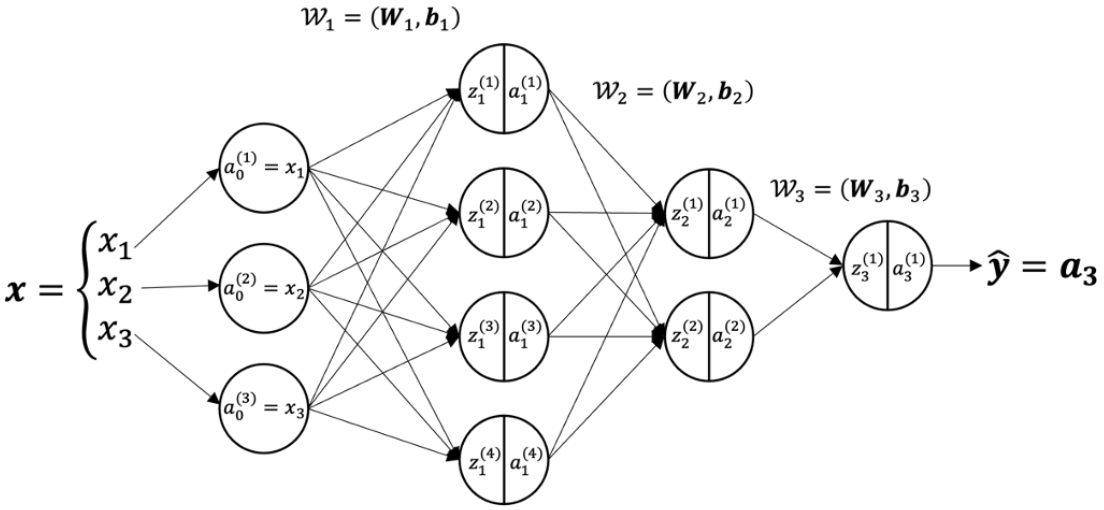


Figure 2.2: Diagram of simple MLP showing several fully connected layers. In each layer l, first the weighted sum of an input plus a bias term is calculated, resulting in $\boldsymbol{z}_l$, after which the activation, $\boldsymbol{a}_l$, is calculated after passing through an activation function $\phi(\cdot)$. The superscripts in round brackets indicate the elements of a vector.

2.2.2.2 Non-Linear Activations

Non-linear activations enable neural networks (NNs) to model nonlinear relationships in the provided training data. This is important for the approximation of complex, non-linear functions, which a purely linear network cannot achieve. They facilitate hierarchical feature learning, allowing networks to build high-level representations from lower-level features, enhancing the network's expressive power and generalisation to unseen data. They also support better optimisation during

training and allow for more accurate modelling of real-world phenomena which are inherently nonlinear. The Rectified Linear Units (ReLU) [231] (Equation 2.2.2.4) is a widely used non-linear function thanks to its simplicity and effectiveness, while the Gaussian Error Linear Unit (GELU) [232] (Equation 2.2.2.5) is a more recent development, which introduces a smooth approximation to the rectifier function, providing non-linearity while mitigating the vanishing gradient problem (Figure 2.3). This is the phenomenon where gradients of the loss function become too small for the network to learn effectively, often occurring in very deep networks, leading to longer or stalled training processes. GELU has gained attention recently due to its use in Transformers [233]. Another nonlinearity which is utilised in Transformers is the Softmax (Equation 2.2.2.6). In Equations 2.2.2.4-2.2.2.6, i, j are the indices of elements x_i, x_j of vector $\mathbf{X}$, and n is the total number of $\mathbf{X}$'s elements. Section 2.2.3 provides more information on Transformers.

$$\mathrm{ReLU}(x_i) = \max(0, x_i) \tag{2.2.2.4}$$

$$\mathrm{GELU}(x_i) = 0.5x_i \left(1 + \tanh\left(\sqrt{\frac{2}{\pi}}\left(x_i + 0.044715x_i^3\right)\right)\right) \tag{2.2.2.5}$$

$$\mathrm{Softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{2.2.2.6}$$

2.2.2.3 Convolutional Layers

Most training data used in this work consists of 3D voxel intensity arrays converted to tensors. Convolutions extract information from 3D tensor data with a grid-like structure, $\boldsymbol{I}$, through cross-correlation, by using a kernel filter composed of multiple weights $\mathbf{W}$ and a bias b (Equation 2.2.2.7) [230, 234]. Here i, j, k are output tensor position indices showing where the convolution result is being stored, and traversing the tensor in x, y and z dimensions respectively. Similarly m, n, p are indices that are used to traverse the kernel in x, y and z-dimensions respectively. A convolution layer is composed of multiple kernels, with each kernel producing

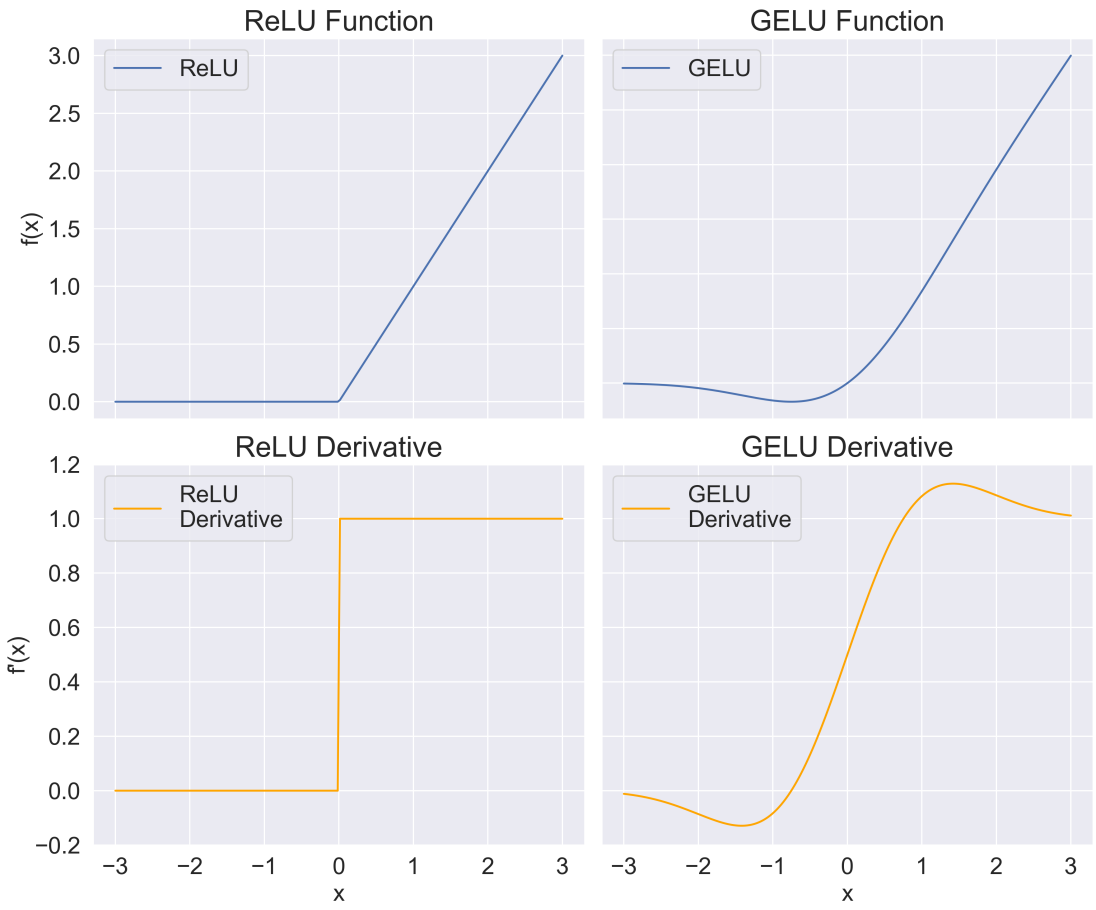


Figure 2.3: ReLU and GELU activation functions and their derivatives.

an output channel. A larger number of output channels, resulting from the use of multiple kernels, aids in capturing multiple features from the input data [235]. Kernels shift across the input tensor $\boldsymbol{I}$ by steps known as strides, s_i, s_j, s_k (Equation 2.2.2.8). Prior to the convolution, zero-padding, with sizes $S_{\text{pad},x}, S_{\text{pad},y}, S_{\text{pad},z}$, are added to $\boldsymbol{I}$ to ensure that its spatial dimensions are preserved following the convolution. Equation 2.2.2.9 presents the calculation of the padding sizes given a kernel of size $d \times h \times w$, where d, h, and w are the depth, height, and width of the kernel. Equation 2.2.2.10 defines the padded input tensor $\boldsymbol{I}_{\text{pad}}$, the convolution operation being applied it (Equation 2.2.2.11).

Pooling layers down-sample the feature maps post-convolution. Max-pooling and average-pooling are common types, taking the maximum and average value from a specified window, respectively, as shown in Equations 2.2.2.12-2.2.2.13. Here

d_p, h_p, w_p are the dimensions of the pooling window, $s_{i,p}, s_{j,p}, s_{k,p}$ denote the pooling strides in the x, y, and z dimensions, and m_p, n_p, p_p are indices used to traverse the local region in the input tensor over which the pooling operation is performed. Figure 2.4 provides a visualisation of Equations 2.2.2.11 and 2.2.2.12.

$$(\mathbf{W} * \boldsymbol{I})(i,j,k) = b + \sum_m \sum_n \sum_p \boldsymbol{I}(i-m, j-n, k-p)\mathbf{W}(m,n,p) \tag{2.2.2.7}$$

$$(\mathbf{W} * \boldsymbol{I})(i,j,k) = b + \sum_m \sum_n \sum_p \boldsymbol{I}(s_i \cdot i - m, s_j \cdot j - n, s_k \cdot k - p)\mathbf{W}(m,n,p) \tag{2.2.2.8}$$

$$S_{\text{pad},x} = \text{int}\left(\frac{d-1}{2}\right), \quad S_{\text{pad},y} = \text{int}\left(\frac{h-1}{2}\right), \quad S_{\text{pad},z} = \text{int}\left(\frac{w-1}{2}\right) \tag{2.2.2.9}$$

$$\boldsymbol{I}_{\text{pad}}(i,j,k) = \begin{cases} \boldsymbol{I}(i - S_{\text{pad},x}, j - S_{\text{pad},y}, k - S_{\text{pad},z}) & \begin{array}{r} i \geq S_{\text{pad},x} \\ \text{if } j \geq S_{\text{pad},y} \\ k \geq S_{\text{pad},z} \end{array} \\ 0 & \text{otherwise} \end{cases} \tag{2.2.2.10}$$

$$(\mathbf{W} * \boldsymbol{I}_{\text{pad}})(i,j,k) = b + \sum_m \sum_n \sum_p \boldsymbol{I}_{\text{pad}}(s_i \cdot i - m, s_j \cdot j - n, s_k \cdot k - p)\mathbf{W}(m,n,p) \tag{2.2.2.11}$$

$$\text{MaxPool}(\boldsymbol{I})(i,j,k) = \max_{m_p=0}^{d_p-1} \max_{n_p=0}^{h_p-1} \max_{p_p=0}^{w_p-1} \boldsymbol{I}(s_{i,p} \cdot i - m, s_{j,p} \cdot j - n, s_{k,p} \cdot k - p) \tag{2.2.2.12}$$

$$\text{AvgPool}(\boldsymbol{I})(i,j,k) = \frac{1}{d_p \cdot h_p \cdot w_p} \sum_{m_p=0}^{d_p-1} \sum_{n_p=0}^{h_p-1} \sum_{p_p=0}^{w_p-1} \boldsymbol{I}(s_{i,p} \cdot i - m, s_{j,p} \cdot j - n, s_{k,p} \cdot k - p) \tag{2.2.2.13}$$

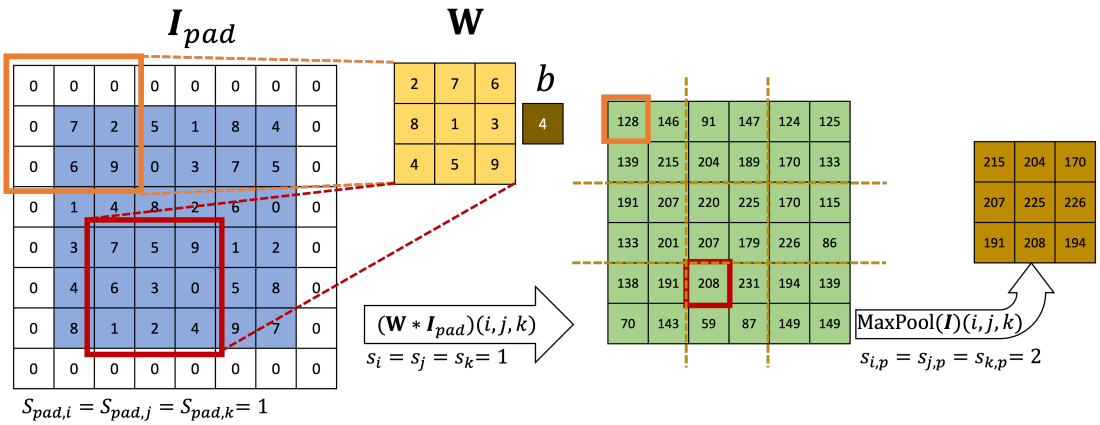


Figure 2.4: Example simple convolution operations. The orange and red squares show different locations where the convolution kernel, composed of a weight $\mathbf{W}$ and a bias b is applied, with the corresponding outputs. $\boldsymbol{I}_{pad}$ is an input array with uniform padding and undergoes a convolution operation with uniform stride of 1. The output of the convolution operation, in green, is then passed through a MaxPool filter with a stride of 2. The path taken by the MaxPool is displayed with dotted brown lines on the green output, the result being shown as a brown matrix.

2.2.2.4 Weight Initialisation

Across this thesis, trainable parameters were initialised using the Kaiming He initialisation, proven to perform best when using ReLU derived activation functions (Equation 2.2.2.14) [236], where layer l's weights $\mathcal{W}_l$ are sampled from a uniform distribution $\mathcal{U}$ based on the input dimensionality n_l.

$$\mathcal{W}_l \sim \mathcal{U}\left[-\frac{\sqrt{6}}{\sqrt{n_l + n_{l+1}}}, \frac{\sqrt{6}}{\sqrt{n_l + n_{l+1}}}\right] \tag{2.2.2.14}$$

2.2.2.5 Normalisation Layers

During training, each layer's input can change due to parameter initialisation stochasticity, input data, and parameter updates, a phenomenon known as "internal covariance shift". This shift in the distribution of inputs to a NN layer can hider the learning process, as the model needs to continuously adapt to changing data distributions. Normalisation layers mitigate the effects of internal covariance shift. By normalising the inputs to, and statistics of each layer, they stabilise

the learning process and enable the training of deep networks with higher learning rates, facilitating faster convergence and better training efficiency [230, 237].

Herein, two normalisation techniques are used: Batch Normalisation (BatchNorm) and Layer Normalisation (LayerNorm) (Figure 2.5). BatchNorm operates across the batch dimension, while Layer Normalisation operates across the feature dimension for each data point, independent of the batch size. In ML, a batch refers to the entire dataset being used for training in one iteration. A mini-batch refers to a small subset of the training dataset used in one iteration, which allows for more frequent model parameters updates and potentially faster convergence.

Equation 2.2.2.15 describes BatchNorm during training, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the input and output tensors for a mini-batch, γ and β are learnable parameters scaling and shifting the data, and $\epsilon = 1e-5$ is a constant to avoid division by zero. The mean, μ, and variance, σ^2, are calculated per feature over a mini-batch, resulting in means and variances vectors. During training, running averages of μ and σ^2, referred to as $\hat{\mu}$ and $\hat{\sigma}^2$, are calculated and used during evaluation (Equation 2.2.2.16). A *momentum* term equal to 0.1 is applied to facilitate the running average calculations for the two statistics over multiple mini-batches, helping stabilise them particularly in cases where the batch size is small and the statistics are noisy [234]. μ_t and σ_t^2 are the mean and variance calculated for the current mini-batch, respectively.

$$\boldsymbol{Y} = \frac{\boldsymbol{X} - \mu(\boldsymbol{X})}{\sqrt{\sigma^2(\boldsymbol{X}) + \epsilon}} * \gamma + \beta \tag{2.2.2.15}$$

$$\begin{aligned} \hat{\mu} &= (1 - \text{momentum}) \times \hat{\mu} + \text{momentum} \times \mu_t \\ \hat{\sigma}^2 &= (1 - \text{momentum}) \times \hat{\sigma}^2 + \text{momentum} \times \sigma_t^2 \end{aligned} \tag{2.2.2.16}$$

Equation 2.2.2.17 describes LayerNorm. It improves model generalisation and training stability, especially in recurrent networks, scenarios with variable batch sizes, and ViTs [238]. In LayerNorm, the mean, μ_{layer}, and variance, σ^2_{layer}, are computed across all the activations in a given layer.

$$\boldsymbol{Y_{LN}} = \frac{\boldsymbol{X} - \mu_{\text{layer}}(\boldsymbol{X})}{\sqrt{\sigma^2_{\text{layer}}(\boldsymbol{X}) + \epsilon}} * \gamma_{\text{layer}} + \beta_{\text{layer}} \tag{2.2.2.17}$$

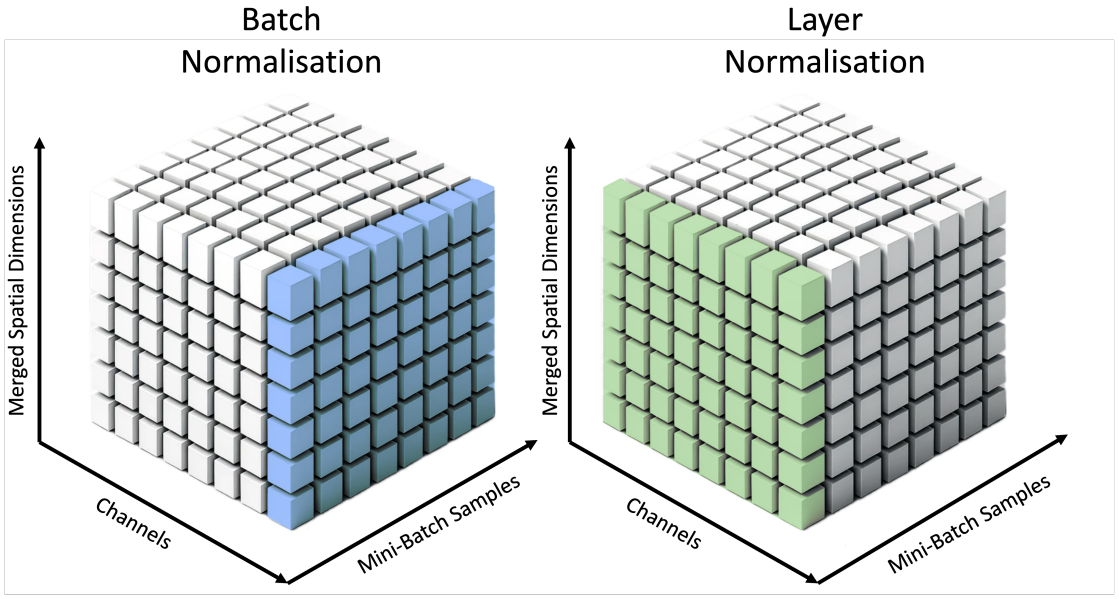


Figure 2.5: Normalisation methods, the left image showing Batch Normalisation, and the right Layer Normalisation. Both subplots show a feature tensor map with three axes: a mini-batch axis, a channel axis, and a merged spatial axis representing the flattened dimensions of an input. The coloured parts in the two subplots show the two normalisations. Given a 3D volumetric input, the flattened dimensions would be of size $lenght \times width \times height$, while the channel axis would be set to 1, increasing after subsequent processing as the latent channel dimensionality increases. This contrasts 2D RGB images, not used in this work, where the flattened dimensions would be of size $width \times height$, while the channel axis would be set to 3, corresponding to the three red, green and blue channels.

2.2.2.6 Network Training

NN training is an iterative process, consisting of passing data forward through the network layers, then calculating the loss, followed by back-propagation, where the gradient of the loss is calculated, and weight updates using an optimiser, which adjusts the weights in the direction that reduces the loss. These steps are repeated for a number of iterations equal to the number of mini-batches into which the training data was split. Collectively, these iterations are known as an epoch. Neural network training typically consists of multiple epochs, continuing until convergence is achieved or a predetermined stopping criterion is met.

During the forward pass, a batch of data which consists of n input features, $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$, is fed into the network. The data passes through the layers of

the network, each composed of weights, $\mathbf{W}_l$, biases, $\mathbf{b}_l$, and activation functions to transform the inputs. Equations 2.2.2.2-2.2.2.3 describe the forward pass for a fully connected layer, and Equation 2.2.2.11 for a convolutional layer.

The final output, $\boldsymbol{a}_L$, from the last layer, L, is used to calculate the loss, $\mathcal{J}$, which measures the overall discrepancy between the predicted outputs and the true values, $\boldsymbol{y}$. Equation 2.2.2.18 showcases a typical loss function, where m is the number of mini-batch samples, $\mathcal{L}$ is a single sample's loss, $\boldsymbol{a}_L^{(i)}$ is the i-th sample's prediction, and $\boldsymbol{y}^{(i)}$ is the corresponding true value. MSE is an example of such a loss function (Equation 2.2.2.1). The loss function is a crucial component that guides the training process, and is chosen based on the specific task.

$$\mathcal{J}(\mathbf{W}_l, \mathbf{b}_l; \boldsymbol{X}, \boldsymbol{y}) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(\boldsymbol{a}_L^{(i)}, \boldsymbol{y}^{(i)}) \tag{2.2.2.18}$$

Back-propagation follows the forward pass. This is the process of calculating the gradient of the loss function with respect to each weight in the network. This is done by applying the chain rule of calculus in reverse order, from the last layer to the first (Equations 2.2.2.19-2.2.2.20). The gradients $\frac{\partial \mathcal{J}}{\partial \mathbf{W}_l}$ and $\frac{\partial \mathcal{J}}{\partial \mathbf{b}_l}$ indicate the direction in which the weights $\mathbf{W}_l$ and biases $\mathbf{b}_l$ should be adjusted to minimise the loss.

$$\frac{\partial \mathcal{J}}{\partial \mathbf{W}_l} = \frac{\partial \mathcal{J}}{\partial \boldsymbol{a}_L} \frac{\partial \boldsymbol{a}_L}{\partial \boldsymbol{z}_L} \cdots \frac{\partial \boldsymbol{z}_{l+1}}{\partial \boldsymbol{a}_l} \frac{\partial \boldsymbol{a}_l}{\partial \boldsymbol{z}_l} \frac{\partial \boldsymbol{z}_l}{\partial \mathbf{W}_l} \tag{2.2.2.19}$$

$$\frac{\partial \mathcal{J}}{\partial \mathbf{b}_l} = \frac{\partial \mathcal{J}}{\partial \boldsymbol{a}_L} \frac{\partial \boldsymbol{a}_L}{\partial \boldsymbol{z}_L} \cdots \frac{\partial \boldsymbol{z}_{l+1}}{\partial \boldsymbol{a}_l} \frac{\partial \boldsymbol{a}_l}{\partial \boldsymbol{z}_l} \frac{\partial \boldsymbol{z}_l}{\partial \mathbf{b}_l} \tag{2.2.2.20}$$

Following gradient computation, an optimiser is used to iteratively update the network's parameters, such as $\mathbf{W}_l$ and $\mathbf{b}_l$, to minimise the loss. To present a unified view of these parameters, they are referred to as θ. The generic update rule for θ is described by Equation 2.2.2.21, where θ_t represents θ at iteration t, η denotes the learning rate, a scalar that determines the size of the step to take in the gradient direction, and Optimiser is a function that takes the current parameters and their gradients, and computes the step to take in parameter space.

The gradient $\nabla_\theta \mathcal{J}(\theta)$ is the vector of partial derivatives of the loss function with respect to each parameter in θ.

$$\theta_{t+1} = \theta_t - \eta \cdot \text{Optimiser}(\theta_t, \frac{\partial \mathcal{J}(\theta)}{\partial \theta}) \tag{2.2.2.21}$$

The choice of the optimiser and η are important for good training performance, as different optimisers have different strategies for navigating the parameter space to find the loss function minima. Classic optimisers, like Gradient Descent, use the gradient to update parameters [239], while more advanced algorithms, like ADAM [240] and RMSprop [241], incorporate mechanisms to adjust the learning rate dynamically, and use the history of gradients to inform the current update.

Out of the hundreds of published optimisation algorithms, ADAM stands out for its versatility in tackling numerous diverse learning problems [242]. For this reason, the CNNs described in this thesis were trained using it. Equations 2.2.2.22-2.2.2.26 describe ADAM, where m_t and v_t are the first and second moments at step t, and β_1, β_2, and ϵ are constants. The first moment represents the exponentially weighted moving average of the gradients, capturing their mean direction, while the second moment represents the exponentially weighted moving average of the squared gradients, capturing their uncentred variance.

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) d\theta_t \tag{2.2.2.22}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) d\theta_t^2 \tag{2.2.2.23}$$

$$\hat{m}_t = \frac{m_t}{1 - {\beta_1}^t} \tag{2.2.2.24}$$

$$\hat{v}_t = \frac{v_t}{1 - {\beta_2}^t} \tag{2.2.2.25}$$

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{2.2.2.26}$$

The ADAM optimiser, however, was observed to sometimes handle weight decay ineffectively, potentially leading to suboptimal generalisation. To account for this, AdamW [243] decouples weight decay from the gradient updates, offering more appropriate regularisation and improved convergence in certain tasks. To do this,

AdamW modifies Equation 2.2.2.26 to 2.2.2.27, where λ is the weight decay. This is the optimiser of choice for Transformer networks.

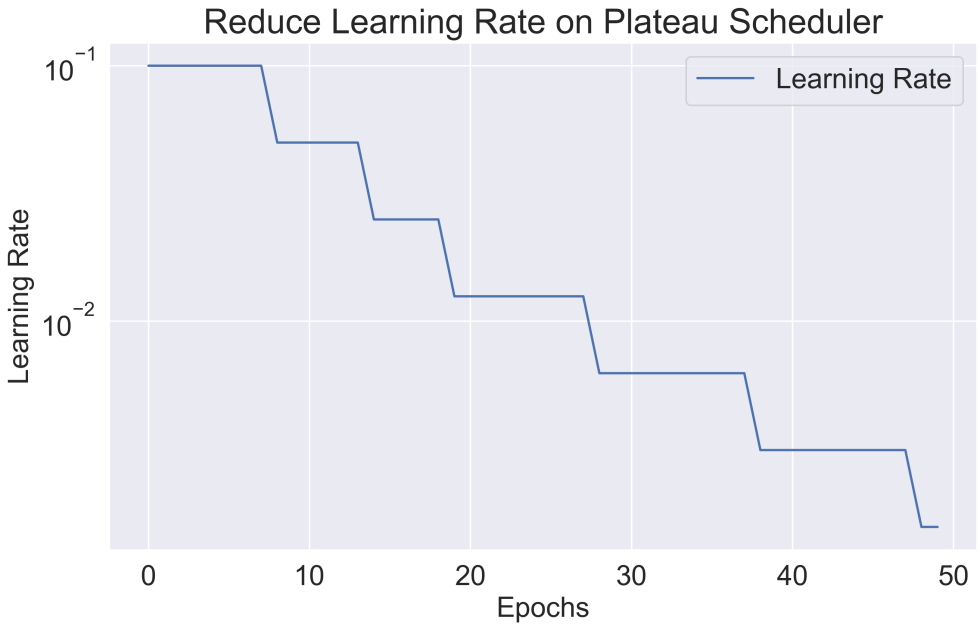
$$\theta_{t+1} = \theta_t - \eta \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda \theta_t \right) \tag{2.2.2.27}$$

Selecting an appropriate learning rate η requires careful experimentation, as it greatly influences model convergence and performance. A static learning rate often proves suboptimal due to the evolving landscape of the loss function during training. Consequently, dynamic learning rate scheduling is used. The 'Reduce Learning Rate on Plateau' (Figure 2.6a) scheduler adjusts the learning rate by monitoring metrics like validation loss, reducing it by a predefined step if no improvement is detected over a designated 'patience' period. Conversely, the 'Cosine Annealing Learning Rate' scheduler [244] (Figure 2.6b) starts with a linear warm-up phase, increasing the learning rate, and then decreases it following a cosine curve, promoting faster and more stable convergence, particularly in complex models such as Transformers [245].

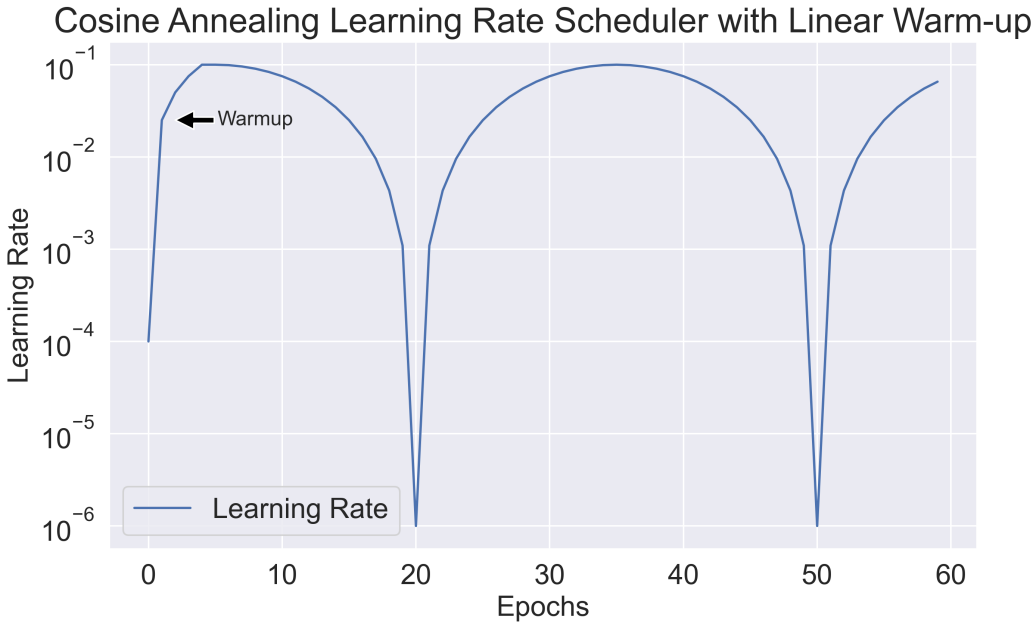
Convergence evaluation is another important aspect of network training. It involves monitoring the loss function and the accuracy of predictions on a separate validation dataset that the model has not seen during training. The use of train, test, and validation datasets ensures that the model not only learns the training data patterns but can also generalise well to new, unseen data. The convergence of the model is typically assessed by observing the stability of the validation loss and accuracy metrics over successive epochs.

2.2.3 Transformers

As above, here the fundamentals of Vision Transformers (ViTs) [245] are presented, with an emphasis on the Shifted Window Transformer (SWIN) [246] (referred to as the vanilla SWIN model below) and its adaptation to the 3D domain [247]. Interested readers are invited to read the original papers, and the literature review by Li et al. [248].



(a) 'Reduce Learning Rate on Plateau' scheduler, with steps occurring at random intervals.



(b) 'Cosine Annealing with Linear Warmup' scheduler, with 5 warmup epochs and a period of 20 epochs.

Figure 2.6: Schematic diagram of different learning rate schedulers with **(a)** showing an example of the 'Reduce Learning Rate on Plateau' scheduler, and **(b)** an example of the 'Cosine Annealing with Linear Warmup' scheduler.

2.2.3.1 Inputs, Patches and Windows

Given a 3D volumetric input of shape $H \times W \times D$ voxels, where $H = D = 160$ and $W = 192$, as those used in this thesis (see Chapter 3 Section 3.3.1), the transformer first splits this into equal-sized patches of size $p \times p \times p$, where the vanilla patch dimension $p = 2$ voxels. This results in a patches (or tokens) volume of size $H/p \times W/p \times D/p = 80 \times 96 \times 80$. In SWIN, the patches are also grouped into windows of size $M \times M \times M$ patches, (or $14 \times 14 \times 14$ voxels for $p = 2$ voxels and $M = 7$ in the vanilla case). For both operations, padding is used to ensure dimensional alignment. In this case, the patches volume is padded from $80 \times 96 \times 80$ to $84 \times 98 \times 84$ (equivalent to padding the input to $168 \times 196 \times 168$), resulting in a windows volume of size $12 \times 14 \times 12$ windows (Figure 2.7).

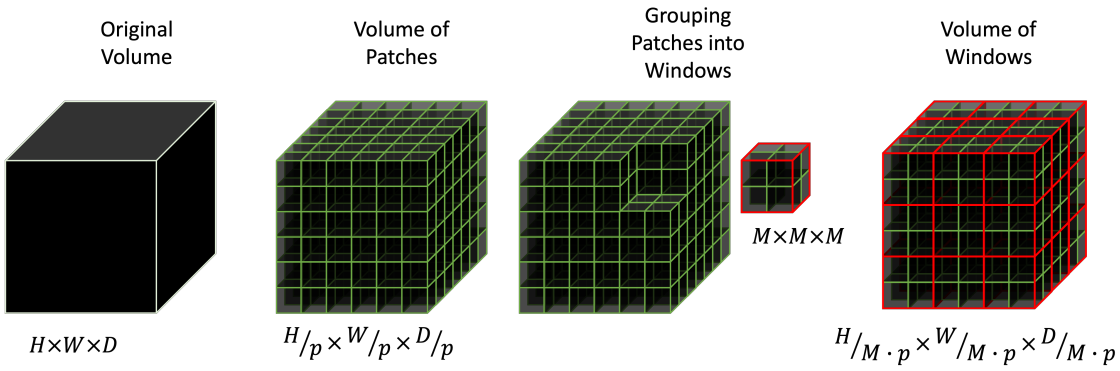


Figure 2.7: Patches and Windows example schematic, showing how an input volume is first split into multiple patches, with $p = 6$ for this example, resulting in a patches volume. After this, the patches can be grouped into windows, with $M = 2$ for this example, resulting in a windows volume.

After these operations, Transformers utilise a linear embedding layer to project the flattened resulting feature volumes (either patches for ViT or windows for SWIN) into an arbitrary embedding feature dimension C. In this work, the vanilla $C = 24$ is maintained.

2.2.3.2 SWIN Hierarchical Feature Representations

Prior to discussing the Transformer's specific operations, the hierarchical feature representation strategy utilised by SWIN must be understood. While in ViTs the number of patches is kept constant, SWIN uses patch merging layers to reduce the dimensionality of latent features as the network gets deeper. To achieve a reduction by a factor of 2, these layers group $2 \times 2 \times 2$ adjacent patches, concatenate them, and apply a linear transformation to the $8C$ concatenated features, reducing them to $2C$. Thus, across a 4-layer deep network, the resolution is progressively reduced from $\frac{HWD}{2^3}$ to $\frac{HWD}{4^3}$, $\frac{HWD}{8^3}$, $\frac{HWD}{16^3}$, and $\frac{HWD}{32^3}$ (Figure 2.8).

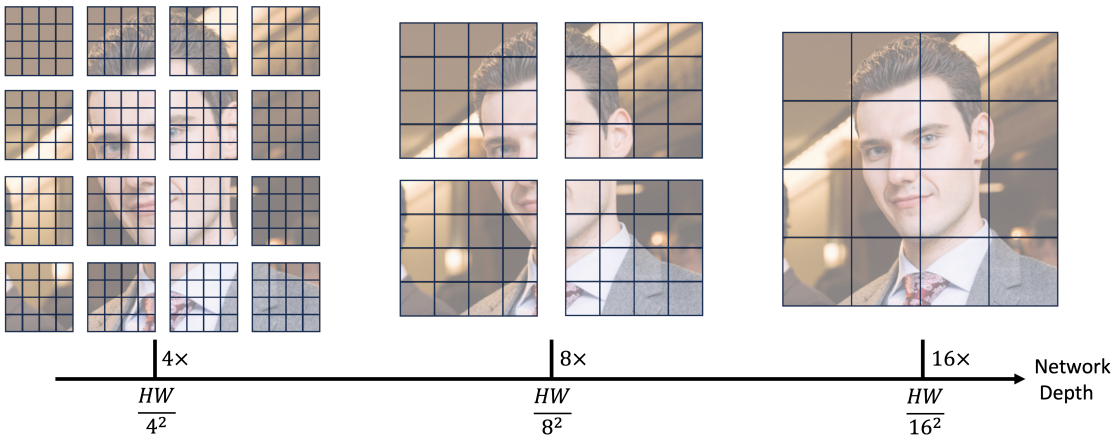


Figure 2.8: 2D hierarchical features example, showing how a SWIN transformer builds hierarchical feature maps. As the network gets deeper, neighbouring smaller image patches are merged progressively. Given that self-attention (SA) is only computed within a local window of 4×4 squares, the SA operation has only linear computation complexity with respect to the input image size. 2D images were used for simplicity and ease of visualisation.

2.2.3.3 Self-Attention

The key advantage of ViTs is the use of the self-attention (SA) mechanisms. This functions by calculating a weighted aggregation of features, capturing internal correlations, filtering, and emphasising the most significant components (Figure 2.9).

SA linearly maps a latent input $\mathbf{X} \in \mathbb{R}^{n \times C}$ into a query $\mathbf{Q} \in \mathbb{R}^{n \times d}$, key $\mathbf{K} \in \mathbb{R}^{n \times d}$, and value $\mathbf{V} \in \mathbb{R}^{n \times d}$ arrays using three learnable weight matrices $\mathbf{W^q}, \mathbf{W^k}, \mathbf{W^v} \in \mathbb{R}^{C \times d}$, where n is the number of patches fed to the transformer block,

C the dimensionality of the embeddings of each token, and d is the dimensionality of each query, key, and value vector for each of the n patches (Equation 2.2.3.1).

$$\begin{aligned} \mathbf{Q} &= \mathbf{X} \cdot \mathbf{W^q} \\ \mathbf{K} &= \mathbf{X} \cdot \mathbf{W^k} \\ \mathbf{V} &= \mathbf{X} \cdot \mathbf{W^v} \end{aligned} \tag{2.2.3.1}$$

Then, the attention mechanism computes the attention activation weight $\mathbf{A} \in \mathbb{R}^{n \times n}$, which is applied to $\mathbf{V}$ to produce the output $\mathbf{Z} \in \mathbb{R}^{n \times d}$ (Equation 2.2.3.2). In calculating $\mathbf{A}$, to account for large magnitudes, the $\mathbf{Q} \cdot \mathbf{K}^\top$ product is normalised by $\sqrt{d}$. Failing this could push the Softmax function into extremely small gradients. Softmax (Equation 2.2.2.6) is used to convert the raw scores (logits) from the SA mechanism into probabilities that sum to one, indicating the relative importance of each input element. In Transformers, $\mathbf{A}$ is the element which provides them with the ability to efficiently work with long-range information, as it connects all tokens. $\mathbf{A}$ also enables the visualisation of how important each key, or patch location, is in the network outputted values. This can be visualised in a map where intensity corresponds to key importance. Appendix C provides practical details on the extraction of these attention activation intesity maps.

$$\begin{aligned} \mathbf{A}(\mathbf{Q}, \mathbf{K}) &= Softmax\left(\frac{\mathbf{Q} \cdot \mathbf{K}^\top}{\sqrt{d}}\right) \\ \mathbf{Z} = SA(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \mathbf{A}(\mathbf{Q}, \mathbf{K}) \cdot \mathbf{V} \end{aligned} \tag{2.2.3.2}$$

Usually, multiple SA blocks, referred to as heads, are utilised in parallel, the output $\mathbf{Z}$ being the concatenation and projection of all the multi-head SA blocks (MSA) (Equation 2.2.3.3), where i is a head index, and $\mathbf{W^o} \in \mathbb{R}^{hd \times C}$ is a linear projection matrix used for aggregation. To maintain a consistent computational complexity and control the growth of the model's parameter count, preventing the exponential increase of parameters involved in multi-head attention with the addition of more heads, the dimension d of each SA head's output is set to C/h.

$$\begin{aligned} \mathbf{Z}_i = SA(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) &= SA(\mathbf{X} \cdot \mathbf{W^q}_i, \mathbf{X} \cdot \mathbf{W^k}_i, \mathbf{X} \cdot \mathbf{W^v}_i) \\ MSA(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= Concat[\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_h] \cdot \mathbf{W^o} \end{aligned} \tag{2.2.3.3}$$

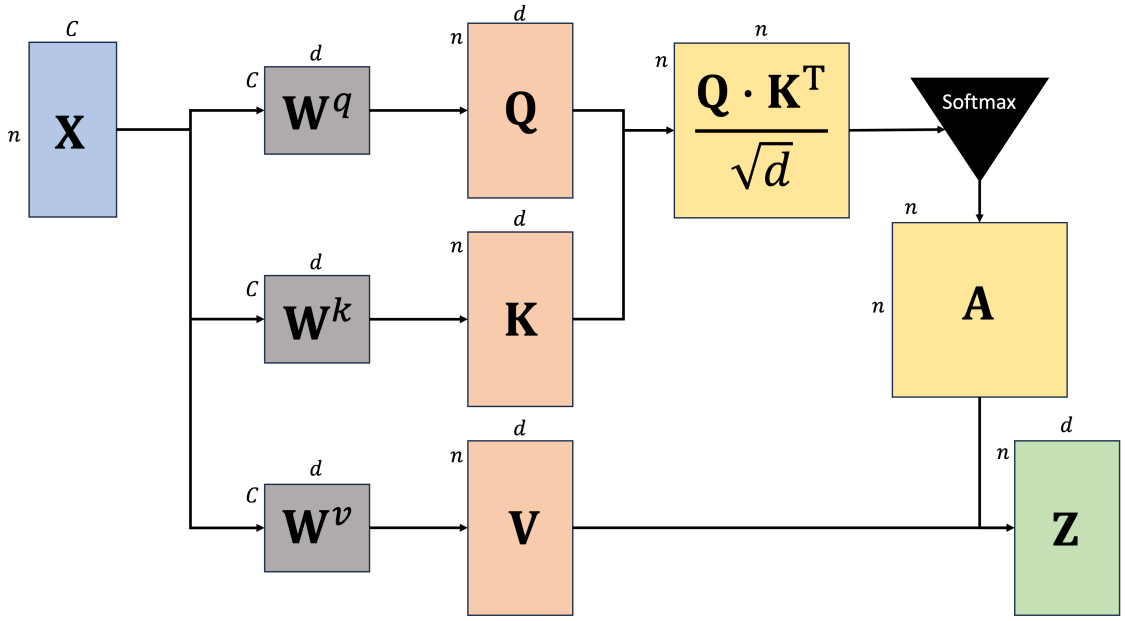


Figure 2.9: Self-Attention diagram, presenting the flow of information within the SA operation. Each block represents an array, with specific dimensions.

2.2.3.4 Window and Shifted-Window MSA

SWIN introduces several modifications to these operations, by replacing each MSA block with a window MSA (W-MSA) followed by shifted windows MSA (SW-MSA). Thus, SWIN computes local, window-based attention rather than global attention. This change reduces the complexity of the operation from quadratic with respect to the number of patches, to linear [246].

However, this approach lacks connections across windows, depriving the model of global context. This issue is resolved by the SW-MSA blocks, which introduce connections between neighbouring non-overlapping windows. The SW approach displaces the windows by $[\lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor] = [\lfloor \frac{7}{2} \rfloor, \lfloor \frac{7}{2} \rfloor, \lfloor \frac{7}{2} \rfloor] = [3, 3, 3]$ voxels from the regularly partitioned windows, performing attention in these new shifted windows. To account for the larger number of windows resulting from this operation, SWIN first cyclic-shifts the features towards the top-left direction together, after which it applies a masking mechanism, which limits the SA to within each sub-window (Figure 2.10). After passing through a SW-MSA block, the shifted window is displaced again to reform the original input.

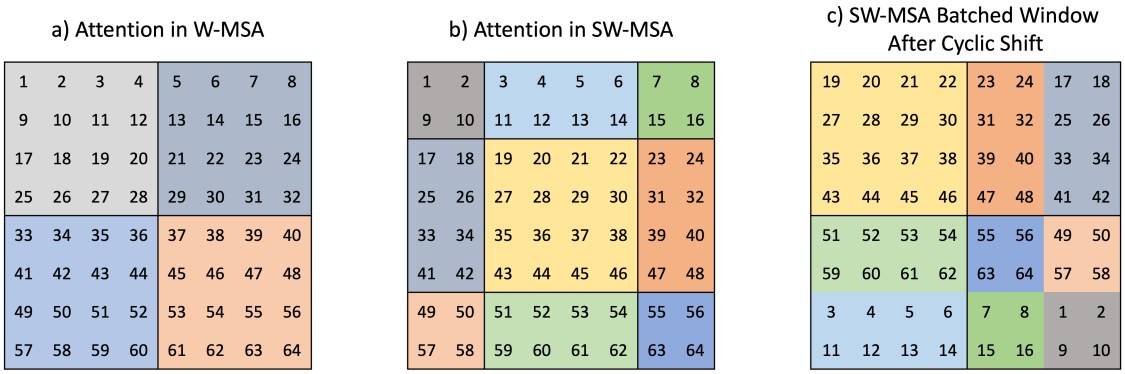


Figure 2.10: Shifted windows diagram, providing intuition into the SW shifting and masking mechanism for an 8×8 example feature map. **(a)** displays the normal version of the feature map, with a window size $M = 4$. W-MSA blocks would perform self-attention within each of the 4 local windows highlighted in different colours. In the SW-MSA block, the feature map is first shifted by $[\lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor] = [2, 2]$ from the regularly partitioned window, resulting in multiple windows, highlighted in different colours. Of these, some are smaller than $M \times M$, which means that padding would be needed to perform attention in that window. To account for this, in **(c)**, a mask is applied, which realigns the image and allows attention to be performed within each coloured area without the need for padding. The figure was inspired by a blog post by A. Arora [249].

SWIN also introduces a modification to Equation 2.2.3.2, by adding an additional relative position bias term $\mathbf{B} \in \mathbb{R}^{M^3 \times M^3}$, where now $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{M^3 \times d}$ account for the use of windows, with M^3 representing the number of patches within a window, and d remaining the dimensionality of the transformed feature space (Equation 2.2.3.4). The values of $\mathbf{B}$ are taken from a smaller size, learnable, bias matrix $\hat{\mathbf{B}} \in \mathbb{R}^{2M-1 \times 2M-1 \times 2M-1}$, accounting for the relative positions along each axis being in the range $[-M + 1, M - 1]$. $\hat{\mathbf{B}}$ encodes the relative positional information in a 3D window, its $2M - 1$ dimensions reflecting all possible pairwise positional relationships in three dimensions, from one window corner to another. $\hat{\mathbf{B}}$ is randomly initialised using a truncated normal distribution $\mathcal{T}$ with mean $\mu = 0$ and standard deviation $\sigma = 0.02$ (Equation 2.2.3.5).

$$\mathbf{Z} = SA(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax\left(\frac{\mathbf{Q} \cdot \mathbf{K}^\top}{\sqrt{d}} + \mathbf{B}\right) \cdot \mathbf{V} \tag{2.2.3.4}$$

$$\hat{\mathbf{B}} \sim \mathcal{T}(\mu = 0, \sigma = 0.02) \tag{2.2.3.5}$$

After each MSA block, both ViT and SWIN employ a local and translation-equivariant MLP operation (Equation 2.2.3.6) which injects inductive bias into the transformer, favouring the learning of spatial hierarchies. The MLP block is based on the assumption that data patterns closer together are more relevant to each other, thus helping the model understand the image structure. The block is composed of two feed-forward network layers, with weights $\mathbf{W}_{l1}, \mathbf{b}_{l1}, \mathbf{W}_{l2}, \mathbf{b}_{l2}$ for layer l and input $\mathbf{Z}_{l-1}$, and a GELU nonlinearity (Equation 2.2.2.5). This structure allows the Transformer to capture both global interactions, via the SA mechanism, and local interactions via the MLPs, enabling it to generalise well to new data.

$$MLP(\mathbf{Z}_{l-1}) = GELU(\mathbf{Z}_{l-1} \cdot \mathbf{W}_{l1} + \mathbf{b}_{l1}) \cdot \mathbf{W}_{l2} + \mathbf{b}_{l2} \tag{2.2.3.6}$$

Equation 2.2.3.7 describes the typical information flow through a SWIN transformer block at layer l, where LN stands for Layer Normalisation (Equation 2.2.2.17). Equation 2.2.3.7 also includes the skip connections utilised in SWIN Transformers.

$$\begin{aligned} \mathbf{Z'}_l &= \textit{W-MSA}(LN(\mathbf{Z}_{l-1})) + \mathbf{Z}_{l-1} \\ \mathbf{Z}_l &= MLP(LN(\mathbf{Z'}_l)) + \mathbf{Z'}_l \\ \mathbf{Z'}_{l+1} &= \textit{SW-MSA}(LN(\mathbf{Z}_l)) + \mathbf{Z}_l \\ \mathbf{Z}_{l+1} &= MLP(LN(\mathbf{Z'}_{l+1})) + \mathbf{Z'}_{l+1} \end{aligned} \tag{2.2.3.7}$$

3

Predicting Brain Age Using Multiple Distinct MRI Modalities and Convolutional Neural Networks

Contents

3.1 Overview

The work presented in this chapter has been previously published in:

- **Roibu, Andrei-Claudiu**, Stanislaw Adaszewski, Torsten Schindler, Stephen M. Smith, Ana IL Namburete, and Frederik J. Lange. "Brain Ages Derived from Different MRI Modalities are Associated with Distinct Biological Phenotypes." In 2023 10th IEEE Swiss Conference on Data Science (SDS), pp. 17-25. IEEE, 2023.

Brain ageing is a highly variable, spatially and temporally heterogeneous process, marked by numerous structural and functional changes. These can cause discrepancies between individuals' chronological age and the apparent age of their brain, as inferred from neuroimaging data. Machine learning models, and particularly Convolutional Neural Networks (CNNs), have proven adept in capturing patterns relating to ageing-induced changes in the brain. The differences between the predicted and chronological ages, or brain age deltas, have emerged as useful biomarkers for exploring those factors which promote accelerated ageing (positive deltas) or resilience (negative deltas), such as pathologies or lifestyle factors. However, the majority of past studies relied only on structural neuroimaging for predictions, overlooking potentially informative functional and microstructural changes.

Here I show that multiple maps derived from different MRI modalities can predict brain age, each encoding bespoke brain ageing information. By using 3D CNNs and UK Biobank data, I found that 57 maps derived from structural, susceptibility-weighted, diffusion, and functional MRI can successfully predict brain

age. For each map, different patterns of association with non-imaging phenotypes were found, resulting in a total of 191 unique, statistically significant associations. Furthermore, I found that ensembling data from multiple maps results in both higher prediction accuracies, achieving state-of-the-art performance for the considered age range (45-82 years old), and stronger correlations to non-imaging measurements. My results demonstrate that other MRI maps and modalities, which have not been considered so far for the task of brain age prediction, encode different information about the ageing brain.

I envision my work as being the starting point for future investigations into the causal links underpinning the observed brain age deltas and non-imaging measurement associations. For instance, drug effects can be monitored, given that certain medications correlated with accelerated brain ageing. Furthermore, continued development of brain age models could facilitate their deployment in clinical trials for recruitment and monitoring, and hospitals for diagnostic and screening tasks.

3.2 Introduction

From a demographic standpoint, the 21^{st} century will bear witness to continuously increasing lifespans. It is predicted that the population aged over 65 years will reach 16% of the global population by 2050 and 22% by 2100. This phenomenon will be ever more pronounced in the developed world, reaching 25% and 31% respectively [29].

As discussed in Chapter 1, increased lifespans have significant implications for both individuals and the healthcare system. This is because increased lifespans are associated with a range of non-fatal, yet debilitating neurodegenerative disorders, such as cognitive decline and various forms of dementia, including Alzheimer's disease [75, 102, 110–112, 114, 120–125].

Brain ageing is a subtle, highly variable, spatially and temporally heterogeneous process even in healthy individuals [30, 31]. This can lead to certain differences between different groups of healthy individuals in regards to apparent acceleration or resilience to brain ageing, with sometimes young individuals showing signs

of early cognitive decline, while some older subjects retaining cognitive health past the age of 100 [8, 9].

Moreover, as discussed in Chapter 1 Section 1.4, multiple contributing factors have been associated with accelerated brain ageing, or resilience to it, such as lifestyle factors (e.g., smoking [35, 53, 56, 81, 143, 146, 147] and alcohol consumption [35, 53, 59, 81, 137, 138, 143–145]), socio-economic factors [56, 109] and physiological factors [21, 35, 140]. This is compounded by the fact that some early stages of neural pathologies could resemble normal or slightly advanced healthy ageing [31–33]. Thus, understanding the associations between accelerated brain ageing and risk factors is important for identifying those at risk for accelerated brain degeneration, but also for developing prevention strategies and treatments.

Despite the existence of several approaches to understanding brain ageing, modern methods explore the "brain age paradigm". This refers to the discrepancy between apparent and chronological brain ages. These methods are discussed at length in Chapter 1 Section 1.2. In summary, by using neuroimaging data, such as Magnetic Resonance Imaging (MRI), statistical models are trained to predict chronological age ($age_{chronological}$), establishing models of healthy ageing. They assume that any post-convergence residual errors, or brain age deltas ($\Delta_{BrainAge}$), between predicted ($\widehat{age}_{predicted}$) and chronological ages (Equation 3.2.0.1) reflect non-standard ageing trajectories. These indicate either accelerated ageing (positive Δ) or resilience to ageing (negative Δ) compared to the population norm. The current state-of-the-art method achieves mean absolute error (MAE) predictions of 2.097 years for a population spanning the $3 - 97$ years age range, using T1 Linear maps derived from structural MRI (sMRI) [250].

$$\Delta_{BrainAge} = \widehat{age}_{predicted} - age_{chronological} \tag{3.2.0.1}$$

Multiple supervised regression algorithms have been utilised for brain-age prediction, including linear regression [35, 60], relevance vector machines [61], support vector machines [62] and artificial neural networks [21, 53, 65, 250]. The majority of these employ user-defined features, such as summary image

measurements extracted from the neuroimaging data, referred to as image-derived phenotypes (IDPs) [35]. To circumvent potentially subjective decisions about which features are important, CNNs trained with voxel-level image data rather than summary features have been utilised successfully for brain age predictions [21]. By using large amounts of data, CNNs learn relevant filters which enable them to discover relevant data patterns and relationships automatically, without the need for handcrafted features [21]. Still, given the black-box nature of CNNs, results must be carefully assessed to ensure that any network architecture and training induced effects are understood and differentiated from structured, biologically-relevant errors [51].

Varying the MRI acquisition parameters produces images (modalities) that are sensitive to different structural and functional properties of the brain. Multivariate modalities, such as diffusion and functional MRI (dMRI and fMRI), can be further processed to produce sets of images (maps) that represent summary measures of interest. Despite the existence of numerous modalities and maps, of 146 previous brain-age studies (summarised in Tables A.1-A.6), 75% utilised images or features derived from just T1-weighted data, with the rest using information derived from dMRI and fMRI [64, 75, 80, 108, 113, 115, 251–253]. In the case of CNN-based methods, using 2D or 3D volumetric neuroimaging inputs, this increases to 95% of the past 39 studies [72]. Only two recent CNN-based studies employed non-sMRI maps, in the form of dMRI derived FA and MD maps [74, 252]. The wide use of the T1-weighted sMRI maps can be partially explained by their wide prevalence in clinical and research datasets. Moreover, sMRI provides simple brain structure and tissue information, making aging-related changes more easily interpretable [21, 41].

Given that brain-ageing is highly heterogeneous and nonlinear, impacting many structural and functional brain aspects, this over-utilisation of sMRI maps can lead to missed opportunities. Certain age-related changes are not captured by sMRI maps, but can be seen in those derived from other modalities, such as increased mineral and iron depositions, captured by susceptibility-weighted MRI (swMRI) maps, white matter structural and microstructural changes, to which dMRI is

sensitive, or the impact of prolonged alcohol consumption on cerebral blood flow, reflected by fMRI [254]. Moreover, the use of IDPs, seen in most non-sMRI based studies, can lead to models missing subtle changes associated with the ageing brain, such as those relating to spatial interactions, which are captured by models using 2D and 3D input maps. In addition, few studies investigated the use of multiple MRI maps simultaneously for brain-age predictions [35, 56, 67, 74, 75, 81]. The richer set of inputs may improve predictions over single-map methods by capturing more age-related variability through the learning of complementary between-map features [81]. Thus, overall, numerous vital brain-ageing information streams are yet unexplored.

Given these facts, in this chapter, rather than solely attempting to obtain the highest accuracy brain age predictions, I seek to address the following questions:

1. What is an appropriate deep learning architecture to explore the brain age paradigm?

2. Can other 3D MRI derived maps, including those which have not yet been explored thus far for this purpose, inform brain age and brain age acceleration and resilience?

3. Which maps lead to the most accurate brain age predictions?

4. For those maps which can predict age, are the predictions biologically meaningful?

5. Also, for those maps which can predict age, do they encode different information relating to the ageing brain?

6. Does ensembling the predictions obtained using multiple maps lead to better and more informative brain age predictions?

7. Can multiple independent modes of ageing be identified within a population of subjects?

To answer these questions, I show that 57 different 3D maps, derived from 5 core MRI modalities, can be used to predict brain age when independently used to train a CNN, and that they form different associations with nIDPs. I first propose a lightweight 3D CNN architecture, developed for brain age prediction tasks. This is then compared against the current state-of-the-art brain age prediction model, known as the Simple Fully-Convolutional Network (SFCN) [67], to establish which of the two is the best method for the outlined tasks. Then, I evaluate the performance of each individual map at predicting brain age, determining which are most accurate. I then investigate if different maps encode different information relating to the ageing brain, by quantifying the associations between their predicted brain age deltas and non-imaging derived measurements, or phenotypes (nIDPs) from UK Biobank [36]. I also test if ensembling the predictions obtained from multiple maps leads to better and more informative brain age predictions, by both increasing prediction accuracies and associations with nIDPs. Finally, I investigate if different independent modes of ageing can be identified within the population of test subjects using the map predicted brain age deltas.

3.3 Common Methods

This section presents a brief description of the datasets, modalities, maps and any pre-processing steps utilised for the purpose of this chapter. It also introduces the proposed CNN architecture, and discusses the steps required for calculating the brain ages, the brain age deltas, and their associations to nIDPs. For a full description of the datasets, please consult Chapter 2, Sections 2.1.2, 2.1.3. An in-depth discussion of the machine learning methods underpinning the proposed CNN can be found in Chapter 2, Section 2.2.2.

3.3.1 Data Description and Pre-Processing

The data from 29331 subjects, aged $45 - 82$ years, from UK Biobank was used [36]. The large number of subjects ensured that as much population variability as possible is captured. The subjects were split by sex into male (13640) and female

(15691) datasets. While this is not common in most brain age studies, which aim to maximise the population utilised for training and inference, several studies have found differences between the biological sexes in terms of nIDP associations [21, 56, 58, 132, 139]. For instance, both Smith et al. [56] and Dinsdale et al. [21] found sex-specific differences in brain ageing, such as acceleration post-menopause in women.

No additional filtering of healthy subjects from subjects with specific diseases was carried out. This is because the majority of UK Biobank subjects were healthy at the time of their scan, and any sub-population with an existing pathology constitutes a minority, alleviating the risk of models overfitting to them [35]. Moreover, the brain age study conducted by Dinsdale et al. [21] using UK Biobank data found no significant differences in either the brain age predictions, or the associations between brain age predictions and nIDPs, between models which were trained using only healthy subjects (i.e. excluding those with chronic, neurological or psychiatric pathologies), and those trained on mixed healthy-unhealthy cohorts.

For each dataset, the subjects were randomly split into training (60%), validation (8%) and two testing subsets (16% each). One test subset was utilised for fitting ensembling models, and another for reporting results. Using two test sets is necessary as the ensembling models require training, during which they will overfit, to a certain degree, the training data. Thus, for consistency and a correct comparison between models, all results will solely be reported using the second test set.

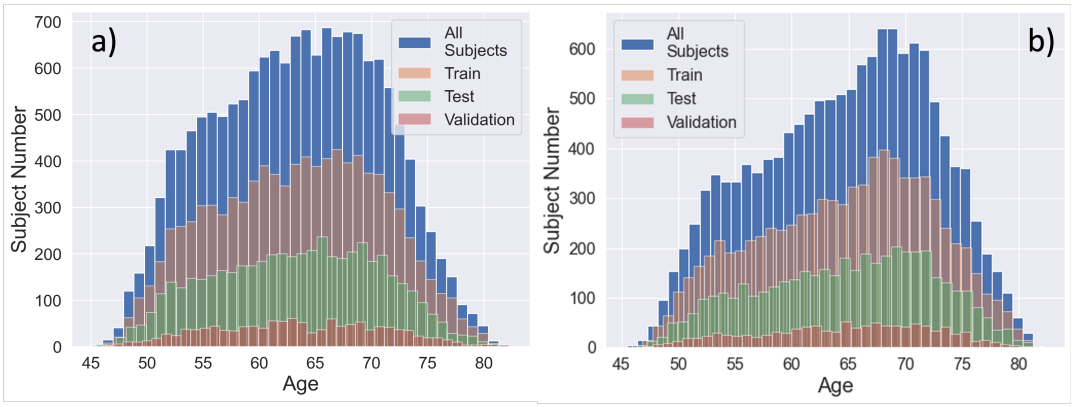


Figure 3.1: Female and male data distribution for the full UK Biobank population, including the train, validation and test datasets histograms. **(a)** indicates female and (**b**) male subjects.

For prototyping and conducting hyperparameter searches, two additional small and tiny female dataset were created utilising 5000 and 2000 subjects respectively. The small dataset allocated 3000 subjects for training and 1000 each for validation and testing. The tiny dataset allocated 1000 and 500 subjects for training, and validation and testing respectively.

Neuroimaging data from UK Biobank was processed with the standard UK Biobank Pipeline, as described by Alfaro et al. [173]. Maps corresponding to 5 core modalities were used (see Table 2.1 in Chapter 2 Section 2.1.2), including:

- 6 sMRI maps, reflecting the gross anatomy of the brain;
- 1 swMRI map, revealing information about compounds distorting local magnetic fields, such as cerebral microbleeds, or iron, calcium or myelin concentration;
- 25 resting-state fMRI (rsfMRI) ICA dual-regressed z-score normalised maps, representing neuronal activity at rest;
- 6 task fMRI (tfMRI) maps, representing brain activations in the presence of a stimulus, split into:
  - 3 z-statistic maps;
  - 3 maps of parameter estimate maps (COPE);
- 19 diffusion MRI maps, providing information about the brain white matter and structural connectivity, including:
  - 9 quantifying aspects of water diffusion and microstructural properties;
  - 9 which represent the same features but projected onto a white matter skeleton using tract-based spatial statistics (TBSS);
  - 1 summation of 27 major axonal tracts obtained with probabilistic tractography [41].

For a description of these maps and how they are obtained, please see Chapter 2 Section 2.1. Different registration techniques were used for the various maps. While most maps utilised non-linear registration with the FMRIB Non-Linear Image Registration Tool (FNIRT) [205], some only employed linear registration via the FMRIB Linear Registration Tool (FLIRT) [178]. Maps such as T1 Linear and SWI were registered using FLIRT with 12 degrees of freedom (DOF) between subject brain-extracted T1 and MNI-152 T1 brain-extracted. The remaining structural maps, as well as the fMRI maps were registered nonlinearly with FNIRT (after initialisation with FLIRT), using 10mm spacing, aligning the entire head (including the skull) from a subject's T1-weighted MRI to the MNI-152 whole head template [175–177]. The diffusion maps utilised a similar approach, however registration was conducted between the subject's FA map and the FMRIB58_FA_1mm [172]. The different registration techniques can cause misalignments to occur between the various maps, however, when these experiments were conducted, tools such as MMORF (FSL's MultiMOdal Registration Framework) [255] were not widely available, so the raw maps processed according to the UK Biobank Pipeline were utilised.

The maps are of either $1mm$ or $2mm$ isotropic spatial resolution. All $2mm$ isotropic spatial resolution maps were upsampled to $1mm$ isotropic resolution using trilinear interpolation. Their dimensions were then cropped from $182 \times 218 \times 182$ to $160 \times 192 \times 160$ voxels. Each map was then normalised. For the sMRI maps, where voxel intensities are not quantitative, the maps were first scaled at the subject-level by mean division. Then, all maps were scaled by a population-wide scaling factor. For quantitative maps, the scaling factor is the largest absolute value between the extreme value means $\pm$ two standard deviations. For the sMRI maps, it was chosen similarly but using the 1^{st} and 99^{th} percentiles instead, as the extreme value distributions can be skewed by outliers.

In addition to these, data corresponding to 3921 IDPs, split into modality and phenotype specific categories (Appendix B, Section B.1) was utilised for training an IDP-based baseline model, as well as a set of 17526 nIDPs from UK Biobank. These were split into 19 thematically defined categories for easier interpretability

(Appendix B, Section B.2). The nIDP-brain age delta associations were later quantified in a phenome-wide association study. Any significant correlations thus obtained enrich our understanding of factors contributing to brain age acceleration and resilience. Yet, given the already wide scope of the work presented in this chapter, the causal mechanisms underpinning each the observed correlations are not investigated in detail.

3.3.2 Deep Learning Architecture and Experimental Setup

For this work, a 3D CNN was constructed, referred to as Happy-Go-Lucky net (HGL) (Figure 3.2), adapted from VGG-16 [87]. The network uses 3D volumes as inputs and returns a single scalar output, representing a subject's predicted age. A 3D architecture was selected as it preserves spatial interactions which could be associated with brain ageing. In addition, 3D architectures have been found to outperform 2D CNNs in certain neuroimaging applications, despite their higher computational costs [256].

HGL represents a reduced version of the VGG network, using a smaller number of trainable parameters. This is because over-parametrisation leads to longer training times and an increased number of training subjects [257]. Large parameter spaces enable the learning of complex tasks, such as the one VGG-16 was originally trained for. Yet, a case can be made that regression tasks, such as the one at hand, require the learning of a reduced set of features. Reducing the parameter space using techniques such as network pruning have been found to improve performance on small medical datasets while reducing the number of trainable parameters by up to 85% [257]. These are, however, outside the scope of this work.

Using these considerations, HGL is composed of 5 convolutional blocks, each consisting of a 3D convolution operation, batch normalisation, 3D max pooling and a ReLU nonlinearity. The number of convolution filters ranges from 32 to 64 per block, with 3D kernels of size 3, and strides of 1. The max pooling 3D kernel size is 2 and the stride 2. Following the convolutional blocks, the output is flattened and passed through three fully-connected layers of sizes 96×1, 32×1

and 1×1. The first two fully-connected layers are followed by ReLU nonlinearities, and the final by a linear activation.

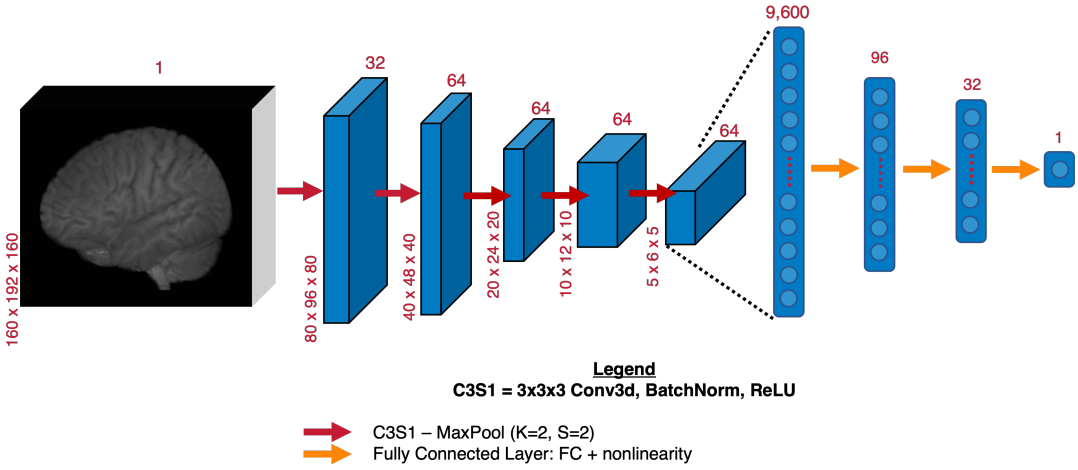


Figure 3.2: HGL network architecture utilised in this chapter.

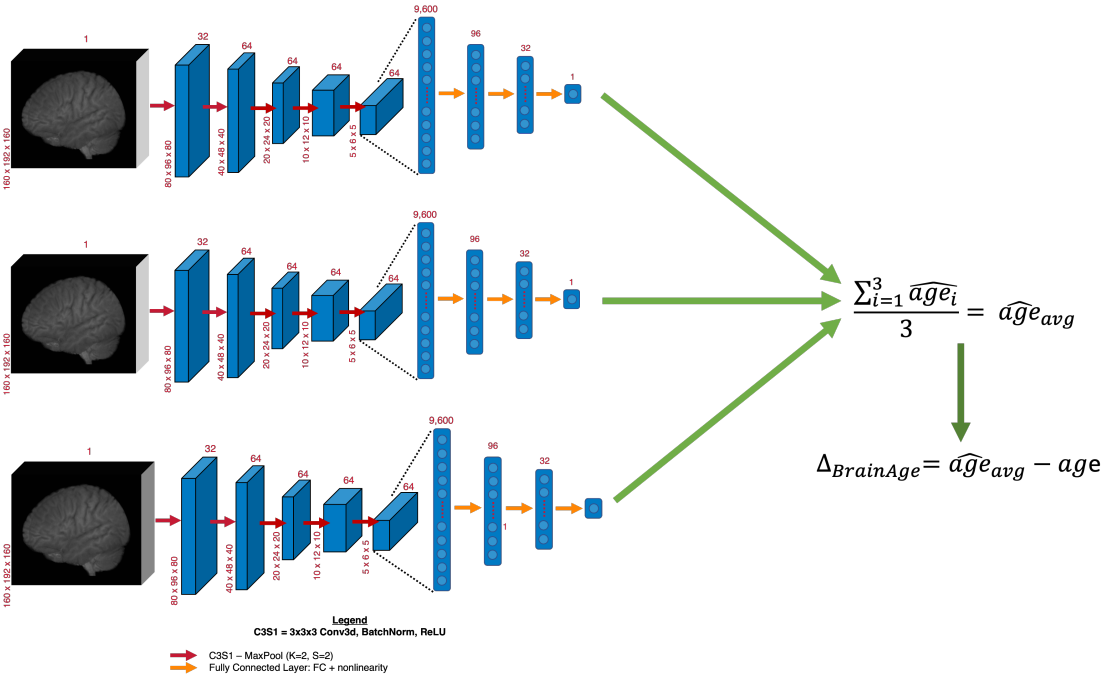


Figure 3.3: HGL network ensembling process for 3 identical runs of the same network, where brain age predictions are averaged at a subject level post-training.

For each map, and for both female and male subjects, a set of 3 identical networks were trained separately to convergence. Test set predictions were averaged

post-training at the subject level across the 3 runs (Figure 3.3). Thus, noise induced by random weight initialisation and stochastic training mechanisms was decreased, leaving more structured errors.

The networks were trained using the mean squared error (MSE) loss function and the ADAM optimiser with default parameters [240], over a maximum of 200 epochs. Training and validation batch sizes were set to 12, and testing batch sizes were set to 1. To reduce overfitting, training was stopped if the validation loss did not improve for 40 epochs, the network being saved at the epoch with the best validation loss. The initial learning rate was set to $1e-2$ and multiplied by 0.5 each time the validation loss did not improve for 15 epochs. During training, to improve generalisability, voxels were randomly shifted by $0-2$ voxels along each axis [67]. No other augmentations were used due to their possible detrimental effects on the biological meaning of various maps. This is largely because all maps have been aligned, either rigidly or nonlinearly, to the same space. Thus, other augmentation techniques are likely to have detrimental effects.

The network was implemented using PyTorch 1.8.1 [258] and trained using NVIDIA A100 and RTX8000 GPUs. The two GPU architectures were utilised interchangeably in order to maximise hardware availability.

3.3.3 Obtaining the Brain Age Deltas

As described above, each network in the 3-network HGL ensemble creates a non-linear mapping $f_i(\cdot)$ from the input brain images to chronological ages. At inference, the brain imaging data $\boldsymbol{B}_j$ of each j^{th} test subject is used to predict that subject's brain age $\hat{a}_{ij}$ using each network i in the ensemble (Equation 3.3.3.1). The mean brain age prediction $\overline{\hat{a}_j}$ is then calculated at the subject level (Equation 3.3.3.2). The brain age delta is then calculated as the difference between $\overline{\hat{a}_j}$ and the subject's chronological age a_j (Equation 3.3.3.3).

$$\hat{a}_{ij} = f_i(\boldsymbol{B}_j),\ i \in \{1,2,3\} \tag{3.3.3.1}$$

$$\overline{\hat{a}_j} = \frac{1}{3}\sum_{i=1}^{3}\hat{a}_{ij} \tag{3.3.3.2}$$

$$\Delta_j = \overline{\hat{a}_j} - a_j \tag{3.3.3.3}$$

The calculated deltas tend to show a degree of dependence on chronological age, manifesting as an overestimation of age in younger subjects and underestimation in older subjects [21, 259]. This has been observed in numerous studies [8, 21, 35, 53, 67, 260–262].

This is primarily because of the regression towards the mean phenomenon [263]. This occurs in processes impacted by stochasticity, where observations tend to shrink and cluster around the mean of a predicted distribution rather than remaining as extreme as in their original sampling [264, 265]. Here, stochasticity is induced by the various external factors impacting healthy brain ageing, leading to acceleration or resilience.

In the context of assessing associations between brain age deltas and nIDPs, this relationship between brain age deltas, or regression residuals, and chronological age is detrimental. Uncorrected, it amounts to the study of the relationship between nIDPs and a transformation of chronological age [35]. This behaviour can be corrected using one of two methods, described in [265], which use either the chronological ages $\boldsymbol{a}$ (Method 1), or the predicted ages $\hat{\boldsymbol{a}}$ (Method 2).

Method 1 assumes that the brain age deltas $\boldsymbol{\Delta}$ can be explained as a linear model of chronological ages $\boldsymbol{a}$, where α and β are the regression slope and intercept (Equation 3.3.3.4). This enables the calculation of the corrected predicted ages $\hat{\boldsymbol{a}}_{corr}$ (Equation 3.3.3.5), and the corrected deltas then being calculated using Equation 3.3.3.3. A mathematically identical solution can be obtained by assuming the predicted ages $\hat{\boldsymbol{a}}$ can be explained as a linear model of chronological ages $\boldsymbol{a}$ (Equations 3.3.3.6-3.3.3.7). These methods can be further extended to include higher order terms, accounting for potential non-linear relationships [21]. Yet, they do have several limitations, as found by Butler et al. [265]. Firstly, the reported model accuracy, following correction, is inflated relative to the true model

accuracy. In addition, the degree of inflation is much greater for models with low correlations between predicted and chronological age, than for models with higher correlation values. Thus, in an extreme case where the predicted and chronological ages have 0 correlation, the corrected predicted age would be perfectly correlated to chronological age.

$$\boldsymbol{\Delta} = \alpha \times \boldsymbol{a} + \beta \tag{3.3.3.4}$$

$$\hat{\boldsymbol{a}}_{\boldsymbol{corr}} = \hat{\boldsymbol{a}} - (\alpha \times \boldsymbol{a} + \beta) \tag{3.3.3.5}$$

$$\hat{\boldsymbol{a}} = \alpha \times \boldsymbol{a} + \beta \tag{3.3.3.6}$$

$$\hat{\boldsymbol{a}}_{\boldsymbol{corr}} = \hat{\boldsymbol{a}} + [\boldsymbol{a} - (\alpha \times \boldsymbol{a} + \beta)] \tag{3.3.3.7}$$

Method 2 assumes *a priori* that the debiassed, or corrected, predicted ages and chronological brain ages belong to the same space and distribution. It assumes that the original predicted ages were skewed from this distribution by factors impacting brain ageing, regression dilution and non-Gaussian data distributions [265]. Given this assumption, it proposes to correct the brain age predictions using Equations 3.3.3.8-3.3.3.9, where α and β are the regression slope and intercept between predicted, $\hat{\boldsymbol{a}}$, and chronological, $\boldsymbol{a}$, ages [140, 266]. This approach debiases the predictions while retaining the original correlations between chronological and predicted brain ages, given the invariance of linear operations [265].

The downside of this method is that adding higher order terms is non trivial, as complex values might arise when performing the calculations. This can be somewhat circumvented by means of a pseudo-invertible operation. Here, the higher order regression parameters are calculated normally and used to define a discretised approximation of predicted age, $\hat{\boldsymbol{a}}_{approx}$, using a function of the chronological ages, $\boldsymbol{a}$, (Equation 3.3.3.10). The corrected brain age predictions are then found as the values in the chronological age space of the function, for which the absolute differences

between each uncorrected predicted age a_i and the discrete approximated $\hat{\boldsymbol{a}}_{approx}$ values is minimum (Equation 3.3.3.11). This approach has obvious limitations, particularly for non-monotonically increasing functions which can lead to artefactual components emerging at the edges of the predicted age range where the fitted function is not extrapolated sufficiently. Thus, this higher order approach should only see limited utilisation, for demonstrative purposes only.

Nevertheless, given that Method 2 debiases predictions while retaining the original correlation between the predicted and chronological ages, this method was chosen as the default brain age prediction debiasing method. In addition, preliminary experiments using both Method 1 and the invertible operation described above for Method 2, found that higher order coefficients are generally very small. This indicates that in most cases, the first order bias correction operation is sufficient.

$$\hat{\boldsymbol{a}} = \alpha \times \boldsymbol{a} + \beta \tag{3.3.3.8}$$

$$\hat{\boldsymbol{a}}_{\boldsymbol{corr}} = \frac{\hat{\boldsymbol{a}} - \beta}{\alpha} \tag{3.3.3.9}$$

$$\hat{\boldsymbol{a}}_{\boldsymbol{approx}} = f(\boldsymbol{a}) \tag{3.3.3.10}$$

$$\hat{a}_i^{corr} = a_k, \text{ where } k = argmin\left(|\hat{a}_i - \hat{\boldsymbol{a}}_{\boldsymbol{approx}}|\right) \tag{3.3.3.11}$$

3.3.4 Correlating Brain Age Deltas with Biological Phenotypes and Lifestyle Factors

To investigate whether different maps encode unique information relating to brain ageing, the correlations between the debiased brain age deltas and UK Biobank nIDPs were calculated for both female and male left-out test subjects separately. To ensure that the observed correlations are solely due to differences in encoded signals specific to each modality, and linearly independent of chronological age and confounding factors, both debiased brain age deltas and nIDPs were linearly

deconfounded using 613 UK Biobank confounds (Appendix B, Section B.3). Deconfounding was carried out using the same approach as described by Elliott et al. [229] utilising linear regression, where the confounds represented the various regressors. Given the large number of confounds considered, an independent analysis of the variance explained by each regressor was not carried out. Then, only those nIDPs with valid data for at least 10% of test subjects were retained, leaving $13,809$ nIDPs. Finally, both the brain age deltas and nIDPs were standardised.

The Pearson-r correlations were computed for each brain age delta-nIDP pair. To assess the significance of these correlations, t-statistics were calculated based on the Pearson correlation coefficients, from which p-values were derived. Given the very small numerical values of the p-values, the map-specific $-log(p)$ values were calculated and then grouped based on nIDP categories. To quantify statistical significance, the $-log(p)$ values were sorted and compared to map and sex-specific false discovery rates (FDR) and Bonferroni (Bonf.) thresholds. The Bonferroni threshold is equal to ≈ 5.442 for the 13809 nIDPs.

This approach of estimating p-values is, however, limited by the assumption that the data is normally distributed, which may not be the case for all nIDPs. To address this, two-sided paired permutation tests were employed to verify the accuracy of the predicted parametric p-values. As permutations are not predicated on data distribution assumptions [267], the non-parametric p-values derived from permutation testing can be seen as more accurate. However, permutation tests are computationally demanding, especially for large datasets where enumerating all possible permutations becomes impractical. To mitigate this, the minimum number of permutations, n_{perm}, for a specific p-value, was approximated using the resolution of the achievable p-value [268] (Equation 3.3.4.1). It should be noted that while this method provides a theoretical minimum, the resulting p-values can be sensitive to the specific set of permutations used. Achieving robust resolutions for all possible nIDP correlation p-values remains computationally challenging, limiting their extensive use in this work.

$$p = \frac{1}{n_{perm}} = \frac{1}{10^{-log(p)}} \tag{3.3.4.1}$$

3.4 Comparing HGL with an Established Brain Age Prediction Architecture

This section provides an in-depth comparison between the state-of-the-art brain age predicting model at the start of this project (late 2021), the Simple Fully-Convolutional Network (SFCN) [67], and HGL, my lightweight deep learning architecture. Although not the state-of-the-art anymore at the time of writing (late 2022), SFCN still represents the model of choice for numerous brain age studies [53, 114, 252, 269–271]. The aim of this comparison is to establish the best method for the outlined objectives in this chapter. First, a detailed description of SFCN is provided, together with any experimental setup particularities. Then, the performances of HGL and SFCN are compared across several experiments using a subgroup of maps. Finally, the observed results are discussed, and the best CNN method is selected.

3.4.1 Methods

3.4.1.1 Simple Fully-Connected Network (SFCN)

The SFCN network (Figure 3.4) is based on VGG-16 [87], with several modifications. It is composed of five convolutional blocks, similarly organised to HGL. These are followed by a sixth convolutional block, composed of a $1 \times 1 \times 1$ 3D convolution, a 3D batch normalisation layer and ReLU nonlinearity. The seventh and final block is composed of a global average pooling layer, a 50% dropout layer, another $1 \times 1 \times 1$ convolution and a final softmax activation layer. Rather than returning a single scalar output, SFCN outputs a vector corresponding to a probability distribution for the chronological age. The filter number for each convolutional layer is as follows: $[32, 64, 128, 256, 256, 64, 40]$. The final brain age prediction, $\widehat{age}$, is calculated by means of a weighed average between the predicted probabilities $p_{age\ bin}$ and the age at each bin centre $age_{age\ bin}$ (Equation 3.4.1.1).

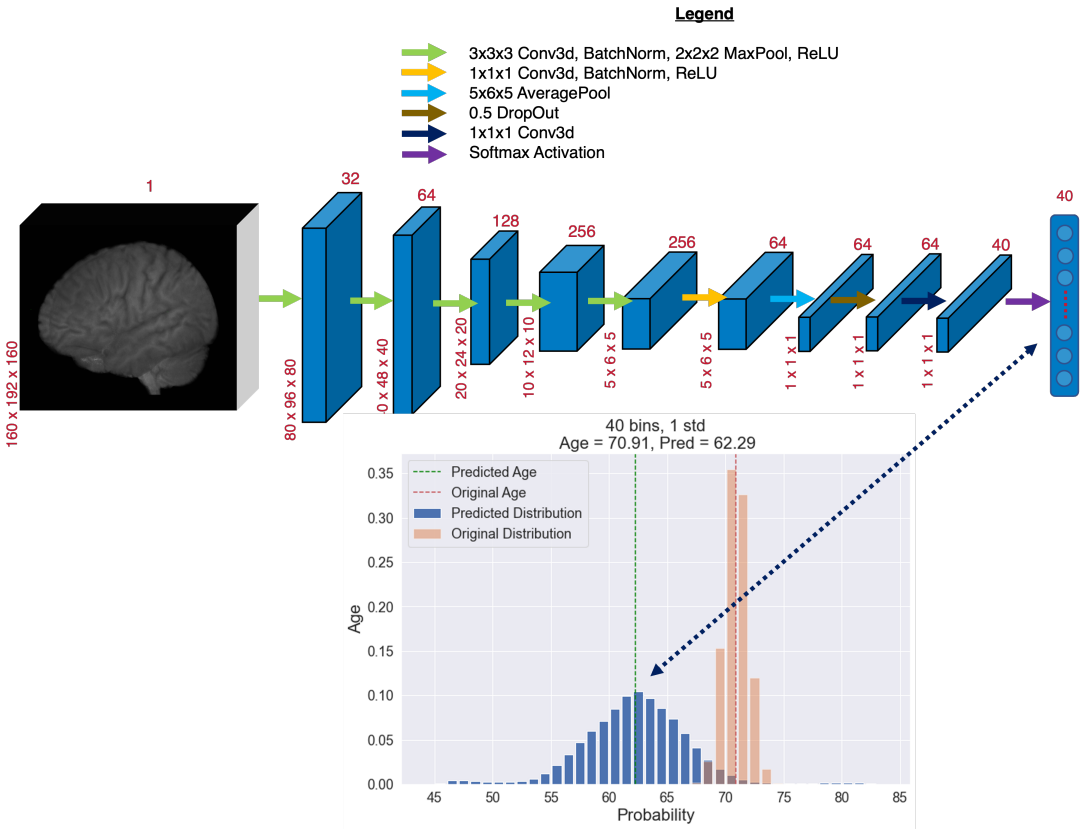


Figure 3.4: SFCN Architecture as proposed in [67] together with an example of soft labels and predicted probabilities, which are used to compute the KL-divergence loss. The output of the network is a 40-element long vector corresponding to output age probabilities. This relationship is indicated by the dotted arrow.

$$\widehat{age} = \sum_{bin=1}^{40} p_{age\ bin} \times age_{age\ bin} \tag{3.4.1.1}$$

Given its output, SFCN essentially performs a soft classification. The target vectors used during training are generated using a discrete normal distribution, where the mean is the ground truth chronological age, and the standard deviation (σ) is equal to 1 year. The distribution is discretised over 40 age bins corresponding to the age interval between 44 and 84 years. Such example distributions can be observed in Figure 3.4 and the left column of Figure 3.7. As this age range truncates the distributions for very young or old subjects (top-left and middle-left rows of Figure 3.7), experiments were also run using a 50-bin vectors with age ranges between 39 and 89.

Rather than using a conventional regression loss function, such as the mean squared error (MSE), SFCN employs the Kullback-Leibler divergence (KLD) with batch mean reduction. Compared to MSE, which incentivises the model to match the target value, KLD aims to match the prediction and target value distributions. Narrow target distributions, marked by small σ values, do not necessarily incentivise narrow predictions, as there might be little incentive for the output to be Gaussian (bottom row of Figure 3.7). This could lead to training difficulties. These could theoretically be alleviated by using wider target distributions, as these ensure at least some overlap between predicted and target distributions. This consideration prompted the running of several experiments where σ was increased from 1 to 10 years (right column of Figure 3.7).

3.4.1.2 Experimental Setup

Experiments were carried out using the large female dataset and the T1 Linear map, as well as the small female dataset (see Section 3.3) with the T1 Nonlinear, T1 Linear, T2 FLAIR Nonlinear, and dMRI Summed Probabilistic Tracts maps. HGL and SFCN were compared in terms of prediction accuracy, measured using the mean absolute error (MAE), and the density distributions of the predicted ages when compared against the ground truth chronological ages. The latter comparison was carried out using density scatter plots and kernel density estimates (KDEs). KDEs represent the data using a continuous probability density curve. For all experiments, separate linear and cubic debiasing approaches were used.

For the large female dataset, the associations between the predicted brain age deltas and nIDPs were also compared between the two methods. To validate the observed correlations, two-sided paired permutation tests were also performed, testing the null-hypothesis that the observations within each sample are drawn from the same underlying distribution, and that the pairings between observations are random [272].

SFCN's experimental setup was largely similar to that described for HGL, with several differences. Firstly, no network ensembling was carried out for either

SFCN or HGL. This is because, if any significant differences exist between the two architectures, they can be assumed to be large in comparison to any beneficial denoising effect of ensembling. In addition, to focus the comparison solely on architecture-induced differences, no data augmentation techniques were applied to either network.

In terms of training, the process described by the original SFCN paper [67] was followed. Network training was carried out with the stochastic gradient descent (SGD) optimiser. The initial learning rate of $1e-2$ was multiplied by 0.3 every 30 epochs. The network was trained for a fixed 200 epochs, with the epoch producing the best validation loss being saved for evaluation.

3.4.2 Results

3.4.2.1 Brain Age Prediction Accuracy

The performance of HGL and SFCN were first compared in terms of brain age prediction accuracy across a subgroup of maps (Table 3.1). It can be observed that HGL outperforms SFCN in most of the considered summary statistics, with one exception for SFCN (50 bins, 10 std), which performs marginally better in terms of the linearly debiased MAE and correlation metrics, despite still performing worse in terms of MAE.

3.4.2.2 Predicted Brain Age Distributions

Probably the most important difference between HGL and SFCN can be observed when comparing them in terms of their predicted brain age KDE distributions and density scatter plots (Figure 3.5). While HGL produced distributions in line with the expectation that these should be close to the ground truth, SFCN shows a truncation of predicted results for the very young or old subjects, and a dip in the predicted KDE distribution close to the target dataset median. This can be observed for both the experiments conducted with the large female dataset and the T1 Linear map (Figure 3.5), but also for the experiments carried out with several other maps

Table 3.1: Age prediction accuracy of HGL and SFCN. The performance of HGL is compared to three versions of SFCN (vanilla, with 50-bin distributions and with both 50-bin distributions and a 10-year standard deviation), across 4 maps and 2 datasets. The MAE is reported for the original, linear and cubic-debiased cases. The Weighted MAE is calculated as in [8], by dividing the MAE by the dataset age range (37.137 years).

Map & Dataset / Network	MAE	MAE (Linear Debias)	MAE (Cubic Debias)	Predicted-Chronological Correlation (r)	Weighted MAE	R^2
T1 Linear - Large Female Dataset						
HGL	**2.699**	3.036	**3.135**	0.886	**0.073**	0.784
SFCN	2.852	3.208	3.485	0.872	0.077	0.76
SFCN (50 bins)	2.850	3.221	3.414	0.871	0.077	0.759
SFCN (50 bins, 10 std)	2.881	**2.986**	3.206	**0.890**	0.078	**0.792**
T1 Linear - Small Female Dataset						
HGL	**2.943**	**3.394**	**3.419**	**0.864**	**0.079**	**0.747**
SFCN	3.273	3.805	4.084	0.835	0.088	0.697
SFCN (50 bins)	3.367	3.837	4.098	0.840	0.091	0.706
SFCN (50 bins, 10 std)	3.493	3.653	3.949	0.851	0.094	0.724
T1 Nonlinear - Small Female Dataset						
HGL	**2.858**	**3.240**	**3.251**	**0.881**	**0.077**	**0.776**
SFCN	2.994	3.423	3.576	0.864	0.081	0.747
SFCN (50 bins)	3.310	3.471	3.706	0.863	0.089	0.745
SFCN (50 bins, 10 std)	3.327	3.368	3.619	0.870	0.090	0.756
T2 FLAIR Nonlinear - Small Female Dataset						
HGL	**2.707**	**3.011**	**3.096**	**0.891**	**0.073**	**0.793**
SFCN	3.297	3.347	3.541	0.867	0.089	0.752
SFCN (50 bins)	3.108	3.532	3.846	0.857	0.084	0.734
SFCN (50 bins, 10 std)	3.384	3.462	3.641	0.862	0.091	0.743
Summed Tracts - Small Female Dataset						
HGL	**3.876**	**4.954**	**4.704**	**0.767**	**0.104**	**0.589**
SFCN	4.310	5.629	5.898	0.728	0.116	0.53
SFCN (50 bins)	4.225	5.482	5.843	0.740	0.114	0.548
SFCN (50 bins, 10 std)	4.538	5.715	5.531	0.728	0.122	0.529

and the small female dataset (Figure 3.6). As there is no clear explanation as to what might be causing this behaviour, several hypotheses were tested.

Firstly, it was hypothesised that the 40-bin age probability soft labels are too narrow and truncate the distributions of very young and old subjects, which prompted the running of experiments with 50-bin vectors. The σ standard deviation used to generate the soft labels was also increased from 1 to 10-years for the 50-bin distributions, as the narrow σ could disincentivise the KLD loss from training the network to output narrow Gaussian predictions. Finally, a cubic-bias correction was attempted, as the linear bias correction might be insufficient for correction of method-induced biases. The summary statistics for these modified SFCN networks are presented in Table 3.1, and examples of the various soft labels and predicted distributions in Figure 3.7.

Of these, only the SFCN (50 bin, 10 std) network produced results which alleviate to some extent the observed issues for the T1 Linear map and the large

female dataset (Figure 3.5). Yet, this behaviour was not reproduced for the other maps and the small female dataset (Figure 3.6). The cubic-debiasing algorithm also leads to some alignment improvements between the debiased predicted age KDE distributions and the ground truth (Figure 3.8). However, this correction does not seem to be sufficient to match the underlying ground truth distribution. In addition, as the fitted function is not monotonically increasing, the cubic-debiasing algorithm generates artefactual components at the edges of the predicted age range.

Given these observations, it is likely that the observed behaviour is caused by a complex interplay of several factors, such as the regression towards the mean and effects induced by the KLD loss and network architecture design decisions.

Regarding the KLD loss, as mentioned previously, this gives a measure of the divergence between two probability distributions. As can be observed in Figure 3.7, narrow target distributions do not necessarily incentivise the network to also make narrow predictions. Moreover, if there is little or no overlap between the prediction and target, this will result in a bad, potentially unstable KLD loss value. Such instability can be observed when comparing the validation-dataset training curve of SFCN with that of HGL (Figure 3.9). Thus, as seen on the bottom row of Figure 3.7, this can actually lead to the network predicting wide distributions to get at least some similarity to the narrow target. This also explains why the higher σ distributions achieve somewhat better results. However, this does not explain the full extent of the observed behaviour.

The final element contributing to the observed results is the presence of dropout in the final layer of SFCN. Though a useful tool for machine learning problems performing classification tasks, Dropout can be detrimental when utilised in regression problems. During training, dropout randomly zeros a set of inputs using a user-defined probability, after which it scales the outputs to preserve their mean value. The operation, however, does not preserve the original variability, which, when the activations are transformed using a non-linear activation, leads to changes in the mean of the activations themselves [258, 273]. When the signal reaches the final fully-connected layer in the network, the weights are fine tuned

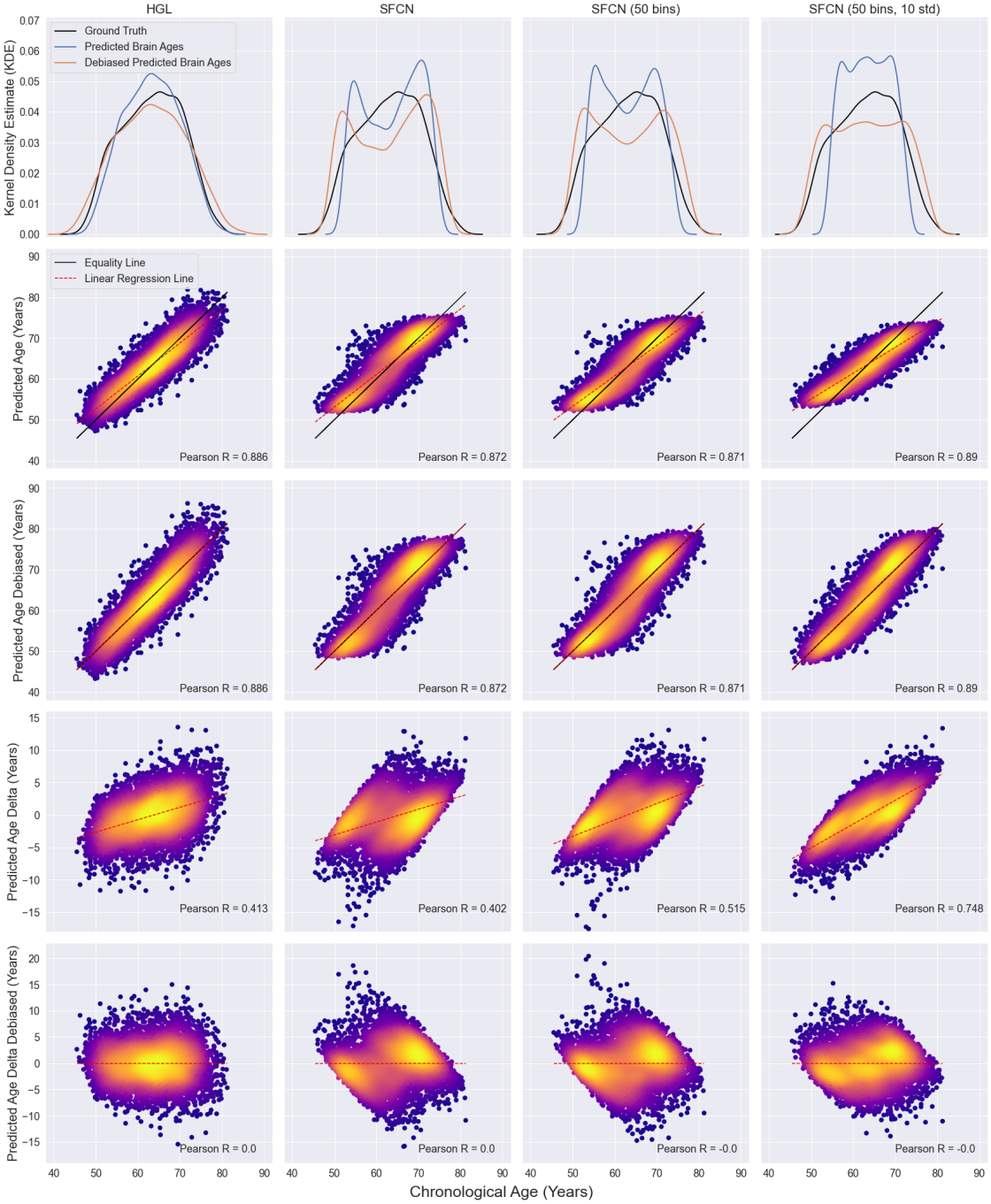


Figure 3.5: Kernel density estimates (KDE) and density plots (DP) of HGL and SFCN with T1 Linear data for the large female dataset. Each column of the figure corresponds to a network configuration, and each row to a plot with respect to chronological age: KDE of predicted and debiased predicted age distributions for the left-out subjects, DPs of predicted age and predicted age linearly debiased, and DPs of predicted age deltas and predicted age deltas debiased.

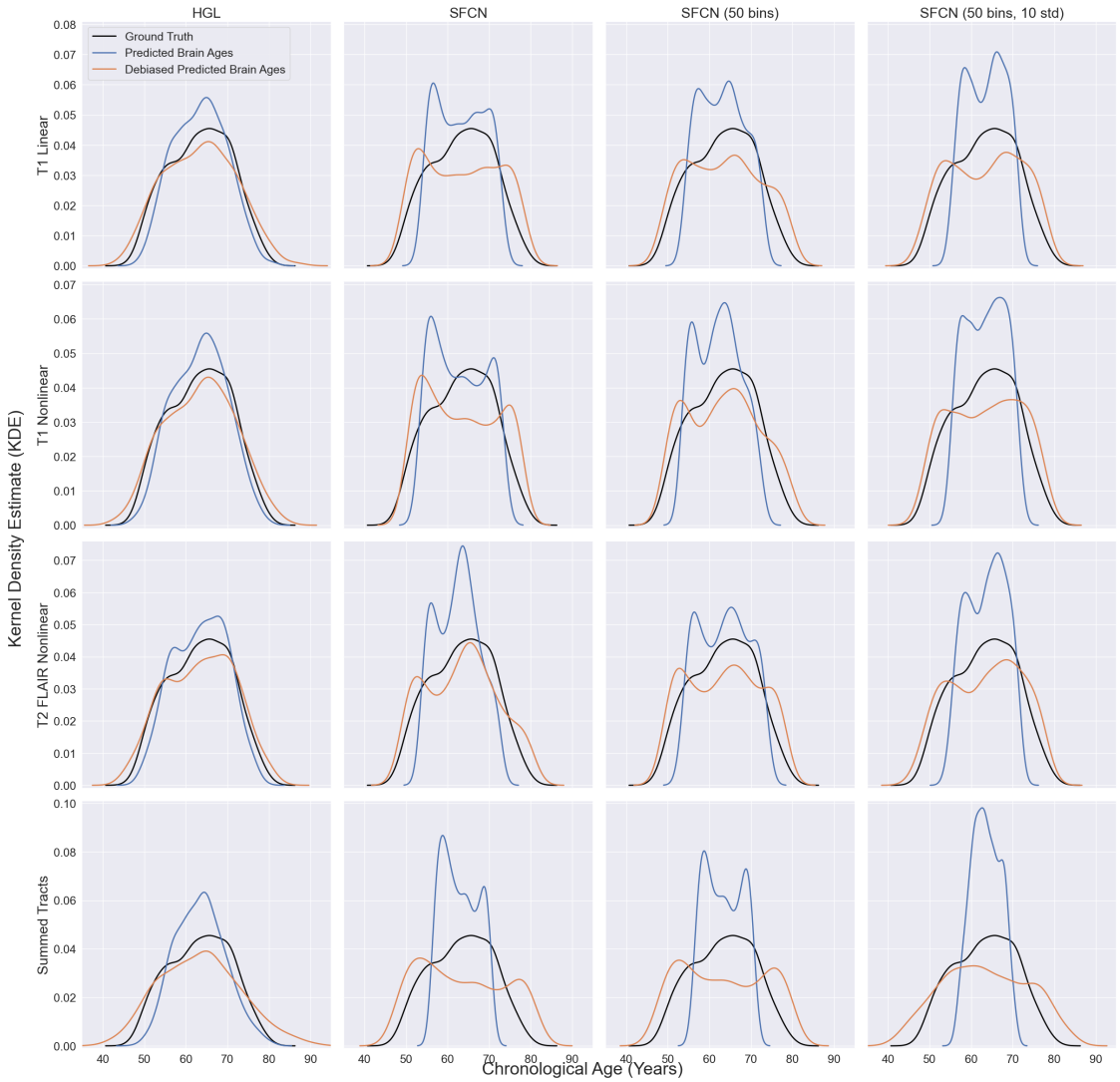


Figure 3.6: Kernel density estimates (KDE) of several maps with the small female dataset with linear debiasing. Each column of the figure corresponds to a network configuration, and each row to the KDE of a different map with respect to chronological age: T1 Linear, T1 Non-Linear, T2 FLAIR Nonlinear, and dMRI Summed Tracts.

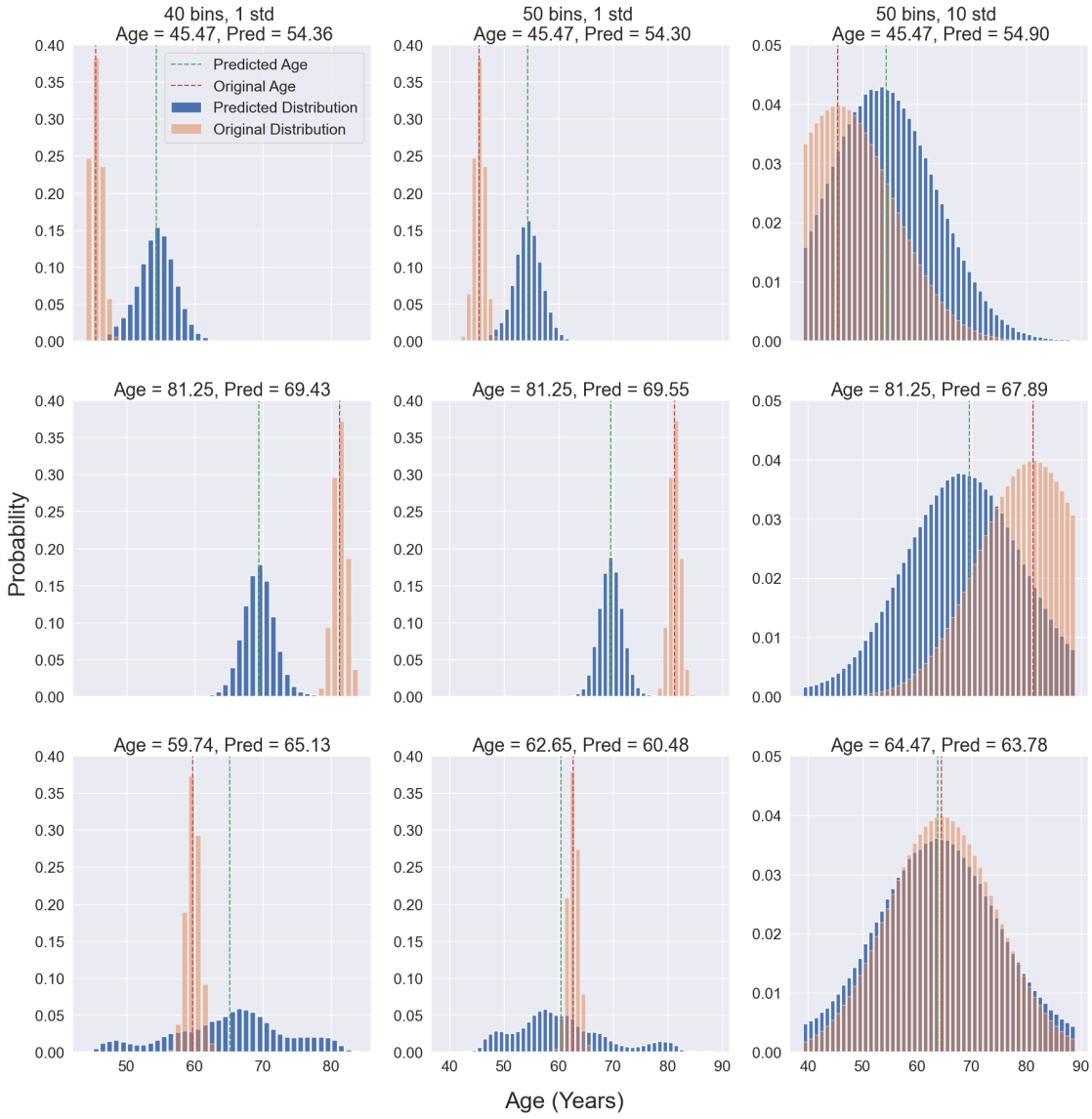


Figure 3.7: Examples of SFCN-specific soft labels and output probabilities. The top and middle row display extreme cases, referring to the youngest and oldest female subjects in UK Biobank, while the bottom row displays failure cases for the 40 and 50 bin experiments, where the output (blue) distribution is non-Gaussian, and a typical mid-age range output for the 50 bin 10 std experiment. The top two rows show how subjects with chronological ages close to the tails of the distribution are biased towards the mean when having their brain ages predicted.

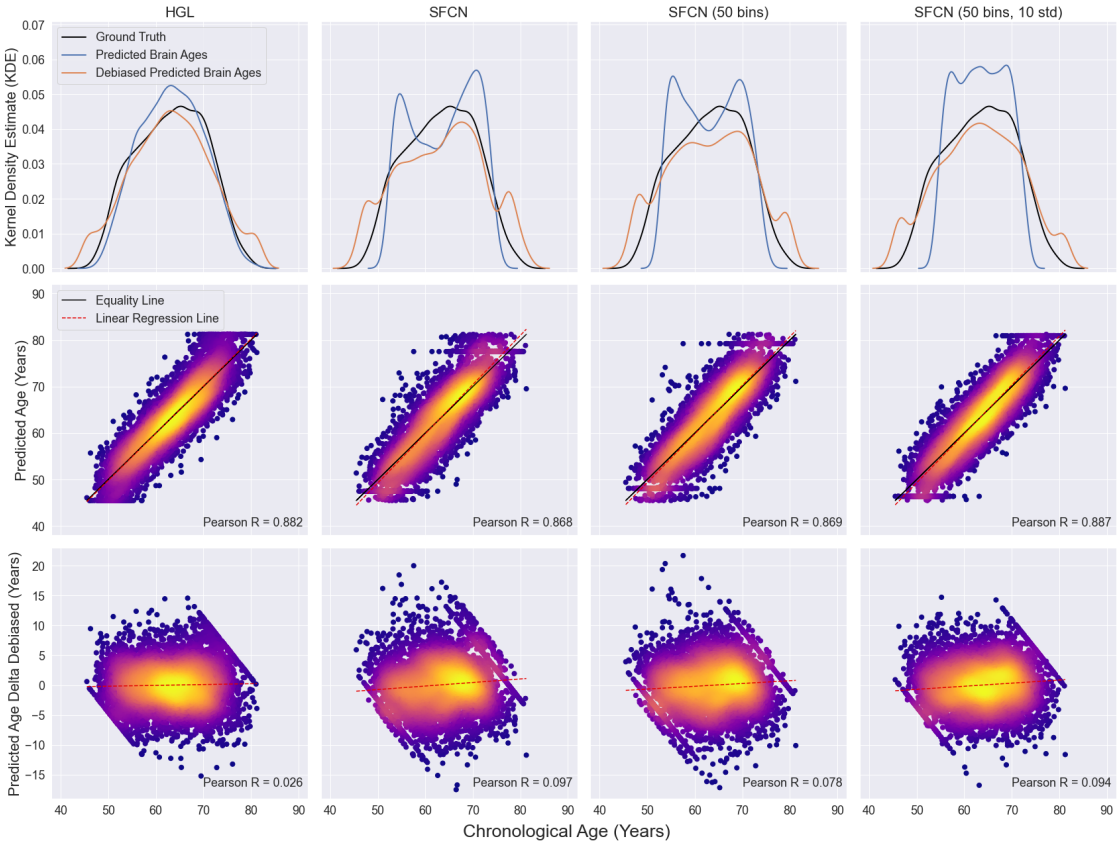


Figure 3.8: Kernel density estimates (KDE) and density plots (DP) of HGL and SFCN of T1 Linear data for the large female dataset with cubic debiasing. Each column of the figure corresponds to a network configuration, and each row to a plot with respect to chronological age: KDE of predicted and debiased predicted ages distributions for the left-out subjects, DP of predicted age with cubic debiasing, and DP of predicted age deltas with cubic debiasing.

to fit the training dataset statistics. This is not problematic in classification tasks where the relative scale of results is not as important, dropout being an effective regularisation technique, preventing the co-adaptation of neurons. However, this is not the case in regression, where the output scale is relevant [274].

When adding to HGL's final layer (32 parameters) an equivalent amount of dropout ($p = 0.5$) to that used in SFCN's final layer (40 parameters), a similar truncation to that seen in SFCN can be observed in HGL (Figure 3.10). However, while the two peaks in SFCN are distributed evenly about the median, this is not the case for HGL. Moreover, the truncation appears more aggressive for HGL, and affects only the lower part of the distribution. This can be caused by the

previously discussed impact of KLD for SFCN, as well as different architecture decisions between the two networks.

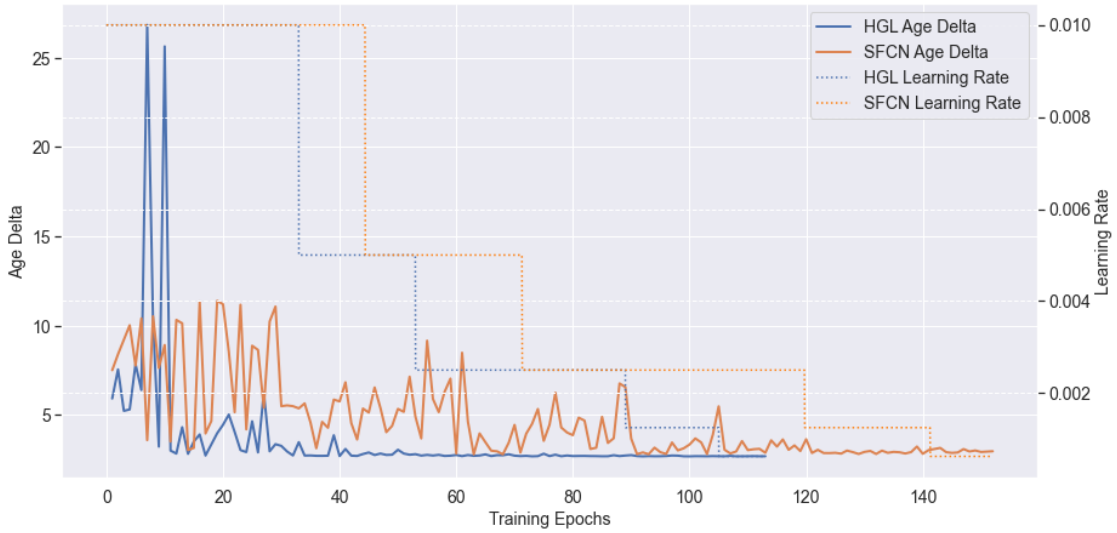


Figure 3.9: HGL and SFCN validation curves, obtained during training by passing the validation dataset through the network at the end of each training epoch. Rather than expressing the y-axis values in terms of MAE and KLD losses for each network respectively, the epoch-wise Mean Age Delta was calculated at the end of each training epoch. For a complete view, the learning rates for both networks are also added on the secondary y-axis.

3.4.2.3 nIDP Associations Comparison

Finally, the associations between the predicted brain age deltas (linearly debiased and deconfounded) and nIDPs were calculated for both HGL and the three SFCN versions (Figure 3.11). When comparing the associations obtained with the two methods, it can be seen that despite some similarity, SFCN produces strong associations passing the Bonferroni threshold in several additional nIDP classes than HGL. These include Alcohol, Lifestyle Measurements, Diet or Blood Assays. In addition, several differences can be seen between different SFCN models as well.

These differences raise questions regarding the validity of the calculated p-values. To address this, two-sided paired permutation tests were performed, under the assumption that the parametric and non-parametric p-values should be identical, or very close, in order to validate original parametric ones. Two permutation tests were

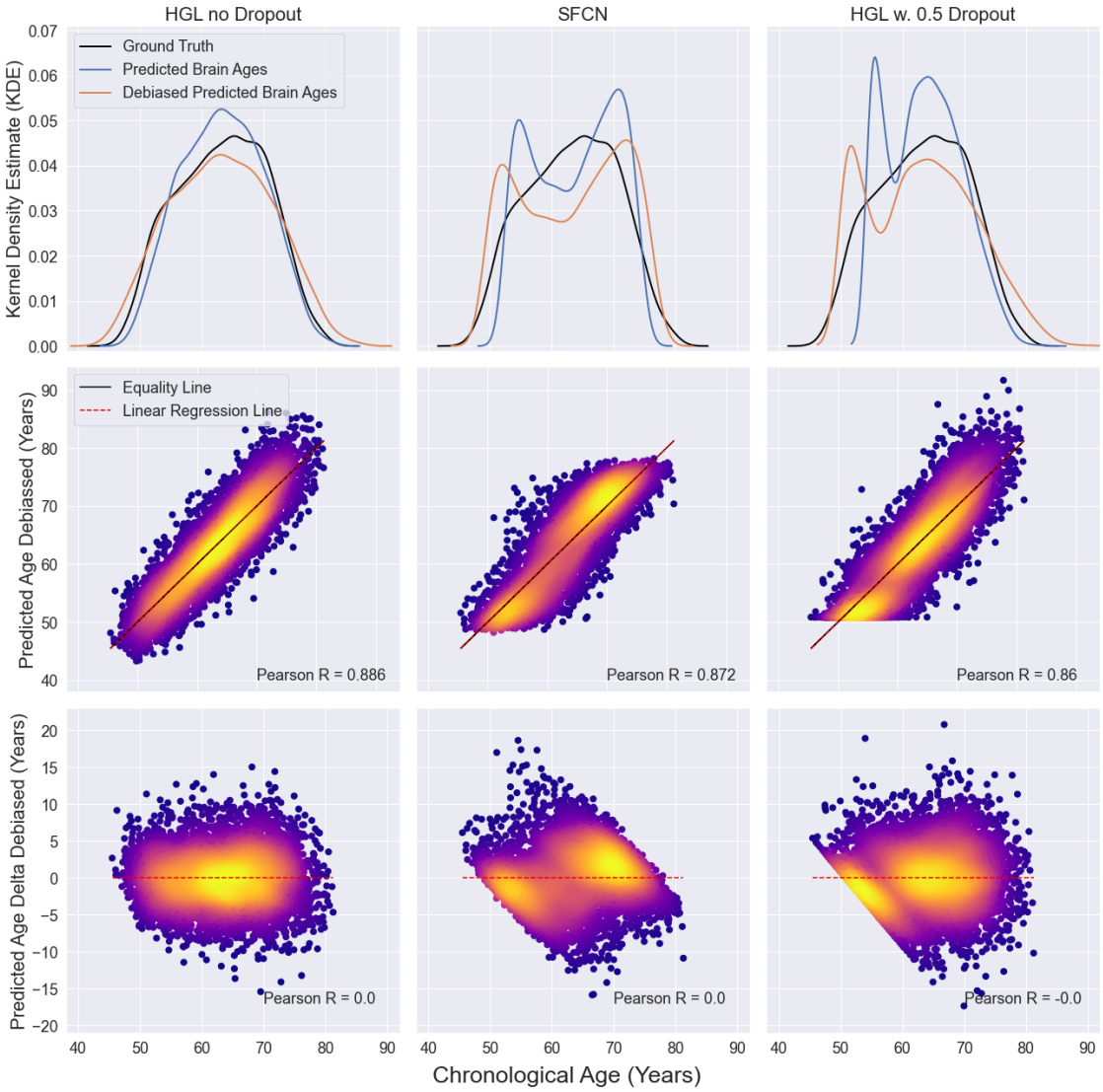


Figure 3.10: Kernel density estimates (KDE) and density plots (DP) of HGL and SFCN of T1 Linear data for the large female dataset with and without dropout in the final layer. Each column of the figure corresponds to a network configuration, and each row to a plot with respect to chronological age: KDE of predicted and debiased predicted ages distributions for the left-out subjects, DPs of predicted age linearly debiased, and DPs of predicted age deltas debiased. The first two columns are identical to those in Figure 3.5, and are added for comparison purposes.

performed, with the number of permutations (n_{perm} - Equation 3.3.4.1) being set at two resolutions corresponding to $-log(p) = 3.69$ and $-log(p) = 5.44$, the latter being equal to the Bonferroni threshold. The number of permutations (i.e. the resolution) was limited by computational constraints. In addition, for the higher resolution, only nIDPs with $-log(p) \geq 2.0$ were considered. The results are presented in Figure 3.12.

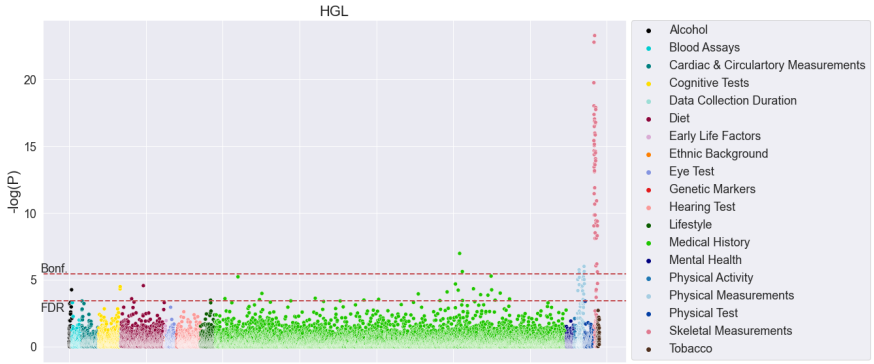
Under the resolution constraints, determining which method produced the best results is difficult. For the lower resolution, SFCN and SFCN (50 bins) show larger deviations from the identity line when compared to HGL (Figure 3.12a). This suggests that they are violating to a larger degree the assumptions of the parametric p-value calculations. This, however, cannot be said for the SFCN (50 bins, 10 std) method at lower resolution. Similar observations can be made for the higher resolution (Figure 3.12b). Yet, here it can be seen that the parametric and non-parametric nIDPs for HGL show better agreement than any of the SFCN networks.

3.4.3 Discussion

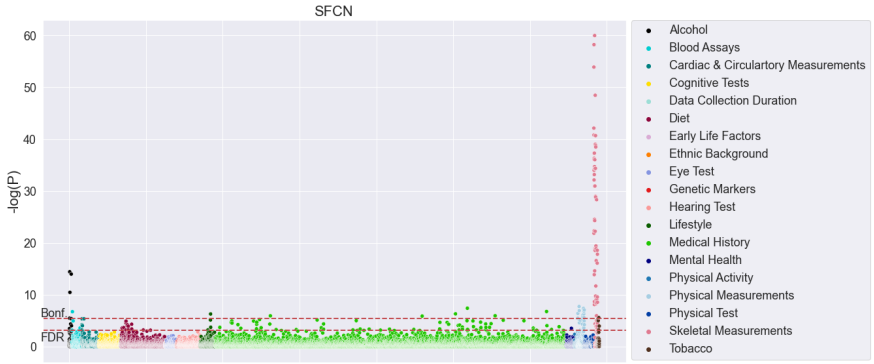
In this section, the performance of HGL, a new lightweight CNN for brain age prediction, was compared against that of SFCN [67], an established model. The comparison was carried out for two datasets and a subgroup of maps.

The comparison was initially designed to provide an established baseline for HGL. However, following the observations made for the predicted age distributions obtained with SFCN, a decision was made to investigate the properties and reliability of the SFCN method. This is of interest, given that SFCN has become a popular method utilised for investigating the "brain age gap" phenomenon [53, 114, 252, 269–271].

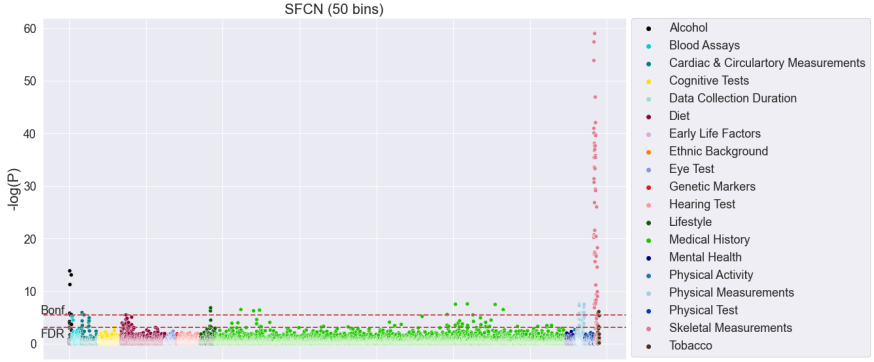
SFCN was found to produce test-data distributions which diverge significantly from the ground truth, showing a truncation of predictions for subjects at the edges of the age distribution, and a dip close to the median (Figures 3.5 and 3.6). A similar behaviour can be observed in Figure 2(b) of the work carried out by Leonardsen et al [53], though those figures do not contain density information. Several tests revealed that this behaviour can be attributed to the interplay between several factors,



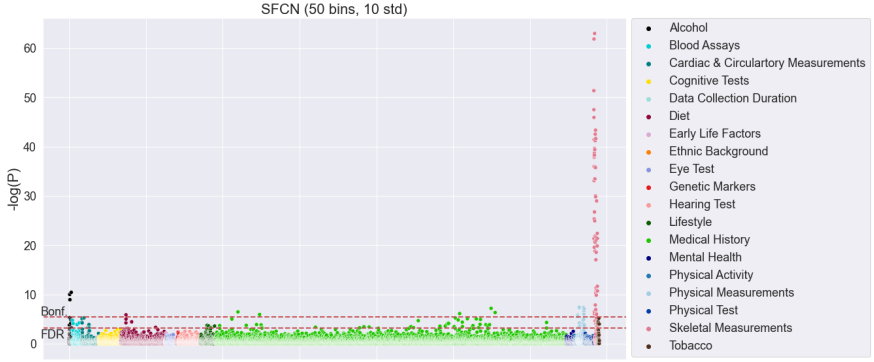
(a) Manhattan plot of for HGL deltas.



(b) Manhattan plot of for default SFCN deltas.

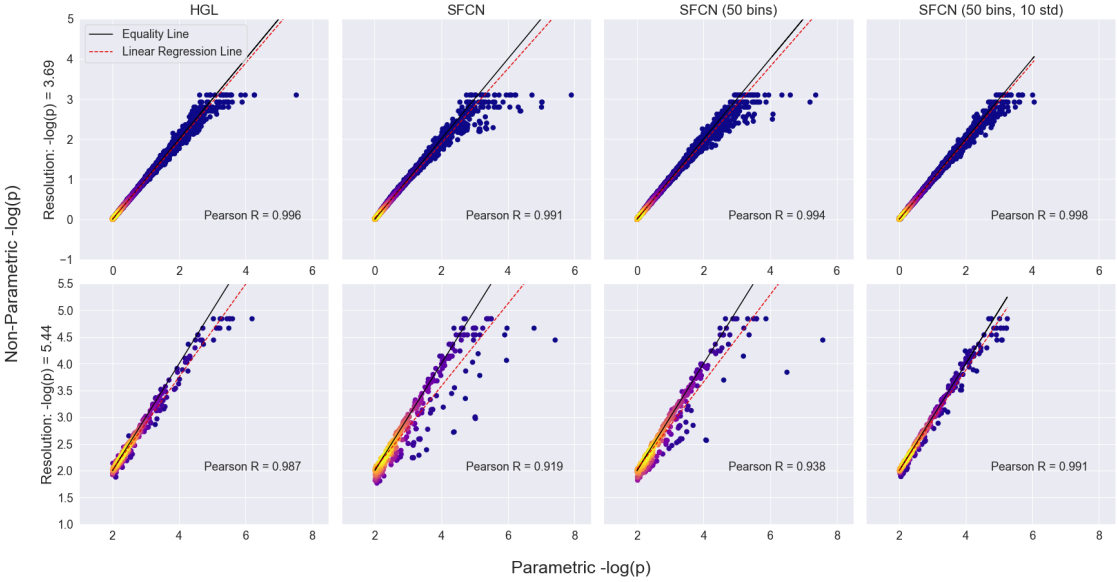


(c) Manhattan plot of for SFCN (50 bins) deltas.

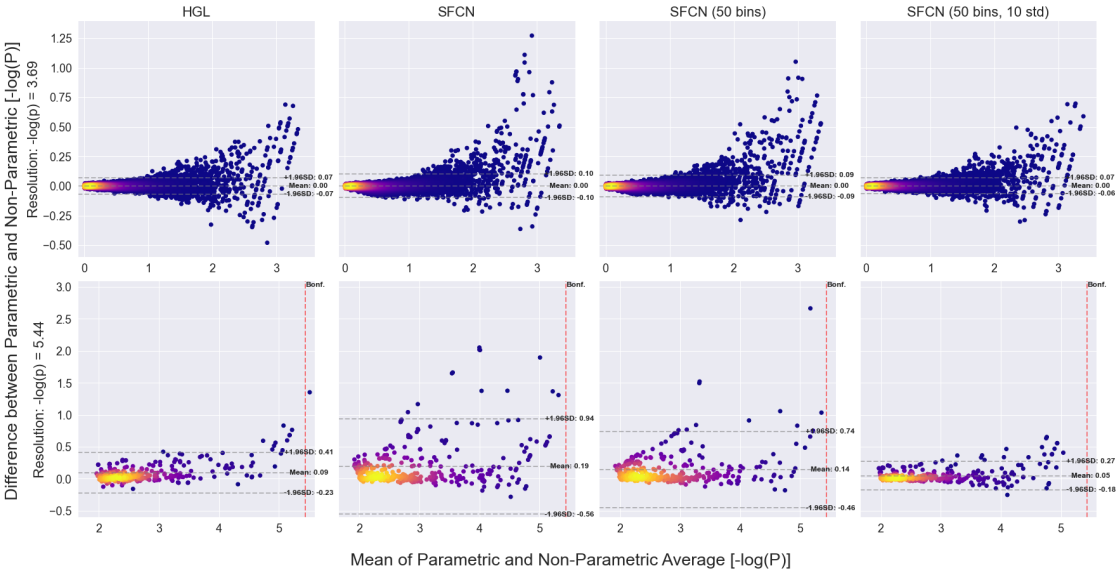


(d) Manhattan plot of for SFCN (50 bins, 10 std) deltas.

Figure 3.11: Manhattan plots relating UK Biobank nIDPs to SFCN and HGL predicted brain age deltas. The results are obtained for the T1 Linear map and the large female dataset, with brain age deltas being linearly debiased and having had confounds removed. Each dot corresponds to the statistical significance of the respective correlation.



(a) Density plots between the parametric and non-parametric p-values.



(b) Bland-Altman plots between the parametric and non-parametric p-values

Figure 3.12: Permutation testing results for SFCN and HGL. Each column corresponds to a different network. These tests were run for two resolutions, corresponding to $-log(p) = 3.69$ and $-log(p) = 5.44$. For easier visualisation, all points with values above these resolutions were eliminated. The higher resolution tests were only conducted for those nIDPs with $-log(p) \geq 2.0$. The results are presented in terms of **(a)** density and **(b)** Bland-Altman plots of the non-parametric (i.e. permutation-tested) and parametric (i.e. original) correlations. The Bonferroni threshold is displayed in the second row of **(b)** as a red dotted line. The mean of the difference and 95% limits of agreement (mean difference ± 1.96 standard deviation of the differences) are displayed as dotted grey lines.

including regression towards the mean, effects induced by the KLD loss, and the use of dropout close to the network output. Of these, the largest contributing factor is probably the use of dropout, as seen in Figure 3.10. This seems to be confirmed by Leonardsen's results, where the described effect is still manifesting after replacing the original soft classification with a regression operation [53]. It is also likely that the anomalous SFCN distributions have knock-on effects for any subsequent analysis, such as associations with nIDPs. When tested using permutation testing, possible violations of the assumptions underpinning the parametric p-values were found. These manifested as a lower degree of agreement between the parametric and non-parametric $-log(p)$ values than for HGL (Figure 3.12).

Given that the majority of subjects in UK Biobank are healthy, any regression method would be expected to produce predicted distributions which resemble in shape the ground truth distributions. This is the case with HGL, which also consistently produced more accurate brain age predictions. Therefore, HGL is better suited for the subsequent experiments in this chapter, as it does not suffer from the possible negative side-effects of SFCN.

3.5 Age Prediction for Individual Maps

Building on the finding that HGL represents a more appropriate method for predicting brain ages than SFCN, in this section the question of which of the 57 3D MRI maps considered for this work can be utilised for brain age predictions was addressed. This was achieved by utilising the methods described in Section 3.3.2 to conduct experiments for both the female and male subject groups. The obtained results were compared against a baseline model proposed by Smith et al [35], referred to as a PCA-IDP Regression Model, representing a successful IDP-based linear method at the time of writing (late 2022). First, a detailed description of this model and its implementation is provided. After this, the brain age predictions obtained by HGL and the PCA-IDP model are presented and compared.

3.5.1 Methods: the PCA-IDP Regression Model

Rather than using 3D volumetric inputs, the PCA-IDP regression method proposed by Smith et al [35] utilises image derived phenotypes (IDPs). The IDPs are first linearly deconfounded using all confounds available in UK Biobank, except those related to age (Appendix B, Section B.3). Then, the model carries out dimensionality reduction using principal component analysis (PCA), after which it feeds the results to a multiple linear regression model. In this work, while keeping to this general framework proposed by Smith et al, several modifications were introduced.

Firstly, to allow for better comparisons between the results obtained with volumetric maps and those obtained with IDPs, brain age predictions were made with linear models trained with both the full set of 3921 IDPs, and groups of IDPs corresponding to their thematic categories (Appendix B, Section B.1). For consistency, all IDPs were verified to contain valid data for at least 99% of the test subjects. In addition, within each IDP category, any subject not containing valid data for all IDPs contained within that category were discarded. This was the case for only a handful of subjects. Following these checks, the training and test datasets were demeaned using the training set feature means.

The optimal PCA dimensionality was determined by decomposing the data using several hand-chosen principal components, and then passing it through an unregularised multiple linear regression algorithm trained to predict chronological age [35]. In their paper, Smith et al. propose the Q metric to select the optimal PCA dimensionality. This is a measure of significance, being calculated as the 99^{th} percentile of the $-\log(p)$ values obtained for the correlation between the linearly-predicted brain age deltas and 5792 nIDPs from UK Biobank (more in Section 3.3.4). Rather than using this approach, the optimal PCA dimensionality was chosen using the best averaged MAE value calculated across 5 cross-validation folds. This maintains experimental consistency with the HGL and SFCN networks.

Then, to boost the method's performance, the linear regression method (Equation 3.5.1.1, where $\hat{\mathbf{w}} = \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}$ are the regression coefficients, $\mathbf{y}$ the n predicted brain ages and $\mathbf{X}$ the m predictor IDP variables) was augmented by the addition of an

ElasticNet penalty. This regularises the ordinary least squares linear regression loss function L_{OLS} (Equation 3.5.1.2, where $\tilde{\mathbf{X}}$ is $\mathbf{X}$ augmented with a column of ones to account for b, and $\tilde{\mathbf{x}}_i$ and y_i are the ith rows of $\tilde{\mathbf{X}}$ and $\mathbf{y}$) by penalising both the sum of squared coefficients (L2 penalty) and the sum of absolute coefficient values (L1 penalty). It achieves this by utilising two additional hyperparameters: the λ constant which multiplies the penalty terms, and the L1 mixing ratio parameter α (Equation 3.5.1.3). For an overview of these linear methods, please see Chapter 2 Section 2.2.1. The optimal values of these hyperparameters were found using a grid search with 5-fold cross-validation. The final results are then reported on the left-out test dataset.

$$\mathbf{y} = \mathbf{w} \cdot \mathbf{X} + b \tag{3.5.1.1}$$

$$L_{OLS}(\hat{\mathbf{w}}) = \sum_{i=1}^{n} (y_i - \hat{\mathbf{w}}^\top \tilde{\mathbf{x}}_i)^2 = \|\mathbf{y} - \tilde{\mathbf{X}}\hat{\mathbf{w}}\|^2 \tag{3.5.1.2}$$

$$L_{EN}(\hat{\mathbf{w}}) = \frac{\sum_{i=1}^{n} (y_i - \hat{\mathbf{w}}^\top \tilde{\mathbf{x}}_i)^2}{2n} + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^{m} \hat{w}_j^2 + \alpha \sum_{j=1}^{m} |\hat{w}_j| \right) \tag{3.5.1.3}$$

The PCA-ICA models were trained locally on a MacBook Pro equipped with a 2.3 GHz 8-Code Intel Core i9 processor.

3.5.2 Results

The results for HGL's brain age prediction, for both the male and female groups, are presented in Table 3.2. Similarly, the results for the PCA-IDP linear method are presented in Table 3.3. For easier visual comparison between the two methods, the prediction errors, reported in terms of Mean Absolute Errors (MAEs), are contained in Figure 3.13.

For both the female and male subject groups, it can be observed that HGL is capable of predicting brain ages for all the considered maps, with varying degrees of success. The female subject group produced slightly better predictions across all maps, which is likely due to the higher number of available training subjects: 9412 vs. 8184 for the male group. Overall, the lowest MAEs were obtained as follows:

- sMRI: T2 FLAIR Nonlinear ($MAE = 2.189$ years), which were also the best performing maps throughout;
- rsfMRI: rsfMRI-0 ($MAE = 4.173$ years);
- tfMRI: tfMRI-1 ($MAE = 3.444$ years);
- dMRI: ICVF ($MAE = 2.631$ years);

When compared to previously published results (Appendix A, Tables A.1-A.6), the predictions for the sMRI, dMRI and rsfMRI maps were similar to other studies in terms of MAE and weighted MAE. The results for the tfMRI maps surpassed those reported in literature.

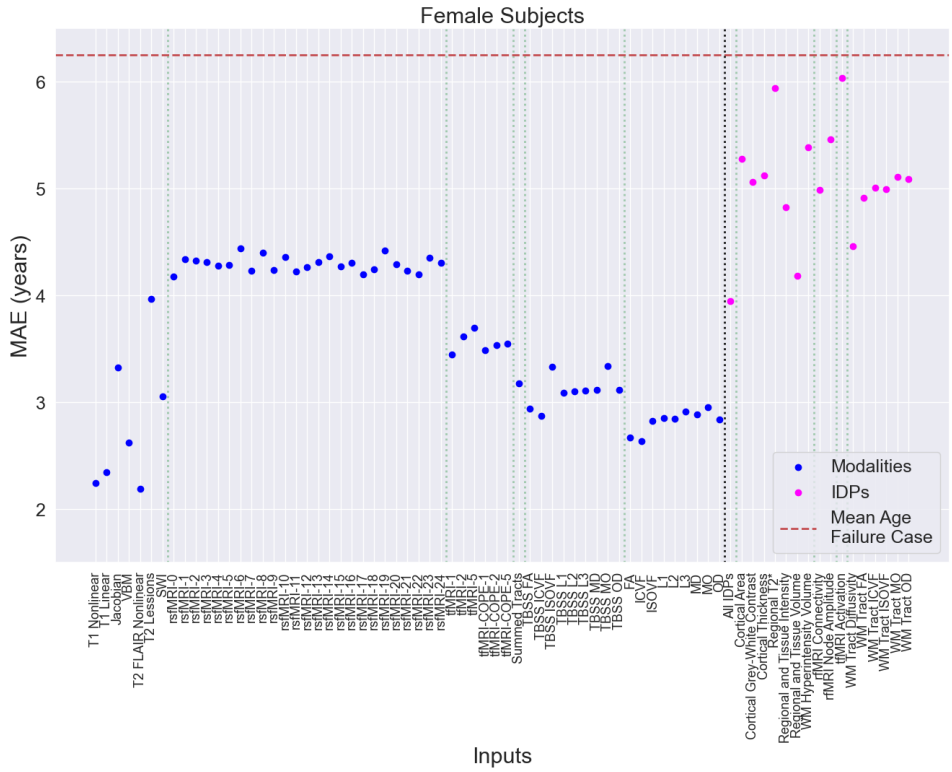
Given the previously observed anomalous brain age prediction distributions observed for SFCN, Kernel Density Estimates (KDE) and density scatter plots were created for the raw and debiassed predictions coming from all tested maps. Figure 3.14 shows these plots for a subset of maps coming from 4 of the core modalities for the female group, demonstrating that the HGL predictions capture the underlying data distributions well. Similar results were obtained for the male subjects.

When comparing the results obtained with HGL to those obtained with the IDP-based linear method, HGL outperforms the latter for all modalities. While the best linear result is obtained for the female subject group when all IDPs are utilised as inputs ($MAE = 3.944$ years), some of the results obtained for IDP subgroups come close to the mean population age failure case (Figure 3.13b). This is the case where the model has been unable to learn any distinguishing features.

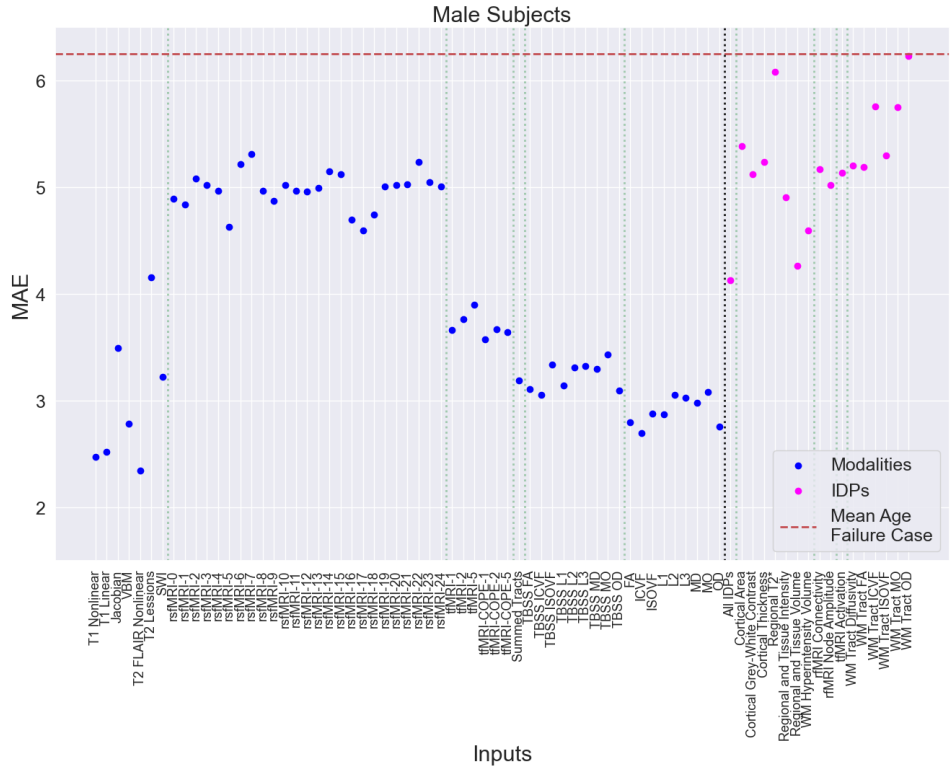
Despite these differences, Figure 3.15 shows a moderate degree of correlation between HGL and IDP-based brain age deltas, after debiasing and deconfounding. Strongly correlated clusters can be observed between the information (maps and IDPs) derived from the sMRI and dMRI modalities, while weaker correlations are present for the two fMRI modalities. This could be caused by the use of cortical parcellation schemes to split fMRI information into a series of nodes and edges when creating the fMRI IDPs, which has been found to sometimes obscure biological

Table 3.2: Ensemble single-map network results split by sex and core modality. Novel maps without prior literature correspondence (Appendix A, Tables A.1-A.6) are marked in green. The Weighted MAE allows for easier comparison between studies, as proposed by Cole et al [8]. The age ranges are $45.12 - 82.26$ for female and $45.45 - 82.18$ for male subjects.

Map	Female				Male			
	MAE	R^2	Predicted-Chronological Correlation (r)	Weighted MAE	MAE	R^2	Predicted-Chronological Correlation (r)	Weighted MAE
	sMRI Modalities - Female Subjects				**sMRI Modalities - Male Subjects**			
T1 Nonlinear	2.242	0.853	0.922	0.060	2.468	0.837	0.913	0.068
T1 Linear	2.344	0.838	0.914	0.063	2.519	0.829	0.909	0.069
Jacobian	3.324	0.678	0.824	0.089	3.493	0.663	0.813	0.096
VBM	2.622	0.798	0.893	0.071	2.783	0.790	0.888	0.077
T2 FLAIR Nonlinear	2.189	0.856	0.926	0.059	2.345	0.849	0.920	0.065
T2 Lesions	3.965	0.531	0.731	0.107	4.153	0.538	0.729	0.114
SWI	3.054	0.719	0.847	0.082	3.224	0.701	0.841	0.089
	rsfMRI Modalities - Female Subjects				**rsfMRI Modalities - Male Subjects**			
rsfMRI-0	4.173	0.507	0.702	0.112	4.891	0.392	0.621	0.135
rsfMRI-1	4.335	0.475	0.676	0.117	4.837	0.379	0.614	0.133
rsfMRI-2	4.322	0.467	0.678	0.116	5.081	0.337	0.578	0.140
rsfMRI-3	4.311	0.469	0.679	0.116	5.018	0.346	0.597	0.138
rsfMRI-4	4.278	0.461	0.677	0.115	4.965	0.362	0.600	0.137
rsfMRI-5	4.285	0.462	0.676	0.115	4.628	0.416	0.658	0.128
rsfMRI-6	4.443	0.442	0.654	0.120	5.218	0.301	0.546	0.144
rsfMRI-7	4.232	0.481	0.690	0.114	5.314	0.247	0.512	0.146
rsfMRI-8	4.401	0.448	0.664	0.118	4.965	0.367	0.600	0.137
rsfMRI-9	4.237	0.478	0.690	0.114	4.876	0.375	0.613	0.134
rsfMRI-10	4.357	0.458	0.667	0.117	5.019	0.341	0.588	0.138
rsfMRI-11	4.221	0.491	0.696	0.114	4.968	0.363	0.600	0.137
rsfMRI-12	4.261	0.470	0.691	0.115	4.962	0.371	0.604	0.137
rsfMRI-13	4.309	0.471	0.678	0.116	4.993	0.360	0.596	0.138
rsfMRI-14	4.368	0.450	0.672	0.118	5.147	0.320	0.565	0.142
rsfMRI-15	4.272	0.464	0.682	0.115	5.124	0.346	0.578	0.141
rsfMRI-16	4.302	0.464	0.680	0.116	4.699	0.414	0.636	0.130
rsfMRI-17	4.197	0.486	0.696	0.113	4.599	0.430	0.662	0.127
rsfMRI-18	4.244	0.484	0.686	0.114	4.744	0.426	0.654	0.131
rsfMRI-19	4.416	0.432	0.660	0.119	5.007	0.356	0.593	0.138
rsfMRI-20	4.290	0.454	0.674	0.116	5.022	0.329	0.579	0.138
rsfMRI-21	4.233	0.483	0.691	0.114	5.026	0.343	0.584	0.139
rsfMRI-22	4.198	0.490	0.696	0.113	5.240	0.306	0.531	0.144
rsfMRI-23	4.350	0.463	0.675	0.117	5.045	0.341	0.574	0.139
rsfMRI-24	4.304	0.458	0.681	0.116	5.007	0.345	0.583	0.138
	tfMRI Modalities - Female Subjects				**tfMRI Modalities - Male Subjects**			
tfMRI-1	3.444	0.651	0.805	0.093	3.663	0.627	0.795	0.101
tfMRI-2	3.616	0.633	0.788	0.097	3.765	0.608	0.784	0.104
tfMRI-5	3.693	0.617	0.777	0.099	3.900	0.593	0.768	0.107
tfMRI-COPE-1	3.483	0.647	0.801	0.094	3.574	0.642	0.805	0.099
tfMRI-COPE-2	3.532	0.636	0.796	0.095	3.666	0.627	0.796	0.101
tfMRI-COPE-5	3.549	0.637	0.798	0.096	3.643	0.631	0.796	0.100
	dMRI Modalities - Female Subjects				**dMRI Modalities - Male Subjects**			
Summed Tracts	3.177	0.719	0.839	0.086	3.191	0.714	0.847	0.088
TBSS FA	2.939	0.743	0.862	0.079	3.107	0.734	0.857	0.086
TBSS ICVF	2.872	0.757	0.871	0.077	3.053	0.749	0.863	0.084
TBSS ISOVF	3.331	0.684	0.823	0.090	3.335	0.699	0.837	0.092
TBSS L1	3.088	0.725	0.847	0.083	3.138	0.724	0.853	0.086
TBSS L2	3.100	0.722	0.847	0.083	3.312	0.704	0.834	0.091
TBSS L3	3.108	0.716	0.843	0.084	3.325	0.702	0.834	0.092
TBSS MD	3.115	0.720	0.845	0.084	3.294	0.704	0.839	0.091
TBSS MO	3.336	0.683	0.824	0.090	3.433	0.687	0.828	0.095
TBSS OD	3.114	0.719	0.846	0.084	3.096	0.730	0.856	0.085
FA	2.668	0.788	0.887	0.072	2.799	0.786	0.886	0.077
ICVF	2.631	0.794	0.891	0.071	2.694	0.795	0.892	0.074
ISOVF	2.822	0.773	0.874	0.076	2.875	0.769	0.878	0.079
L1	2.849	0.766	0.872	0.077	2.867	0.768	0.876	0.079
L2	2.844	0.767	0.872	0.077	3.052	0.745	0.864	0.084
L3	2.911	0.761	0.867	0.078	3.024	0.749	0.865	0.083
MD	2.887	0.762	0.868	0.078	2.978	0.758	0.868	0.082
MO	2.953	0.746	0.860	0.080	3.083	0.737	0.857	0.085
OD	2.835	0.770	0.874	0.076	2.758	0.783	0.885	0.076



(a) MAE predictions for female subjects



(b) MAE predictions for male subjects

Figure 3.13: Mean absolute error (MAE) distributions for single-map predictions obtained with CNNs (blue) and IDP-based regression (magenta), for female **(a)** and male **(b)** subjects. The failure case in which the models predict only a population mean age is also plotted for comparison.

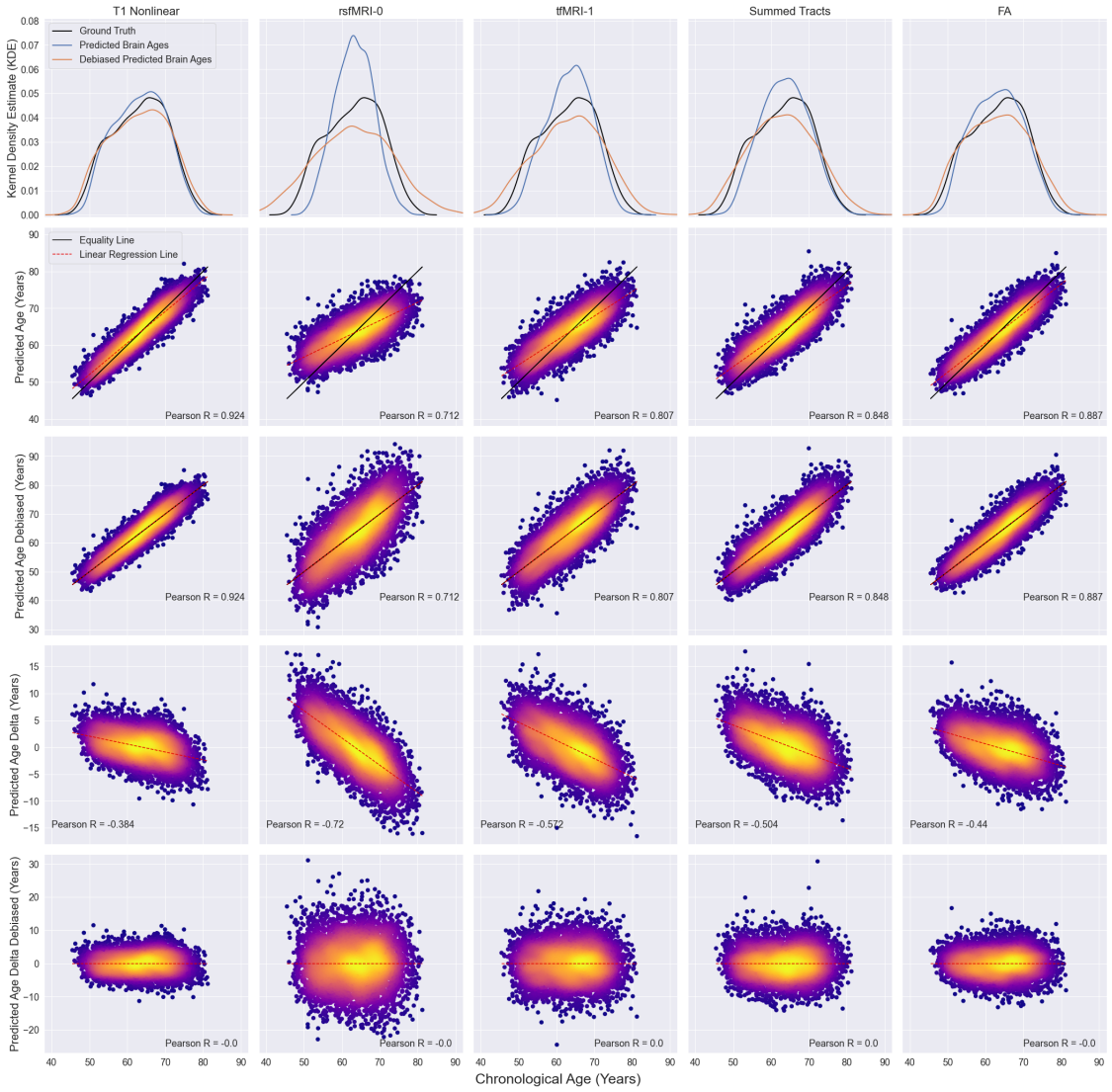


Figure 3.14: Kernel density estimates (KDE) and density plots (DP) of CNN predicted ages for the female dataset. Each column of the figure corresponds to a map, and each row to a plot with respect to chronological age: KDE of predicted and debiased predicted ages distributions for the left-out subjects, DPs of predicted age and predicted age linearly debiased, and DPs of predicted age deltas and predicted age deltas debiased.

Table 3.3: IDP-based single-map regression results split by sex and core modality. The number in brackets following each map represents the number of PCA components used for dimensionality reduction, with 0 representing a case where no reduction was used. The Weighted MAE allows for easier comparison between studies, as proposed by Cole et al [8]. The age ranges are $45.12 - 82.26$ for female and $45.45 - 82.18$ for male subjects.

Map (PCA Components)	Female				Male			
	MAE	R^2	Predicted-Chronological Correlation (r)	Weighted MAE	MAE	R^2	Predicted-Chronological Correlation (r)	Weighted MAE
All IDPs (1000)	3.944	0.542	0.736	0.106	4.129	0.535	0.731	0.114
	sMRI IDPs - Female Subjects				**sMRI IDPs - Male Subjects**			
Cortical Area (0)	5.278	0.204	0.451	0.142	5.385	0.245	0.495	0.148
Cortical Grey-White Map (0)	5.063	0.263	0.513	0.136	5.126	0.312	0.558	0.141
Cortical Thickness (0)	5.126	0.248	0.498	0.138	5.238	0.274	0.523	0.144
Regional T2* (0)	5.940	0.037	0.194	0.160	6.081	0.067	0.258	0.168
Regional and Tissue Intensity (0)	4.824	0.324	0.569	0.130	4.907	0.364	0.604	0.135
Regional and Tissue Volume (0)	4.181	0.491	0.701	0.113	4.261	0.501	0.708	0.117
WM Hyperintensity Volume (2)	5.389	0.185	0.430	0.145	5.759	0.158	0.398	0.159
	rsfMRI IDPs - Female Subjects				**rsfMRI - Male Subjects**			
rfMRI Connectivity (1000)	4.988	0.296	0.544	0.134	5.299	0.27	0.519	0.146
rfMRI Node Amplitude (0)	5.459	0.177	0.421	0.147	5.750	0.156	0.395	0.158
	tfMRI IDPs - Female Subjects				**tfMRI - Male Subjects**			
tfMRI Activation (0)	6.034	0.019	0.138	0.162	6.230	0.038	0.196	0.172
	dMRI IDPs - Female Subjects				**dMRI IDPs - Female Subjects**			
WM Tract Diffusivity (0)	4.457	0.417	0.646	0.120	4.595	0.43	0.655	0.127
WM Tract FA (0)	4.911	0.304	0.552	0.132	5.168	0.3	0.547	0.142
WM Tract ICVF (0)	5.006	0.275	0.524	0.135	5.021	0.327	0.571	0.138
WM Tract ISOVF (0)	4.992	0.284	0.533	0.134	5.140	0.303	0.550	0.142
WM Tract MO (0)	5.107	0.252	0.502	0.138	5.205	0.288	0.536	0.143
WM Tract OD (0)	5.088	0.262	0.512	0.137	5.194	0.293	0.541	0.143

details [275, 276]. The importance of spatial information is also revealed when observing that the volumetric T2 Lesions map, despite being a binary mask of white matter hyperintensities (WMH), performs better than the equivalent IDP (WMH Volumes). WMHs tend to increase with a subject's age [20, 183], suggesting their volume could be an accurate predictor of brain age. Yet, this finding suggests that spatial information plays a part in accurate brain age prediction.

Brain age deltas, following debiasing and deconfounding, were also found to form clusters according to their core modality, for both males and females (Figures 3.16a for HGL and 3.18a for the linear method). For volumetric data, an additional cluster is present between sMRI and dMRI maps, which could be due to the sharing of certain underlying structural information between maps derived from the two modalities. These observations are complemented by those seen in Figure 3.17, which shows the results of a PCA decomposition applied to the matrix of predicted brain age deltas [277]. The first two principal components explain $\approx 83\%$ of the total variance in the dataset. The 2D PCA projection of the first two principal components reveals the existence of three clusters: one defined by the rsfMRI maps,

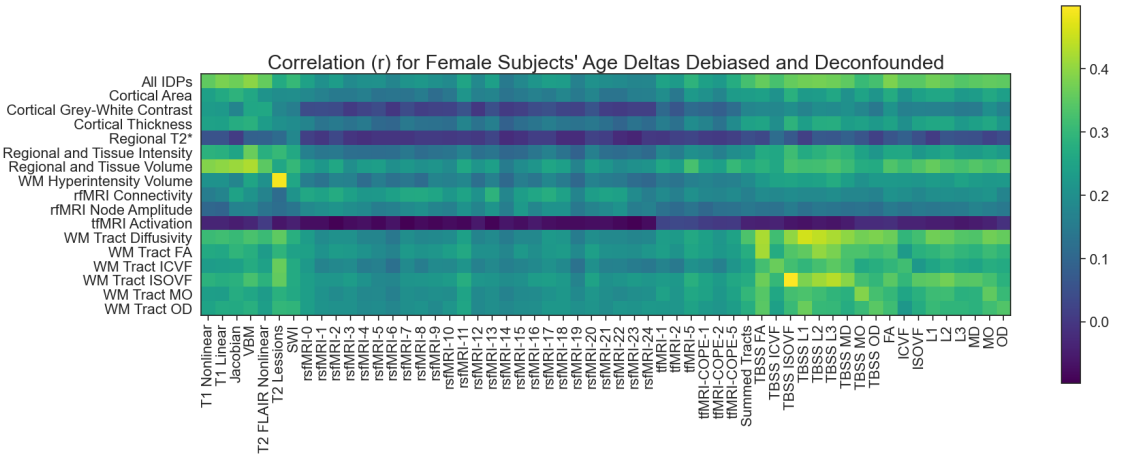


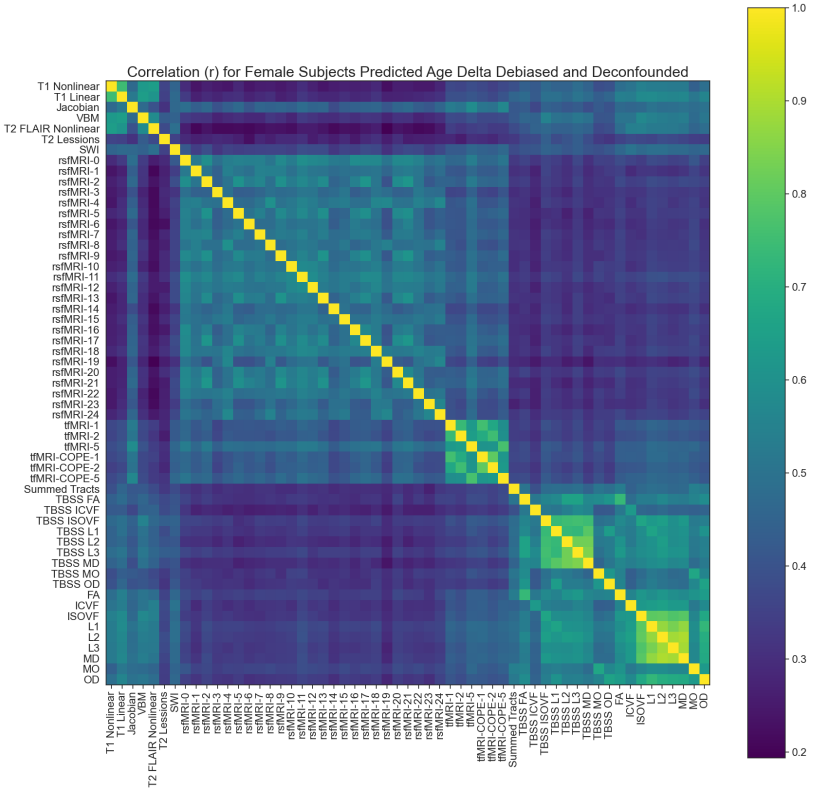
Figure 3.15: The relationship between CNN and IDP-regression brain age deltas, for female subjects, calculated using the Pearson (r) correlation. The results for male subjects are similar.

another by the sMRI and dMRI maps, and a third one containing the tfMRI and the Jacobian maps, with the SWI and T2 Lesions maps not belonging to any clusters.

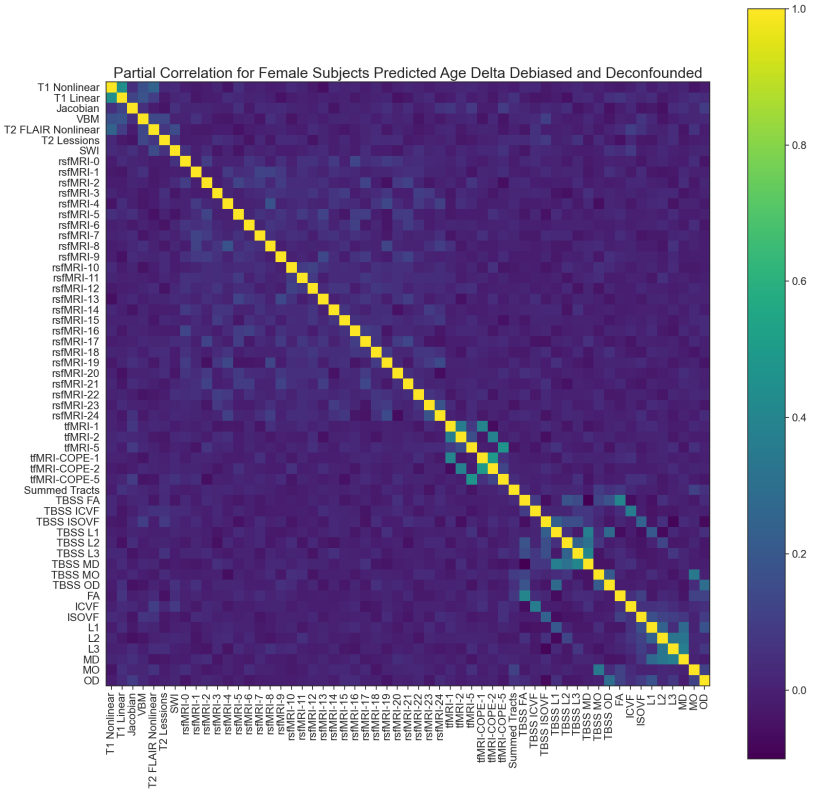
The clusters observed in the correlation matrices and PCA suggest that a certain degree of shared underlying information exists between the various maps. To determine which maps contain relatively unique information, partial correlations were also computed for both the volumetric maps and the IDPs (Figures 3.16b and 3.18b). These reveal that only a small number of map pairs share some underlying information. These include those which have a common root such as T1 Linear and T1 Nonlinear map, the skeletonised and non-skeletonised dMRI maps, and the z-stat and COPE tfMRI maps. Similar observations can be made for the IDPs, with underlying shared information existing between the model trained with all IDPs and the other models as well as between the predictions based on the rsfMRI and white matter (WM) IDP groups.

3.5.3 Discussion

In this section, the HGL CNN model and data from UK Biobank were used to predict brain ages from 57 different 3D maps derived from 5 core MRI modalities. sMRI derived maps, which have been traditionally utilised in brain age investigations, obtained the best predictive performances, with T2 FLAIR Nonlinear maps

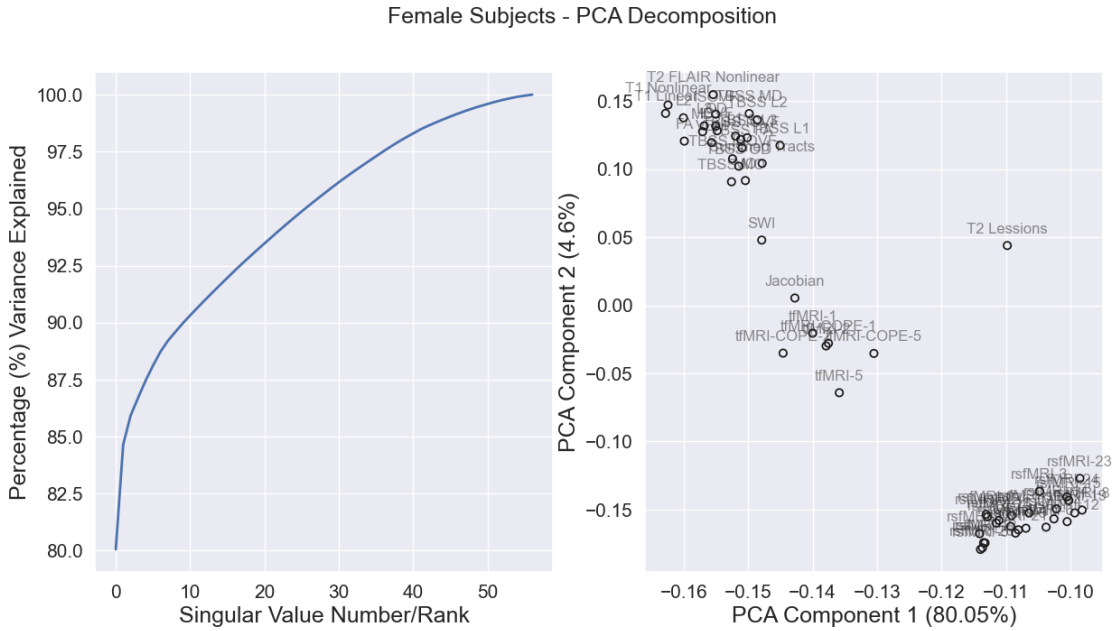


(a) Correlation matrix for female subjects

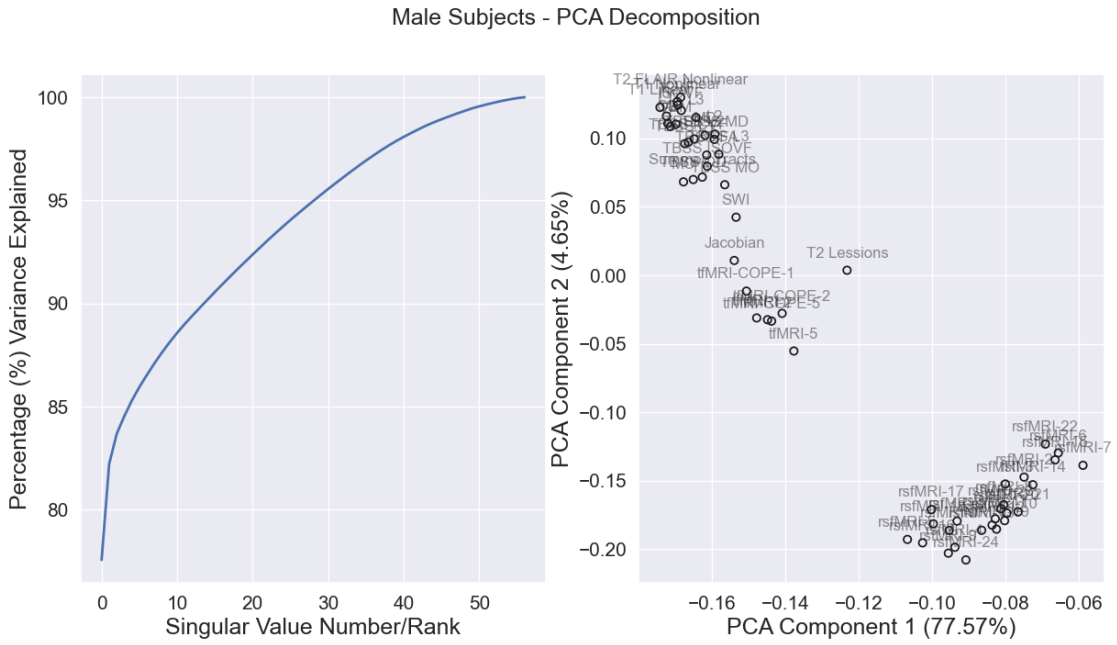


(b) Partial correlation matrix for female subjects

Figure 3.16: The relationship between CNN-predicted brain age deltas for the utilised maps, for female subjects, **(a)** calculated using Pearson (r) correlation and **(b)** partial correlation. The results for male subjects are similar.

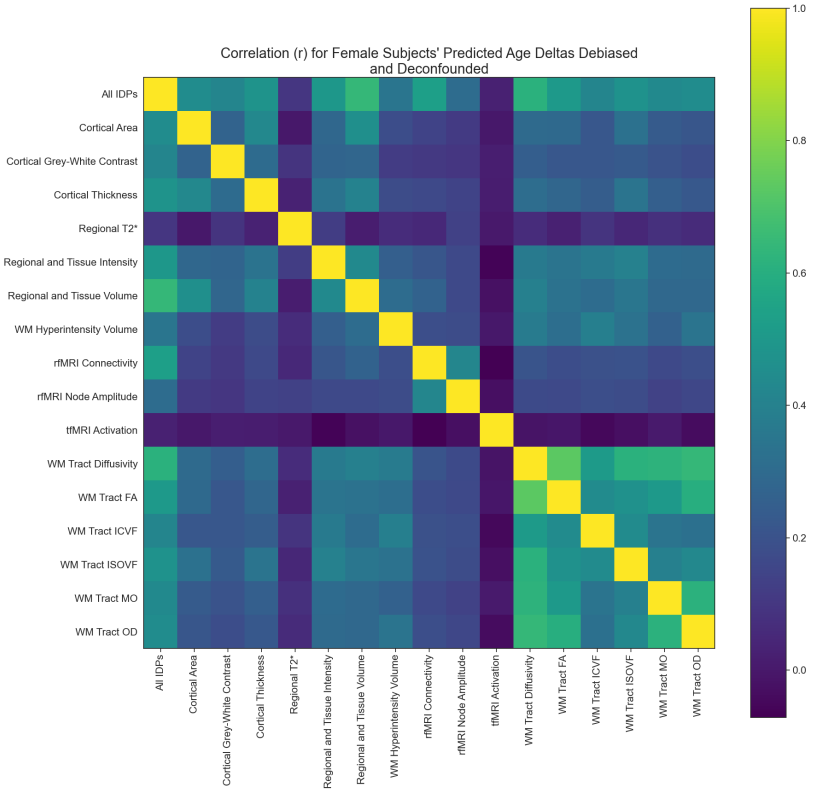


(a) PCA decomposition of brain age deltas for female subjects group

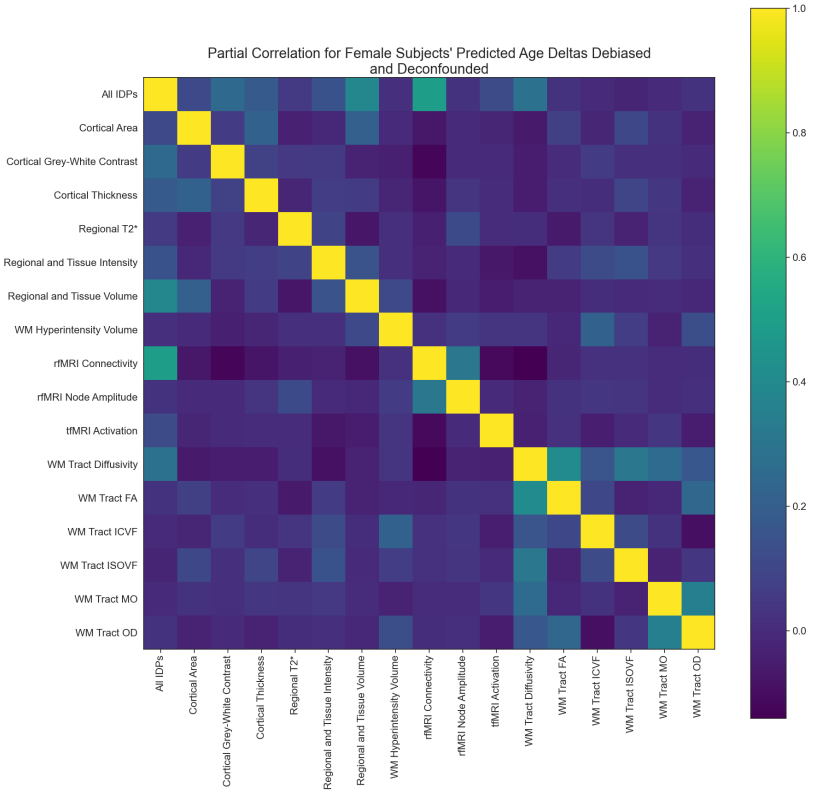


(b) PCA decomposition of brain age deltas for the male subjects group

Figure 3.17: PCA decompositions of predicted subject brain age deltas by maps for **(a)** female and **(b)** male subjects. The left-hand plots show the percentage variance explained by the various components, while the right side plots project the maps in the plane described by the first two principal components.



(a) Correlation matrix for female subjects



(b) Partial correlation matrix for female subjects

Figure 3.18: The relationship between IDP-regression-predicted brain age deltas for the utilised maps, for female subjects, **(a)** calculated using Pearson (r) correlation, and **(b)** partial correlation. The results for male subjects are similar.

producing the best single-map MAEs: 2.186 years for female subjects and 2.345 years for male subjects respectively. These, and the MAEs of other sMRI maps, are in line with those reported in literature for similar age ranges ($45 - 82$ years) [9, 21, 53, 67, 250, 271]. These findings validate the initial hypothesis that other 3D maps carry age-associated characteristics and information, thereby making it possible to use them to deduce either acceleration or resilience to brain ageing.

The results described in this section also confirmed the hypothesis that volumetric data contains spatial information relevant for accurately capturing brain ageing. As illustrated in Figure 3.15, although moderately strong correlations exist between predictions derived from maps and IDP groups with similar roots (such as T2 Lesions and WM Hyperintensity Volume, TBSS ISOVF and WM Tract ISOVF, and TBSS L1 and WM Tract Diffusivity), volumetric data typically yields better prediction accuracies. This has the potential to open new avenues into more bespoke investigations of the ageing brain, circumventing the necessity for the additional assumptions needed for preprocessing image data to extract certain derived phenotypes.

An interesting observation in this section arises from the patterns shown across different maps based on their brain age deltas (Figures 3.16b and 3.17a). While clusters form indicating relationships among maps, only a few of them display a direct or unique relationship when accounting for the influence of other maps, as evidenced in the partial correlation matrix (Figure 3.16b). This observation suggests that different maps, especially those from different core modalities, may encode distinct information pertaining to brain ageing. For maps derived from the same core modality, it remains uncertain whether the unique relationships observed are genuine or merely results of stochasticity. Building upon this insight, the subsequent section will explore further the hypothesis that each map encodes unique information related to ageing.

3.6 Determining if Maps Encode Information Relating to Different Mechanisms of Brain Ageing

In the preceding section, the HGL model was demonstrated to effectively predict brain ages across a multitude of 3D MRI maps, supporting the assertion that each map carries information relating to ageing. A crucial observation was that only a fraction of the maps shared underlying information. This suggests that different maps could encode unique aspects of brain ageing, which was particularly evident for maps derived from different modalities. However, this is not yet clear for map groups belonging to the same core modality. Thus, this section delves further into answering the question of whether different maps do indeed encode distinct information related to the ageing brain. This is accomplished through the correlation analyses described in Section 3.3.4, by looking at the associations between brain age deltas and 13809 non-imaging measurements (nIDPs) from UK Biobank.

3.6.1 Results

Associations were mapped for each map, and for male and female subjects separately. A subset of the relationships established in this way is presented in Figures 3.19-3.20 for female subjects, and Figures 3.21-3.22 for male subjects. Each Manhattan plot in these figures indicates the statistical significance of the correlations between one nIDP and the brain age deltas derived from a particular map. Two tests of significance were included in each plot: the False Discovery Rate (FDR), calculated independently for each map, and the Bonferroni (Bonf.) Threshold. Those associations whose $-log(p)$ value is greater than the Bonferroni Threshold are considered to be statistically significant. These figures are supplemented by Figures 3.23 and 3.24, which show the proportion of nIDPs passing the Bonferroni threshold for each nIDP category for female and male subjects.

As presenting all the association studies in this thesis would be infeasible, the full correlation results between all maps with all nIDPs can be accessed freely on *Zenodo* or by going to this URL: `https://doi.org/10.5281/zenodo.8110876`.

The variable descriptions and IDs are identical to those utilised in UK Biobank. More information on them can be found by accessing the *UK Biobank website* or by going to this URL: `https://biobank.ctsu.ox.ac.uk/showcase/search.cgi`.

When analysing the results, three sets of differences can be observed:

- Between maps within each sex.
- For maps derived from the same modality.
- For the same map between sexes.

Firstly, differences can be observed between the distinct map results within each sex. For instance, for female subjects, all plots show significant correlations with Skeletal Measurements, with the exception of T2 Lesions (Figure 3.19c). Another example can be seen in Figure 3.23, with sMRI and swMRI maps not having any significant hits for nIDP categories such as Alcohol, Blood Assays and Physical Measurements, compared to all other groups which do. Similar observations can be made for the male group. On closer inspection, traditional sMRI-maps, typically utilised in brain age prediction studies (T1 Linear and Nonlinear, and T2 FLAIR Nonlinear), account for 51 significant hits (49 for the female group and 2 for the male group), out of a total of 191 observed significant associations across both subject groups. Thus, 140 represent correlations which would not have been discovered by solely using traditional maps. In addition, the sMRI maps only correlated significantly with variables from the Medical History and Skeletal Measurements categories, with variables from the Alcohol, Blood Assays, Cardiac & Circulatory Measurements, Physical Measurements, Diet and Tobacco correlating only with the non-traditional maps.

Differences can also be observed for correlations obtained with maps derived from the same modality. For instance, in the female subject group, rsfMRI-0 had 1 association passing the Bonferroni threshold for Alcohol consumption and none passing it for Physical Measurements, in opposition to rsfMRI-3 (Figures 3.19e-3.19f). This difference could be caused by the fact that each of the rsfMRI maps

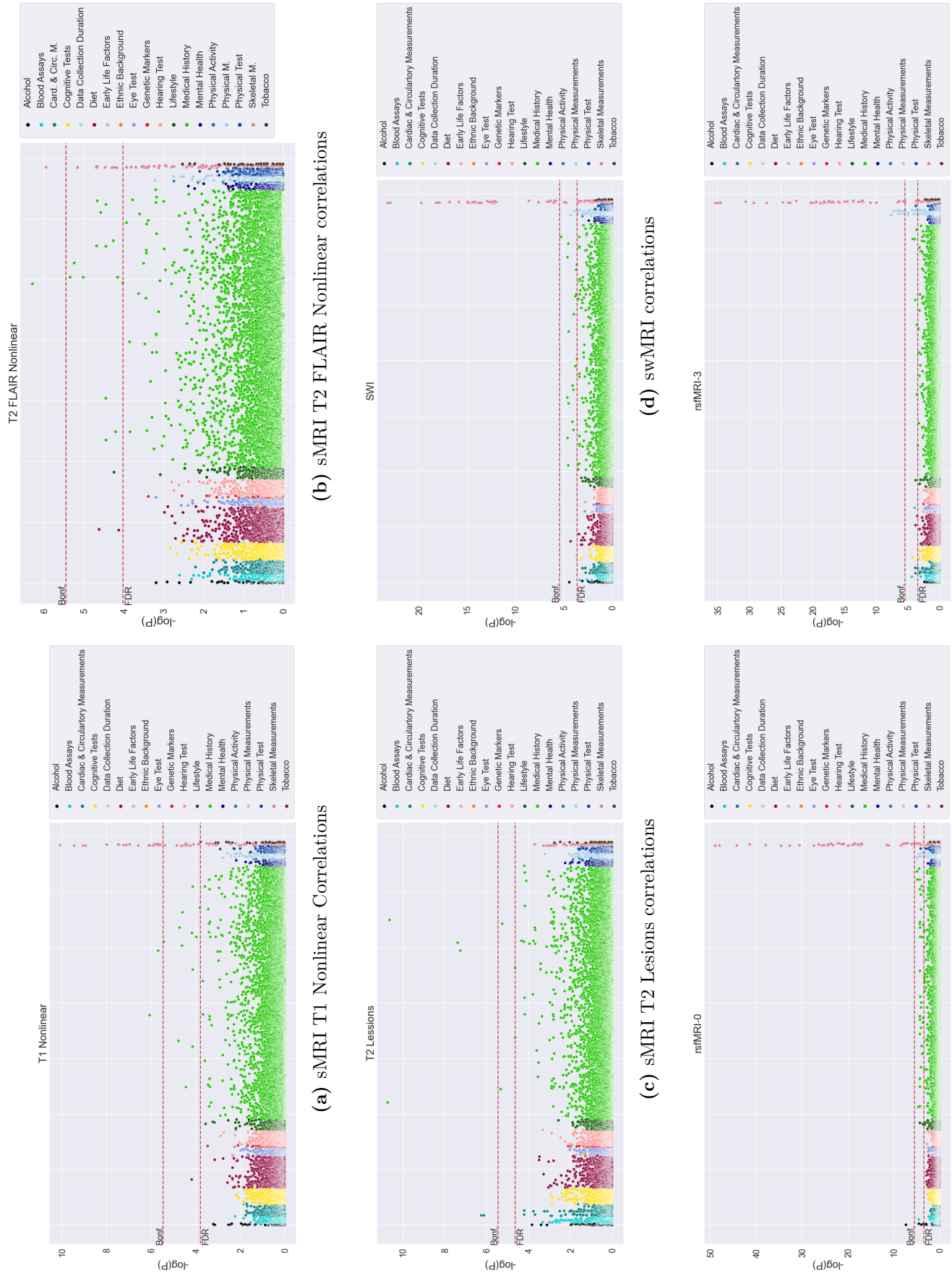
represents a major resting-state network, with rsfMRI-0 generally being associated with the Default Mode Network (DMN), and rsfMRI-3 corresponding to the network group processing visual information [41, 224]. A similar case is observed for FA and TBSS FA, where the latter has no correlations passing Bonferroni in the Cardiac & Circulatory Measurements but does for Physical Measurements, while the former is the exact opposite, despite the Skeletonised TBSS FA map being derived from the FA map (Figures 3.20e-3.20f). These differences between maps derived from the same core modalities can be observed even more clearly in Figures 3.23-3.24. As mentioned in the previous section, some of these differences could be due to inherent network stochasticity which is translated into the brain age predictions. This phenomenon will be explored further in Chapter 4 Section 4.6.1.1.

Finally, as suggested by literature [21, 35, 56, 58, 132, 139], differences can also be observed between sexes. The largest difference can be observed in the absence of any significant hits in the Skeletal Measurements category between the female and male subject groups. Other differences can be observed, for instance, in the case of the Summed Tracts (Figures 3.20b-3.22a) and tfMRI derived maps (Figures 3.23b-3.24b), where numerous associations pass the Bonferroni threshold for the female subjects, but few-to-none do so for the male subjects.

The strongest associations for each variable category are presented in Tables 3.4-3.5. Given the large number of associations passing the Bonferroni threshold for some variable categories, only the top-10 map-variable associations for each category are presented. If a variable is positively correlated with brain age deltas, indicating an accelerated ageing process, it is marked in red, while negatively correlated variables, suggesting a resilience to ageing mechanisms, are left in black.

3.6.2 Discussion

To verify whether maps encode unique information relating to brain ageing acceleration or resilience to ageing, UK Biobank nIDPs were correlated with the predicted brain age deltas of male and female test subjects.



(a) sMRI T1 Nonlinear Correlations

(b) sMRI T2 FLAIR Nonlinear correlations

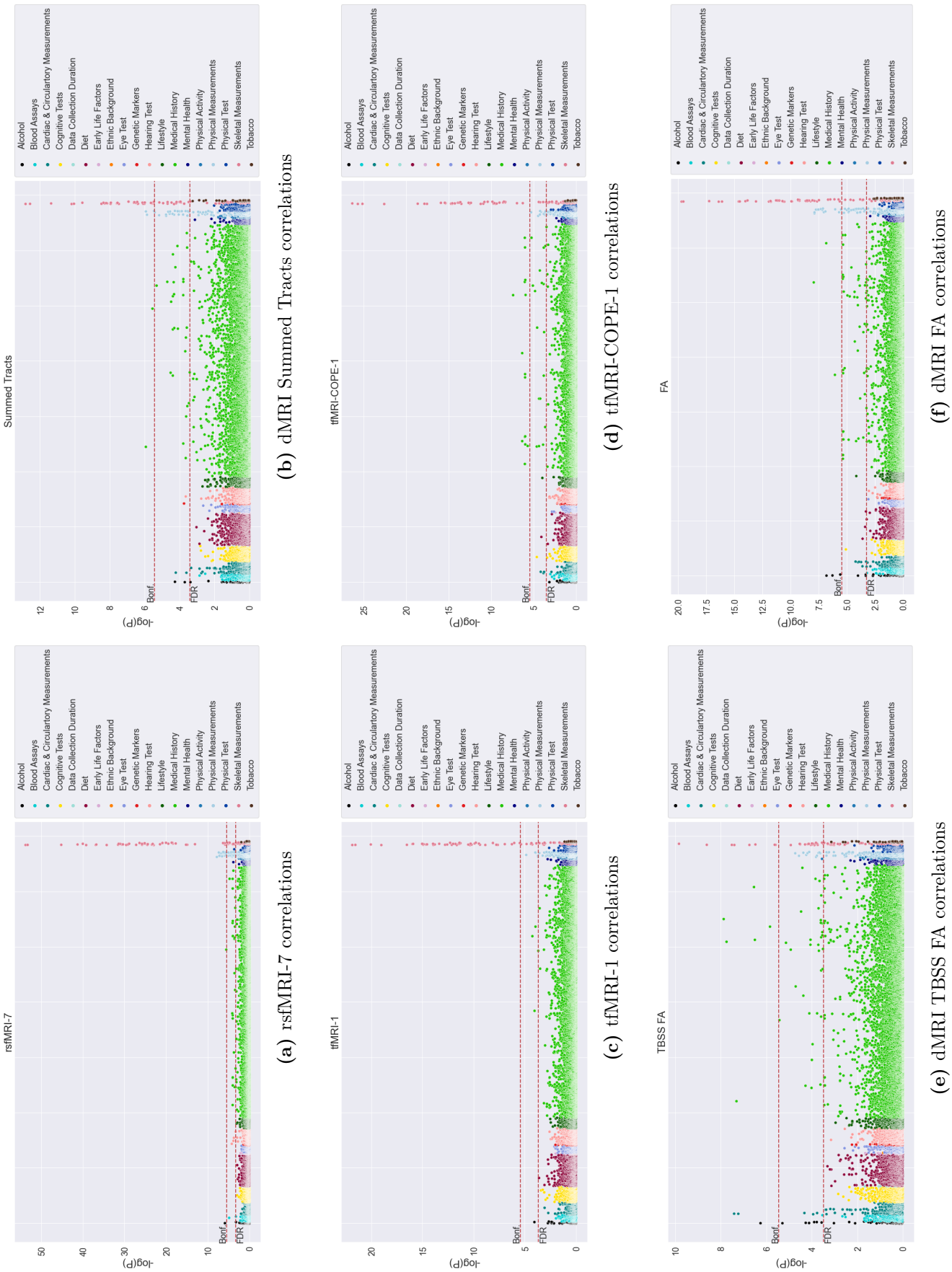
(c) sMRI T2 Lesions correlations

(d) swMRI correlations

(e) rsfMRI-0 correlations

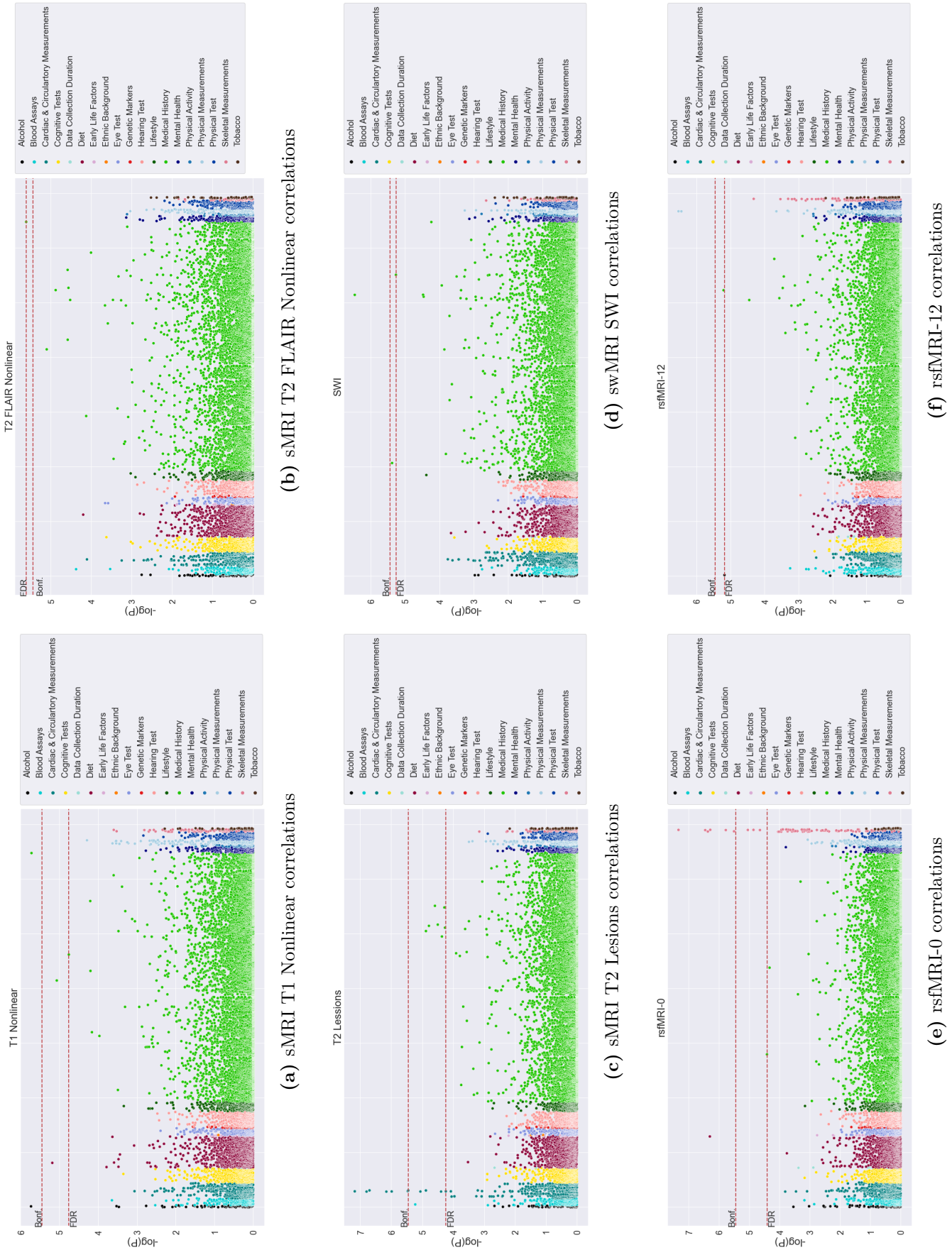
(f) rsfMRI-3 correlations

Figure 3.19: Manhattan plots relating brain age deltas to UK Biobank nIDPs for female subjects (1/2) for a subset of maps, with each dot representing the statistical significance of the correlation. The False Discovery Rate (FDR) and Bonferroni thresholds are also plotted.



(a) rsfMRI-7 correlations (b) dMRI Summed Tracts correlations (c) tfMRI-1 correlations (d) tfMRI-COPE-1 correlations (e) dMRI TBSS FA correlations (f) dMRI FA correlations

Figure 3.20: Manhattan plots relating brain age deltas to UK Biobank nIDPs for female subjects (2/2) for a subset of maps, with each dot representing the statistical significance of the correlation. The False Discovery Rate (FDR) and Bonferroni thresholds are also plotted.



(a) sMRI T1 Nonlinear correlations

(b) sMRI T2 FLAIR Nonlinear correlations

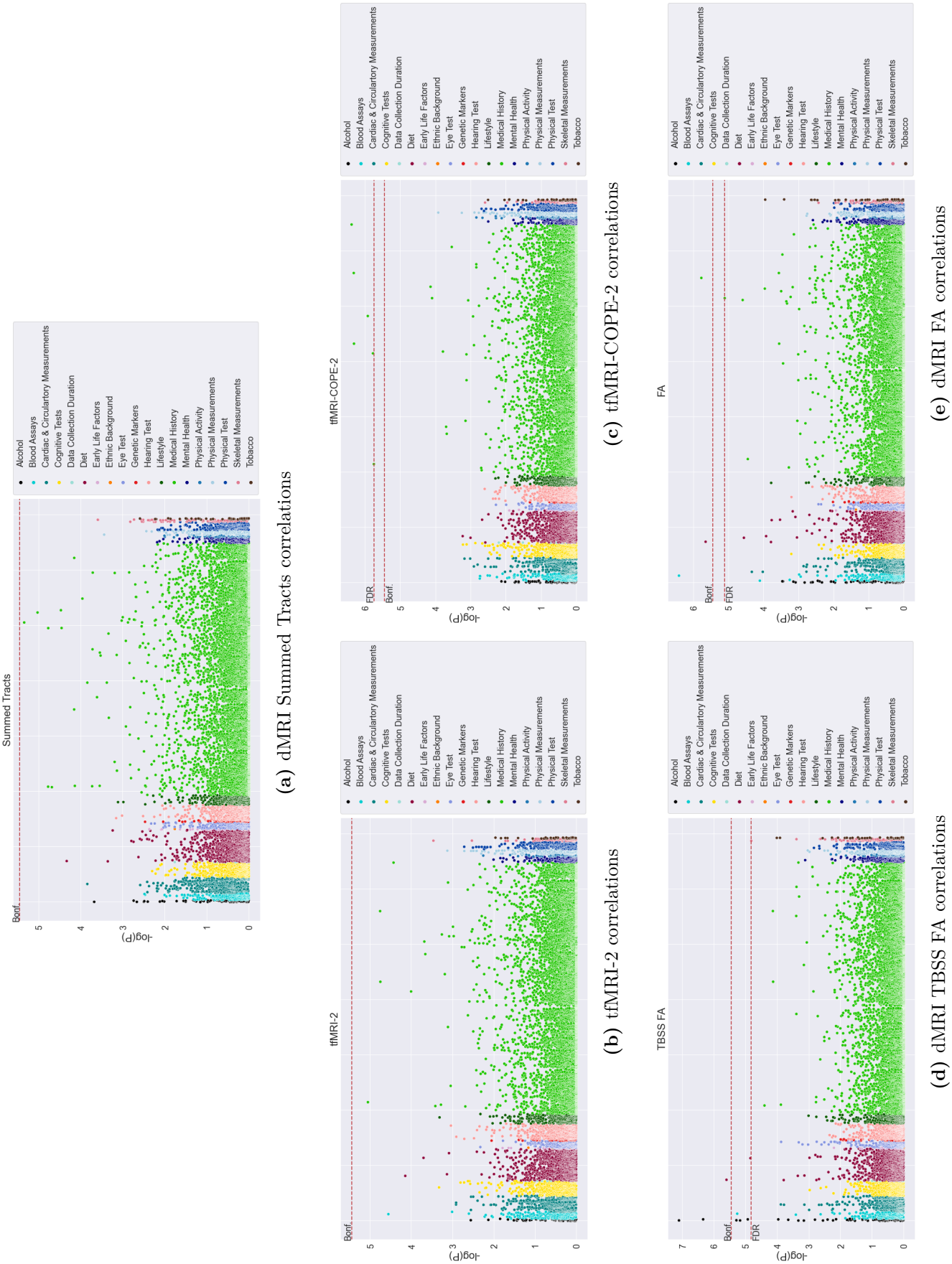
(c) sMRI T2 Lesions correlations

(d) swMRI SWI correlations

(e) rsfMRI-0 correlations

(f) rsfMRI-12 correlations

Figure 3.21: Manhattan plots relating brain age deltas to UK Biobank nIDPs for male subjects (1/2) for a subset of maps, with each dot representing the statistical significance of the correlation. The False Discovery Rate (FDR) and Bonferroni thresholds are also plotted.



(a) dMRI Summed Tracts correlations

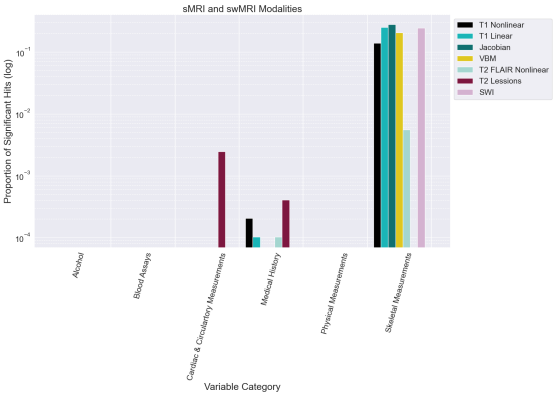
(b) tfMRI-2 correlations

(c) tfMRI-COPE-2 correlations

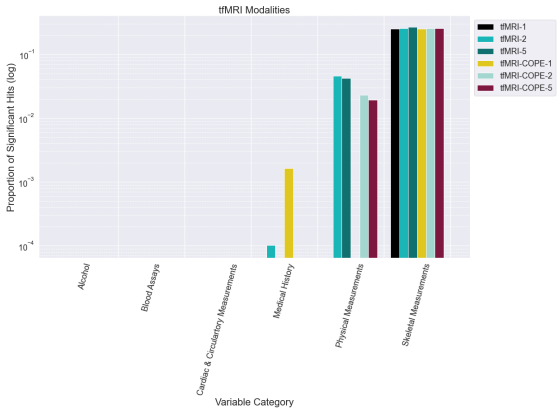
(d) dMRI TBSS FA correlations

(e) dMRI FA correlations

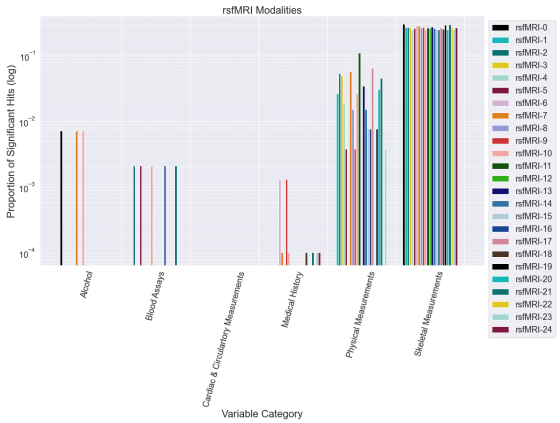
Figure 3.22: Manhattan plots relating brain age deltas to UK Biobank nIDPs for male subjects (2/2) for a subset of maps, with each dot representing the statistical significance of the correlation. The False Discovery Rate (FDR) and Bonferroni thresholds are also plotted.



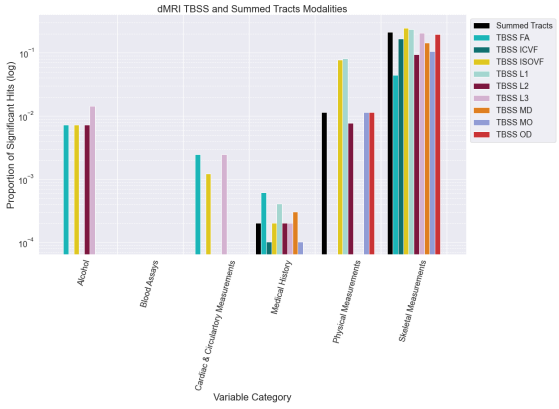
(a) sMRI and swMRI maps



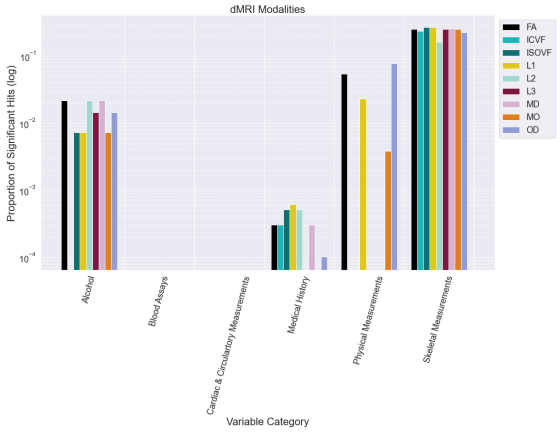
(b) tfMRI maps



(c) rsfMRI maps

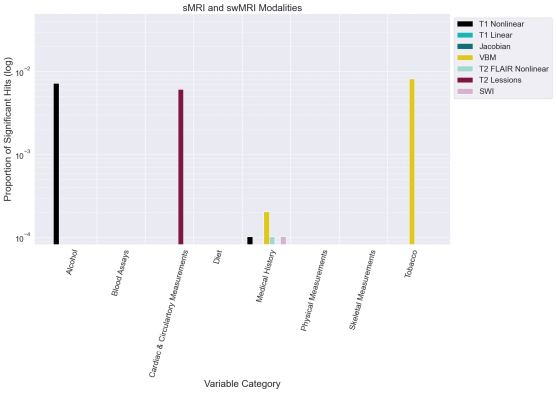


(d) dMRI maps, skeletonised

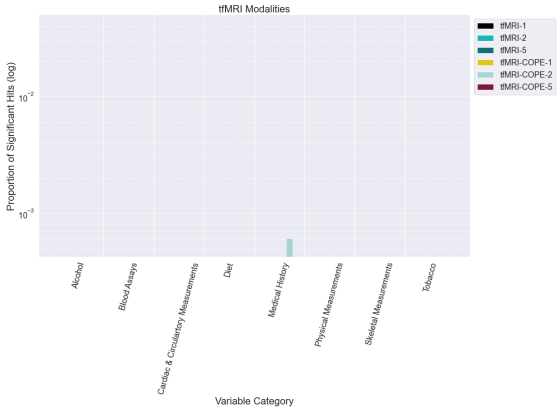


(e) dMRI maps

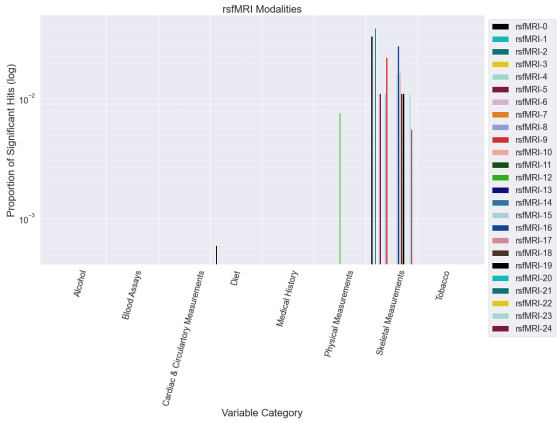
Figure 3.23: Proportion of significant hits per UK Biobank nIDP category for the female subjects group, calculated by dividing the number of variables passing the Bonferroni threshold by the total number of variables in that category.



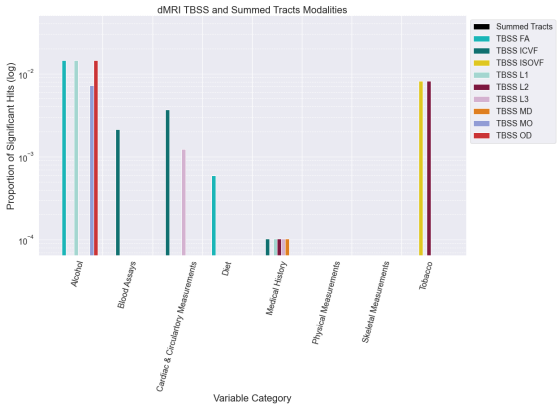
(a) sMRI and swMRI maps



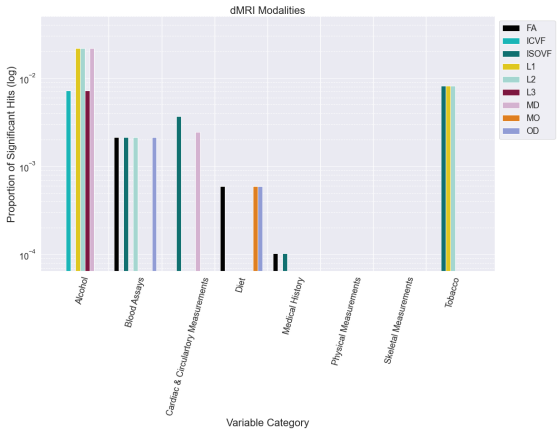
(b) tfMRI maps



(c) rsfMRI maps



(d) dMRI maps, skeletonised



(e) dMRI maps

Figure 3.24: Proportion of significant hits per UK Biobank nIDP category for the male subjects group, calculated by dividing the number of variables passing the Bonferroni threshold by the total number of variables in that category.

Table 3.4: Strongest associations between UK Biobank nIDPs and brain age deltas for female subjects, with *positive correlations* suggesting accelerated brain ageing. For brevity, the following abbreviations have been used: M. for Measurements, and Card. & Circ. for Cardiac & Circulatory.

Map	-log(p)	Correlation (Pearson r)	Variable Category	Variable Description	Variable ID
MD	7.802	0.139	Alcohol	*Frequency of consuming six or more units of alcohol (0.0)*	20416-0.0
MD	7.338	0.129	Alcohol	*Frequency of drinking alcohol (0.0)*	20414-0.0
rsfMRI-0	7.321	0.129	Alcohol	*Frequency of drinking alcohol (0.0)*	20414-0.0
MO	7.18	0.127	Alcohol	*Frequency of drinking alcohol (0.0)*	20414-0.0
L2	7.152	0.133	Alcohol	*Frequency of consuming six or more units of alcohol (0.0)*	20416-0.0
TBSS ISOVF	7.14	0.127	Alcohol	*Frequency of drinking alcohol (0.0)*	20414-0.0
L2	7.125	0.127	Alcohol	*Frequency of drinking alcohol (0.0)*	20414-0.0
L2	6.991	0.131	Alcohol	*Amount of alcohol drunk on a typical drinking day (0.0)*	20403-0.0
FA	6.799	0.124	Alcohol	*Frequency of drinking alcohol (0.0)*	20414-0.0
TBSS L3	6.563	0.121	Alcohol	*Frequency of drinking alcohol (0.0)*	20414-0.0
rsfMRI-2	6.825	-0.108	Blood Assays	C-reactive protein (0.0)	30710-0.0
rsfMRI-10	6.739	-0.108	Blood Assays	C-reactive protein (0.0)	30710-0.0
rsfMRI-5	6.716	-0.107	Blood Assays	C-reactive protein (0.0)	30710-0.0
rsfMRI-21	5.982	-0.101	Blood Assays	C-reactive protein (0.0)	30710-0.0
rsfMRI-16	5.683	-0.098	Blood Assays	C-reactive protein (0.0)	30710-0.0
TBSS FA	7.409	0.113	Card. & Circ. M.	*Systolic blood pressure, automated reading (0.0)*	4080-0.0
TBSS FA	7.226	0.113	Card. & Circ. M.	*Systolic blood pressure, automated reading (0.1)*	4080-0.1
TBSS L3	6.914	0.109	Card. & Circ. M.	*Systolic blood pressure, automated reading (0.0)*	4080-0.0
T2 Lesions	6.235	0.104	Card. & Circ. M.	*Systolic blood pressure, automated reading (0.1)*	4080-0.1
T2 Lesions	6.114	0.102	Card. & Circ. M.	*Systolic blood pressure, automated reading (0.0)*	4080-0.0
TBSS ISOVF	5.541	0.103	Card. & Circ. M.	*Cardiac index during PWA (2.0)*	12702-2.0
TBSS L3	5.492	0.097	Card. & Circ. M.	*Systolic blood pressure, automated reading (0.1)*	4080-0.1
T2 Lesions	10.696	0.133	Medical History	*Diagnoses - ICD10 (I10 - I10 Essential (primary) hypertension)*	41270-0.4052
T2 Lesions	10.616	0.133	Medical History	*Diagnoses - secondary ICD10 (I10 - I10 Essential (primary) hypertension)*	41204-0.4052
FA	7.931	0.114	Medical History	*Treatment/medication code (1140868080 - cyclizine)*	20003-0.1140868080
TBSS FA	7.868	0.113	Medical History	*Diagnoses - secondary ICD10 (I10 - I10 Essential (primary) hypertension)*	41204-0.4052
TBSS FA	7.742	0.112	Medical History	*Non-cancer illness code, self-reported (1065 - hypertension)*	20002-0.1065
tfMRI-COPE-1	7.41	0.109	Medical History	*Treatment/medication code (1140872198 - sodium valproate)*	20003-0.1140872198
T2 Lesions	7.36	0.109	Medical History	*Non-cancer illness code, self-reported (1065 - hypertension)*	20002-0.1065
TBSS FA	7.306	0.109	Medical History	*Diagnoses - ICD10 (I10 - I10 Essential (primary) hypertension)*	41270-0.4052
T2 Lesions	7.236	0.108	Medical History	*Non-cancer illness code, self-reported (1261 - multiple sclerosis)*	20002-0.1261
rsfMRI-9	7.233	0.108	Medical History	*Diagnoses - ICD10 (F320 - F32.0 Mild depressive episode)*	41270-0.2912
rsfMRI-11	11.823	-0.149	Physical M.	Legs total mass (2.0)	23277-2.0
rsfMRI-11	11.724	-0.148	Physical M.	Gynoid total mass (2.0)	23265-2.0
rsfMRI-11	11.18	-0.145	Physical M.	Total mass (2.0)	23283-2.0
rsfMRI-11	10.431	-0.132	Physical M.	Weight (pre-imaging) (2.0)	12143-2.0
rsfMRI-17	10.325	-0.131	Physical M.	Weight (pre-imaging) (2.0)	12143-2.0
rsfMRI-11	9.999	-0.131	Physical M.	Whole body fat mass (2.0)	23100-2.0
TBSS ISOVF	9.903	-0.13	Physical M.	Whole body fat mass (2.0)	23100-2.0
TBSS ISOVF	9.528	-0.128	Physical M.	Trunk fat mass (2.0)	23128-2.0
rsfMRI-17	9.523	-0.133	Physical M.	Total mass (2.0)	23283-2.0
TBSS ISOVF	9.485	-0.127	Physical M.	Leg fat mass (right) (2.0)	23112-2.0
rsfMRI-2	62.341	-0.344	Skeletal M.	Total BMD (bone mineral density) T-score (2.0)	23239-2.0
rsfMRI-2	61.69	-0.342	Skeletal M.	Total BMD (bone mineral density) (2.0)	23236-2.0
rsfMRI-21	57.467	-0.331	Skeletal M.	Total BMD (bone mineral density) T-score (2.0)	23239-2.0
rsfMRI-13	57.466	-0.331	Skeletal M.	Total BMD (bone mineral density) T-score (2.0)	23239-2.0
rsfMRI-2	57.24	-0.33	Skeletal M.	Head BMD (bone mineral density) (2.0)	23226-2.0
rsfMRI-21	56.922	-0.329	Skeletal M.	Total BMD (bone mineral density) (2.0)	23236-2.0
rsfMRI-13	56.517	-0.328	Skeletal M.	Total BMD (bone mineral density) (2.0)	23236-2.0
rsfMRI-9	56.054	-0.327	Skeletal M.	Total BMD (bone mineral density) T-score (2.0)	23239-2.0
rsfMRI-9	55.65	-0.326	Skeletal M.	Total BMD (bone mineral density) (2.0)	23236-2.0
rsfMRI-5	54.407	-0.322	Skeletal M.	Total BMD (bone mineral density) T-score (2.0)	23239-2.0

In total, 191 unique significant correlations passing the Bonferroni threshold were found, spread across 8 of the 19 nIDP categories. 140 of these were achieved with maps other than those traditionally used for brain age investigations (T1 Linear, T1 Nonlinear, T2 FLAIR Nonlinear). Moreover, associations with traditional maps were only found in 2 nIDP categories: Skeletal Measurements and Medical History. The associations with nIDPs from the Alcohol, Blood Assays, Cardiac & Circulatory Measurements, Physical Measurements, Diet and Tobacco categories came from non-traditional maps.

Table 3.5: Strongest associations between UK Biobank nIDPs and brain age deltas for male subjects, with *positive correlations* suggesting accelerated brain ageing. For brevity, the following abbreviations have been used: M. for Measurements, and Card. & Circ. for Cardiac & Circulatory.

Map	-log(p)	Correlation (Pearson r)	Variable Category	Variable Description	Variable ID
TBSS FA	7.099	0.141	Alcohol	*Frequency of drinking alcohol (0.0)*	20414-0.0
TBSS L1	6.717	0.137	Alcohol	*Frequency of drinking alcohol (0.0)*	20414-0.0
TBSS OD	6.706	0.137	Alcohol	*Frequency of drinking alcohol (0.0)*	20414-0.0
MD	6.631	0.136	Alcohol	*Frequency of drinking alcohol (0.0)*	20414-0.0
L1	6.577	0.11	Alcohol	*Alcohol intake frequency. (2.0)*	1558-2.0
TBSS L1	6.439	0.109	Alcohol	*Alcohol intake frequency. (2.0)*	1558-2.0
L1	6.398	0.133	Alcohol	*Frequency of drinking alcohol (0.0)*	20414-0.0
TBSS FA	6.335	0.108	Alcohol	*Alcohol intake frequency. (2.0)*	1558-2.0
MD	6.21	0.107	Alcohol	*Alcohol intake frequency. (0.0)*	1558-0.0
L2	6.209	0.131	Alcohol	*Frequency of drinking alcohol (0.0)*	20414-0.0
OD	7.038	-0.118	Blood Assays	IGF-1 (0.0)	30770-0.0
TBSS ICVF	6.616	-0.114	Blood Assays	IGF-1 (0.0)	30770-0.0
FA	6.411	-0.112	Blood Assays	IGF-1 (0.0)	30770-0.0
ISOVF	5.952	-0.108	Blood Assays	IGF-1 (0.0)	30770-0.0
L2	5.676	-0.105	Blood Assays	IGF-1 (0.0)	30770-0.0
T2 Lesions	7.177	0.122	Card. & Circ. M.	*Central systolic blood pressure during PWA (2.0)*	12677-2.0
T2 Lesions	6.796	0.118	Card. & Circ. M.	*End systolic pressure during PWA (2.0)*	12683-2.0
T2 Lesions	6.701	0.117	Card. & Circ. M.	*Systolic brachial blood pressure (2.0)*	12697-2.0
T2 Lesions	6.156	0.114	Card. & Circ. M.	*End systolic pressure during PWA (2.1)*	12683-2.1
T2 Lesions	5.983	0.11	Card. & Circ. M.	*Mean arterial pressure during PWA (2.0)*	12687-2.0
TBSS ICVF	5.87	0.113	Card. & Circ. M.	*Cardiac index during PWA (2.0)*	12702-2.0
ISOVF	5.839	0.109	Card. & Circ. M.	*Central systolic blood pressure during PWA (2.0)*	12677-2.0
TBSS ICVF	5.76	0.11	Card. & Circ. M.	*Cardiac output during PWA (2.0)*	12682-2.0
MD	5.682	0.107	Card. & Circ. M.	*Central systolic blood pressure during PWA (2.0)*	12677-2.0
MD	5.659	0.107	Card. & Circ. M.	*Mean arterial pressure during PWA (2.0)*	12687-2.0
OD	6.543	0.199	Diet	*Alcohol (2.0)*	100022-2.0
rsfMRI-0	6.301	0.185	Diet	*Leek intake (3.0)*	104230-3.0
MO	5.965	-0.105	Diet	Cereal intake (2.0)	1458-2.0
FA	5.651	0.184	Diet	*Alcohol (2.0)*	100022-2.0
TBSS FA	5.598	0.183	Diet	*Alcohol (2.0)*	100022-2.0
TBSS MD	6.876	0.113	Medical History	*Treatment/medication code (1140860806 - ramipril)*	20003-0.1140860806
SWI	6.465	0.109	Medical History	*Vascular/heart problems diagnosed by doctor (4 - High blood pressure)*	6150-0.4
tfMRI-COPE-2	6.38	-0.108	Medical History	External causes - ICD10 (W014 - W01.4 Street and highway)	41201-0.15435
tfMRI-COPE-2	6.32	-0.108	Medical History	Diagnoses - main ICD10 (S761 - S76.1 Injury of quadriceps muscle and tendon)	41202-0.13440
tfMRI-COPE-2	6.312	-0.107	Medical History	Diagnoses - ICD10 (S761 - S76.1 Injury of quadriceps muscle and tendon)	41270-0.13440
TBSS L2	6.246	0.107	Medical History	*Treatment/medication code (1140860806 - ramipril)*	20003-0.1140860806
TBSS L3	6.187	0.106	Medical History	*Treatment/medication code (1140860806 - ramipril)*	20003-0.1140860806
VBM	5.989	0.105	Medical History	*Diabetes diagnosed by doctor (2.0)*	2443-2.0
tfMRI-COPE-2	5.92	0.104	Medical History	*Antibiotic codes for last 3 months (1141180228 - amoxicillin product)*	20199-0.1141180228
tfMRI-COPE-2	5.773	-0.102	Medical History	Diagnoses - ICD10 (W014 - W01.4 Street and highway)	41270-0.15435
rsfMRI-12	6.526	-0.111	Physical M.	Hand grip strength (right) (2.0)	47-2.0
rsfMRI-12	6.44	-0.11	Physical M.	Hand grip strength (left) (2.0)	46-2.0
rsfMRI-15	8.728	-0.135	Skeletal M.	Head BMC (bone mineral content) (2.0)	23306-2.0
rsfMRI-9	8.633	-0.134	Skeletal M.	Head BMC (bone mineral content) (2.0)	23306-2.0
rsfMRI-16	8.455	-0.133	Skeletal M.	Head BMC (bone mineral content) (2.0)	23306-2.0
rsfMRI-15	7.827	-0.127	Skeletal M.	Head BMD (bone mineral density) (2.0)	23226-2.0
rsfMRI-9	7.751	-0.126	Skeletal M.	Head BMD (bone mineral density) (2.0)	23226-2.0
rsfMRI-2	7.588	-0.125	Skeletal M.	Head BMD (bone mineral density) (2.0)	23226-2.0
rsfMRI-0	7.336	-0.122	Skeletal M.	Head BMD (bone mineral density) (2.0)	23226-2.0
rsfMRI-16	7.255	-0.122	Skeletal M.	Head BMD (bone mineral density) (2.0)	23226-2.0
rsfMRI-19	7.218	-0.121	Skeletal M.	Head BMD (bone mineral density) (2.0)	23226-2.0
rsfMRI-17	6.988	-0.12	Skeletal M.	Head BMC (bone mineral content) (2.0)	23306-2.0
ISOVF	6.651	0.111	Tobacco	*Smoking status (0.0)*	20116-0.0
TBSS L2	6.367	0.108	Tobacco	*Smoking status (0.0)*	20116-0.0
L2	5.945	0.104	Tobacco	*Smoking status (0.0)*	20116-0.0
TBSS ISOVF	5.768	0.102	Tobacco	*Smoking status (0.0)*	20116-0.0
L1	5.617	0.101	Tobacco	*Smoking status (0.0)*	20116-0.0
VBM	5.444	0.099	Tobacco	*Smoking status (0.0)*	20116-0.0

Something which is not immediately obvious from the results observed thus far is how biologically sensible the obtained associations are. Carrying out this type of analysis for all the map-nIDP pairs passing the FDR or Bonferroni thresholds is outside the scope of this work, yet, some consideration still needs to be given to this topic by looking at some of the more statistically significant associations.

For instance, for the female subject group, strong negative correlations were found between Skeletal Measurements, in particular those related to bone mineral

density (BMD) and brain age deltas, suggesting a decelerating effect (Table 3.4). Past studies have found positive correlations between BMD and oestrogen levels in women pre-menopause [278–280], as well as negative correlations between oestrogen levels and inflammation in the body and apparent brain age [141]. Moreover, reductions in BMD have also been associated with an older appearing brain [35]. Thus, it can be seen how BMD can form strong correlations with brain age deltas. Yet, this example serves as a caution, as all the results presented in this work represent correlations, with no causal links being inferred.

In addition to the above example, several other strong association patterns can be found, as shown by Tables 3.4-3.5. Alcohol consumption, for instance, is associated with accelerated brain ageing in both dMRI and rsfMRI maps. Past studies have found associations between alcohol consumption and structural changes in the brain [35, 53, 59, 81, 137, 138, 143–145, 281], but also that alcohol consumption impacts aspects such as blood flow in the brain [282], arterial vessels and blood pressure [283] and the presence of blood clot emboli [284]. These support this study's findings of strong positive correlations between alcohol consumption and dMRI and fMRI maps.

Other strong positive correlations with correspondences in previous brain age literature have been found for systolic blood pressure [21, 35, 56, 81, 138, 139], chronic conditions such as hypertension [21, 137] and diabetes [21, 35, 53, 56, 81, 143], mental health conditions such as the presence of mild depressive episodes, which correlates positively with accelerated ageing in rsfMRI-9 [8, 35, 56, 93, 96, 102, 106–109, 143], neurological problems such as multiple sclerosis [56, 102, 119], the amounts of IGF-1, with alterations to the insulin growth factor-1 leading to structural and functional alterations and ageing in the brain [35, 53, 151], and smoking [35, 53, 56, 81, 143, 146, 147, 285].

Similarly, strong negative correlations have been found for physical measurements, and in particular for higher body or body part mass and fat percentage. The negative correlations are indicative of a resilience to the ageing process, and have correspondence in literature [56]. However, these results are contradicted partially by other findings [81, 100, 150], which observed that obesity correlates positively to

higher brain age deltas. There are also several other results which bring some form of contradiction to previous work, such as the level of C-reactive protein in female subjects, which although being a measure of inflammation in the body [286] correlates negatively with brain age delta. Literature generally suggests that inflammatory processes are typically associated with an older apparent brain [56, 143, 169]. In this case, what is being observed might potentially be influenced, for instance, by anti-inflammatory medication taken by these subjects [21, 135]. This could be an example of "Berkson's bias", a phenomenon often encountered in biomedical observational studies conducted on small patient cohorts, which erroneously infer causal links that cannot be reproduced in the general population [287]. In situations such as this, more in depth, longitudinal investigations are required for determining the causality underpinning the observed map-nIDP associations [288]. This is why the full datasets containing all the observed correlations from this work are made available to the wider community.

All these findings support the hypothesis that some maps encode unique information about brain ageing acceleration and resilience. This potentially opens up the possibility of future bespoke investigations into better understanding the causal links underpinning the observed brain age deltas-nIDP associations. The benefits of understanding these include:

- The possibility of early detection and management of age-related cognitive decline, neurodegenerative diseases and other neurological disorders whose early signs might resemble accelerated ageing;

- The development of preventive therapies and interventions for at-risk populations for accelerated brain ageing and neurodegenerative conditions;

- Identifying traits or genes associated with resilience or accelerated ageing, and linking those to target identification for pharmacological interventions aimed at reducing the risk of accelerated brain ageing and age-associated degeneration;

- The creation of data analysis pipelines for clinical setting deployment, aiding clinicians in their diagnostic work by providing brain age and brain age delta information from multiple MRI modalities;

- Guiding policy-making towards the design of ageing-friendly environments, services and programmes that promote healthy brain ageing and overall better quality of life;

- Using brain ageing models in pharmacological clinical trials at various stages, from the initial recruitment and screening of subjects for any pre-existing conditions, to monitoring for potential off-target drug effects;

The potential use-case concerning the identification of unintended effects of medication is supported by several of the observed significant associations. For instance, the consumption of Cyclizine, an anti-sickness and antihistamine medication, was associated with accelerated brain ageing in dMRI maps. This association is interesting, as preliminary literature searches did not reveal any indications as to the presence of any obvious underlying causal mechanisms. In the case of Cyclizine, one potential explanation might come from the fact that, in clinical practice, it is often prescribed in cases of increased intracranial pressure [289, 290], which could indicate inflammation as the root cause for accelerated brain ageing. Yet, there is no medical consensus in this regard, with recommendations for use being based on clinical experience [291, 292]. In addition, Cyclizine acts by blocking the vestibular pathways [292], with vestibular dysfunctions being associated with accelerated cognitive decline [293] and a degradation of spatial memory and spatial navigation abilities [294]. This could provide an alternative explanation to the positive correlation with accelerated brain ageing.

Several other associations between the consumption of certain medications and accelerated brain ageing were also found. These include, for male subjects, the use of Amoxicillin (positive correlations with task fMRI maps), Ramipril (dMRI), Aspirin (dMRI) and Gliclazide (sMRI and dMRI). For female subjects, significant positive correlations were found for Cyclizine (dMRI) Buprenorphine (task and resting fMRI)

Sodium Valproate (task and resting fMRI) Insulin (dMRI), Codeine phosphate and Kaolin (resting fMRI). Metformin (sMRI, dMRI, fMRI) use showed positive correlations for both sexes. For reference, Ramipril is a drug prescribed for blood pressure, Buprenorphine is an opioid, Sodium Valproate is generally prescribed as an anti-epileptic, Codeine is another opioid for pain, and Gliclazide and Metformin are diabetes medications. Similar associations have been found by previous studies, such as associations between brain age and Ramipril [35] and Aspirin [21].

3.7 Post-Training Ensembling

So far, this work has shown that HGL can predict brain ages using 3D map volumetric information derived from several different MRI modalities as inputs, and that most of these maps encode different information streams relating to the ageing brain. Building on these results, this section seeks to address the question of whether ensembling the predictions from multiple maps leads to better and more informative predictions. The driving hypothesis behind this work is that, through ensembling, complementary between-map features can be learned [81]. By capturing more age-related variance, these ensemble models should result in improved predictions. This section first presents the various post-training ensembling strategies utilised for this work, after which it shows and discusses the obtained results.

3.7.1 Methods: Multimodal Ensembling Techniques

Prior to discussing any potential ensembling technique, the question of what constitutes *"improved"* or *"more informative"* brain age predictions, and how these can be measured, needs to be addressed. Given the methods used in previous sections, these could refer to either improvements in the prediction accuracies, measured as lower mean absolute errors (MAEs), or improvements in ageing-related associations, as measured through correlations with nIDPs. The first explanation assumes that the incorporation of complementary brain age information streams provides a richer, more comprehensive picture of the ageing mechanisms involved, allowing for better prediction accuracies. The latter explanation, on the other hand, is based on the

hypothesis that utilising multimodal inputs can filter and enhance any ageing-related pieces of information, resulting in stronger correlations with a particular nIDP. The two explanations are not mutually exclusive, however, the choice of either one or the other serves as a guide for choosing the most appropriate ensembling strategy.

Considering the above, several ensembling methodologies were tested. To address the first explanation, ensembles were created using either all the 57 maps, or groups of maps based on the clusters formed by hierarchical clustering guided by the correlation matrix between map-specific predicted and debiased brain age deltas [35]. To test the second explanation, two approaches were considered. The first consisted of exhaustive multi-dimensional space searches, but these quickly proved to be computationally unfeasible. Given this, an empirical approach was devised and employed to determine which maps to ensemble. Here, several of the statistically significant correlations between the map-predicted brain age deltas and nIDPs, known *a priori*, were used to guide the map selection. Thus, subsets of maps which correlated strongly with a particular nIDP were ensembled to determine if this improves correlations with the target nIDP. A total of 10 experiments were carried out this way, with 5 each for the male and female subject groups.

Ensembling was performed with several methods. The simplest employed approach was naive averaging, where the predictions from multiple maps were averaged [67, 160]. Naive averaging, however, allocates equal weights to each ensembled map. This is potentially undesirable, as some maps might have more predictive power than others. Thus, more advanced data-blending techniques are required which automatically assign higher weights to more informative maps (Figure 3.25). In this work, three linear and one non-linear ensembling models were trained using the first left-out testing dataset, as described in Section 3.3.1, and evaluated on the second half of the testing dataset, similar to the experiments in the previous sections.

The simplest linear data-blending technique used in this work is a multiple-linear regression model, trained using the brain age predictions from multiple maps to predict chronological age. A natural extension of this method is given

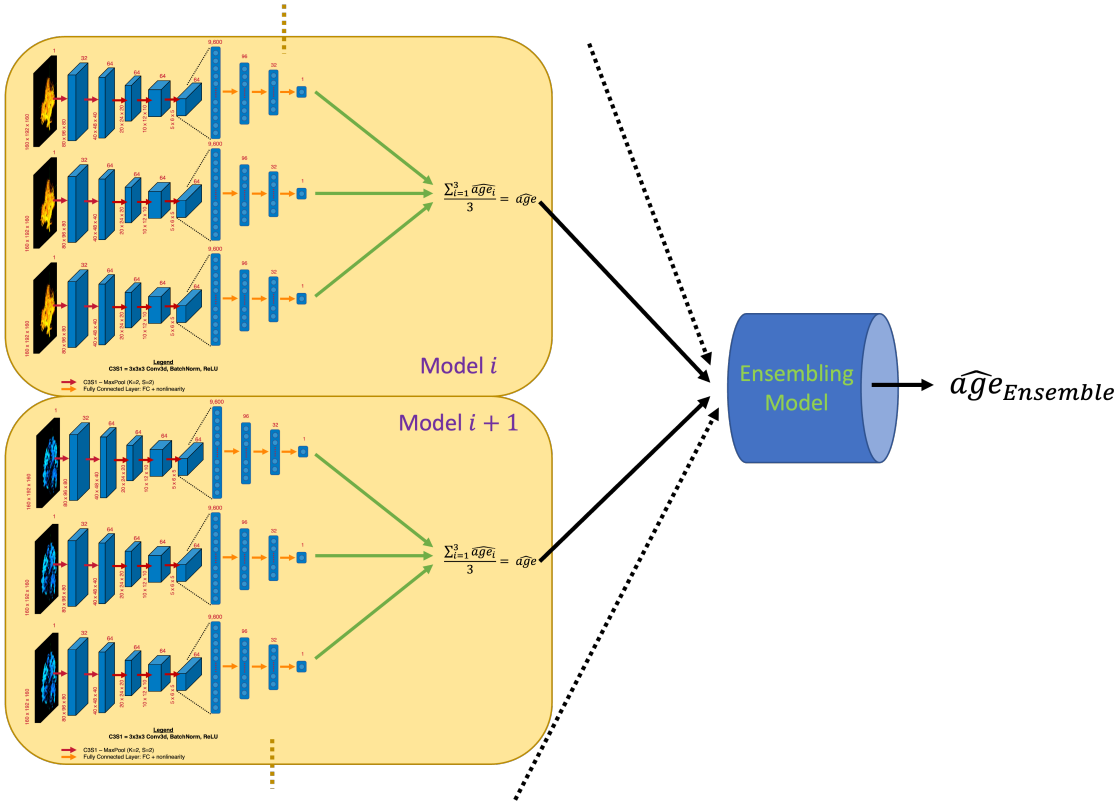


Figure 3.25: Post-training ensembling illustration. Each map-specific model is obtained by ensembling the predictions of 3 identical deep neural networks for the left-out test subjects, as shown in Figure 3.3. Then, half of the test subjects are used to fit an ensembling model, while the other half are utilised for evaluating the ensemble predictions.

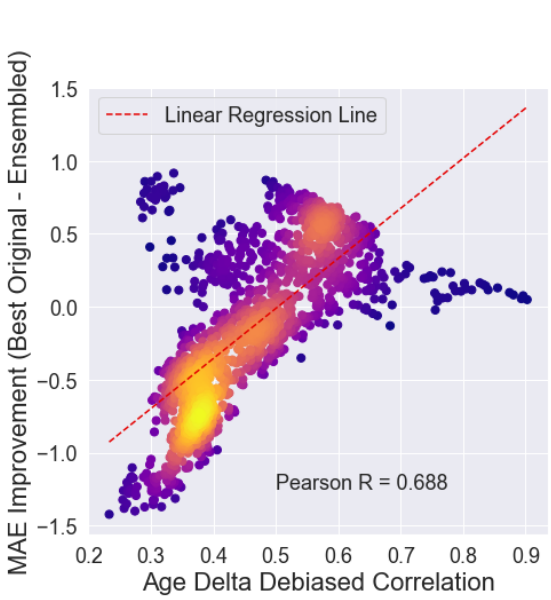
by ElasticNet, which adds additional regularisation penalties to the loss function during training (Equation 3.5.1.3). The λ and α hyperparameters of ElasticNet were fine-tuned using 5-fold cross-validation. A further extension of ElasticNet is provided by the Repeated ElasticNet (RENT) [295] model. This carries out an additional feature selection step by means of a series of Generalised Linear Models (GLMs) with ElasticNet regularisation, retaining only those features which minimise the predicted MAE while also matching a set of stability criteria. The selected features are then fed into an ElasticNet model, as described above. In this work, the original author-prescribed RENT hyperparameters were retained. The number of RENT GLM models fitted to each data pair was determined by means of a stability analysis investigation, as suggested in the original paper, with 50 models being utilised for both the male and female datasets.

A non-linear ensembling method was also tested, in the form of a multi-layer perceptron (MLP). As the number of hyperparameters to be refined is considerably higher, only a subset of them were tuned. These included the depth and width of the hidden layers, the chosen non-linear activation functions, and the learning rate. To determine the best hyperparameter combination, an exhaustive grid search with cross-validation was performed, with the model achieving the lowest MAE being selected. This approach yielded a model with 4 hidden layers of sizes $[100, 50, 50, 25]$, ReLU nonlinearities, trained with the Adam optimiser and an adaptive learning rate. This means that the original learning rate ($1e-3$) is kept constant unless two consecutive training epochs fail to decrease the training loss by at least $1e-4$, when the learning rate is divided by 5 [296]. The final results were obtained by training 5 identical MLP models and averaging the predictions obtained using the left-out second test dataset at the subject level, thus reducing stochastic effects induced by weight initialisation or training.

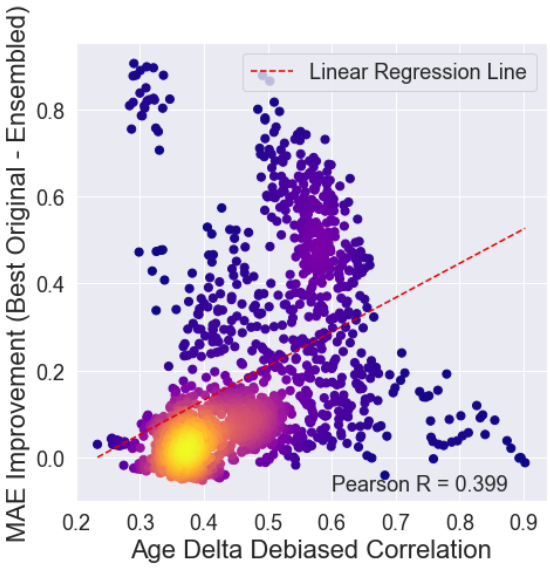
3.7.2 Results

Prior to testing the several methodologies described in the previous section, the most common approach proposed by literature was investigated. This suggests that ensembles of maps which have less correlated brain age delta vectors will lead to the best performance improvements, as it makes the best use of complementary information [67]. However, Figure 3.26 shows that this is not necessarily the case. In general, the less correlated two brain age delta vectors are, the lower their ensembled MAE improvement versus the best performing single-map in that pair. This appears to hold true for both naive and linear regression ensembles. However, Figure 3.26b shows the existence of two clusters which diverge from this observed trend. The first cluster forms in the upper left corner for brain age delta vector correlations of ≈ 0.3. This group is formed of the ensembles composed of the T2 Lesions map and the various rsfMRI maps. The second cluster forms in the centre, for correlations of ≈ 0.6, and is formed of ensembles composed of pairs of rsfMRI maps. These high improvements could be explained by the fact that the T2

Lesions and rsfMRI map start off with the highest MAEs of all investigated maps, and as such, there is a larger margin for performance improvement. As all these maps encode some information about cerebral blood circulation, there is also the possibility of certain complementary information being present, yet this observation does not appear to be present in other map groups.



(a) Naive ensemble MAE improvement vs. component map brain age delta vector correlation



(b) Linear regression ensemble MAE improvement vs. component map brain age delta vector correlation

Figure 3.26: Random ensembles of any two maps for female subjects, using either naive averaging **(a)** or linear regression **(b)**. Results for male subjects are almost identical. The improvements in ensemble MAE are decoupled from the correlation between the two component maps' brain age delta vectors.

This observations confirm that map selection for the creation of ensembles is not a straight-forward process, and requires the definition of a set of rules. Knowing this, post-training ensembling was carried out using two different approaches. The first approach ensembled the brain age predictions of either all 57 maps, or clusters of maps formed using the correlation matrix (e.g. Figure 3.16a for female subjects) and hierarchical clustering, while the second approach focused on utilising associations with nIDPs as a guide for selecting the maps to ensemble.

3.7.2.1 First Ensembling Approach: Improving Prediction MAEs

The results of the first ensembling approach are displayed in Table 3.6 and Figure 3.28. Global ensembling, using all available maps, was carried out using several methods, as described in the Methods section. Cluster ensembling was carried out solely with ElasticNet, as this method achieved the best MAE in the all-map experiments. Figure 3.27 shows the various map clusters obtained with hierarchical clustering.

The clusters are identical to those previously observed in Figure 3.16, with two large map-clusters forming, one containing the maps derived from rsfMRI, and the other containing the remaining maps, which can then be further broken down into sub-clusters. Similarities can also be observed with the PCA decomposition data in Figure 3.17, particularly in terms of the allocation of the tfMRI, Jacobian and SWI maps to either one or the other clusters identified with the cluster dendrogram, depending on the distance between these maps and the two main clusters.

Overall, all the ensembles which employ data blending techniques (i.e., not naive averaging) achieve better results than any of their single-map components. The best results were obtained by ensembling the volumes in Cluster 11 with ElasticNet. With an MAE of 1.982 years, this achieves state-of-the-art performance for the $45 - 82$ age interval. This cluster contains all maps except for those derived from rsfMRI, which were the overall worst single-map performers.

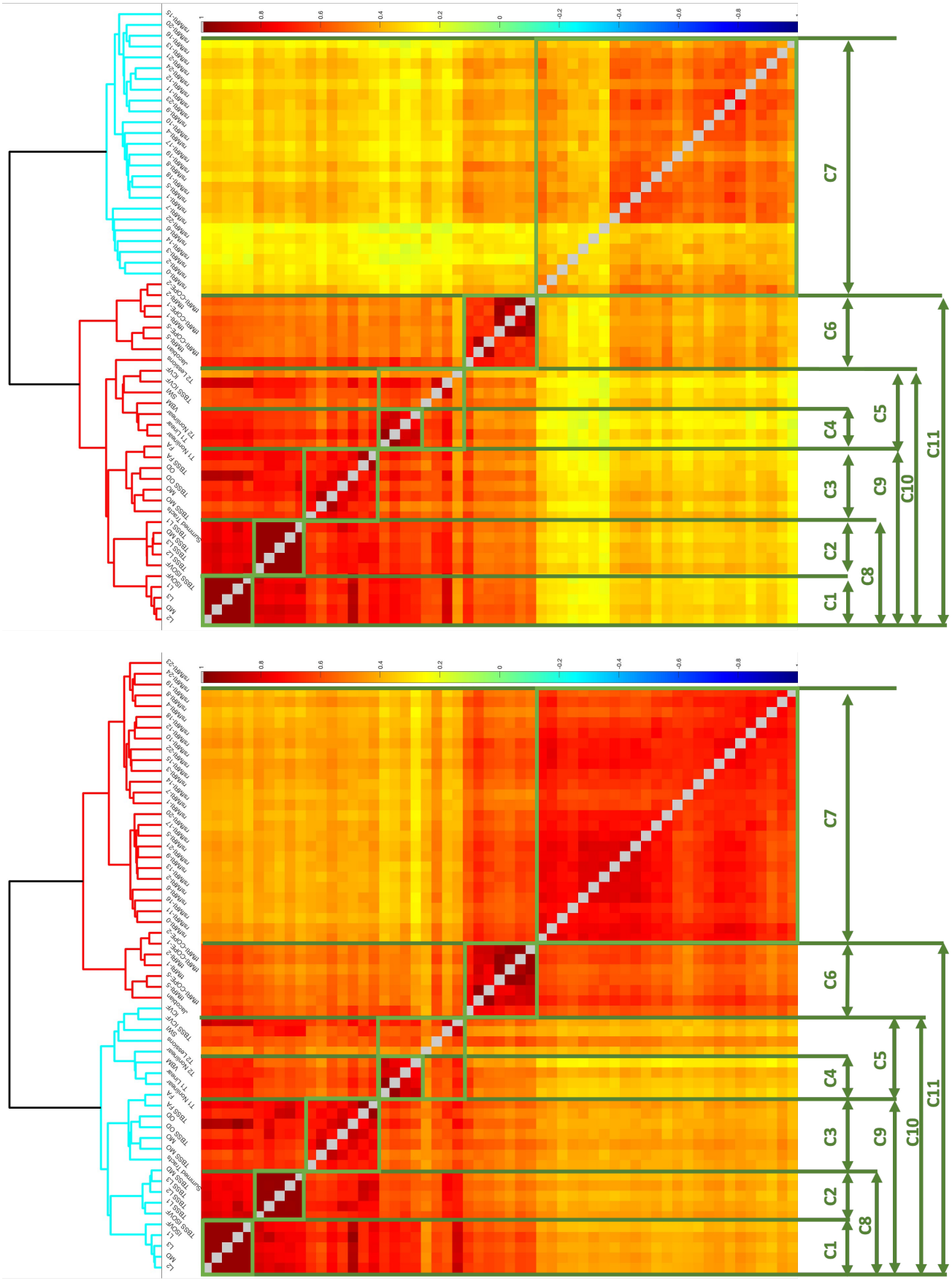
By analysing the regression coefficients assigned to each map, insight can be gained into their relative contribution to the ensemble. Figures 3.29-3.30 show that the largest weights are assigned to those maps which achieve the best individual accuracies and have traditionally been used in brain age prediction studies (T1 Linear, T1 Nonlinear, and T2 FLAIR Nonlinear), while the lowest absolute weights are assigned to rsfMRI maps. These low weights are consistent with the lack of rsfMRI maps in Cluster 11, as observed in the cluster-based experiments, as well as with the RENT results, which drop the rsfMRI map features as they promote neither model stability nor robustness (Figure 3.30c).

In regards to the other modalities, it can be seen that both the tfMRI and dMRI maps have non-zero contributions, suggesting that the ensembles are making use of complementary information to improve the overall brain age predictions. Yet, some degree of overlap does exist between the information contained in the various modalities and modality-derived maps. This can be deduced from Figures 3.30a-3.30b, which show that progressively adding maps to an ensemble, in order of their absolute weights from the all-map global ensemble, has diminishing returns after some point. Thus, adding more than the top 14 maps for females subjects, and 19 for male subjects, has little impact on the overall ensemble MAE.

The weights allocated to each map by ElasticNet in the all-map ensemble were validated using cross-validation with 100-random permutation shuffle splits. Figures 3.29c-3.29d show that the variation of the cross-validated weights is small, suggesting that the learned weights (marked with a black star) are adequate. In addition, the originally predicted MAE appears to be close to the mean of the narrow distributions obtained using the cross-validated predictions, which increases confidence in the results already presented.

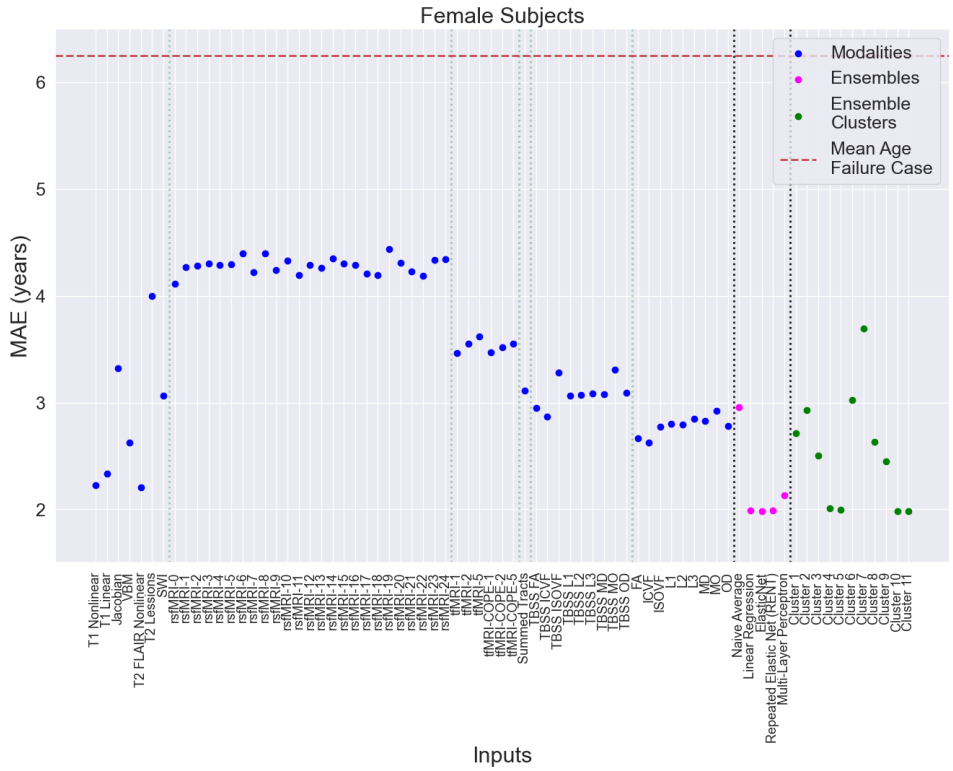
The relationships between nIDPs and the debiased and deconfounded brain age deltas obtained for the various ensembling techniques were also investigated. This was done in a similar manner to the work conducted previously using single map predictions. A subset of results for the all-map ensembling experiments are presented in Figure 3.31, with results for several map clusters being shown in Figure 3.32. Figure 3.33 then presents the proportion of significant hits (i.e., those passing the Bonferroni Threshold) for both the all-map ensembles, and those for the map clusters. All results are displayed for both male and female subjects.

Similar to the single-map results, differences can be observed between the male and female groups, but also between the various clusters (Figure 3.32). This is expected, as correlations are influenced by the maps composing each cluster. For instance, Cluster 7 is composed of rsfMRI maps, while Cluster 1 is composed of several dMRI maps. As such, the general association trends for Cluster 7 follow those seen for individual rsfMRI maps, for both male and female subjects (Figures

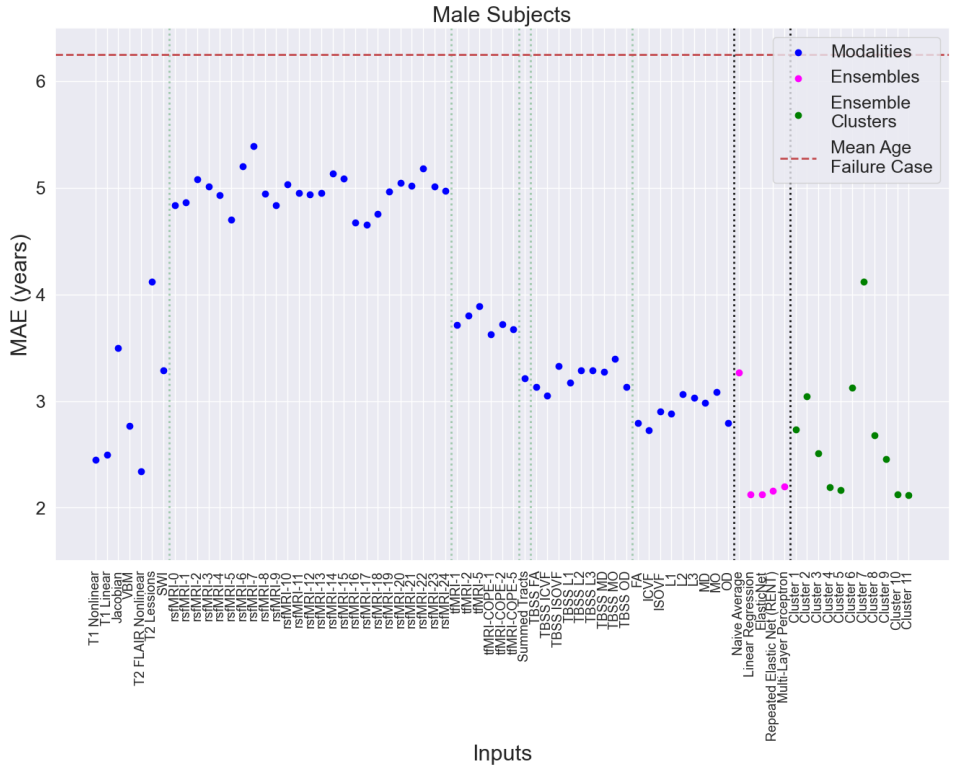


(a) Female subjects dendrogram **(b)** Male subjects dendrogram

Figure 3.27: Hierarchical clustering of the 57 maps for **(a)** female and **(b)** male subjects. The clustering was carried out based on the correlations between single-map CNN-predicted brain age deltas. Cluster dendrograms display the two main clusters, while the smaller sub-clusters are indicated using green markings, with C1 corresponding to Cluster 1, and so on.

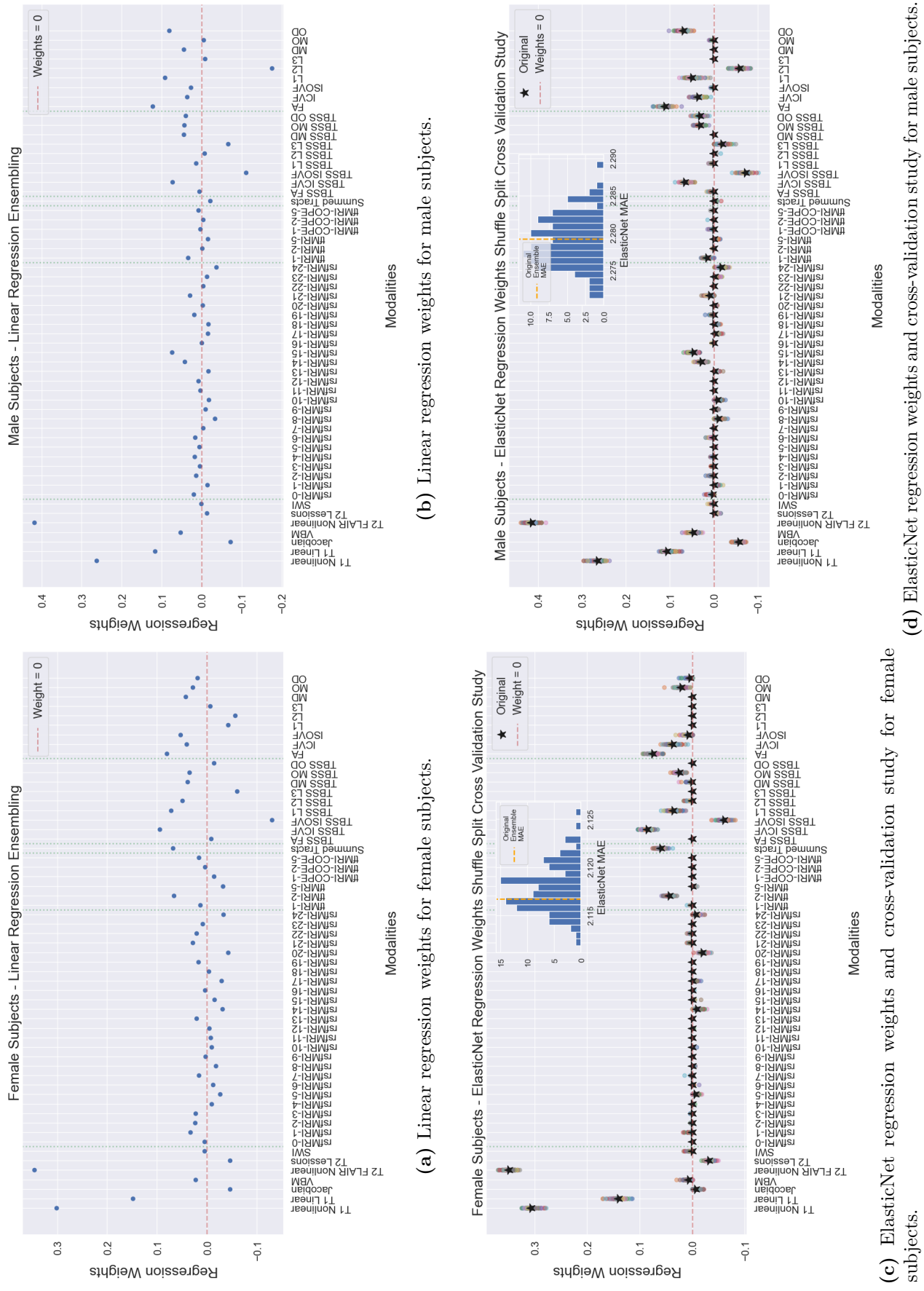


(a) MAE predictions for female subjects



(b) MAE predictions for male subjects

Figure 3.28: Mean absolute error (MAE) distributions for multiple-map ensemble predictions, showing the original single-map predictions (blue), the predictions of ensembles using all 57 maps as inputs (magenta), and those ensembles using sub-map clusters (green) for **(a)** female and **(b)** male subjects. The failure case in which the models predict only a population mean age is also plotted for comparison.



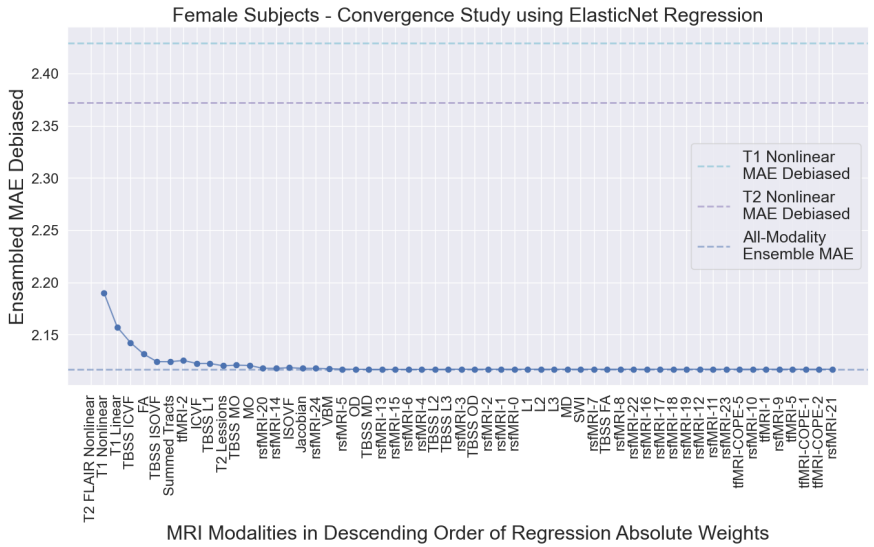
(a) Linear regression weights for female subjects.

(b) Linear regression weights for male subjects.

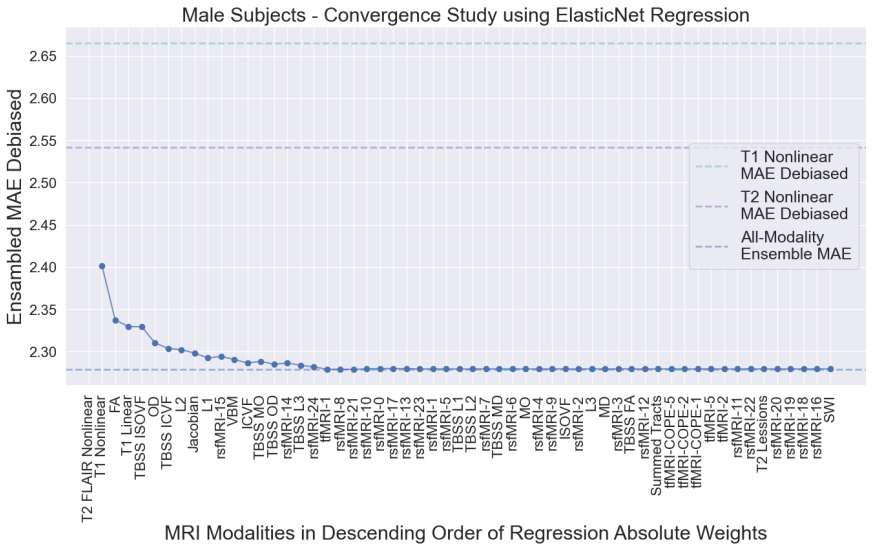
(c) ElasticNet regression weights and cross-validation study for female subjects.

(d) ElasticNet regression weights and cross-validation study for male subjects.

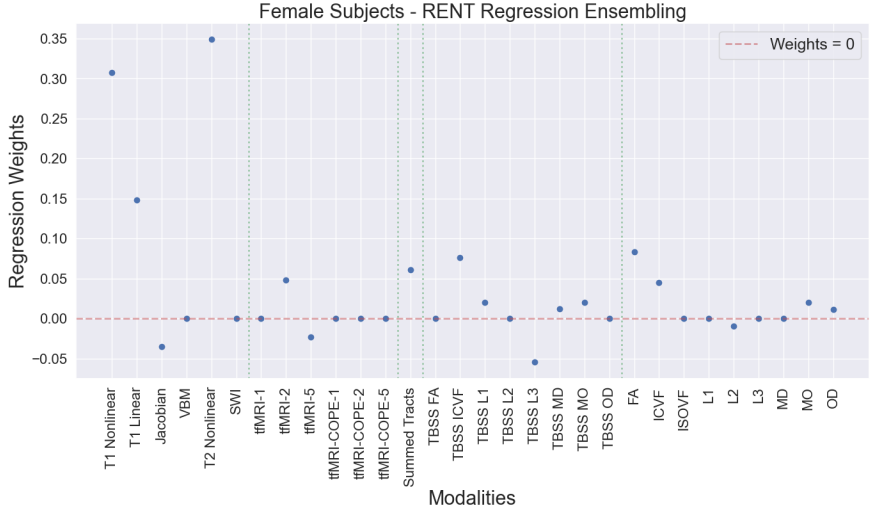
Figure 3.29: Learned weight distributions for all-map ensembles (1/2) for **(a)-(c)** female and **(b)-(d)** male subjects. **(a)-(b)** indicate results for the Linear Regression models, and **(c)-(d)** for the ElasticNet models. For the latter, the original weights are marked with black stars, while the remaining dot marks represent the range of values taken by the weights during the split cross validation study. The MAE distributions for these studies are also shown, with the original MAE marked as an orange dashed line.



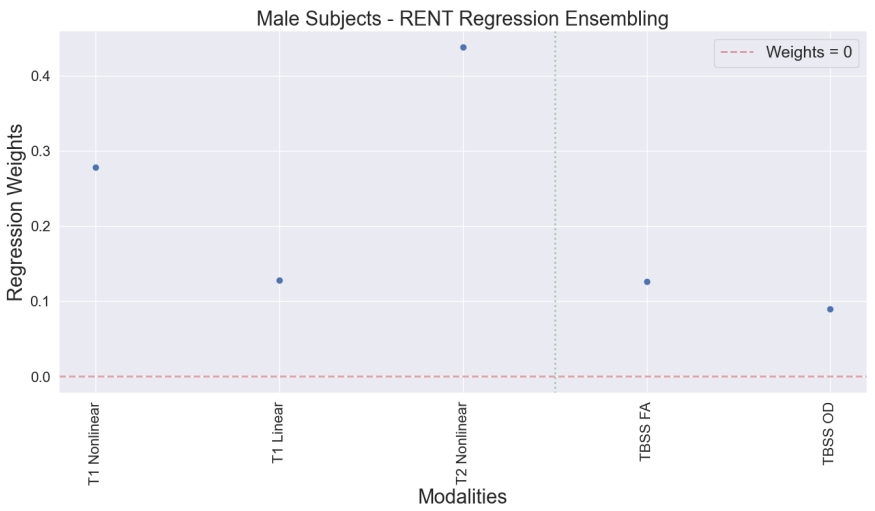
(a) ElasticNet convergence for female subjects



(b) ElasticNet convergence for male subjects



(c) RENT regression weights for female subjects.



(d) RENT regression weights for male subjects.

Figure 3.30: Learned weight distributions for all-map ensembles (2/2) for **(a)-(c)** female and **(b)-(d)** male subjects. **(a)-(b)** show the results of a convergence study, where maps were added to ElasticNet in order of their absolute weights from the all-map ensemble, and **(c)-(d)** indicate the weight distributions obtained with the RENT model. The difference in map numbers between male and female subjects is due to different maps being selected by the RENT GLM in order to meet the required stability and robustness criteria.

Table 3.6: Multiple-map ensemble results split by sex and whether they utilise all available maps, or a subgroup/cluster of them. For each cluster, the component maps corresponding to Figure 3.27 are listed. In the case of RENT, all maps were initially presented to the model, from which only a subset matching the required stability and robustness criteria were retained. The Weighted MAE allows for easier comparison between studies, as proposed by Cole et al [8]. The age ranges are $45.12 - 82.26$ for female and $45.45 - 82.18$ for male subjects.

EnsembleName	maps	Female				Male			
		MAE	R^2	Predicted-Chronological Correlation (r)	Weighted MAE	MAE	R^2	Predicted-Chronological Correlation (r)	Weighted MAE
Ensembles of All maps - Female Subjects						**Ensembles of All maps - Male Subjects**			
Naive Average	All	2.958	0.78	0.883	0.080	3.265	0.785	0.886	0.090
Linear Regression	All	1.991	0.882	0.939	0.054	**2.124**	**0.873**	**0.934**	**0.059**
ElasticNet	All	**1.986**	**0.883**	**0.939**	**0.053**	2.127	0.872	0.934	0.059
RENT	All* (select subset)	1.987	0.882	0.939	0.054	2.162	0.869	0.932	0.060
MLP	All	2.134	0.866	0.930	0.057	2.202	0.862	0.929	0.061
Ensembles of Groups of maps - Female Subjects						**Ensembles of Groups of maps - Male Subjects**			
Cluster 1	L2, MD, L3, L1, ISOVF	2.713	0.783	0.885	0.073	2.733	0.789	0.888	0.075
Cluster 2	TBSS (ISOVF, L1, L2, L3, MD) Summed Tracts	2.927	0.747	0.864	0.079	3.047	0.743	0.862	0.084
Cluster 3	TBSS (FA, MO, OD) FA, MO, OD	2.504	0.814	0.902	0.067	2.510	0.822	0.906	0.069
Cluster 4	T1 Nonlinear, T1 Linear VBM, T2 FLAIR Nonlinear Cluster 4 +	2.010	0.878	0.937	0.054	2.194	0.865	0.930	0.060
Cluster 5	T2 Lesions, SWI ICVF, TBSS ICVF	1.998	0.881	0.938	0.054	2.164	0.868	0.932	0.060
Cluster 6	Jacobian + tfMRI (All)	3.023	0.729	0.854	0.081	3.127	0.73	0.854	0.086
Cluster 7	rsfMRI (All)	3.695	0.592	0.770	0.100	4.121	0.546	0.739	0.114
Cluster 8	Cluster 1 + Cluster 2	2.634	0.796	0.892	0.071	2.678	0.796	0.892	0.074
Cluster 9	Cluster 3 + Cluster 8	2.447	0.824	0.908	0.066	2.457	0.828	0.910	0.068
Cluster 10	Cluster 5 + Cluster 9	1.986	0.883	0.940	0.053	2.127	0.872	0.934	0.059
Cluster 11	Cluster 6 + Cluster 10	**1.982**	**0.883**	**0.940**	**0.053**	**2.118**	**0.873**	**0.934**	**0.058**

3.19e-3.19f and 3.21e-3.21f), while the same is true for Cluster 1. These inter-dependencies were expected, given how the ensembling clusters were selected. A comparison of Figures 3.33c-3.33d for the ensembling clusters correlations and Figures 3.23-3.24 for single-map associations reinforces this observation. Similarities can also be observed between those associations obtained with Cluster 11 and those obtained for all maps with ElasticNet, given that the latter allocates high weights to those maps also composing Cluster 11.

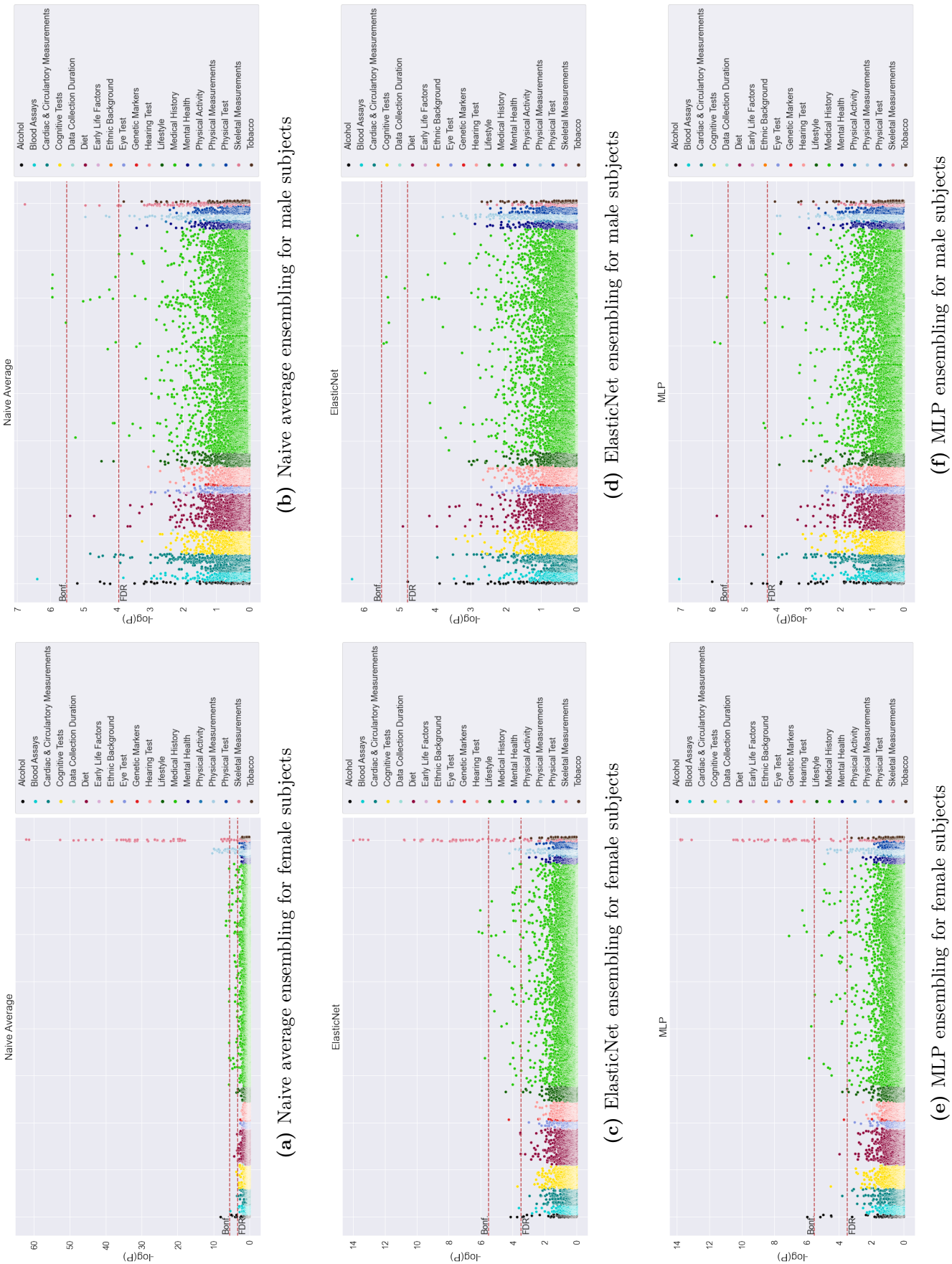
Despite these, probably the most interesting observations can be made when analysing the results obtained for ensembles of all maps using different ensembling techniques. Firstly, due to the large proportion of rsfMRI-maps in the overall group, naive averaging ensembling results are dominated by the brain age deltas obtained with those maps. Then, when considering the associations passing the Bonferroni threshold that are obtained with linear and nonlinear ensembling techniques post-training, it can be observed that no strong associations can be found in several variable categories, such as Blood Assays or Physical Measurements in the case of

the female subject group (Figure 3.33a), or Tobacco for the male subject group (Figure 3.33b), which are all nIDP categories in which prior strong associations were observed for the single-map results (Figures 3.23-3.24). Thus, while achieving improvements in the overall brain age predictions, these methods lead to the loss of some meaningful associations. This appears to be mediated to some extent by the non-linear MLP ensembling technique when compared to the three linear techniques, with the former showing significant associations in categories such as Alcohol (for both female and male subjects) and Diet, where the linear methods have none. In addition, for those variable categories where the linear methods do have significant associations, the non-linear method leads to a slightly larger number of correlations passing the Bonferroni threshold. All these are occurring despite the fact that the MLP method achieves an overall MAE which is marginally worse than those obtained by the linear methods.

Similar to the single-map results, all nIDP associations with the predicted brain age deltas obtained with the various ensembling methods and methodologies can be accessed freely on *Zenodo* or by going to this URL: `https://doi.org/10.5281/zenodo.8110876`.

3.7.2.2 Second Ensembling Approach: Improving Correlations to nIDPs

So far, the results focused on the first ensembling approach, which aims to increase the overall accuracy of the brain age predictions, measured in terms of MAE. Yet, when evaluating the results obtained for different contrast clusters more closely, it was observed that, for some associations, both their absolute correlations and their $-log(p)$ values increased when compared to the single maps. For instance, in the case of female subjects (Figure 3.32c), it can be seen that the significance of Skeletal Measurements associations achieves values of $-log(p) \geq 80$ for some Skeletal Measurements, which is not the case for the individual rsfMRI-maps composing Cluster 7 (Figures 3.19e, 3.19f, 3.20a). Case in point, the nIDP *Total BMD (bone mineral density) T-score (2.0)*, part of the Skeletal Measurements group, achieves a $-log(p)$ value of 82.5 and a Pearson correlation of -0.392 for the Cluster



(a) Naive average ensembling for female subjects

(b) Naive average ensembling for male subjects

(c) ElasticNet ensembling for female subjects

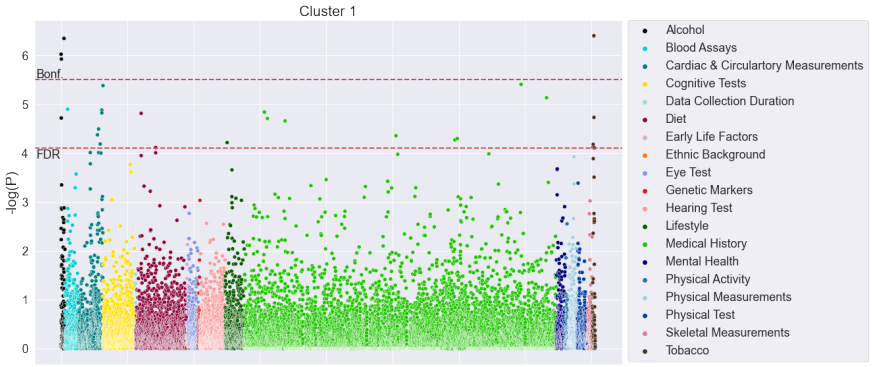
(d) ElasticNet ensembling for male subjects

(e) MLP ensembling for female subjects

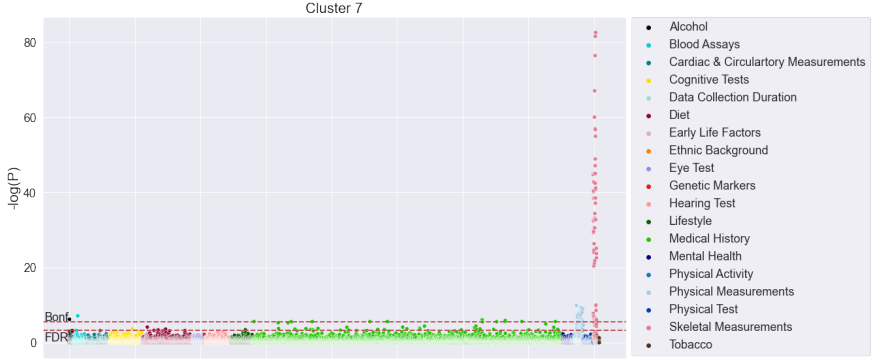
(f) MLP ensembling for male subjects

Figure 3.31: Manhattan plots relating all-map ensembles brain age deltas to UK Biobank nIDPs for Naive Averaging, ElasticNet and the Multi-Layer Perceptron (MLP). **(a),(c),(e)** indicate results for female subjects, while **(b),(d),(f)** for male subjects. The False Discovery Rate (FDR) and Bonferroni thresholds are also plotted.

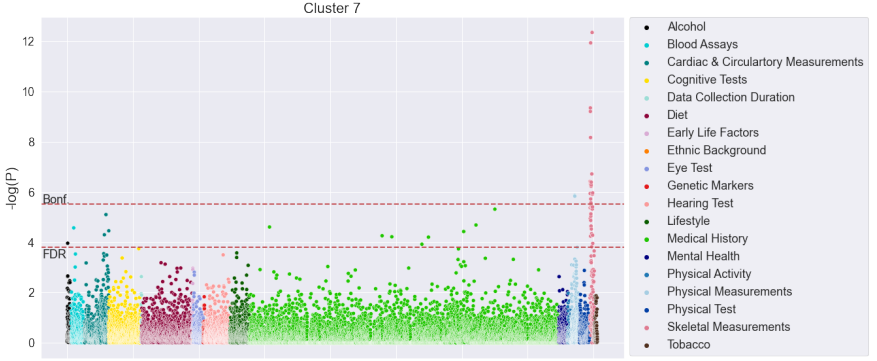
(a) Cluster 1 correlations for female subjects



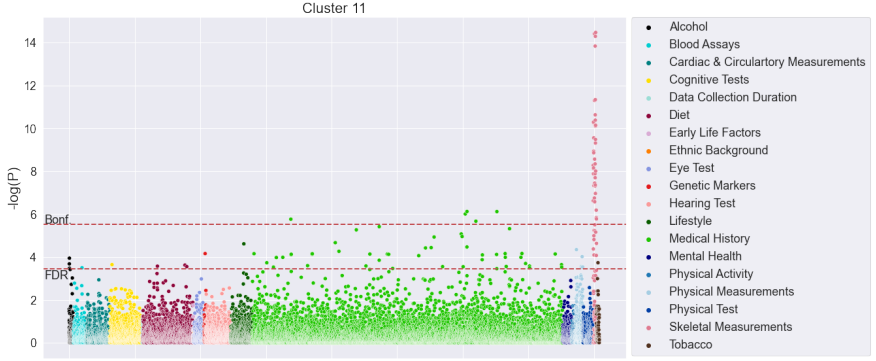
(b) Cluster 1 correlations for male subjects



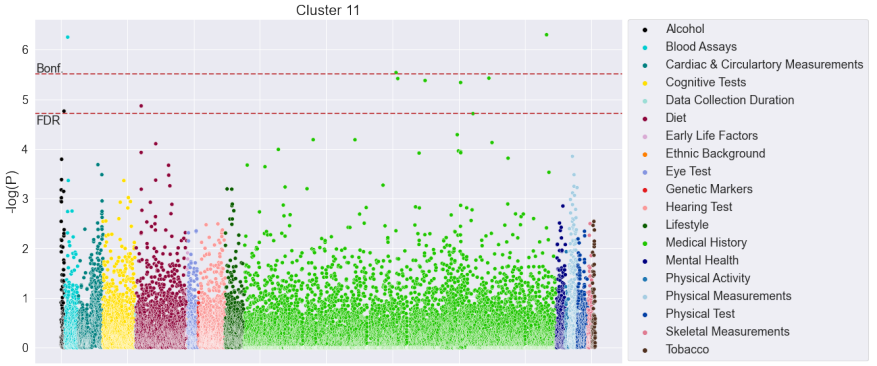
(c) Cluster 7 correlations for female subjects



(d) Cluster 7 correlations for male subjects

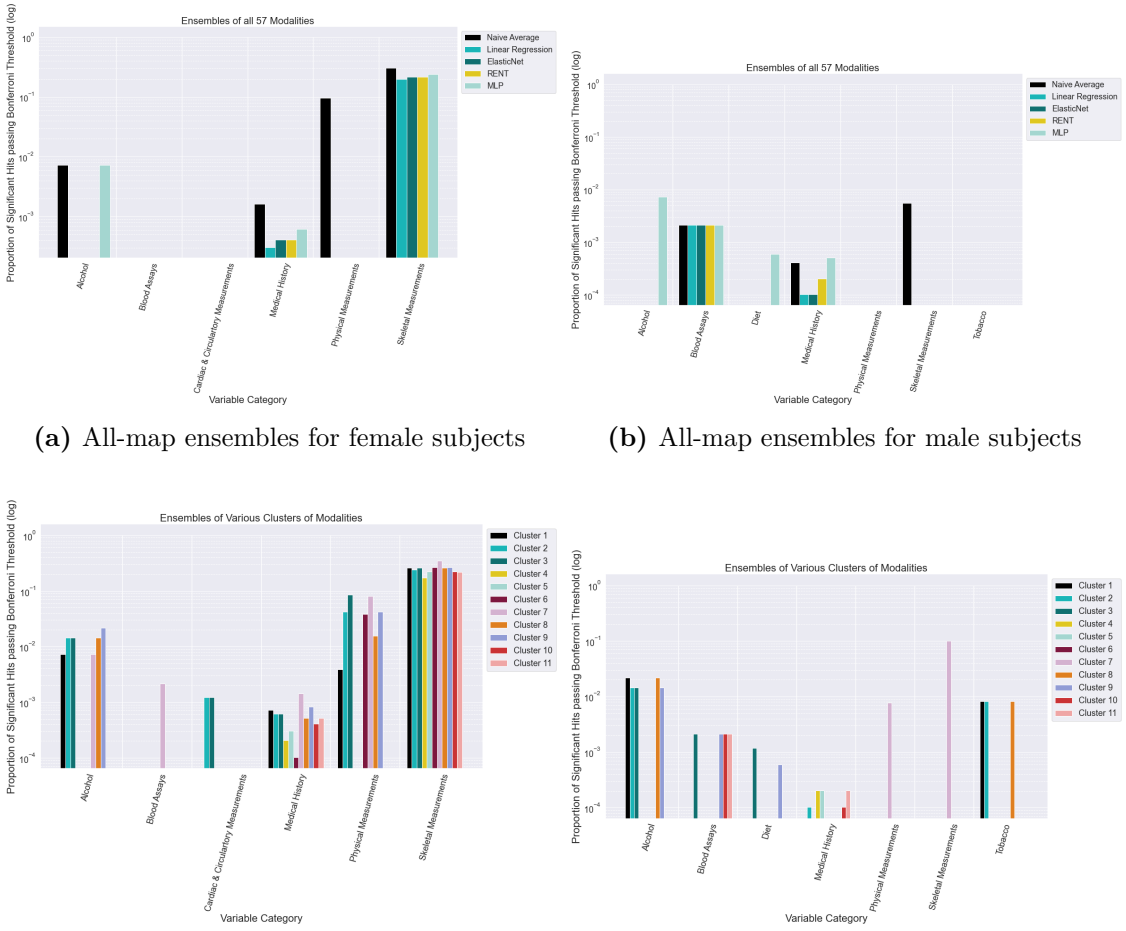


(e) Cluster 11 correlations for female subjects



(f) Cluster 11 correlations for male subjects

Figure 3.32: Manhattan plots relating cluster-based ensembles brain age deltas to UK Biobank nIDPs. All clusters are ensembled using ElasticNet. **(a),(c),(e)** indicate results for female subjects, while **(b),(d),(f)** for male subjects. The False Discovery Rate (FDR) and Bonferroni thresholds are also plotted.



(a) All-map ensembles for female subjects

(b) All-map ensembles for male subjects

(c) Cluster ensembles for female subjects

(d) Cluster ensembles for male subjects

Figure 3.33: Proportion of significant hits per UK Biobank nIDP category for the multi-map ensemble, calculated by dividing the number of variables passing the Bonferroni threshold by the total number of variables in that category. **(a)-(c)** shows results for female subjects, while **(b)-(d)** indicate results for male subjects. **(a)-(b)** The results for the all-map ensembles are presented separately from those for **(c)-(d)** contrast clusters.

7 ensemble. At the same time, when analysing the same nIDP's relationships with the results obtained using single-maps, its highest Pearson r value is equal to -0.344 when correlated with brain age deltas obtained with rsfMRI-2, and with an associated $-log(p)$ value of 62.3.

These observations support the hypothesis governing the second ensembling approach: that the strength of various nIDP associations can be amplified by using bespoke contrast groups created based on prior knowledge of single-contrast associations with that nIDP. To test this, a series of 10 experiments were carried

out, 5 for each subject group. In these, for each nIDP category, several nIDPs which had significant associations and high correlations with at least three maps were selected. Then, the maps were ensembled using ElasticNet, which was chosen based on it being the best performing in the previous ensembling experiments (Table 3.7).

The original hypothesis was confirmed for 9/10 ensembles, with the absolute correlation between the guide-nIDP and the ensemble brain age deltas being larger than the correlation with any of the individual ensemble component maps. The one case where this was not confirmed was for the *Diagnoses - secondary ICD10 (I10 - I10 Essential (primary) hypertension)* nIDP for female subjects. This could be due to T2 Lesions being more common in subjects with symptomatic cardiovascular and cerebrovascular diseases, such as hypertension [184, 185], which would support the idea of stronger correlations between T2 Lesions predicted brain ages and cardiovascular pathologies such as hypertension. This observation reinforces the fact that a careful assessment of the maps composing a cluster remains of paramount importance prior to any ensembling.

Table 3.7: UK Biobank based ensembles, where groups of maps were selected and ensembled using ElasticNet based on their high correlations with a specific UK Biobank nIDP. Groups of maps were selected for each major nIDP group by taking the top 3-5 most highly correlated maps with the particular nIDP of interest. For all groups, the resulting absolute ensemble correlation was higher than that of the components, except in the case of the final female subject group relating to a diagnosis of hypertension, where the information encoded in the T2 Lesions ends up dominating the information from the other maps.

Female Subjects					Male Subjects				
Measure Group	**Measure Name**	**Component maps**	**Individual Correlations (r)**	**ElasticNet Ensemble Correlation (r)**	**Measure Group**	**Measure Name**	**Component maps**	**Individual Correlations (r)**	**ElasticNet Ensemble Correlation (r)**
Alcohol	Frequency of drinking alcohol (0.0)	MD	0.129	0.149	Alcohol	Frequency of drinking alcohol (0.0)	TBSS FA	0.141	0.164
		rsfMRI-0	0.129				TBSS L1	0.137	
		MO	0.127				TBSS OD	0.137	
		TBSS ISOVF	0.127				MD	0.136	
		L2	0.127				L1	0.133	
Blood Assays	C-reactive protein (0.0)	rsfMRI-2	-0.108	-0.130	Blood Assays	IGF-1 (0.0)	OD	-0.118	-0.137
		rsfMRI-10	-0.108				TBSS ICVF	-0.114	
		rsfMRI-5	-0.107				FA	-0.112	
		rsfMRI-21	-0.101				ISOVF	-0.108	
							L2	-0.105	
Cardiac & Circulatory Measurements	Systolic blood pressure, automated reading (0.0)	TBSS FA	0.113	0.130	Cardiac & Circulatory Measurements	Central systolic blood pressure during PWA (2.0)	T2 Lesions	0.122	0.127
		TBSS L3	0.109				ISOVF	0.109	
		T2 Lesions	0.102				MD	0.107	
Medical History	Diabetes diagnosed by doctor (0.0)	ICVF	0.107	0.121	Medical History	Treatment/medication code (1140860806 - ramipril)	MD	0.113	0.114
		L2	0.103				TBSS L2	0.107	
		ISOVF	0.103				TBSS L3	0.106	
		TBSS ICVF	0.103						
	Diagnoses - secondary ICD10 (I10 - I10 Essential (primary) hypertension)	T2 Lesions	0.133	0.128	Tobacco	Smoking status (0.0)	ISOVF	0.111	0.119
		TBSS FA	0.113				TBSS L2	0.108	
		TBSS L3	0.103				L2	0.104	
		TBSS ISOVF	0.100				TBSS ISOVF	0.102	
							L1	0.101	

3.7.3 Discussion

Answering the question of whether using information from multiple maps simultaneously can lead to improved brain age predictions is anything but trivial. The first major challenge consists in devising a strategy for how to group maps into ensembles, with previous work either ensembling all available information [81, 160] or grouping maps using some derived heuristics [67]. On this topic, Peng et al. proposed that the less correlated two maps are, the better the performance of the resulting ensemble, as this enables the ensembling model to take advantage of complementary information in order to enhance brain age predictive performance [67]. The results presented in Figure 3.26 suggest that this might be only partially true, and generally for maps derived from the same core modality. This is based on the observation that ensemble MAE improvements seemed to mainly occur for either groups of rsfMRI maps, or for groups composed of one rsfMRI contrast and T2 Lesions, where both component maps encode cerebrovascular information. This observation is in line with the results presented by Peng et al, who utilised several maps derived from sMRI, all sharing a considerable amount of information between them: T1 Linear, T1 Nonlinear, Gray Matter and White Matter Segmentation. While further analysis is required in addressing this matter, Figure 3.26 suggest that the heuristic proposed by Peng et al. might not be applicable to maps derived from different core modalities. This led to the definition of two ensembling approaches for this work: one which focused solely on finding the best ensembling method and contrast group to achieve the best possible MAE, and another which sought to find a guided ensembling approach which utilised *a priori* knowledge of nIDP-brain age delta associations to form ensembles amplifying the signal underpinning the nIDP association.

Overall, the poorest results were obtained when using naive averaging for ensembling, while the best when employing adaptive data-blending techniques, and in particular ElasticNet (Table 3.6). Yet, despite the improvement in MAE, a reduction in the number of significant nIDP associations was observed, with some nIDP categories having no significant associations for the ensembles, despite doing so for the single-maps (Figures 3.33). These effects were somewhat mitigated by using

either a non-linear ensembling method, or by using clusters of maps created through hierarchical clustering. These experiments obtained poorer MAE results however, but retained more significant associations in more nIDP categories. Moreover, they were shown to amplify signals of interest, such as correlations to a specific nIDP.

This behaviour can be understood by considering the core assumption underpinning the "brain age paradigm": that observed errors are a combination of random Gaussian noise and structured, biologically relevant errors. Single-contrast regression models trained with mean squared error (MSE), or similar loss functions which encode a Gaussian prior in the distribution of noise [297], filter out the noise component leaving only the biologically relevant error. This goal of eliminating apparently random noise is also maintained when fitting ensembling models. Yet, what is classed as random noise in this case is less clear. For instance, situations might arise where ensembled maps have opposing sensitivities to a particular nIDP, which is reflected in their deltas. This would be similar to the results presented by Smith et al. [56], where multiple distinct modes of brain ageing reflected different input modalities and had distinct patterns of association with the full set of nIDPs. In trying to improve brain age predictions over the ensemble inputs, these opposing sensitivities, and associated delta differences, might be perceived as random noise, and thus be eliminated during ensemble training. Thus, weighted ensembling improves overall brain age prediction, albeit at the cost of eliminating possibly interesting biophysical information. In this case, the ensemble brain age deltas correlate only with nIDPs which represent core factors associated with accelerated ageing, or resilience, in all ensemble components.

This can be somewhat mitigated by careful selection of the ensemble components. If these are chosen, for instance, based on having similar correlations to a metric of interest, then the ensemble process harnesses the denoising effect of agreeing maps without losing information of potential interest. This leads to an increase in the structured information encoded by the deltas, which relates to a particular nIDP of interest, at the same time as denoising the deltas themselves. This increases the overall confidence in the obtained results.

Post-training ensembling with nonlinear methods also appears to prevent the loss of information in certain nIDP categories, such as Alcohol consumption (Figure 3.33). They are potentially mitigating some of the effects induced by using conflicting maps by learning complementary features across maps. Thus, to an extent, they are harnessing the denoising effect of aggregating maps while preventing the loss of biologically meaningful information. This topic will be further explored in Chapter 4.

3.8 Identifying Independent Biological Processes using Predicted Brain Ages

In this final section, the question of whether different modes of ageing, akin to underlying biological processes showing distinct patterns of brain changes, can be identified within the considered test populations using the brain age delta predictions was addressed. To this end, a generative model was adapted for this work using the methods described by Smith et al. [56]. It uses a FastICA-based algorithm [219] to perform independent component analysis (ICA). First, the methodology is discussed, after which, similar to the previous sections, the results are presented and discussed.

3.8.1 Methods: PCA-ICA

For a more detailed description of the ICA algorithm, please see Chapter 2 Section 2.2.1.2. In brief, ICA decomposes a multivariate set of signals, organised as an input matrix, into a set of additive and statistically independent components having non-Gaussian distributions.

For an input matrix $\boldsymbol{M}$ with dimensions test-subjects-by-maps, consisting of the standardised debiassed brain age deltas, the decomposition can be carried out in two directions:

- **Subject-direction**, leading to the identification of independent components in the subject space, representing groups of subjects which have a characteristic which co-varies across the maps;

- **Map-direction**, leading to the identification of independent components in the map space, which are patterns in which maps co-vary across subjects.

Given that each map is the result of processing designed to produce summary measures of interest from the core modalities, it is possible that no interesting new patterns are discovered in the map direction. At the same time, discovering underlying biological processes enabling the clustering of subjects across maps is probable given previous literature findings [56]. Thus, the main effort in this section focused mostly on understanding the results obtained from ICA in the subject direction (or subject-ICA). However, for completeness, some consideration was also given to ICA in the map direction (map-ICA). This can be easily performed by transposing the $\boldsymbol{M}$ matrix before performing the following operations.

Prior to any decomposition, the matrix $\boldsymbol{M}$ was deconfounded, with all 613 confounds in UK Biobank (see Section 2.1.3) being removed through multiple regression [56, 229]. Then, $\boldsymbol{M}$ was normalised (mean subtracted, division by standard deviation) in the subject-direction (i.e., across maps).

For fitting the method, Smith et al. [56] employed a combination of principal component analysis (PCA), for dimensionality reduction, and FastICA, to estimate independent underlying processes. This would decompose $\boldsymbol{M}$ of size subjects-by-maps $(n \times m)$ into two matrices, $\boldsymbol{B}$ of size $(n \times b)$ and $\boldsymbol{G}$ of size $(b \times m)$. $\boldsymbol{B}$ is the biological processes matrix, containing information relating to b underlying biological processes which are independent of each other, while $\boldsymbol{G}$ is a mapping matrix between the biological information and the various maps. At this point, $\boldsymbol{B}$ represents an optimal set of features which can be fed into regression to predict age. This can be modelled as the second part of Equation 3.8.1.1, where $\boldsymbol{C}$ of size $(n \times c)$ can be considered an input matrix to a regression operation, and $\boldsymbol{F}$ of size $(c \times b)$ as being the matrix mapping onto biology, or, the weights in a regression operation.

$$\underset{(n\times m)}{\boldsymbol{M}} = \underset{(n\times b)}{\boldsymbol{B}}\,\underset{(b\times m)}{\boldsymbol{G}} \; ; \; \underset{(n\times b)}{\boldsymbol{B}} = \underset{(n\times c)}{\boldsymbol{C}}\,\underset{(c\times b)}{\boldsymbol{F}} \tag{3.8.1.1}$$

The main hyperparameter in ICA operations is represented by the number of components the operation is asked to estimate. Although, traditionally, this

parameter is identical for both PCA and FastICA, representing at the same time both the initial PCA dimensionality reduction and the number of independent components to be identified, Smith et al. propose a method by which these measures are determined independently of each other. Firstly, both PCA and ICA were run at dimensionalities from 2 to ψ, where ψ represents the number of maps, and evaluated on randomly split-half reproducibility. This means that, for each PCA dimensionality, the ICA is run on the full dataset and on the two randomly split halves. The split-half components are then ordered so that they best match the original component obtained with the full-data ICA run. This allows the calculation of correlations between the split-halves paired ICA components' source vectors. Using these correlations, only the most similar components, with Pearson $r \geq 0.7$, were retained. This process was repeated 10-times for subject-ICA (also referred to as vertical-ICA), and 100-times for the map-ICA (horizontal-ICA). The results for map-ICA proved to have a higher variability than those for subject-ICA, which is why a higher number of repetitions was utilised.

When carrying out subject-ICA, before ordering the split-half components, to ensure compatibility between the original ICA decomposition and those carried out on the split-halves, the decompositions are projected back into the PCA input space, so that they match the ICA mixing matrix. The PCA dimensionality chosen is that which results in the highest number of highly reproducible components. Following this, the number of ICA components was found by fixing the PCA dimensionality, and re-running the ICA operation 100-times utilising the split-half method proposed by Smith et al., retaining the most robust run in terms of reproducibility. The code for carrying out these steps is freely made available together with the main codes covering this project.

At the end of the above operation, 5 matrices are returned: $\boldsymbol{A}_{ICA}$, $\boldsymbol{S}_{ICA}$, $\boldsymbol{S}_{PCA}$, $\boldsymbol{U}_{PCA}$, $\boldsymbol{V}_{PCA}$. The latter 3 matrices correspond to the PCA-decomposition of $\boldsymbol{M}$ (Equation 3.8.1.2) while the former 2 to the FastICA operation (Equation 3.8.1.3). Thus, the original $\boldsymbol{M}$ can be rewritten as in Equation 3.8.1.4, where $\boldsymbol{U}_{PCAICA}$ is

described by Equation 3.8.1.5 and corresponds to $\boldsymbol{B}$, and $\boldsymbol{S}_{ICA}$ to $\boldsymbol{G}$ from above for subject-ICA, and vice-versa for horizontal-ICA.

$$\boldsymbol{M} = \boldsymbol{U}_{PCA}\boldsymbol{S}_{PCA}\boldsymbol{V}_{PCA}^{T} \tag{3.8.1.2}$$

$$\boldsymbol{V}_{PCA}^{T} = \boldsymbol{A}_{ICA}\boldsymbol{S}_{ICA} \tag{3.8.1.3}$$

$$\boldsymbol{M} = \boldsymbol{U}_{PCA}\boldsymbol{S}_{PCA}\boldsymbol{A}_{ICA}\boldsymbol{S}_{ICA} = \boldsymbol{U}_{PCAICA}\boldsymbol{S}_{ICA} \tag{3.8.1.4}$$

$$\boldsymbol{U}_{PCAICA} = \boldsymbol{U}_{PCA}\boldsymbol{S}_{PCA}\boldsymbol{A}_{ICA} \tag{3.8.1.5}$$

Utilising $\boldsymbol{B}$, the contribution of each map to each independent component can then be analysed, while $\boldsymbol{G}$ can be used to establish correlations with the various nIDPs in UK Biobank, utilising the methods described above.

3.8.2 Results

For the results in this section, the full subject datasets were used, rather than using only the second half as in the previous sections. This was done as the larger population should yield better results by providing the PCA-ICA algorithm with a larger number of samples for signal decomposition. The algorithm was run in both the subject and map directions.

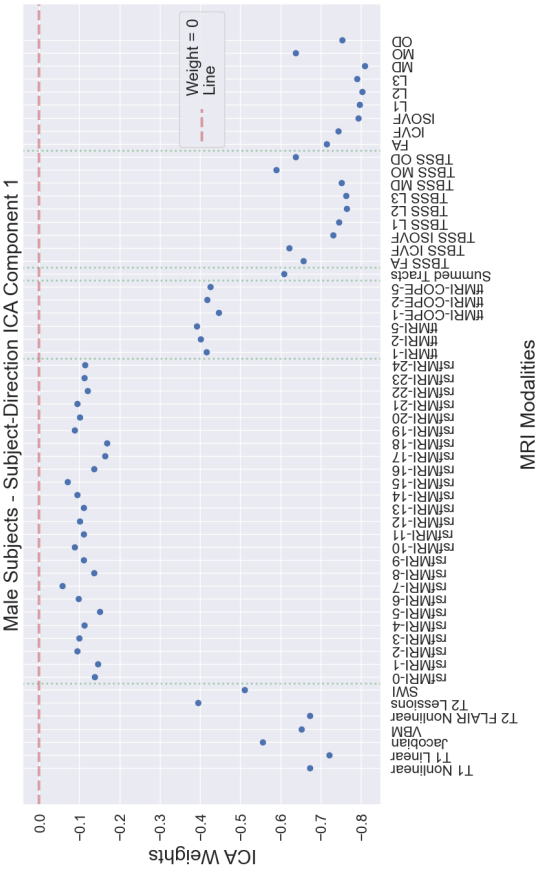
When the input brain age matrix was decomposed in the subject direction, 2 potential modes of population variations were identified, for both the male and female subject groups, with the map-specific weights associated to each component being presented in Figure 3.34. When looking at similarities in the map direction for each of the 2 modes of population variation (i.e., the U_{PCAICA} matrices in Equation 3.8.1.5) between the two sexes, high correlation values were observed, with the absolute Pearson r for the first components being 0.99 ($-log(p) = 48.63$) and for the second components 0.934 ($-log(p) = 25.61$). For both male and female subjects, the first component contains little information from the rsfMRI maps,

with the highest weightings being assigned to dMRI-maps, such as ISOVF, L1-to-3, MD and MO. In the case of the second component, the highest weights are assigned to rsfMRI maps, yet these components contain a mixture of information from all maps taken under consideration.

When considering the associations between the two components for each sex group with nIDPs from UK Biobank (Figures 3.35-3.36), differences can be observed between the significant associations of each component, as well as between the male and female subjects. While the second component shows stronger associations to nIDPs from the Physical and Skeletal Measurements variable categories, the first component correlates strongly with nIDPs from categories such as Diet, Lifestyle, Medical History and Tobacco consumption and Cardiac Measurements.

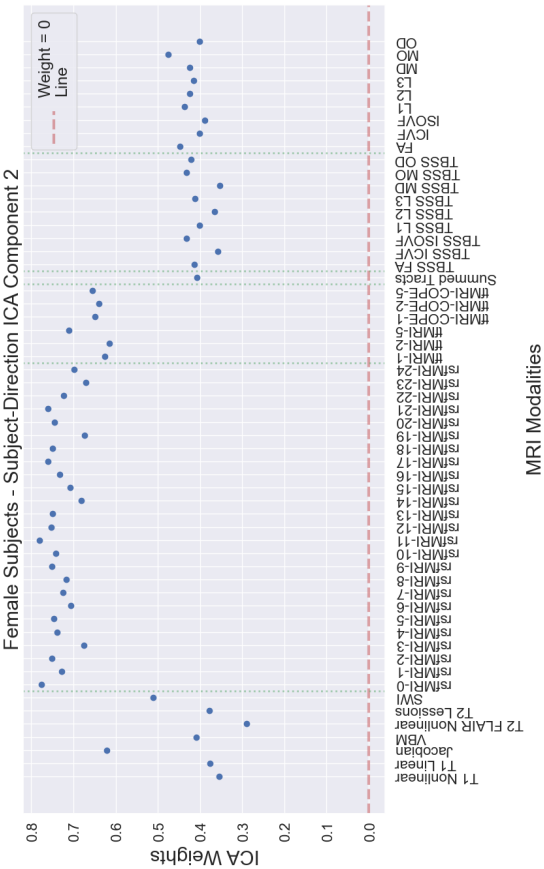
To investigate if these observations are indeed related to different underlying biological processes, the distributions of the separation matrices for the two components were compared for both male and female subjects. Figure 3.37 displays this information, where it can be observed that both curves approach Gaussian distributions. To assess the degree of normality, the Fisher Kurtosis and p-values corresponding to the Shapiro-Wilk test for normality were calculated. For normal distributions, the kurtosis values are expected to be close to 0, while $-log(p)$-values are expected to be lower than 1.3. The calculated kurtosis values were $\{0.152, 0.272\}$ and $\{0.588, 0.195\}$ for female and male subjects, respectively, while the Shapiro-Wilk test p-values were $\{6.446, 3.961\}$ and $\{12.602, 5.979\}$ respectively. Although the Shapiro-Wilk tests indicate that the distributions are not Gaussian, the Fisher Kurtosis reveals only minor differences between the tail weights of the two component distributions relative to the centre of their distributions.

As mentioned above, an identical set of tests were also carried out in the map direction. In this case, the results were not informative, with the independent components identified by the generative model simply corresponding to individual maps in the original brain age predictions array.

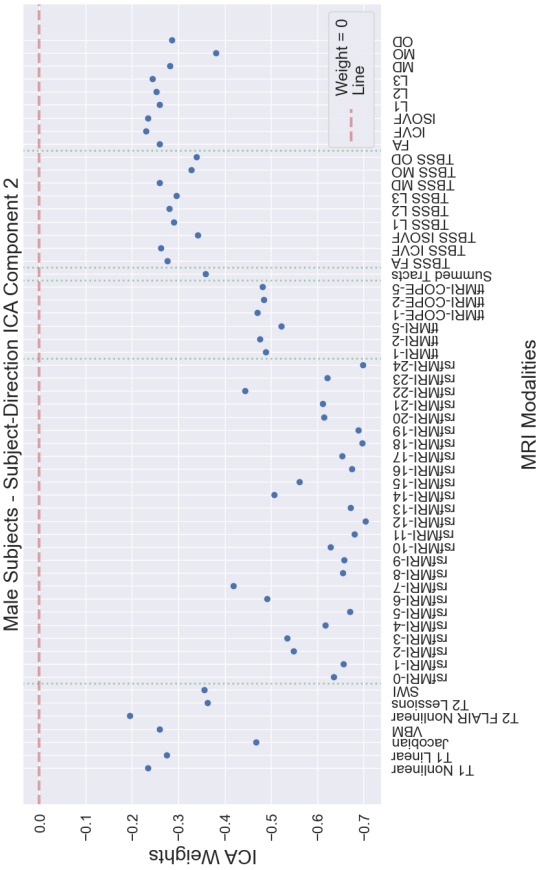


(a) Weights of ICA component 1 for female subjects

(b) Weights of ICA component 1 for male subjects

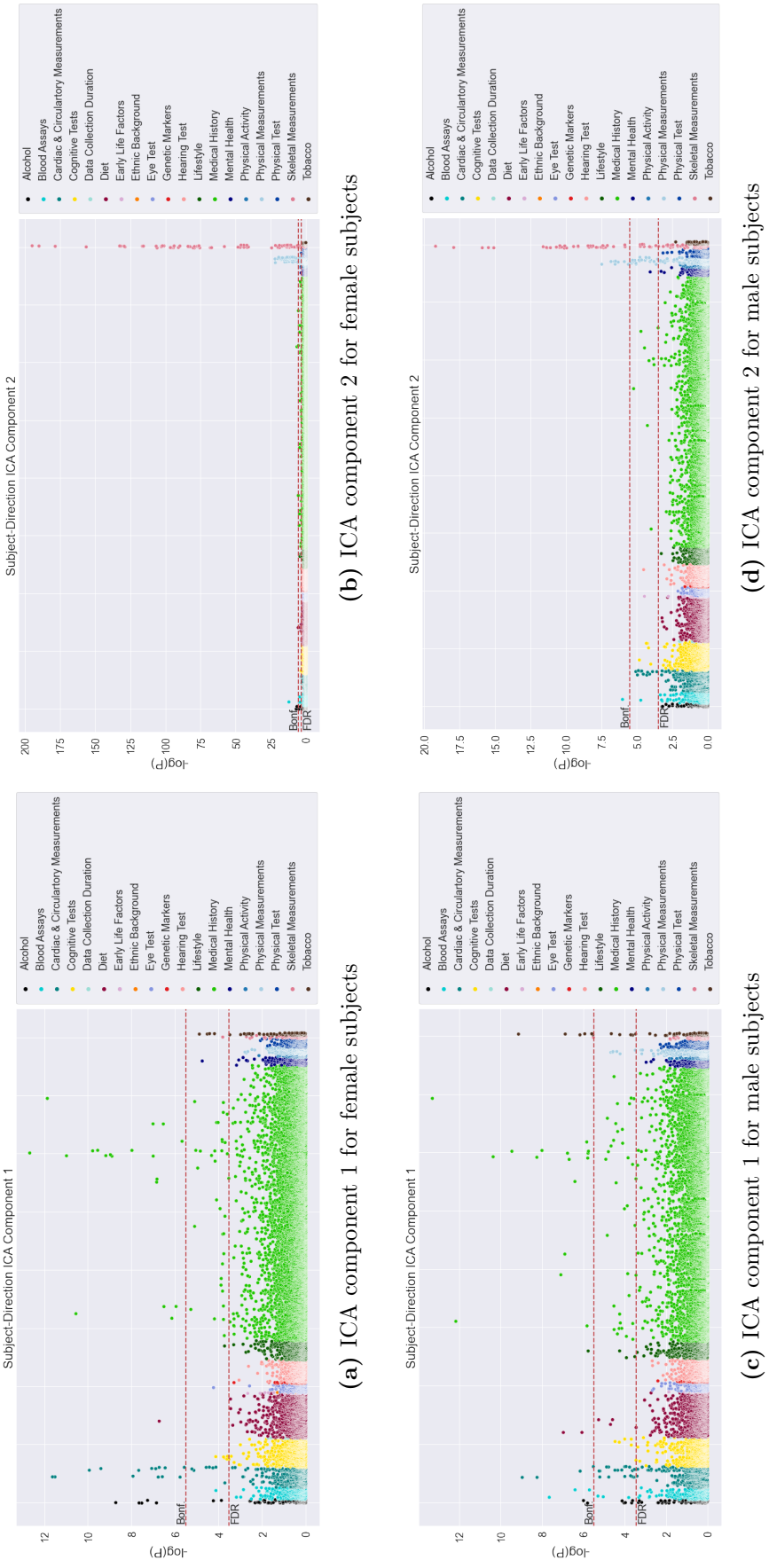


(c) Weights of ICA component 2 for female subjects



(d) Weights of ICA component 2 for male subjects

Figure 3.34: Map-specific weights associated to each subject-direction ICA component for **(a)-(c)** female and **(b)-(d)** male subjects. These weights correspond to the U_{PCAICA} matrix in Equation 3.8.1.5. The signs of the weights are not relevant. The red dashed line indicates the case where weights are equal to 0.



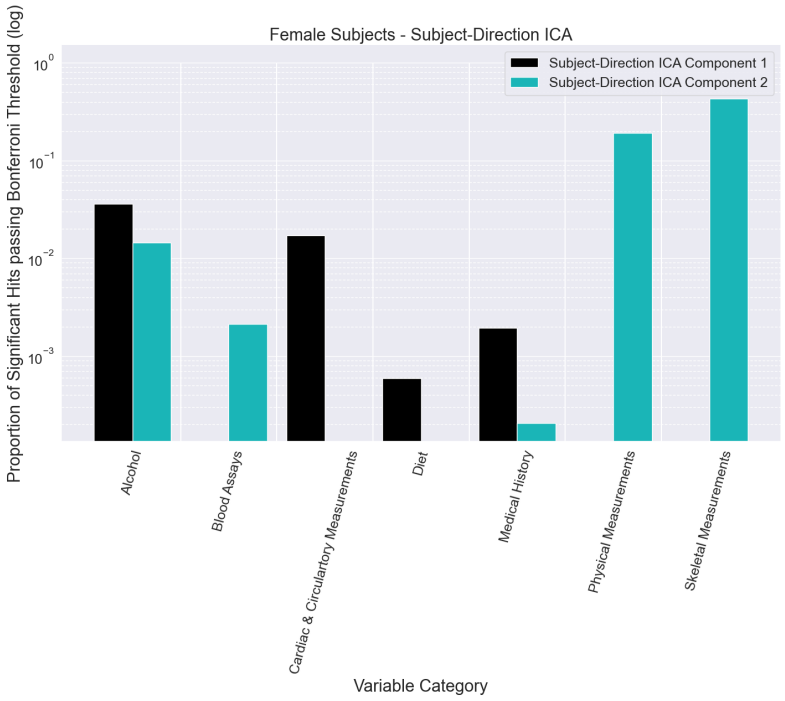
(a) ICA component 1 for female subjects

(b) ICA component 2 for female subjects

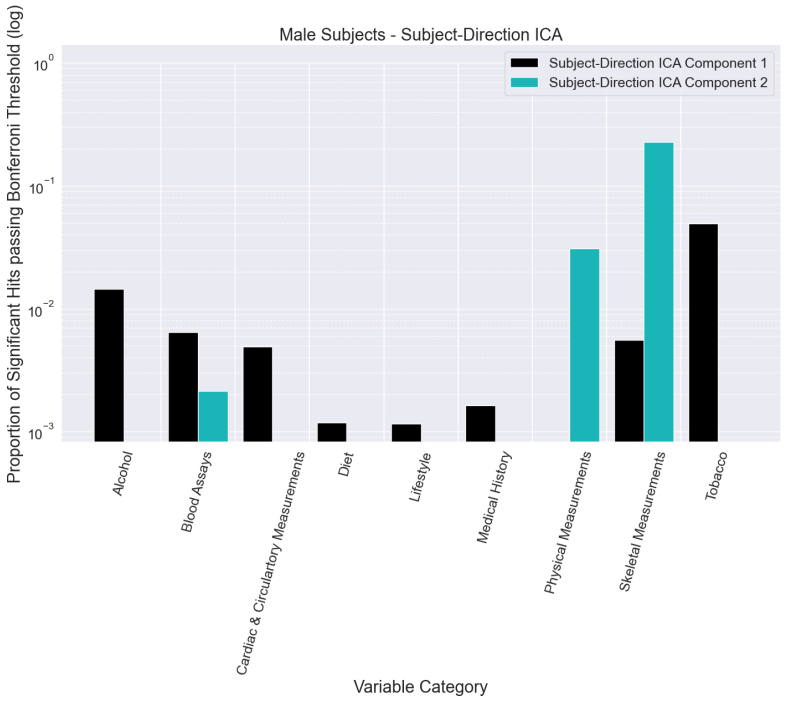
(c) ICA component 1 for male subjects

(d) ICA component 2 for male subjects

Figure 3.35: Manhattan plots relating subject-direction ICA brain age deltas to UK Biobank nIDPs for **(a)-(b)** female and **(c)-(d)** male subjects. These weights correspond to the S_{ICA} matrix in Equation 3.8.1.4. The False Discovery Rate (FDR) and Bonferroni threshold are also plotted. As these results are calculated for the full test set, they should not be directly compared to the rest of the results in this chapter.

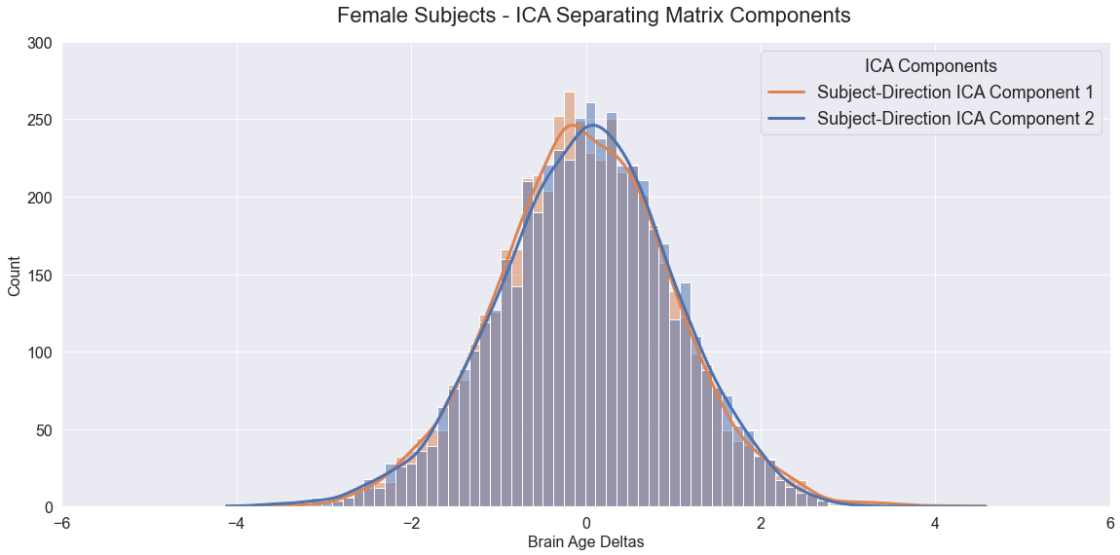


(a) Subject-direction ICA for female subjects

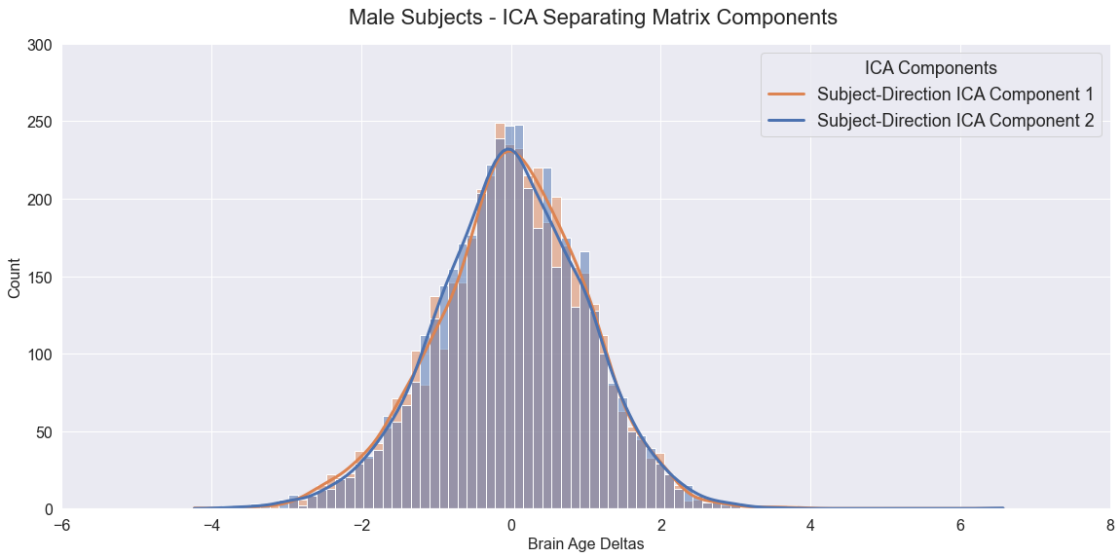


(b) Subject-direction ICA for male subjects

Figure 3.36: Proportion of significant hits per UK Biobank nIDP category for subject-direction ICA components calculated by dividing the number of variables passing the Bonferroni threshold by the total number of variables in that category. **(a)** shows results for female subjects, while **(b)** indicate results for male subjects.



(a) Female subject separation matrix distribution



(b) Male subject separation matrix distribution

Figure 3.37: Separation matrix distributions for subject-direction ICA components for **(a)** female and **(b)** male subjects. The separation matrices correspond to the S_{ICA} matrix in Equation 3.8.1.4.

3.8.3 Discussion

While no interesting results emerged from the map-direction decomposition, two distinct modes of population variability were identified when analysing the data in the subject-direction. Observations suggest these could represent two distinct patterns of brain changes, one correlated more with *"intrinsic"* biological processes, such as those impacting Skeletal and Physical Measurements, while another relating to *"extrinsic"* processes, such as those relating to various pathologies. However, subsequent observations of the separation matrices obtained from subject-direction ICA, together with the results of Shapiro-Wilk and Fisher Kurtosis statistical tests, indicate that, although not Gaussian, there are very small differences between the distributions of the two ICA components.

ICA aims to find the underlying coherent trends in the deltas, by assuming that the deltas associated with each map represent a mixture of some underlying consistent mechanisms. Thus, it is expected that the deltas coming from the individual maps have a Gaussian distribution, but the underlying signals to have a less Gaussian, more tail-heavy distributions, or completely non-Gaussian.

When performing signal decomposition with ICA, if the resulting independent signals appear to be less Gaussian, then this lends credence to them being a true and coherent underlying source signal in the original distribution. The fact that the signals are close to being Gaussian suggests several things: firstly, that ICA might not be the most optimal approach for performing the intended decomposition and dimensionality reduction; and secondly, that the independent components, and by association the correlations, are driven by a small number of outlier subjects.

For a stronger case to be made in favour of the two modes of variation identified, more non-Gaussian distributions would have been expected, with a larger weighting towards the tails, even potentially suggesting the clustering of subjects into distinct sub-populations. By no means does this nullify the obtained results, but it suggests that they might not be as insightful as originally hoped. It is possible that the use of other methods, such as ElasticNet, might represent a better approach for exploring trends across the populations considered in this work.

3.9 Conclusion

This chapter has presented an in-depth exploration of brain ageing, and the relationship between brain age deltas (or deviations from normal ageing trajectories) and lifestyle and biomedical measurements from UK Biobank (nIDPs) in a single-map and multi-map ensemble context. A multi-faceted approach was taken to evaluate various aspects of this complex relationship, focusing on the effect of map selection, ensemble strategy, and population-based variability on the strength and significance of nIDP associations. Overall, different maps were found to encode bespoke information about the ageing brain, with 191 unique statistically significant associations being found between brain age deltas and UK Biobank nIDPs. Of these, 140 came from maps not often used for brain ageing predictions.

In the case of multi-map ensembles, the investigation of the influence of map selection on nIDP-brain age delta associations highlighted the complex and often nuanced nature of these relationships. It was demonstrated that different maps have varying levels of sensitivity towards specific nIDPs. Moreover, the findings stressed the importance of map choice and the potential for selection bias in analyses of this kind, where certain maps might offer strong associations for some nIDPs, while offering weak or no associations for others. The multi-map ensemble analyses further underscored this complexity, showing that the information captured by different maps can both complement and conflict with each other.

The exploration of ensemble strategies offered valuable insights into the balance between predictive accuracy and interpretability. While data-blending techniques, such as ElasticNet, were shown to achieve the best mean absolute error (MAE), these approaches often resulted in a reduction of significant nIDP associations. It was highlighted that weighted ensembling can enhance overall brain age prediction but may come at the expense of eliminating possibly valuable biophysical information. Approaches like non-linear ensembling or carefully guided map selection may alleviate this issue to a certain extent, by leveraging the denoising effect of agreeing maps while retaining biologically meaningful information.

The chapter concluded with an evaluation of population-based variability using Independent Component Analysis (ICA). While two distinct modes of variation were initially identified, further examination indicates these modes may not be sufficiently non-Gaussian to support the claim that two distinct population subgroups were identified. This underscores the need for further research into more robust methods for population trend analysis, such as ElasticNet.

The insights and limitations identified throughout this chapter inform a natural progression into the next area of research, focusing on multi-modal deep fusion learning. This method offers a potential solution to some of the limitations identified in this chapter, particularly in relation to the handling of multi-map data and population variability. Deep fusion learning techniques offer the opportunity to extract richer, more complex representations from neuroimaging data, by blending and fusing information from multiple maps at multiple levels during model training. This could potentially lead to enhanced brain age prediction accuracy while also retaining important biological information across different maps.

In conclusion, while this chapter has uncovered important insights into the complex interplay between nIDPs, brain age delta, and imaging maps, it is evident that further exploration is required. The complexity and richness of neuroimaging data, as well as the intricate biology it reflects, necessitate advanced analytical techniques capable of effectively capturing and interpreting this complexity. Moving into the next chapter, the hope is that multi-modal deep fusion learning may offer such a solution, pushing forward the understanding of brain ageing and the factors influencing it.

4

Multi-Modal Deep Fusion Learning

Contents

4.1 Overview

The previous chapter established that brain ageing is a complex phenomenon affecting multiple structural and functional aspects of the brain. Combining brain age predictions obtained with convolutional neural networks (CNNs) trained on various 3D neuroimaging modalities was observed to improve brain-ageing prediction accuracies and amplify signals of interest, such as associations between brain ageing deltas and neurobiological factors (referred to as non-imaging derived phenotypes - nIDPs). However, it was observed that the chosen ensembling strategy can influence the predictive accuracy and interpretability of results, with linear methods sometimes achieving the highest predictive accuracies at the cost of potentially valuable confound information.

Non-linear ensembling methods, such as multi-layer perceptrons, however, were found to potentially alleviate the reduction in significant nIDP associations seen for linear methods, with only small increases in the reported mean absolute error (MAE). Building on this observation, this chapter tests the hypothesis that more complex non-linear ensembling methods, such as multi-modal deep fusion neural networks, can enhance the richness of information extraction from neuroimaging data by exploiting between-modality complementary features. Rather than ensembling the brain age predictions obtained with single-map networks, deep fusion networks are trained simultaneously with the volumetric information from multiple maps, internally fusing their features before making a single prediction.

In this chapter, several fusion strategies were evaluated in a regression context, differing based on the stage at which the data from different modalities is combined. The results of these fusion networks were then compared to those obtained in the previous chapter by linearly ensembling single-map brain age predictions using ElasticNet. Despite the promising hypothesis that multi-modal fusion might enhance brain age prediction, the fusion experiments reported below fell short of the equivalent ElasticNet linear methods' performance. By manipulating the number of identical networks used in fusion, some improvements in brain age prediction

accuracy were noted, but deep learning fusion results still trailed behind ElasticNet ensembles in terms of prediction accuracies.

These findings led to two distinct further research avenues: an investigation into brain-ageing model convergence and the continuation of multimodal fusion work. I first demonstrate that, to achieve a 95%-convergence rate and stable brain age predictions, averaging across approximately 9 identical networks is necessary for both single- and multi-map experiments. This observation has significant implications for brain age model deployment in clinical settings, where the stability and reliability of predictions are paramount. Yet, despite this advance, the deep fusion networks could still not outperform the ElasticNet ensembles in terms of brain age predictions and nIDP associations.

This unexpected result underlines the importance of balancing complexity with effectiveness. This realisation calls for a broader discussion on the practical implementation of complex deep learning methodologies in neuroimaging studies and emphasises the need for innovative approaches that harmonise computational efficiency with predictive accuracy. I believe that such pragmatic approaches, that prioritize efficiency and applicability, are required for transitioning work such as this in a clinical setting.

4.2 Introduction

In the previous chapter, complex relationships were identified between non-imaging derived phenotypes (nIDPs) and brain age deltas derived from different neuroimaging maps. It was revealed that different maps encode unique brain ageing information and exhibit varying sensitivities to specific nIDPs. When ensembling multiple maps, it was also found that the chosen ensembling strategy can induce certain biases. While linear ensemble strategies like ElasticNet generated the highest accuracies (measured as the lowest mean absolute errors - MAEs), they often reduced significant nIDP associations. This indicated a trade-off between prediction accuracy and the preservation of biologically relevant information. Conversely, non-linear methods such as multi-layer perceptrons alleviated this issue to a degree, albeit with lower

prediction accuracies. To address this discrepancy, this chapter explores the use of multi-modal deep fusion learning for brain age prediction.

As discussed in the previous chapter, the simultaneous use of multiple MRI modalities or maps can be beneficial in the study of brain ageing, as a richer set of inputs could capture more age-related variability than single-input methods. This is due to the fact that multi-input methods are capable of learning complementary between-map features [81]. This consideration has prompted several studies to explore the use of multiple modalities for brain age predictions (Appendix A, Table A.7). Similar to single-map literature, these studies can be divided into two categories based on the form taken by their inputs. The majority of multimodal studies employ as inputs image derived phenotypes (IDPs) obtained from sMRI, dMRI and rsfMRI, which are used to train linear methods for brain age prediction [35, 56, 75, 78, 79, 81, 132, 138, 298, 299]. Non-linear methods were also explored, yet they produced results similar to those reported by the linear methods [260]. These studies found that the use of multi-map inputs led to higher accuracy results than in the case of single-map methods. In addition, they reinforce the findings presented in Chapter 3, which suggest that sMRI derived inputs generally produce the best single-map results. Recently, several studies have also utilised 2D and 3D volumetric inputs and CNNs for multi-modal predictions [74, 82, 83, 160]. These generally follow a similar approach to that employed in Chapter 3, where single-map networks are first trained to convergence, after which their results are ensembled using either naive [74] or linear methods of varying complexity [82, 83, 160]. Despite all of the volumetric studies representing extensive bodies of work into multi-modal brain age prediction, they all suffer from a fundamental limitation: none of them have engaged in deep fusion of modalities or maps. Rather, they have focused on initial individual network training followed by ensembling, potentially missing complex, intrinsic relationships between different maps that might be captured through a more integrated fusion approach.

Deep fusion methods can, however, be found in other biomedical applications of deep learning, such as survival time predictions [300, 301], classification [302–304],

and image segmentation [76, 305–309]. The work done in segmentation allows for the identification of three main strategies for information fusion in networks:

- **Input-level Fusion**, where several volumetric inputs are fused channel-wise at the input to a network, enabling the learning of a fused feature representation. This is the fusion method chosen by most segmentation studies, as it allows the information and correlations between different inputs to be exploited, from an early stage, by all the layers of the network. It also reduces the computational complexity of the network. However, this method could lead to the loss of some imaging modality-specific information and it requires that the inputs have the same spatial resolution [310]. Examples of such networks are given by [311, 312], as well as [313], which, rather than carrying out segmentation, shows how input fusion can be utilised for generating efficient latent embeddings.

- **Layer-level Fusion**, where individual feature extractors are trained for each of the individual inputs, with the learned features being fused at a later, intermediate layer in the network. This approach has the advantage that it can preserve imaging modality-specific information while learning a joint representation that captures the between-input interactions at a later level. This allows both the individual feature extractors, and the overall network, to learn more complex feature representations between the maps. However, this comes at the price of higher network complexity and computational demands, as well as the need for careful consideration on the nature of the fusion strategy [310, 314]. Examples of such fusion strategies can be found in [315, 316].

- **Decision-level Fusion**, where a single network is trained for each input, with the outputs of each network being integrated at the output of the fusion network. This approach has the advantage that it can leverage the power of each network for each input separately, achieving maximum exploitation of the unique information encoded within each input. The disadvantages of this approach, however, are that little room is left for learning complementary

information between the original image inputs, while the computational requirements are the highest of all three fusion strategies [310, 314]. Examples of Decision Fusion can be found in [317, 318].

Yet, given the conceptual and mathematical differences between these methods, previous studies have found it difficult to recommend one method over another [305, 308, 309, 319]. Moreover, no comparison work between the various fusion strategies has yet been carried out for regression problems. Thus, for a holistic picture, in this work, all three fusion strategies are considered.

Based on these observations, in this chapter, the following questions are addressed:

1. Can deep fusion architectures be used effectively for regression tasks?
2. Which fusion methods perform better in a regression context?
3. How do the various deep fusion approaches compare against post-training linear methods in terms of accuracy and associations with nIDPs?
4. Can any augmentations, such as using pre-trained single-map networks, improve the predictive performance of deep fusion networks over conventional end-to-end trained similar networks?

To answer these questions, in this chapter, I test whether, for progressively complex combinations of 18 different 3D maps derived from 5 core MRI modalities, brain age prediction in healthy subjects can be improved by using different deep fusion strategies. I first propose 6 different 3D fusion networks: 2 Input, 3 Layer, and 1 Decision. To facilitate comparisons between the results presented in this chapter and those presented in Chapter 3, the deep fusion network architectures will be based on the HGL CNN architecture. The results from these fusion models are then compared against the ElasticNet ensembling method, which was identified as the best in terms of prediction accuracies in Chapter 3. The fusion and linear methods are evaluated and compared in terms of brain age prediction accuracies

and associations with UK Biobank nIDPs, with the overarching hypothesis being that the deep fusion networks will outperform the equivalent ElasticNets across all map groups. I also test several training augmentation strategies, such as refinement and transfer learning, to determine if using pre-trained single-map networks could help improve the training and prediction performances of the deep fusion networks.

4.3 Common Methods

This section introduces the common methods utilised throughout this chapter, discussing the proposed HGL CNN-based deep fusion architectures and experimental setup. It should be noted that, as this chapter represents a continuation of the previous work introduced in Chapter 3, some overlap exists in the methodologies employed. Thus, readers are invited to consult the relevant information in Section 3.3. Here, details on the following topics can be found: the utilised neuroimaging and nIDP data, the formation of the female and male train, validation and test datasets, the employed preprocessing steps, the baseline HGL architecture, the approach employed for calculating the brain age deltas and the methods used to associate these to nIDPs. Furthermore, Section 3.7.1 introduces the ElasticNet ensembling technique and details how its hyperparameters were determined using cross-validated grid searches. For the purposes of this chapter, only the female subjects group was utilised, given that it is larger than the male subject group (15691 vs. 13640), and the expectation that the obtained results would be very similar between the deep fusion and linear ensembling experiments.

4.3.1 Deep Fusion Networks

For this work, six 3D deep-fusion Convolutional Neural Networks (CNNs) were constructed. This section provides an overview of these fusion methods, which employ a similar architecture to the HGL CNN, with only minor modifications designed to facilitate data fusion.

4.3.1.1 Input-Level Fusion

The input-level fusion strategy dictates that a single multi-channel input is presented to the network. To satisfy this requirement, several 3D maps are first fused channel-wise into a 4D input, which is then presented to a standard HGL architecture (Figure 4.1).

A variation of this architecture was also tested, referred to as *Input Fusion with Filter*, where, before being passed to the standard HGL architecture, the fused 4D input is first transformed to a 3D latent feature using a $1 \times 1 \times 1 \times n_{channels}$ 3D convolution. This transformation aims to condense the information from the multiple n channels into a single, more concise 3D latent feature representation. By applying this convolution before the standard HGL architecture, the network might be guided to focus on the most salient and complementary features from the various maps, potentially reducing noise and enhancing its discriminative ability. This could also improve the interpretability of the learned representations.

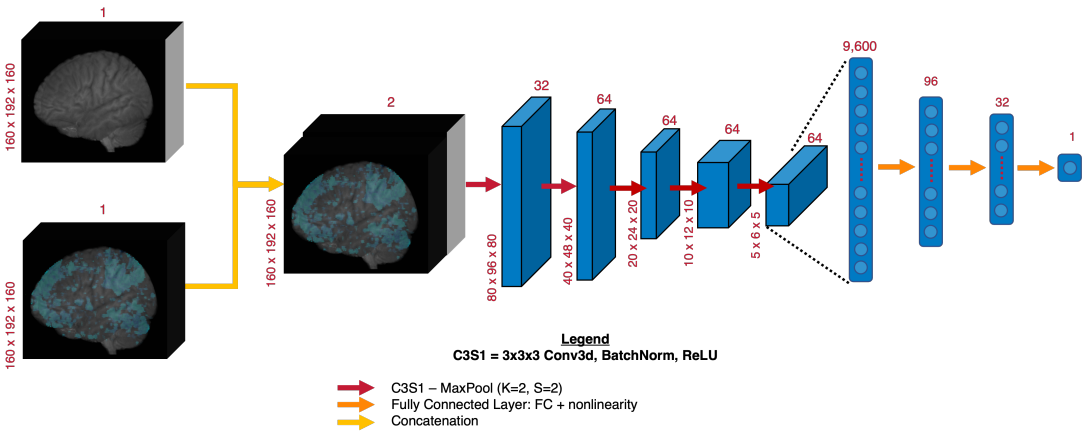


Figure 4.1: Input fusion architecture utilised in this chapter. In the case of the *Input Fusion with Filter* model, the architecture remains the same, with the exception of the addition of a $1 \times 1 \times 1 \times n_{channels}$ 3D convolution operation, without any normalisation, nonlinearity or downsampling, between the 4D input and the HGL architecture.

4.3.1.2 Layer-Level Fusion

In the layer-level fusion strategy, individual maps are passed through independent feature extractors, after which the latent representations are fused at a deeper level

in the network. Segmentation literature generally proposes that fusion should occur between latent features both at the level of the feature extractors and for the dense layers [316]. This works well in segmentation tasks with a fixed number of inputs, however, it can lead to scaling and data bleeding issues when designing regression architectures, which should work with an arbitrary number of maps.

For these reasons, the layer fusion architectures used in this work maintain independent feature extractor paths, and perform fusion only in the dense layers. Figure 4.2 depicts a *Layer Fusion 9600* network, which fuses (concatenates) the latent information before the first fully-connected layer. Two more layer fusion architectures were proposed: *Layer Fusion 96* and *Layer Fusion 32*, which concatenate the latent features before the 2^{nd} and 3^{rd} fully-connected layers, respectively.

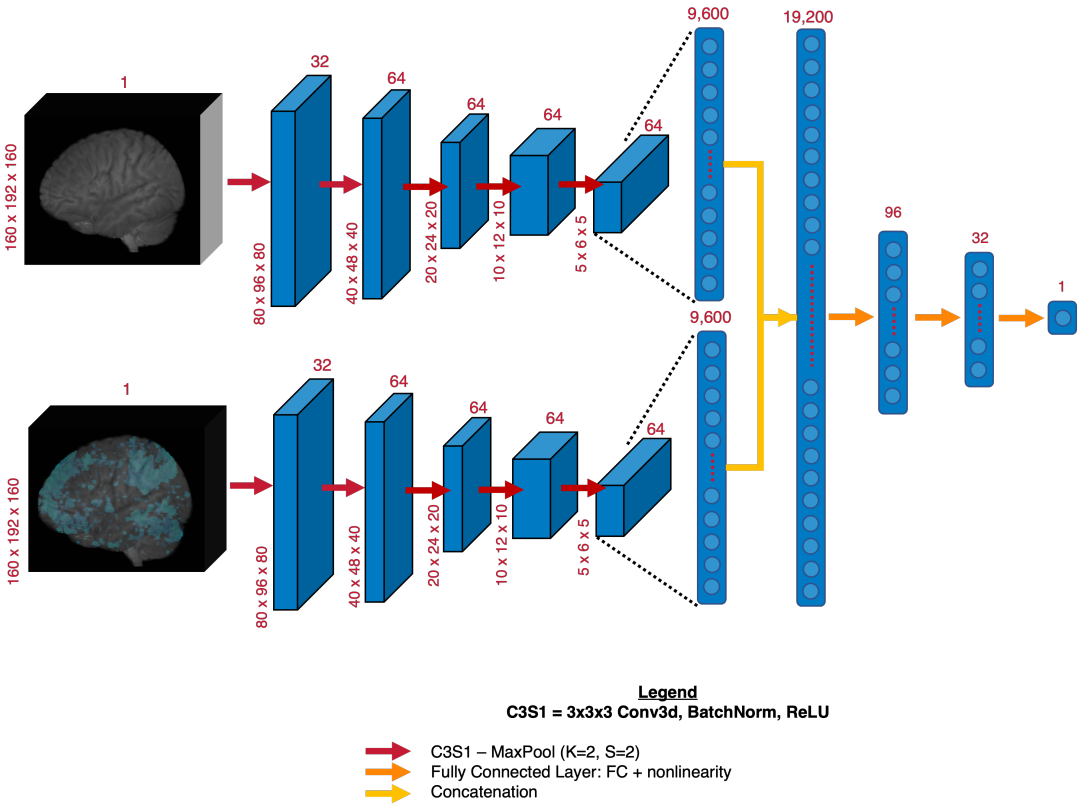


Figure 4.2: Layer Fusion 9600 architecture utilised in this chapter. For the *Layer Fusion 96* and *Layer Fusion 32* networks, the concatenation steps occur later in the network, before the second and third fully-connected layers respectively.

4.3.1.3 Decision-Level Fusion

Finally, the decision-level fusion proposes that each input map trains a single regression network, with feature fusion occurring at the output of these networks. The fused outputs are then either averaged or passed through a linear layer to produce the final fusion network result. Conceptually, this is very similar to the linear ensembling approach taken in Chapter 3, the difference lying in the fact that this fusion network is trained end-to-end. Figure 4.3 shows the proposed decision-level fusion architecture used in this chapter, which uses a fully connected linear layer for the final output operation.

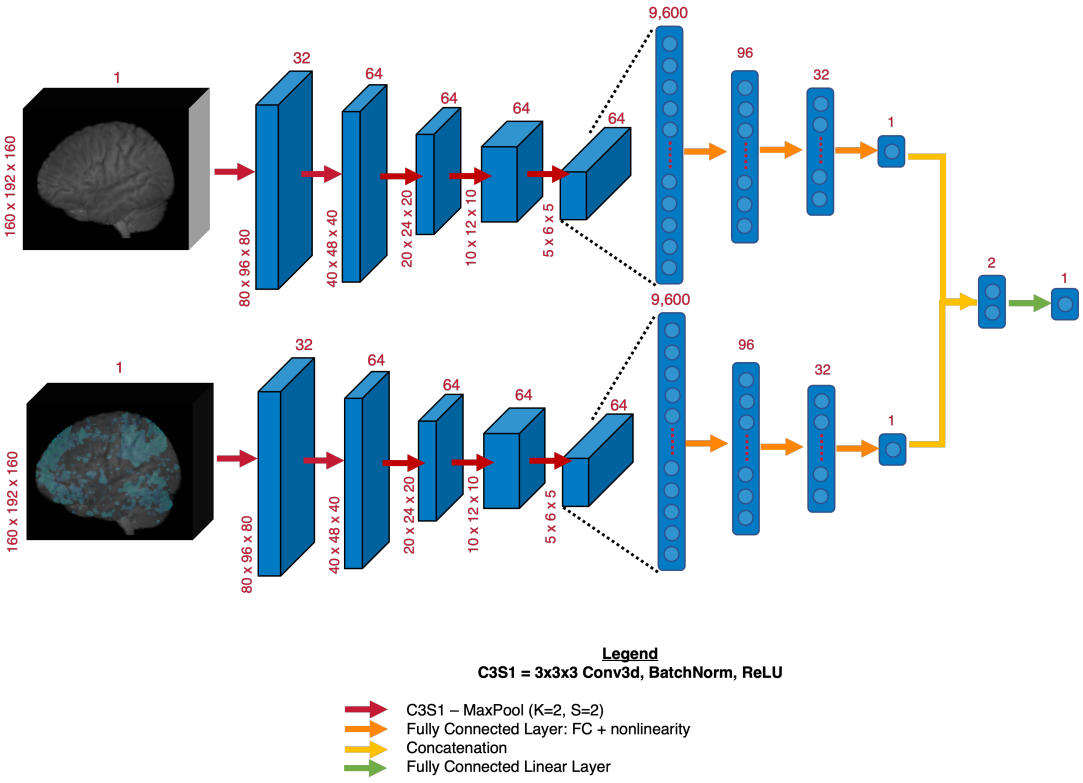


Figure 4.3: Decision Fusion Architecture used in this chapter.

4.3.2 Experimental Setup

To facilitate easier comparison between the HGL and ElasticNet based ensembling methods described in the previous chapter, and the deep fusion methods presented

Table 4.1: Relationship between batch sizes and number of inputs for the various deep fusion networks. These limitations were dictated by the hardware. For all experiments, the largest achievable batch size was always used.

Network Name	Number of Trainable Parameters (2 inputs)	Maximum Number of Inputs (based on batch size)				
		Batch = 12	Batch = 8	Batch = 6	Batch = 4	Batch = 3
Input Fusion	1.31 M	9	12	18	30	30+
Input Fusion w. Filter	1.31 M	9	12	18	27	30+
Layer Fusion 9600	2.62 M	-	2	3	4	5
Layer Fusion 96	2.62 M	-	2	3	4	5
Layer Fusion 32	2.62 M	-	2	3	4	5
Decision Fusion	2.62 M	-	2	3	4	5

herein, a similar experimental setup was maintained to that in Chapter 3 Section 3.3. There were, however, some notable modifications.

Firstly, for each single-map and deep fusion experiment, a set of n identical networks were trained separately to convergence. Originally, and based on literature [21, 67, 74, 82, 83, 160], the value of n was set to 3. Test set predictions were averaged at the subject level post-training, reducing noise induced by random weight initialisation and stochastic training mechanisms.

Then, while the original single-map HGL networks were trained using batches of 12 maps, this is not necessarily possible for the fusion experiments due to the higher computational requirements. Thus, the batch sizes were adjusted as per Table 4.1, depending on the number of input maps.

Compared to Chapter 3, the number of maps used for the various experiments was reduced to 18, corresponding to the 5 core MRI modalities:

- 3 sMRI maps;
- 1 swMRI map;
- 5 resting-state fMRI (rsfMRI) ICA dual-regressed z-score normalised maps;
- 1 z-statistic task fMRI (tfMRI) map;
- 8 diffusion MRI (dMRI) maps, including:
  - 4 quantifying aspects of water diffusion and microstructural properties;
  - 3 representing TBSS skeletonised features;

  - 1 summation of 27 major tracts obtained with probabilistic tractography [41].

To test the various deep fusion architectures, a series of progressively complex experiments was designed. These ranged from presenting the fusion architecture with very similar maps, such as two sMRI maps both containing structural information, to challenging the fusion methods with a group of highly dissimilar maps, such as one composed of a map from each core MRI modality. Following this approach, the following clusters were defined:

- **Cluster 1**: T1 Nonlinear (sMRI) + T2 FLAIR Nonlinear (sMRI);
- **Cluster 2**: T1 Nonlinear (sMRI) + FA (dMRI);
- **Cluster 3**: MD (dMRI) + rsfMRI-0 (rsfMRI);
- **Cluster 4**: T2 FLAIR Nonlinear (sMRI) + SWI (swMRI) + rsfMRI-0 (rsfMRI) + tfMRI-1 (tfMRI) + Summed Tracts (dMRI).

In Section 3.7.1, two different ensembling techniques were proposed: one that ensembled all available maps, and another that utilised associations to nIDPs as a guide to map selection. The latter approach found that careful map selection can lead to an increase in the structured information encoded by the delta, which related to the nIDP used to guide the selection in the first place. To verify if this observation holds true for the deep fusion networks, three additional clusters were defined using Table 3.7 as a guide:

- **Cluster 5**: MD (dMRI) + rsfMRI-0 (rsfMRI) + MO (dMRI) + TBSS ISOVF (dMRI) + L2 (dMRI);
- **Cluster 6**: rsfMRI-2 (rsfMRI) + rsfMRI-10 (rsfMRI) + rsfMRI-5 (rsfMRI) + rsfMRI-21 (rsfMRI);
- **Cluster 7**: T2 Lesions (sMRI) + TBSS FA (dMRI) + TBSS L3 (dMRI) + TBSS ISOVF (dMRI).

4.4 Initial Comparison Between Fusion and Ensembling Results

This section describes a preliminary set of experiments carried out to test the performance of various fusion networks. The working hypotheses were that the fusion networks would outperform both their component maps and the equivalent ElasticNet ensemble in terms of brain age prediction accuracy, measured using the Mean Absolute Error (MAE). Given the preliminary nature of these tests, they were carried out using only Clusters 1 and 2, which contain maps displaying a high and moderate degree of similarity, respectively.

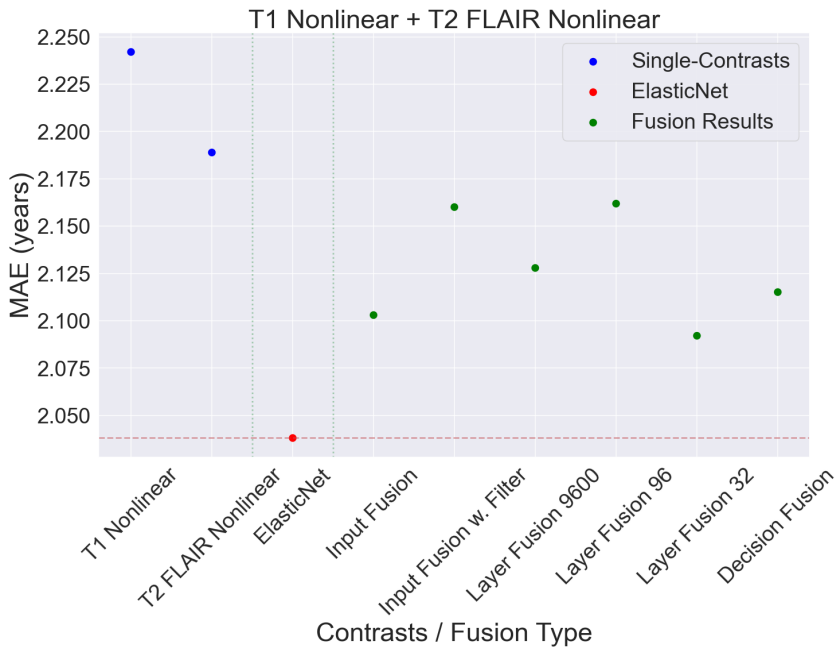
4.4.1 Results

The fusion results for the first two clusters are presented in Figure 4.4. It can be observed that, in all cases, the ElasticNet produces more accurate brain age predictions than any of the fusion results. Moreover, in the case of Cluster 2 (Figure 4.4b), the two Input Fusion networks and the Decision Fusion produce similar or even worse results than the T1 Nonlinear single-map input.

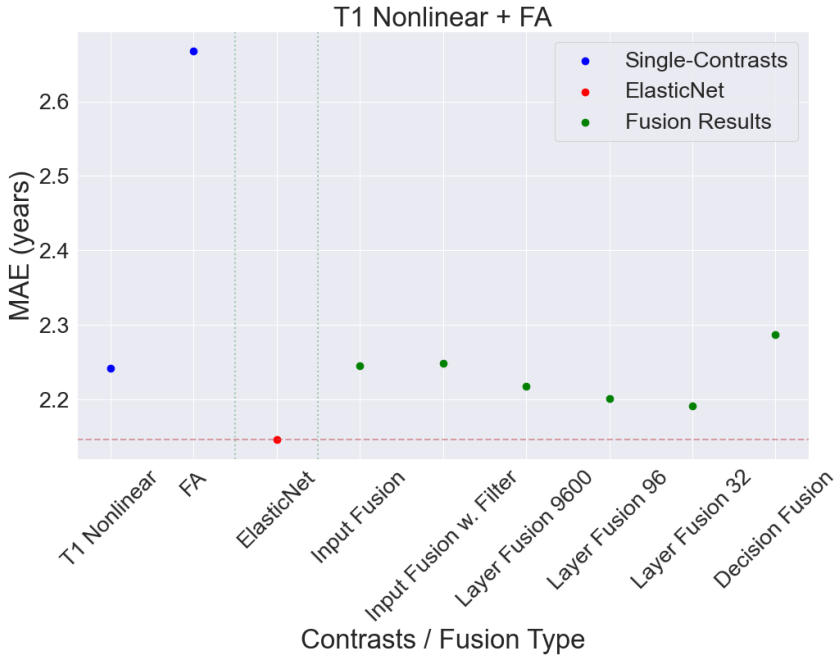
4.4.2 Discussion

Considering the results presented above, the initial comparison between the proposed fusion networks and ElasticNet ensembling has revealed some unexpected findings. Contrary to the hypothesised advantage of fusion networks in terms of prediction accuracy, little to no difference can be observed between the ElasticNet ensembling and the deep fusion networks in the preliminary experiments conducted on Clusters 1 and 2. This indicates that, at least for these specific cases characterised by high and moderate degrees of similarity, the fusion methods were not able to surpass the equivalent linear ensembling technique in terms of MAE.

Furthermore, within the fusion networks themselves, the results do not indicate any clear model as being superior to the others in terms of performance. Moreover, both the Input Fusion networks and Decision Fusion produced similar or even mildly inferior outcomes compared to the T1 Nonlinear single-map input in the



(a) Cluster 1 MAE Distributions



(b) Cluster 2 MAE Distributions

Figure 4.4: Preliminary fusion results MAE distributions. The plots show the MAEs of the single-map HGL networks for each component, the MAE of the ElasticNet post-training linear ensemble utilising the cluster map components, and the results of the various deep fusion algorithms.

case of Cluster 2, contrary to the hypothesis that ensembling and fusion networks always perform better than their components.

These preliminary observations prompt a reassessment of either the architectures themselves or the training strategies employed. Given that the goal of this chapter is to establish a direct comparison to the results obtained in the previous chapter, this limits potential architecture changes. Thus, the focus of the upcoming section will be on enhancing the training strategies for the deep fusion networks.

4.5 Deep Fusion Training Augmentation

Building on the previous observations, this section proposes two training augmentation strategies designed to enhance the performance of the deep fusion networks. The first augmentation strategy, based on the assumption of some overlap between the features learned by the single-map and fusion networks, employs transfer learning to aid the deep fusion network in learning more efficiently. The second strategy assumes that the ElasticNet ensembling method effectively "sees" more data. This assumption stems from the fact that the ElasticNet has access to additional training subjects and uses inputs from more identical networks than the fusion experiments. The first part of this section defines the two augmentation strategies. Afterward, the results for each case are presented and discussed. As this section aims to evaluate general trends, only the maps in Cluster 1 were utilised.

4.5.1 Methods

4.5.1.1 Transfer Learning

Transfer learning is a powerful technique in deep learning, often used for leveraging pre-trained networks as the starting point for a new investigation on a related task. This approach is based on the idea that the features learned by the pre-trained network can be used as a basis for learning new tasks, rather than randomly initialising a new network and training it end-to-end. This usually leads to increased efficiency in terms of data and training resource usage, as well as performance improvements when tasks are closely related [320]. Given the close similarity

between the single-map work presented in the previous chapter and the current fusion work, this last consideration motivates the use of transfer learning to solve the problem of underperforming fusion networks relative to their equivalent ElasticNet.

Given how the single-map networks were trained, transfer learning can only be applied to the Layer and Decision Fusion architectures. Using Figure 4.4 as a guide, transfer learning was attempted using the Layer Fusion 9600 and Layer Fusion 32 networks, representing examples of early and late fusion, respectively. Training was carried out using the left-out test subjects also utilised for fitting the ElasticNet, over a maximum of 100 training epochs, with a stopping patience of 10 epochs and an initial learning rate of $1e-3$. The pre-trained weights were selected from the best-performing single-map networks.

Overall, three different transfer learning strategies were attempted. Firstly, assuming that the fusion and single-map networks learn very similar features, the pre-trained weights were kept frozen during training. Then, assuming that certain changes also occur in the pre-trained weights during fusion training, a test was carried out where they were kept unfrozen throughout the training. Finally, a test was carried out where the transfer learning networks were trained with frozen weights until convergence, after which the weights were unfrozen, and training resumed until convergence again.

4.5.1.2 ElasticNet Alignment

The training augmentation attempted comprises two modifications based on empirical observations regarding the ElasticNet. Firstly, it can be assumed that the ElasticNet "sees" more training subjects. This is because it is fitted using the first half of the left-out test dataset. Thus, the first proposed modification is to refine-train each fusion network using the additional subjects. The larger sample size from the same distribution should lead to improved performance, as the models are exposed to a more detailed view of the underlying data patterns. To test the impact of applying refinement training when fusing at different network depths, this augmentation was applied to the Input Fusion, Layer Fusion 9600, and Layer Fusion 32 architectures.

The second proposed modification is based on the observation that the ElasticNet "sees" the outputs from $3n$ single-map networks, where n represents the number of maps. As described in Section 3.3.2, for each network configuration (be it single- or multi-map), a set of 3 identical networks are trained separately, with predictions being averaged at the subject level post-training (Figure 3.3). These averaged predictions are then used to subsequently train the ElasticNet. Thus, for a 2-map case, it can be said that the ElasticNet receives inputs from 6 networks. Based on this heuristic, the second modification proposes training an additional 3 fusion networks, besides the original 3 already trained. To account for the additional subjects the ElasticNet "sees", all 6 fusion networks were also refine-trained. Given the computational requirements of training additional networks, these modifications were only applied to the Layer Fusion 32 network, which is the best-performing network so far for Cluster 1.

4.5.2 Results

4.5.2.1 Transfer Learning

As shown in Figure 4.5, all of the transfer learning strategies resulted in performance degradation compared to the baseline layer fusion architectures. Of the three strategies, the one in which the weights remained unfrozen performed the best. However, its predicted accuracy was still lower than not only the end-to-end trained layer fusion architectures, but also the individual single-map results.

To try to understand what causes this behaviour, the weights of the single-map networks were compared against those of the end-to-end and transfer learning fusion networks (Figures 4.6-4.7). This comparison involved calculating the Euclidean Distance and Mean Squared Error (MSE) between the fusion networks on one side, and the baseline single-map networks on the other. This was performed separately for the T1 Nonlinear (Figures 4.6a and 4.7a) and T2 FLAIR Nonlinear (Figures 4.6b and 4.7b) data paths. The Euclidean Distance is useful for comparing the weights of different networks when the learned filters maintain their relative positions to each other. This is the case for the experiments performed with

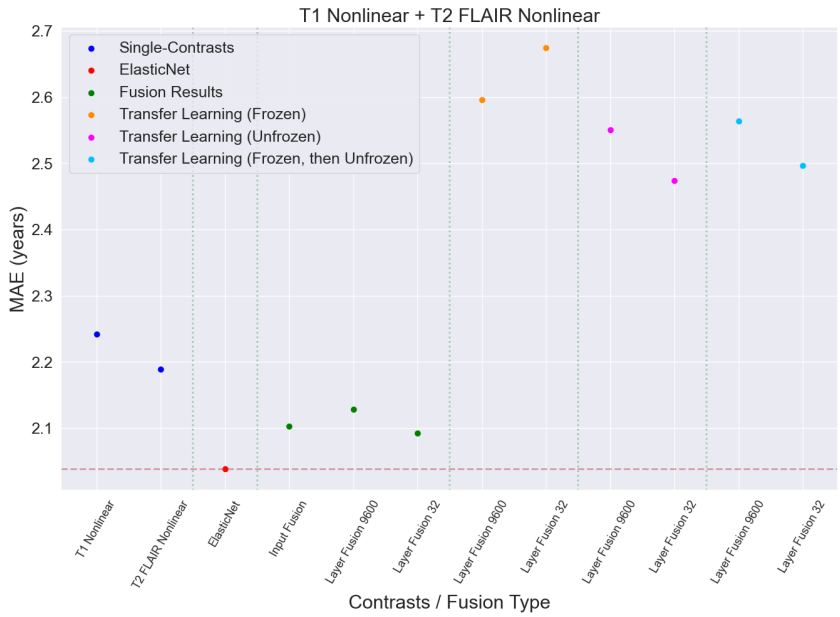


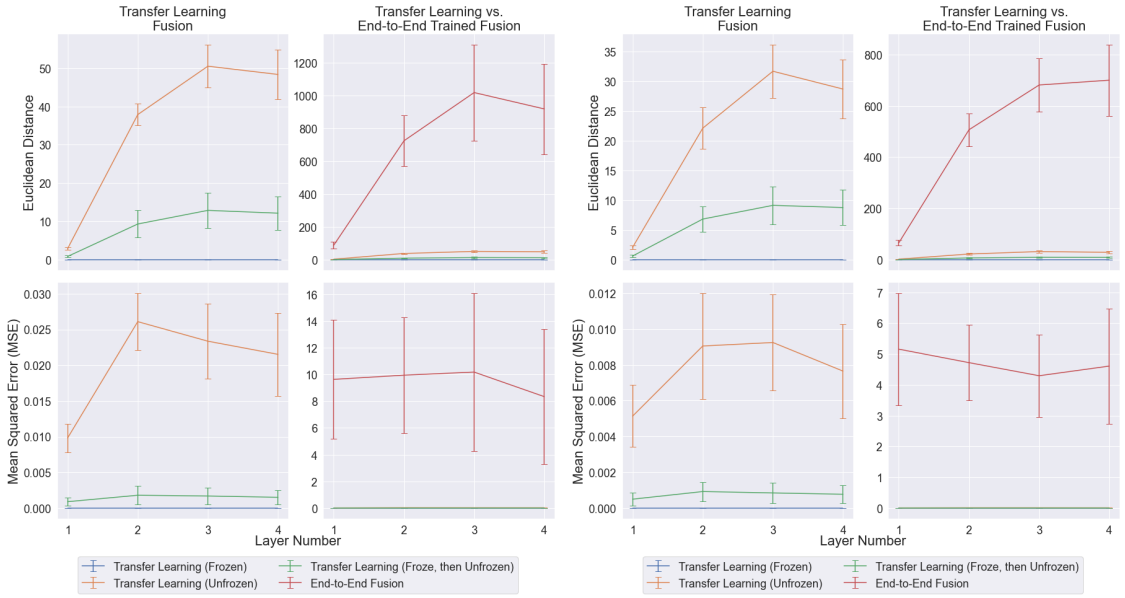
Figure 4.5: Fusion with transfer learning results, in terms of MAE, for the three proposed strategies.

transfer learning. However, this cannot be assumed when comparing the single-map and the end-to-end fusion networks, which is why, for those cases, the MSE plots should primarily be considered.

Overall, significant differences can be observed between the end-to-end fusion network and any of the transfer learning networks. This suggests that the end-to-end fusion networks are learning very different features compared to the single-map and transfer learning approaches. Figure 4.7 suggests that these differences decrease as the parameter space is reduced.

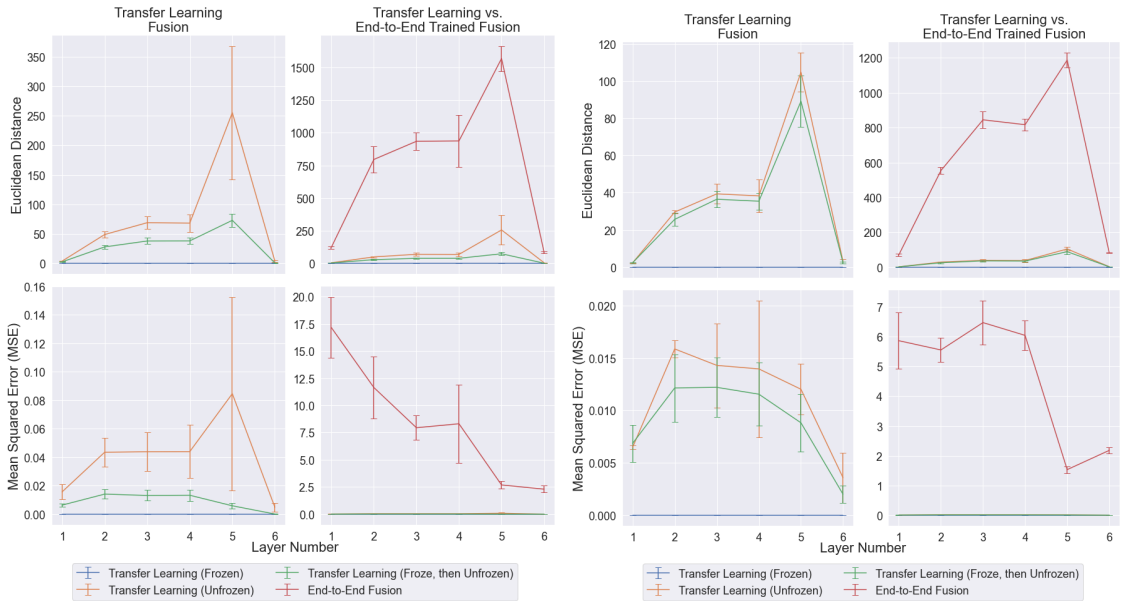
4.5.2.2 ElasticNet Alignment

The results for the ElasticNet-inspired augmentations are presented in Figure 4.8a. While the refinement training seems to lead to minor improvements over the baseline fusion results, the addition of a greater number of identical networks manages to surpass the ElasticNet. Figure 4.8b illustrates the reduction in MAE with the addition of more identical network runs. It also reveals that, even for 6 identical runs, the reported brain age predictions are still not entirely independent of the number of runs performed for a single network configuration. This holds



(a) T1 Nonlinear **(b)** T2 FLAIR Nonlinear

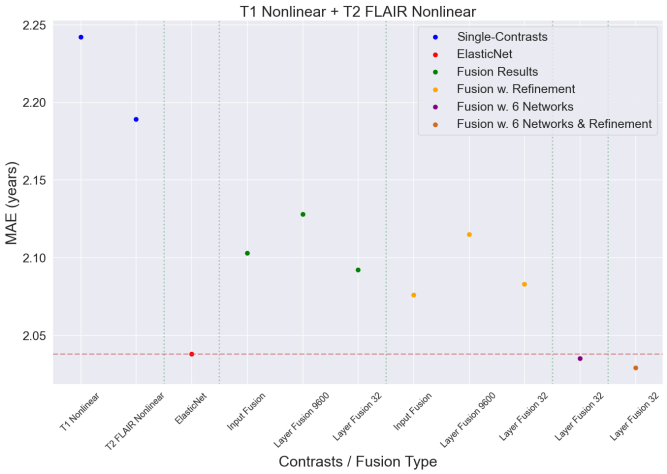
Figure 4.6: Layer Fusion 9600 distance calculations between the baseline single-map network and the various end-to-end and transfer learning fusion experiments. The Euclidean Distance and Mean Squared Error (MSE) were calculated separately for the **(a)** T1 Nonlinear and **(b)** T2 FLAIR Nonlinear data paths.



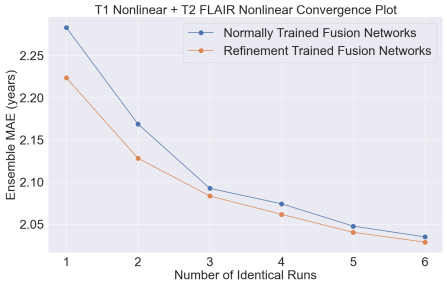
(a) T1 Nonlinear **(b)** T2 FLAIR Nonlinear

Figure 4.7: Layer Fusion 32 distance calculations between the baseline single-map network and the various end-to-end and transfer learning fusion experiments. The Euclidean Distance and Mean Squared Error (MSE) were calculated separately for the **(a)** T1 Nonlinear and **(b)** T2 FLAIR Nonlinear data paths.

true for both the vanilla Layer Fusion 32 network, and the one which benefits from additional refinement training.



(a) ElasticNet augmentation results



(b) MAE convergence plots

Figure 4.8: ElasticNet alignment results showing the **(a)** prediction accuracies achieved using several proposed modifications, and **(b)** a number of runs independence study for the Layer Fusion 32 network.

4.5.3 Discussion

This section's pursuit of improving the predictive capabilities of the proposed deep fusion networks has revealed both promising strategies and unanticipated challenges.

The transfer learning experimental results showed an unexpected degradation in performance compared to end-to-end training. This is counter to the common expectation that, in closely related tasks, transfer learning typically leads to increased predictive performance [320]. This may be due to the features extracted by the single-map networks being either too complex, or not the most relevant for the fusion architecture. Therefore, the model trained end-to-end seems to learn more pertinent features, thus achieving better prediction accuracies. This observation could also lend credence to the idea that the fusion networks are learning different, potentially complementary, features. While the unfrozen case does show some degree of weight modification compared to the frozen baseline, the substantial differences between the end-to-end and transfer learning networks suggest that

entirely new features are being learned. These findings underline the complexity of leveraging pre-trained models across similar, yet distinct tasks.

The most interesting and significant results were obtained from the experiments attempting to align the fusion and ElasticNet methods. The primary finding of this section is that the addition of more identical network runs can enhance the performance of the fusion networks, enabling them to surpass the ElasticNet. This provides insight into the importance of model convergence in brain age prediction.

However, this also introduces a fundamental question: *"At what point do brain age prediction results become independent from the number of trained identical networks?"*

Answering this question is crucial, as operating with non-converged models has significant implications, particularly in a clinical setting. Predictions influenced by random noise may lead to inconsistencies between similar methods, potentially undermining confidence in results. Past studies [21, 53, 67, 74, 82, 83, 160] reveal that no clear consensus exists in terms of how many identical networks should be trained for each sub-ensemble, with numbers ranging from 3, for most authors, to 5. Furthermore, none of these past studies provide a justification for selecting a particular number over another. This, coupled with the observation that even 6 identical runs do not completely eliminate this random noise component, underscores the need for a convergence study, to determine the optimal number of identical networks to ensemble.

4.6 Brain Age Prediction Convergence

In light of the results observed in the previous section, before we continue the investigation of deep fusion networks, the question of prediction convergence needs to be addressed. Answering this question has implications for both the current work and the validation of the results presented in the previous chapter. In this section, the following questions will be addressed:

- How many identical network runs are necessary to achieve result convergence? Are these numbers similar for single-map and deep fusion networks?
- How consistent are the results from single-map and deep fusion when comparing between converged and non-converged experiments?
- What impact does convergence have on result stability?

To answer these questions, the first part of this section introduces the methodology for evaluating convergence and comparing results obtained from converged and non-converged experiments. Following this, a convergence investigation is carried out for several single-map networks, one fusion network, and several ElasticNet ensembles using the single-map results. Converged and non-converged results were compared both in terms of their summary statistics and their associations with nIDPs.

4.6.1 Methods

4.6.1.1 Identical Network Independence Study

To determine at what point brain age prediction results become independent from the number of trained identical networks, achieving what is referred to as "MAE convergence", experiments were carried out with three single-maps (T1 Nonlinear, T2 FLAIR Nonlinear, and FA - the latter to observe if non-sMRI maps display any different behaviours), as well as with the map Cluster 1 (T1 Nonlinear + T2 FLAIR Nonlinear) and the Layer Fusion 32 architecture. The latter was selected as, based on the preliminary findings, it appears to be the best-performing fusion network so far. As the fusion experiments were deemed more complex, 15 identical runs were made with the Layer Fusion 32 network, while only 12 were carried out with the single-map networks.

Three different metrics were utilised to evaluate network convergence:

- **Simple ensemble convergence**, where the number of results averaged at the subject level is progressively increased each time a new network is run;

- **Combinatorial convergence**, where rather than sequentially adding outputs to the ensemble for averaging, at each stage, corresponding to the number of networks in that group, all possible combinations are considered (Equation 4.6.1.1, where n is the number of identical runs, and k the number runs chosen at each step). The reported result is the average and standard deviation of all the combinations;

$$C(n,k) = \frac{n!}{k!(n-k)!} \tag{4.6.1.1}$$

- **Bootstrapping convergence**, which is similar to Combinatorial convergence, but with the difference that when selecting sample runs from the set of identical runs, replacement is allowed. For each step, $10k$ repetitions were used to determine the statistical distribution and estimate the mean and standard deviation.

In addition to determining the point of result independence, consideration also needs to be given to the computational requirements of running so many identical jobs for each experiment. Given this, the number of training epochs was also tracked, and a convergence threshold was established at 95% of the best value. This threshold is used to both prevent diminishing returns in terms of computational resource usage and to acknowledge the fact that, due to the stochastic nature of CNNs, some result variability will always be expected.

In addition to the independence study above, for the experiments with the single-map networks, the stability benefits of model convergence were also investigated. This was done by generating all the viable $n - by - n$ group pairs of identical runs, without overlap between the elements comprising each group, and ensembling them using simple averaging. Given the 12 identical runs available, this means that n was in the range $(1, 6)$. For each of the group pairs, to quantify the stability benefits coming from progressively larger ensembles of identical networks, the following metrics were calculated:

- The Pearson correlation between the ensemble brain age deltas;

- The number of subjects which change their predicted brain age delta sign between the ensembles within the pair;
- The average absolute value of the delta change for those subjects exhibiting a sign change.

4.6.1.2 Comparing Original and Converged Results

Once an acceptable answer to the question of convergence is found, the subsequent step is to compare a subset of the results presented in Chapter 3, referred to as "original" results, to the new identical network independent results, referred to as "converged". This comparison was conducted both in terms of the summary statistics surrounding the brain age predictions and the associations with UK Biobank nIDPs.

These comparisons were undertaken for all the 18 single maps utilised in this chapter, as well as the 7 ElasticNet ensembles which correspond to the map clusters defined earlier. A list of the single maps grouped by modality can be found in Table 4.3, while the cluster groups are defined in Section 4.3.2 above.

The purpose of these comparisons is to identify any differences between the original and converged results and to validate the findings of Chapter 3. If any major differences or discrepancies are found between the converged and original results, they will be further investigated, using methods such as permutation testing.

4.6.2 Results

4.6.2.1 Identical Network Independence Study

The results of the Layer Fusion 32 network's independence study are presented in Figure 4.9, with the single-map networks yielding similar outcomes. As shown in Table 4.2, which presents the numerical results for the four experiments, $\approx$ 95% convergence is usually obtained after $7 - 9$ identical runs, depending on the experiment and chosen metric. Beyond these points, any additional runs produce diminishing returns in terms of expended computational resources and improvements in prediction accuracy. Additionally, around this point, as shown by Figure 4.10, the performance improvement brought about by refinement learning

becomes negligible. This indicates that, when using 9 identical runs per experiment, a point has been reached where the predictions can be assumed to be independent of the number of identical networks trained.

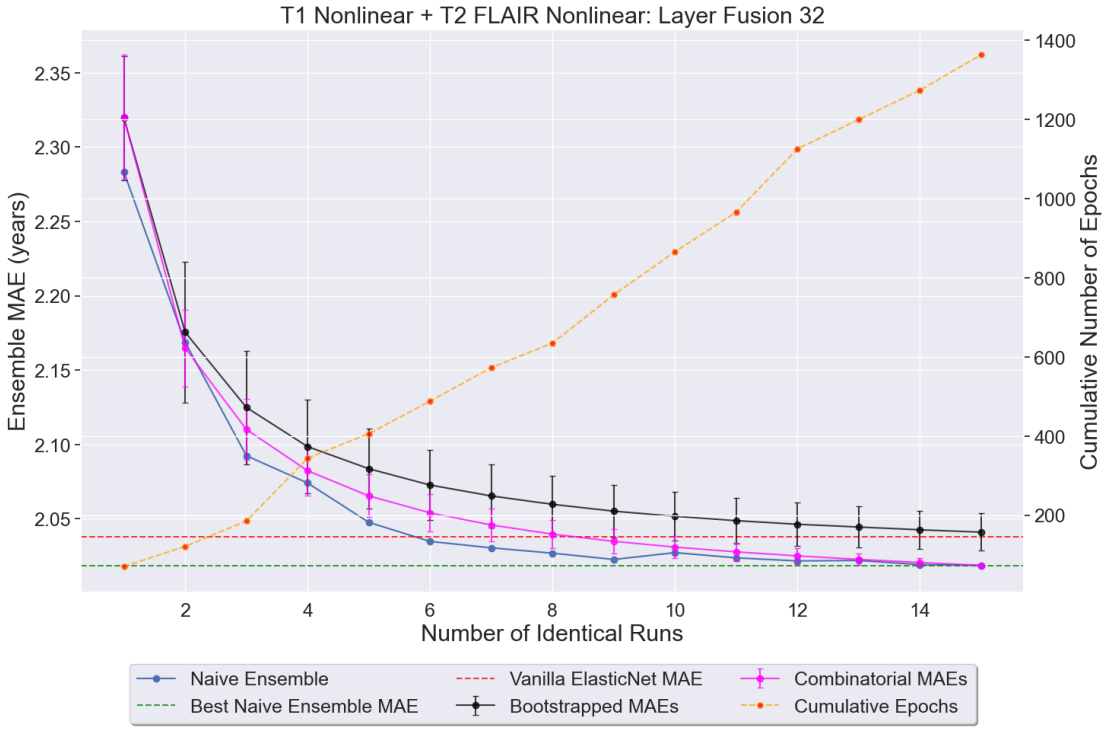
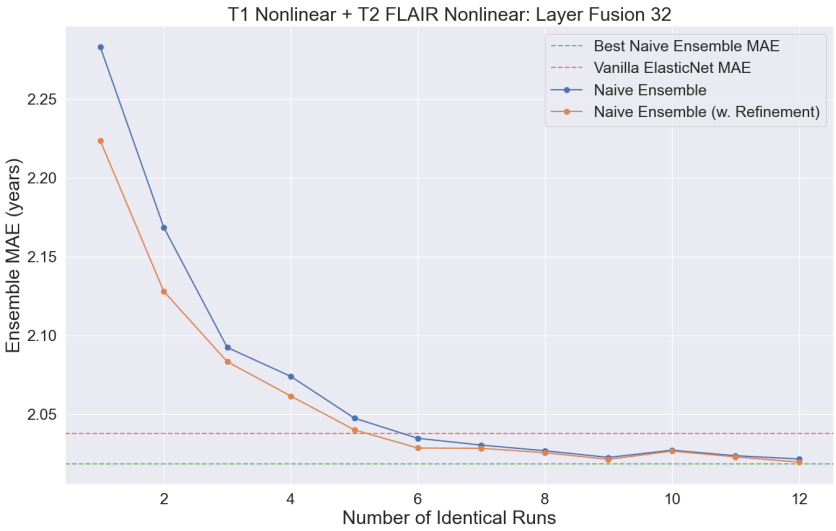


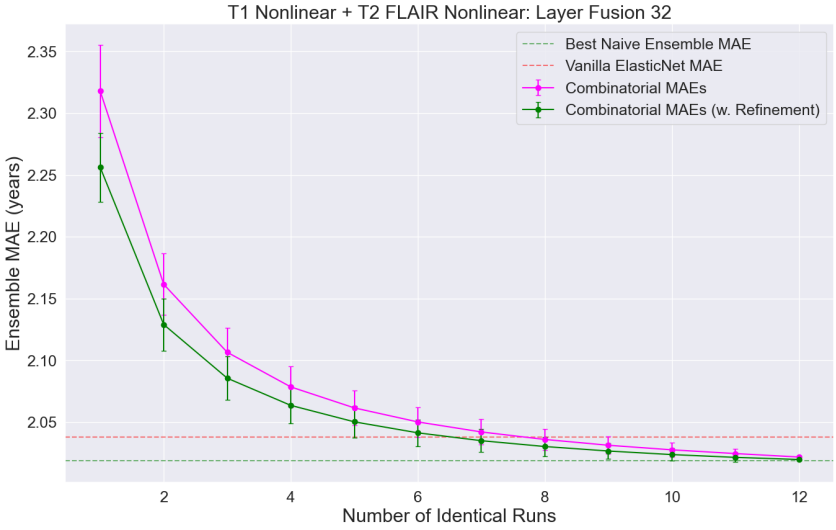
Figure 4.9: Convergence curves of the Layer Fusion 32 network obtained during the independence study. The various MAE curves represent different measures of convergence, while the dotted orange curve indicates the number of training epochs required for that particular number of independent runs. The horizontal dotted lines indicate the best MAE achieved during the independence study, and the MAE obtained with the 3-run ElasticNet from previous sections.

The stability benefits of convergence can be readily observed in Figure 4.11, which shows how larger ensembles of identical networks become progressively more consistent in their predictions. While experiments were limited to identical network ensembles composed of a maximum of 6 networks due to computational overheads, it can be seen that:

- The Pearson correlation improves from ≈ 0.74 for a single network, to ≈ 0.89 for 3-network ensembles, and ≈ 0.94 for 6-network ensembles (Figure 4.11a);



(a) Simple ensemble convergence curves



(b) Combinatorial ensemble convergence curves

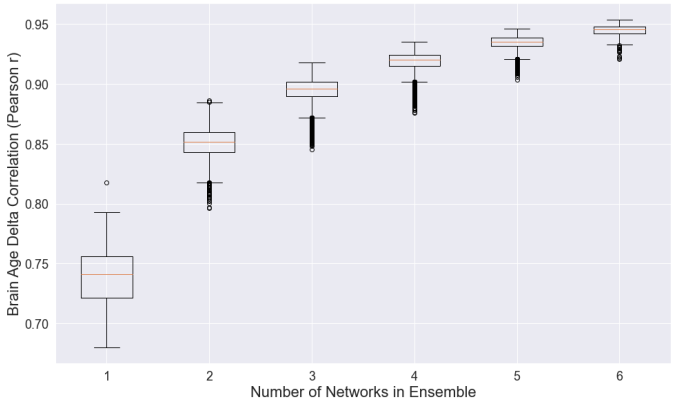
Figure 4.10: Refinement learning convergence curves of the Layer Fusion 32 network obtained during the independence study. The plots represent the various MAE curves obtained with **(a)** naive and **(b)** combinatorial ensembling. The horizontal dotted lines indicate the best MAE achieved during the independence study (red), and the MAE obtained with the 3-run ElasticNet from previous sections (green).

Table 4.2: Identical network independence study numerical results. For each experiment, and each convergence metric, both the MAE and Convergence (Conv.) results are presented. The Convergence results represent, in percentages, how close an MAE value is to the best MAE achieved by that experiment. These, together with the curves presented in Figure 4.9, are an indication of the diminishing returns of running additional jobs. The point where each convergence metric reaches a convergence of ≈ 95% is displayed in bold.

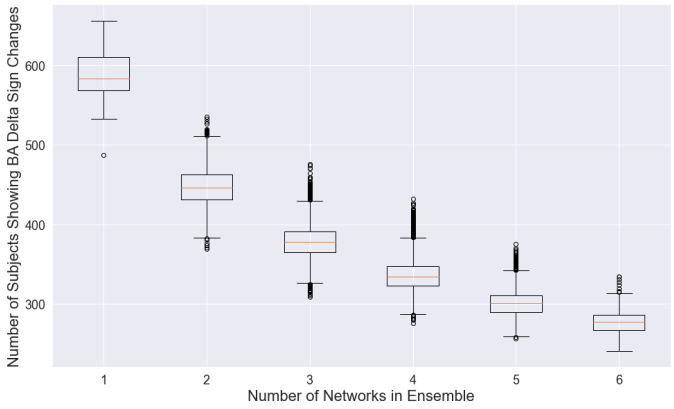
Number of Networks in Ensemble / Ensemble Type	T1 Nonlinear + T2 FLAIR Nonlinear						T1 Nonlinear					
	Naive		Combinatorial		Bootstrapped		Naive		Combinatorial		Bootstrapped	
	MAE (years)	Conv. (%)	MAE (years)	Conv. (%)	MAE (years)	Conv. (%)	MAE (years)	Conv. (%)	MAE (years)	Conv. (%)	MAE (years)	Conv. (%)
1	2.283	0.0	2.32	0.0	2.319	0.0	2.49	0.0	2.457	0.0	2.458	0.0
2	2.169	0.433	2.165	0.515	2.175	0.517	2.344	0.47	2.311	0.528	2.322	0.534
3	2.092	0.721	2.11	0.697	2.126	0.695	2.263	0.731	2.259	0.713	2.276	0.717
4	2.074	0.791	2.082	0.789	2.099	0.793	2.255	0.756	2.233	0.808	2.253	0.807
5	2.048	0.891	2.065	0.845	2.083	0.848	2.246	0.786	2.217	0.865	2.238	0.865
6	2.035	0.939	2.054	0.883	2.073	0.884	2.237	0.816	2.206	0.903	2.228	0.905
7	**2.031**	**0.955**	2.046	0.91	2.066	0.911	2.218	0.877	2.199	0.931	2.221	0.931
8	2.027	0.969	2.04	0.93	2.06	0.932	2.203	0.926	**2.193**	**0.952**	**2.216**	**0.952**
9	2.023	0.985	**2.035**	**0.946**	**2.055**	**0.95**	**2.19**	**0.967**	2.188	0.968	2.212	0.968
10	2.027	0.968	2.031	0.959	2.052	0.961	2.185	0.983	2.185	0.981	2.209	0.981
11	2.024	0.981	2.028	0.97	2.049	0.971	2.185	0.981	2.182	0.992	2.206	0.99
12	2.022	0.989	2.025	0.979	2.047	0.98	2.179	1.0	2.179	1.0	2.204	1.0
13	2.022	0.988	2.023	0.987	2.044	0.988	-	-	-	-	-	-
14	2.019	0.998	2.021	0.994	2.042	0.995	-	-	-	-	-	-
15	2.019	1.0	2.019	1.0	2.041	1.0	-	-	-	-	-	-
Number of Networks in Ensemble / Ensemble Type	**T2 FLAIR Nonlinear**						**FA**					
	Naive		**Combinatorial**		**Bootstraped**		**Naive**		**Combinatorial**		**Bootstraped**	
	MAE (years)	**Conv. (%)**	**MAE (years)**	**Conv. (%)**	**MAE (years)**	**Conv. (%)**	**MAE (years)**	**Conv. (%)**	**MAE (years)**	**Conv. (%)**	**MAE (years)**	**Conv. (%)**
1	(years)	0.0	(years)	(years)	2.329	0.0	2.891	0.0	2.856	0.0	2.857	0.0
2	2.233	0.394	2.197	0.541	2.208	0.539	2.739	0.472	2.703	0.534	2.716	0.54
3	2.157	0.701	2.151	0.727	2.165	0.727	2.673	0.68	2.651	0.718	2.669	0.718
4	2.127	0.824	2.128	0.819	2.145	0.817	2.638	0.788	2.624	0.812	2.644	0.813
5	2.118	0.863	2.115	0.874	2.133	0.873	2.622	0.838	2.608	0.868	2.629	0.869
6	2.11	0.893	2.106	0.911	2.124	0.909	2.606	0.889	2.597	0.905	2.619	0.907
7	2.098	0.943	2.099	0.937	2.118	0.936	2.592	0.931	2.59	0.932	2.613	0.933
8	**2.09**	**0.976**	**2.095**	**0.956**	**2.114**	**0.955**	2.592	0.93	**2.584**	**0.952**	**2.607**	**0.954**
9	2.086	0.989	2.091	0.971	2.111	0.97	**2.587**	**0.948**	2.579	0.968	2.603	0.969
10	2.086	0.992	2.088	0.982	2.108	0.983	2.576	0.98	2.576	0.98	2.6	0.981
11	2.083	1.001	2.086	0.992	2.106	0.993	2.57	1.001	2.573	0.991	2.597	0.992
12	2.084	1.0	2.084	1.0	2.104	1.0	2.57	1.0	2.57	1.0	2.595	1.0
13	-	-	-	-	-	-	-	-	-	-	-	-
14	-	-	-	-	-	-	-	-	-	-	-	-
15	-	-	-	-	-	-	-	-	-	-	-	-

- The number of subjects displaying brain age delta sign changes reduces from ≈ 588 for a single network, to ≈ 379 for 3-network ensembles, and ≈ 277 for 6-network ensembles, out of a total of 2511 test subjects (Figure 4.11b);

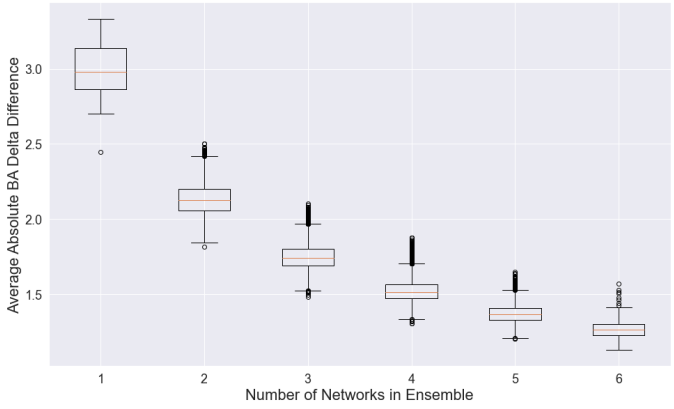
- The average absolute delta change, for those subjects exhibiting a delta sign change, reduces from ≈ 2.99 years for a single network, to ≈ 1.75 years for 3-network ensembles, and ≈ 1.27 years for 6-network ensembles (Figure 4.11c);



(a) Brain age delta correlations



(b) Number of subjects displaying a brain age (BA) delta sign change



(c) Average absolute brain age (BA) delta difference for subjects displaying a brain age delta sign change

Figure 4.11: Stability benefits of convergence for the T1 Nonlinear network, with improvements being observed in **(a)** the correlations between brain age deltas, **(b)** the number of subjects displaying a brain age delta sign change, and **(c)** the average value of that sign change between ensemble pairs. Similar results were obtained for the other single-map and fusion experiments.

Table 4.3: Ensemble single-map network convergence results split by the number of identical runs utilised, into an Original category, which employed 3-identical run ensembles for each experiment, and a Converged category, where 9-identical run ensembles were utilised. The final two columns indicate the correlations between the Original and Converged brain age deltas in their raw and linearly debiased forms. The Weighted MAE, which allows for easier comparison between this and other studies, was calculated using the method proposed by Cole et al [8]

Contrast	3-Runs (Original)				9-Runs (Converged)				Original-Converged	
	MAE	R^2	Predicted-Chronological Correlation (r)	Weighted MAE	MAE	R^2	Predicted-Chronological Correlation (r)	Weighted MAE	Delta Correlation (r)	Delta (Linear Debias) Correlation (r)
sMRI Contrasts										
T1 Nonlinear	2.242	0.853	0.922	0.060	2.190	0.862	0.926	0.059	0.925	0.911
T2 FLAIR Nonlinear	2.189	0.865	0.926	0.059	2.086	0.867	0.933	0.056	0.935	0.926
T2 Lesions	3.965	0.531	0.731	0.107	3.936	0.540	0.733	0.106	0.985	0.973
SWI	3.054	0.719	0.847	0.082	2.939	0.749	0.862	0.079	0.946	0.925
rsfMRI Contrasts										
rsfMRI-0	4.137	0.507	0.702	0.112	4.220	0.518	0.700	0.114	0.964	0.925
rsfMRI-2	4.322	0.467	0.678	0.116	4.159	0.507	0.705	0.112	0.949	0.896
rsfMRI-5	4.285	0.465	0.676	0.115	4.160	0.502	0.700	0.112	0.949	0.898
rsfMRI-10	4.357	0.458	0.667	0.117	4.243	0.487	0.690	0.114	0.952	0.893
rsfMRI-21	4.233	0.483	0.691	0.114	4.060	0.520	0.716	0.109	0.958	0.911
tfMRI Contrasts										
tfMRI-1	3.444	0.651	0.805	0.093	3.407	0.674	0.812	0.092	0.926	0.889
dMRI Contrasts										
Summed Tracts	3.177	0.719	0.839	0.086	3.070	0.742	0.847	0.083	0.946	0.928
TBSS FA	2.939	0.743	0.862	0.079	2.903	0.752	0.866	0.078	0.965	0.953
TBSS ISOVF	3.331	0.684	0.823	0.090	3.286	0.700	0.828	0.088	0.968	0.962
TBSS L3	3.108	0.716	0.843	0.084	3.132	0.724	0.841	0.084	0.968	0.957
FA	2.668	0.788	0.887	0.072	2.587	0.801	0.895	0.070	0.944	0.930
L2	2.844	0.767	0.872	0.077	2.806	0.782	0.877	0.076	0.958	0.947
MD	2.887	0.762	0.868	0.078	2.804	0.780	0.876	0.075	0.954	0.941
MO	2.935	0.746	0.874	0.080	2.873	0.801	0.869	0.077	0.946	0.927

4.6.2.2 Comparing Original and Converged Results - Single Maps

After determining an adequate number of identical runs for MAE convergence, comparisons were conducted between the original (i.e., Chapter 3) and the converged single-map results. The numerical results of this comparison are presented in Table 4.3. When considering the summary statistics for each experiment, it can be seen that the converged experiments obtain slightly better results than the original runs, although the differences between the two are minor. Moreover, when considering the rightmost two columns of the table, it is evident that very high correlations are achieved for all maps between the original and converged runs for both the predicted brain age deltas and the debiased brain age deltas.

The high correlations between the predicted brain age deltas suggest that similar associations with nIDPs are to be expected. This is indeed the case, as evidenced by Figures 4.12 and 4.13, which allow for a qualitative visual comparison between the Manhattan plots for a subset of maps. As in the previous chapter, each

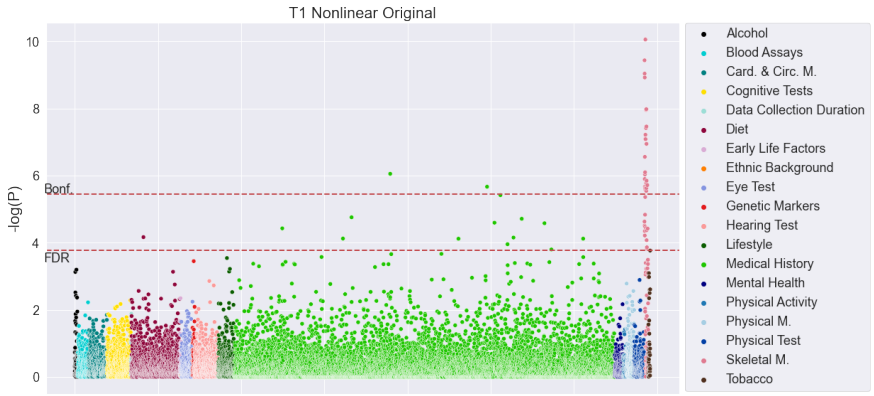
Table 4.4: $-log(p)$ correlations between original and converged single-map ensembles. These were calculated for all nIDPs, as well as for only those nIDPs passing the False Discovery Rate (FDR) thresholds calculated for the original and converged datasets.

Map	-log(p) Original-Converged Correlations (r)		
	All	Passing Original FDR Threshold	Passing Converged FDR Threshold
sMRI Maps			
T1 Nonlinear	0.898	0.907	0.948
T2 FLAIR Nonlinear	0.878	0.372	0.419
T2 Lessions	0.966	0.993	0.990
SWI	0.965	0.990	0.993
rsfMRI Maps			
rsfMRI-0	0.982	0.991	0.990
rsfMRI-2	0.990	0.998	0.998
rsfMRI-5	0.986	0.998	0.998
rsfMRI-10	0.973	0.992	0.992
rsfMRI-21	0.988	0.997	0.998
tfMRI Maps			
tfMRI-1	0.945	0.978	0.985
dMRI Maps			
Summed Tracts	0.936	0.957	0.955
TBSS FA	0.942	0.819	0.824
TBSS ISOVF	0.974	0.983	0.981
TBSS L3	0.958	0.967	0.961
FA	0.963	0.981	0.980
L2	0.938	0.872	0.879
MD	0.963	0.980	0.980
MO	0.955	0.983	0.988

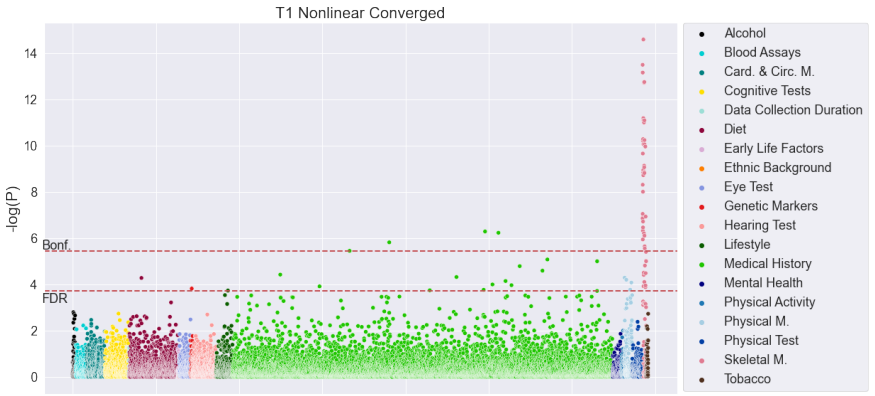
Manhattan plot indicates the statistical significance of the correlation between one nIDP and the brain age deltas associated with that particular map. Figure 4.14 and Table 4.4 provide a quantitative comparison between the original and converged networks in terms of their nIDPs.

Generally, similar associations are observed across all maps, with converged networks usually yielding similar or higher $-log(p)$ values, suggesting that the converged results are better than their original counterparts. This similarity is further confirmed by the results presented in Table 4.4, which displays very high correlations between the $-log(p)$ values obtained with the original and converged networks.

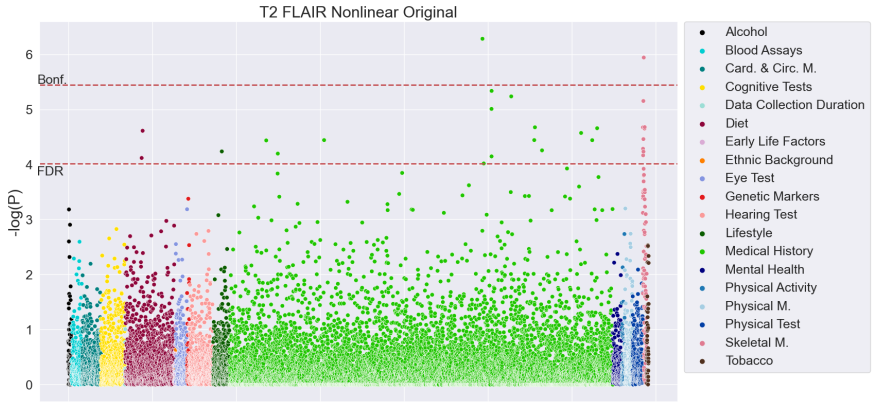
These observations appear to hold true for all examined maps, with the exception of T2 FLAIR Nonlinear. Figures 4.12c-4.12d illustrate significant differences between the original (left) and converged (right) nIDP associations. As depicted in these figures, as well as in Figure 4.14b, these differences are primarily driven by the



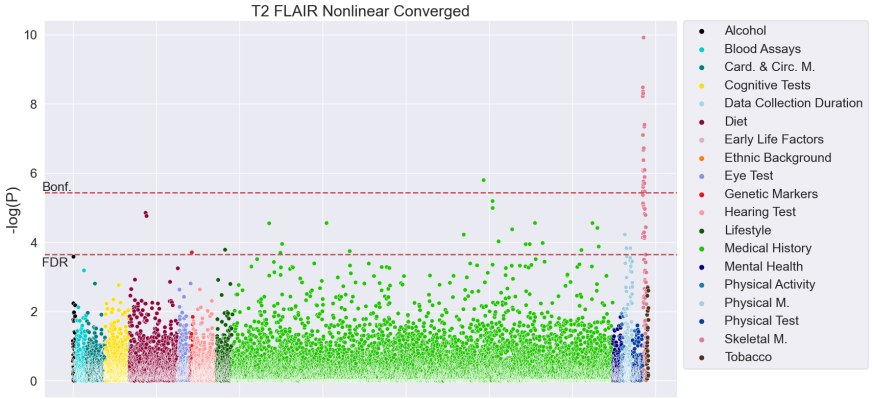
(a) sMRI T1 Nonlinear (original) correlations



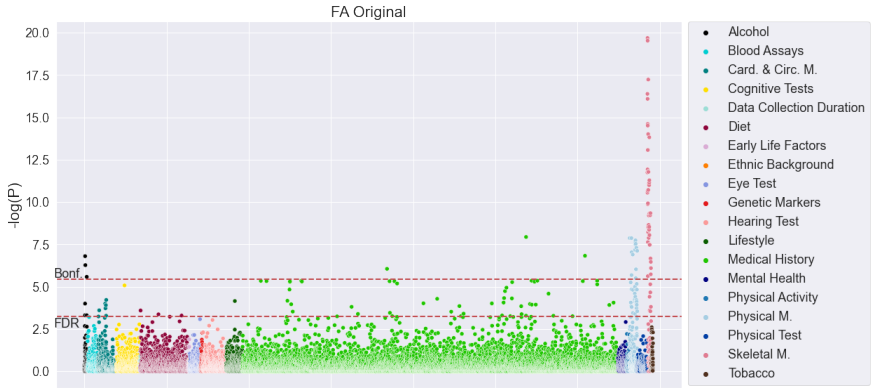
(b) sMRI T1 Nonlinear (converged) correlations



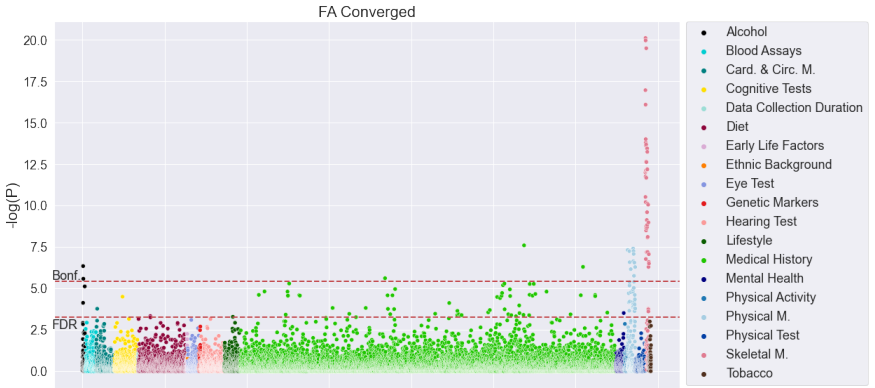
(c) sMRI T2 FLAIR Nonlinear (original) correlations



(d) sMRI T2 FLAIR Nonlinear (converged) correlations

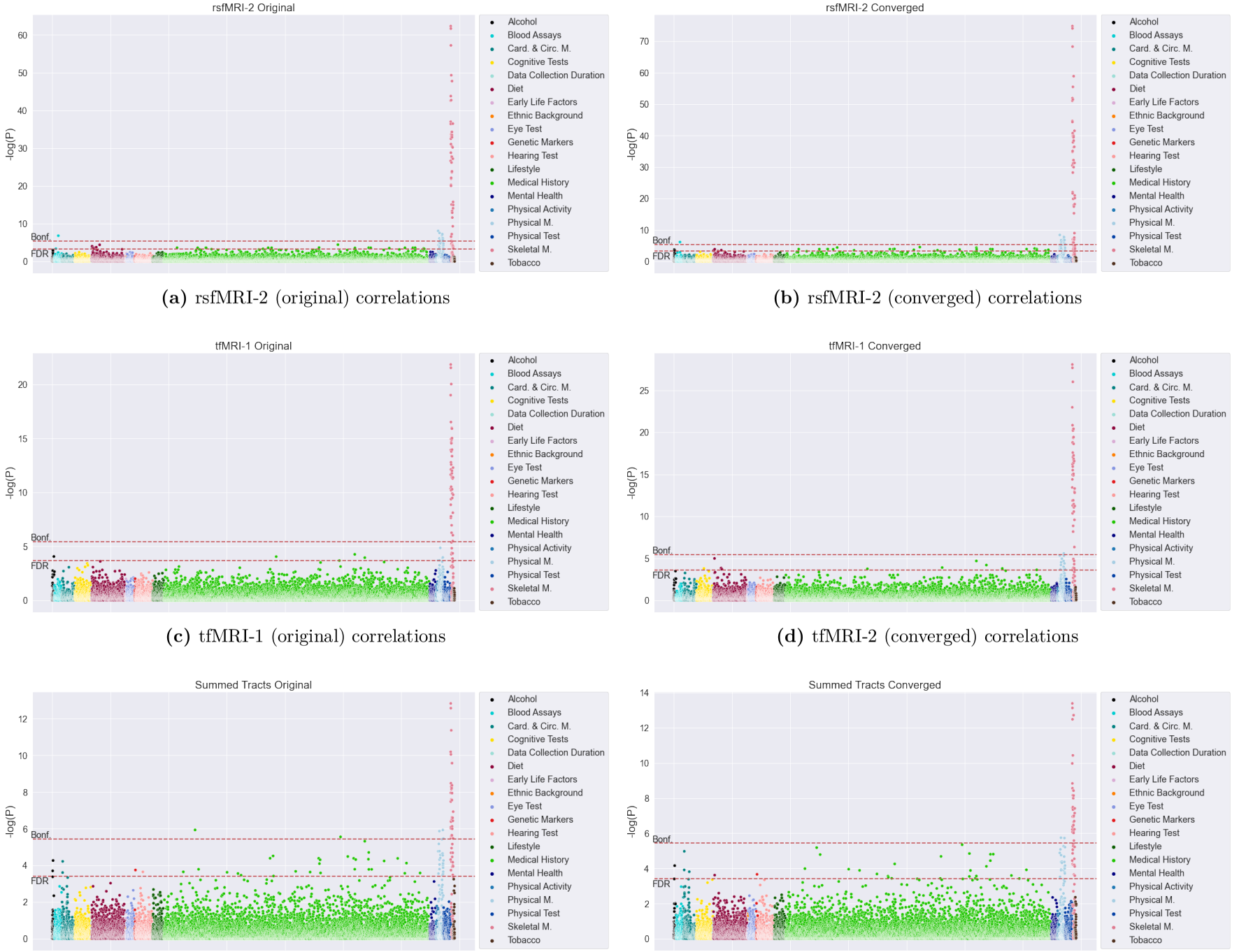


(e) dMRI FA (original) correlations



(f) dMRI FA (converged) correlations

Figure 4.12: Manhattan plots relating brain age deltas to UK Biobank nIDPs for original and converged experiments (1/2) for a subset of maps, each dot representing the statistical significance of the correlation. The False Discovery Rate (FDR) and Bonferroni thresholds are also plotted.



(a) rsfMRI-2 (original) correlations

(b) rsfMRI-2 (converged) correlations

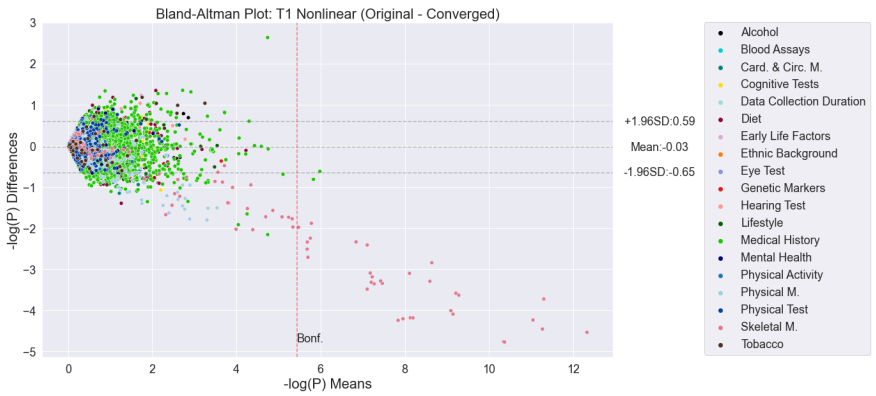
(c) tfMRI-1 (original) correlations

(d) tfMRI-2 (converged) correlations

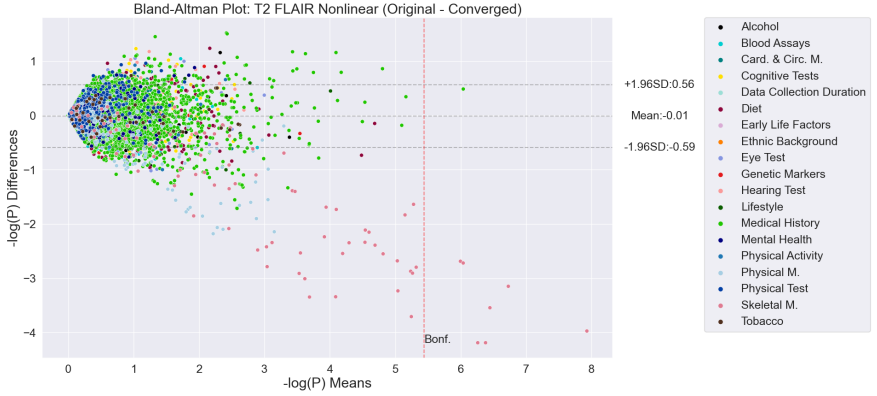
(e) dMRI Summed Tracts (original) correlations

(f) dMRI Summed Tracts (converged) correlations

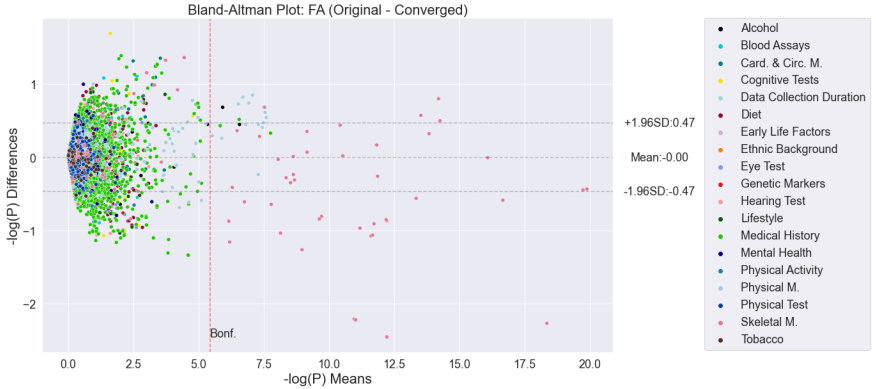
Figure 4.13: Manhattan plots relating brain age deltas to UK Biobank nIDPs for original and converged experiments (2/2) for a subset of maps, each dot representing the statistical significance of the correlation. The False Discovery Rate (FDR) and Bonferroni thresholds are also plotted.



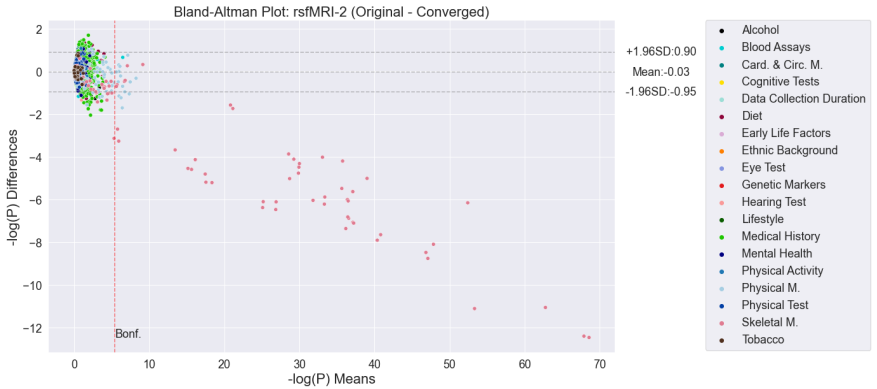
(a) sMRI T1 Nonlinear original/converged Bland-Altman plot



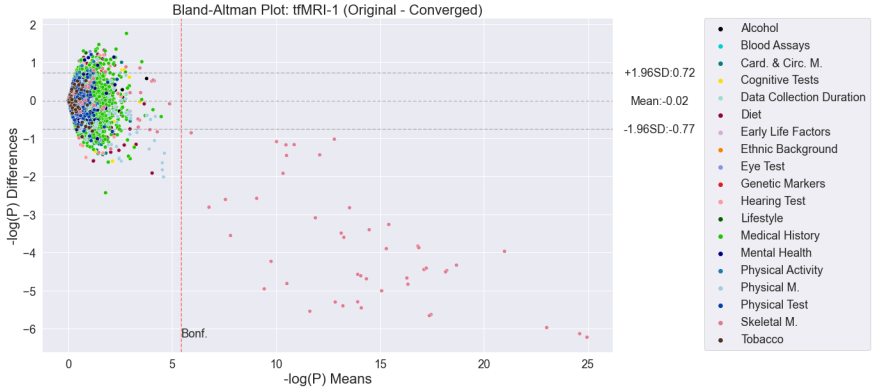
(b) sMRI T2 FLAIR Nonlinear original/converged Bland-Altman plot



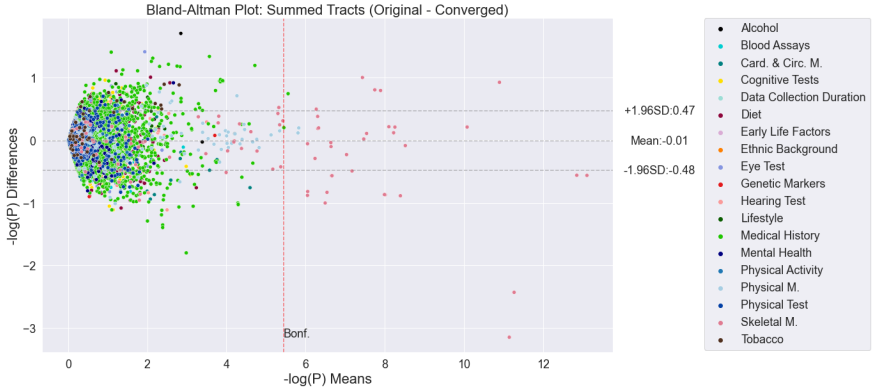
(c) dMRI FA original/converged Bland-Altman plot



(d) rsfMRI-2 original/converged Bland-Altman plot



(e) tfMRI-1 original/converged Bland-Altman plot



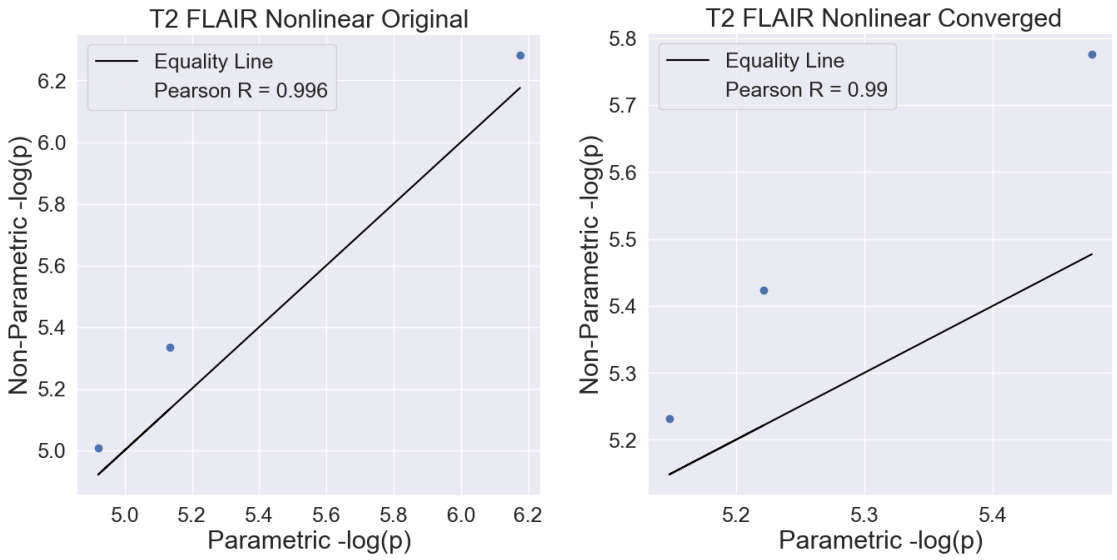
(f) dMRI Summed Tracts original/converged Bland-Altman plot

Figure 4.14: Bland-Altman plots comparing $-log(p)$ for original and converged single-map results for a subset of maps. The $-log(p)$ correspond to the statistical significance of an nIDP-brain age delta association. The Bonferroni threshold is displayed as a red dotted line. The mean of the difference and the 95% limits of agreement are displayed as grey lines.

Skeletal Measurements, with many associations surpassing the Bonferroni threshold for the converged experiment. The figures also show that, while substantial $-log(p)$ improvements are observed for some nIDPs, in the case of other associations, such as several Medical History nIDPs, there appears to be little or no change.

To understand the potential reasons for this behaviour, a two-sided paired permutation test was performed for three Medical History nIDPs. This was done under the assumption that, for validating the original parametric observations, the parametric and non-parametric p-values should be identical, or at least very similar. The permutation tests were conducted in accordance with the methods presented in Chapter 3 Section 3.3.4, using a resolution corresponding to $-log(p) = 7.78$, which is sufficiently high given the original parametric $-log(p)$ for the three nIDPs range from 5.14 to 5.47, while for the converged experiments they range from 5.00 to 6.28. For reference, the Bonferroni threshold lies at approximately 5.442. The results of the permutation test are presented in Figures 4.15a-4.15b for the original and converged networks respectively.

Interestingly, the original networks demonstrate better alignment between the parametric and non-parametric values than the converged networks do. Despite both results displaying high correlation values (≥ 0.99), this contradicts the expectation that the converged networks should have performed better relative to the original networks. This behaviour can be further understood by considering the distributions of the three nIDPs both before and after deconfounding (Figures 4.16a-4.16b), as well as the scatter plots demonstrating the relationship between the nIDPs and the debiased brain age deltas of various subjects (Figures 4.16c-4.16d). Here, it can be seen that all three nIDPs are binary and follow a non-Gaussian distribution when deconfounded. This could cause distortions when calculating the p-values, as the parametric t-values assume the data is normally distributed. Thus, while the Skeletal Measurement nIDP associations are likely influenced by potentially high levels of noise in the original T2 FLAIR Nonlinear predicted deltas, the nIDPs whose associations remain constant could either be less sensitive to noise, or be the result of spurious correlations.



(a) Scatter plot between parametric and non-parametric values of the original data.

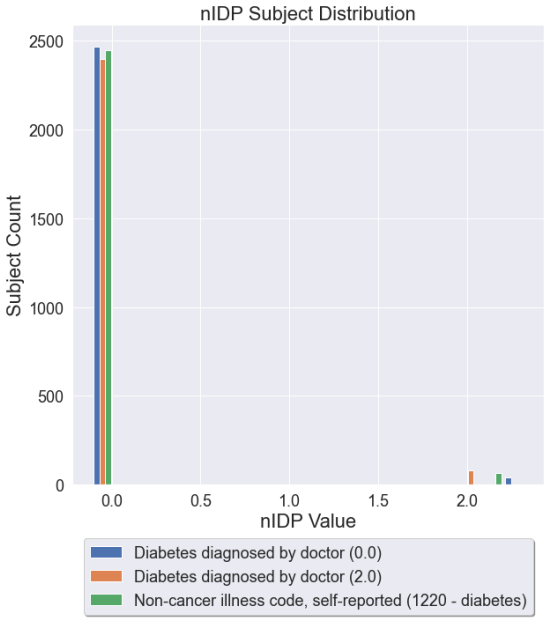
(b) Scatter plot between parametric and non-parametric values of the converged data.

Figure 4.15: Two-sided paired permutation testing results for original and converged T2 FLAIR Nonlinear. **(a)** presents the scatter plot for the original data while **(b)** presents it for the converged results.

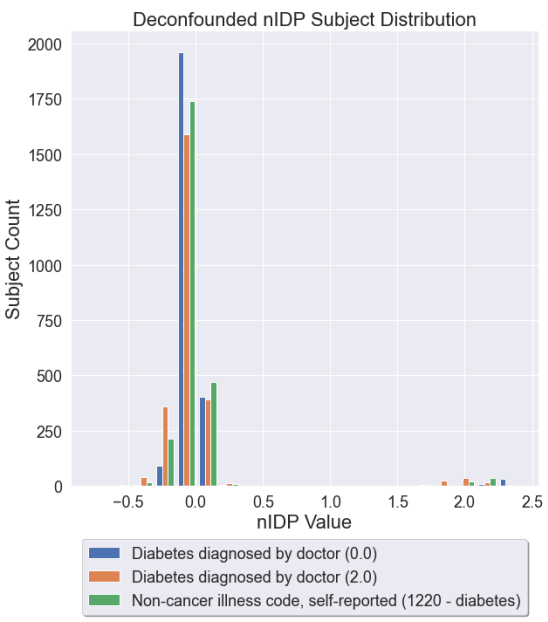
4.6.2.3 Comparing Original and Converged Results - ElasticNet Ensembles

Subsequent to the single-map results, a comparison was also conducted between original and converged ElasticNet ensembles using the 7 clusters defined for this chapter. Table 4.5 presents the summary statistics for this comparison, while Table 4.6 displays the correlations between the computed $-log(p)$ for the nIDP associations. Figures 4.17 and 4.18 provide a qualitative and quantitative comparison between the results obtained with the original and converged datasets for a subset of clusters. The former illustrates the Manhattan plots demonstrating the statistical significance of the associations between brain age deltas and nIDPs, while the latter exhibits the Bland-Altman plots for the $-log(p)$ values.

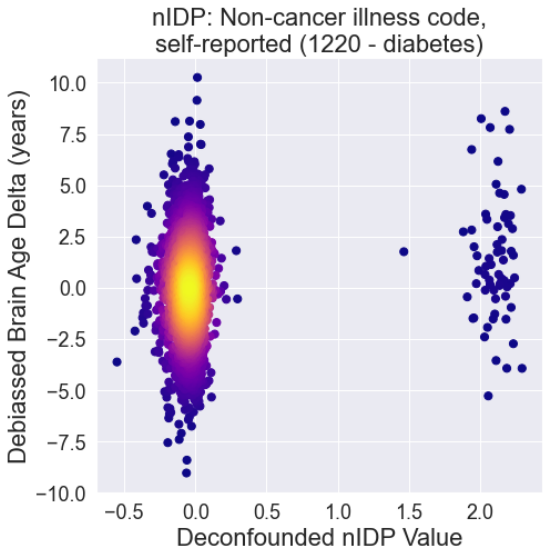
Similar to the single-map results, although the converged ensembles attain slightly improved results in terms of their summary statistics, high correlation values are noted for the brain age deltas, both raw and debiased (Table 4.5), as well as the $-log(p)$ values (Table 4.6). Furthermore, the converged ensembles typically



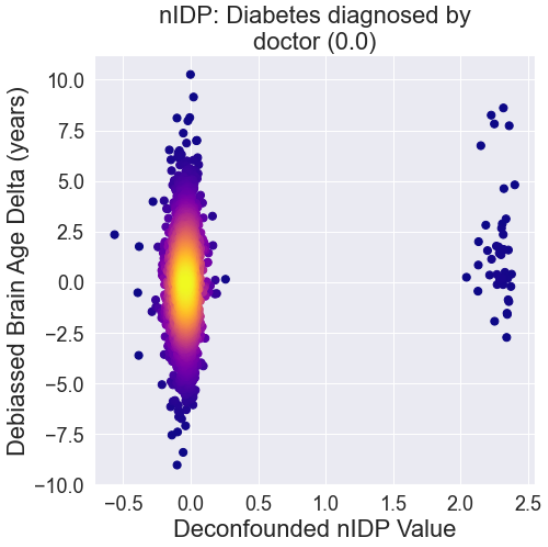
(a) Raw nIDP data distribution.



(b) Deconfounded nIDP data distribution.



(c) Density plot of deconfounded nIDP data against debiased brain age deltas.



(d) Density plot of deconfounded nIDP data against debiased brain age deltas.

Figure 4.16: Investigation of original and converged T2 FLAIR Nonlinear nIDP associations showing the nIDP data distributions for the **(a)** raw and **(b)** deconfounded versions of the dataset, as well as **(c)**-**(d)** scatter plots showing two of the nIDPs' distributions against debiased brain age deltas.

Table 4.5: ElasticNet convergence results split by the number of identical runs utilised for each map component, into an Original category, which employed 3-identical run ensembles for each experiment, and a Converged category, where 9-identical run ensembles were utilised. The final two columns indicate the correlations between the Original and Converged brain age deltas in their raw and linearly debiased forms. The Weighted MAE, which allows for easier comparison between this and other studies, was calculated using the method proposed by Cole et al [8]

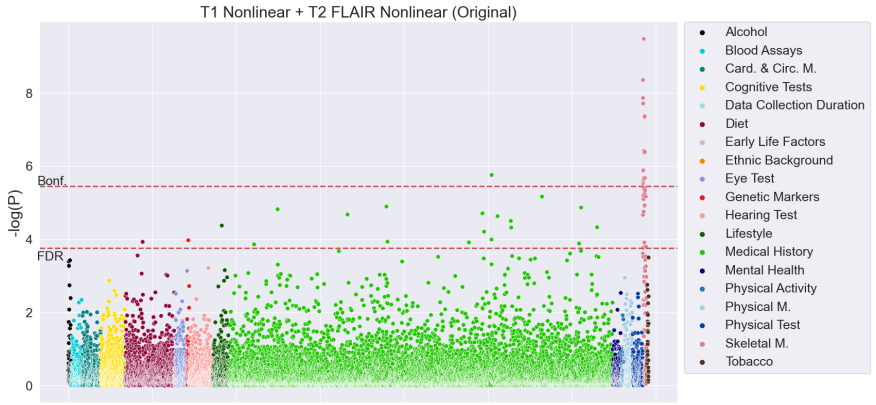
Ensemble Components	3-Runs (Original)				9-Runs (Converged)				Original-Converged	
	MAE	R^2	Predicted-Chronological Correlation (r)	Weighted MAE	MAE	R^2	Predicted-Chronological Correlation (r)	Weighted MAE	Delta Correlation (r)	Delta (Linear Debias) Correlation (r)
Cluster 1	2.038	0.875	0.936	0.055	1.987	0.875	0.939	0.054	0.955	0.949
Cluster 2	2.146	0.862	0.929	0.058	2.103	0.862	0.930	0.057	0.949	0.941
Cluster 3	2.867	0.757	0.870	0.077	2.792	0.757	0.876	0.075	0.961	0.949
Cluster 4	2.117	0.865	0.930	0.057	2.055	0.865	0.934	0.055	0.958	0.952
Cluster 5	2.653	0.792	0.890	0.071	2.620	0.792	0.893	0.071	0.981	0.977
Cluster 6	3.908	0.545	0.738	0.105	3.789	0.545	0.754	0.102	0.977	0.958
Cluster 7	2.788	0.768	0.876	0.075	2.776	0.768	0.878	0.075	0.984	0.980

generate higher $-log(p)$ values than those created with original data (Figures 4.18a-4.18b), but this cannot be generalised, as shown by Figure 4.18c.

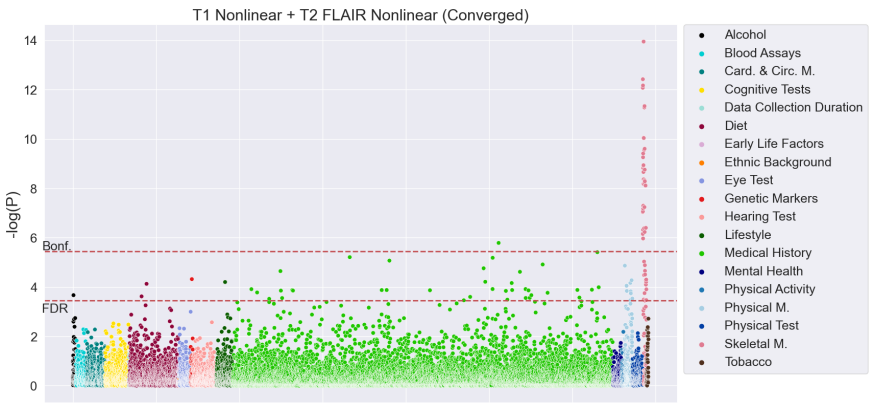
All the ElasticNet ensembles discussed above were created using the approach established in Chapter 3, where the ElasticNet is trained with the averaged results of the single-map networks. For instance, in a 2-map scenario, where each single-map experiment employed 3-identical networks, the results from the 3-identical networks were first averaged at the subject level, after which the ElasticNet is trained with the resulting 2 sets of outcomes. An alternative to this approach is to feed all 6 individual network results at once to the ElasticNet. Doing so produced results very similar to the established method, with the correlations for the brain age deltas and debiased brain age deltas, across all clusters, ranging from 0.981 to 0.997 and from 0.977 to 0.997 respectively. Thus, negligible differences can be assumed to exist between the two approaches for training the ElasticNet ensembles.

4.6.3 Discussion

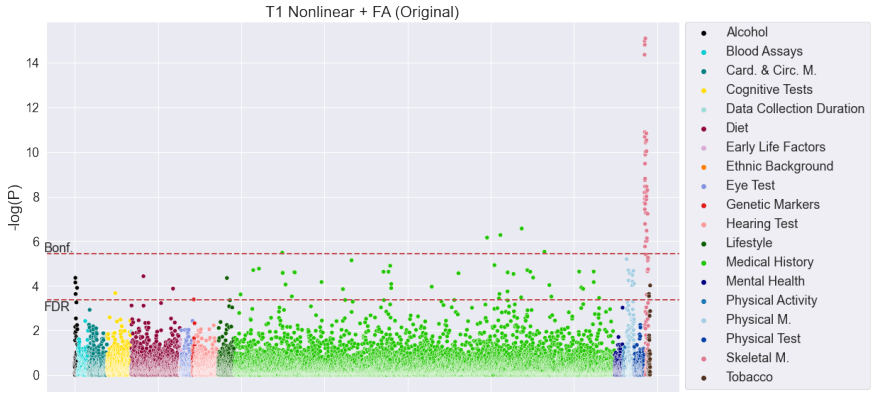
In this section, the concept of network prediction convergence was explored, building on the observations made in the previous section. These observations suggested that 3 identical runs of a network might not be sufficient for prediction convergence, which is essential for stable and reliable predictions.



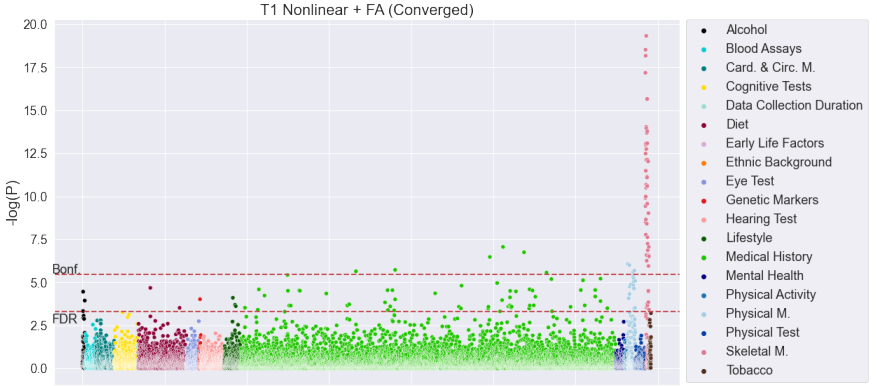
(a) ElasticNet Cluster 1 (original) correlations



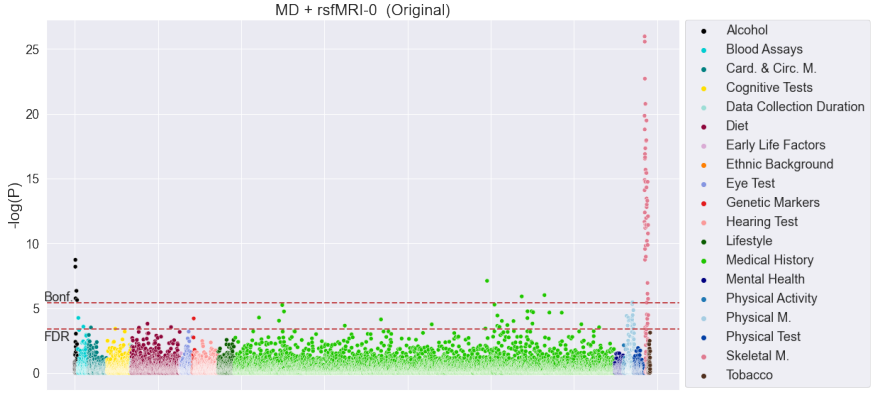
(b) ElasticNet Cluster 1 (converged) correlations



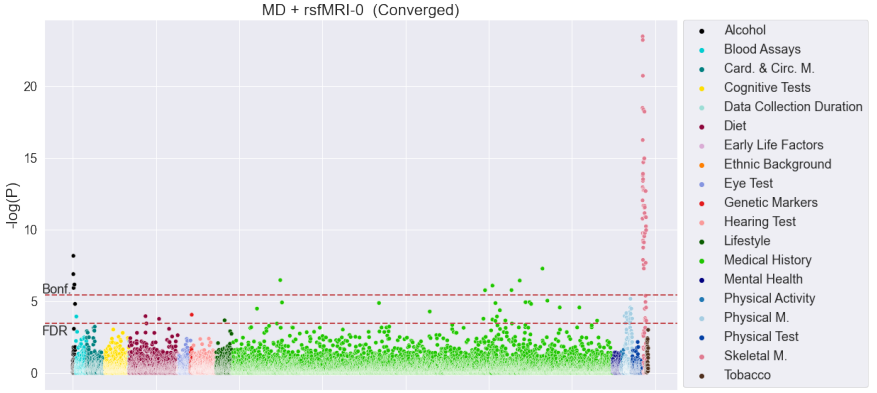
(c) ElasticNet Cluster 2 (original) correlations



(d) ElasticNet Cluster 2 (converged) correlations

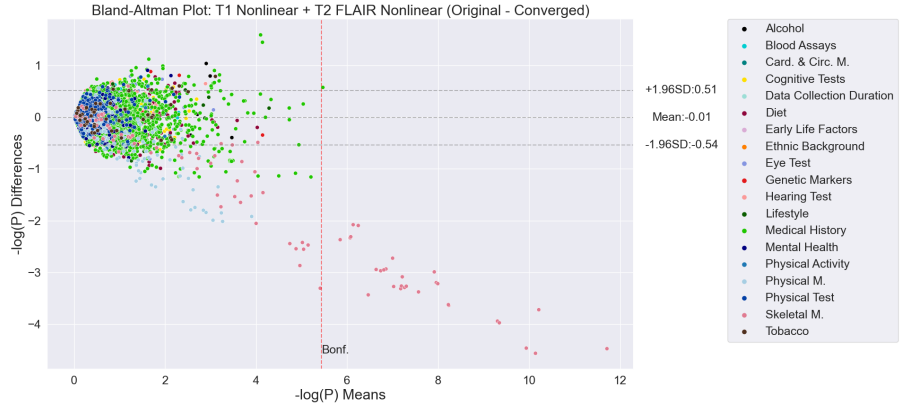


(e) ElasticNet Cluster 3 (original) correlations

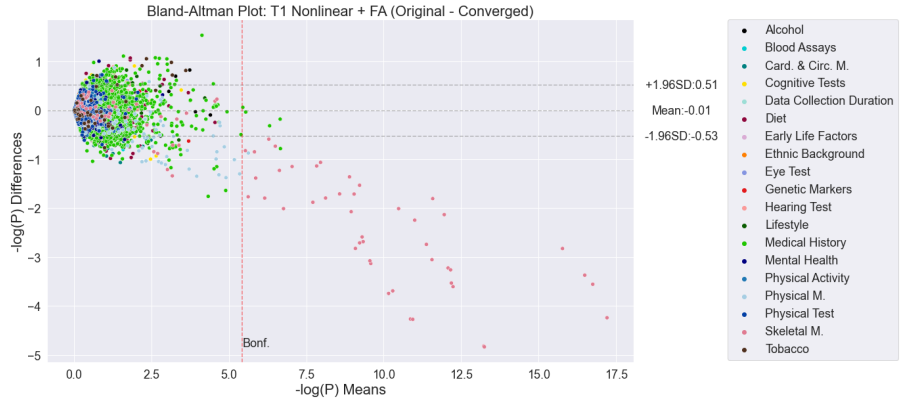


(f) ElasticNet Cluster 3 (converged) correlations

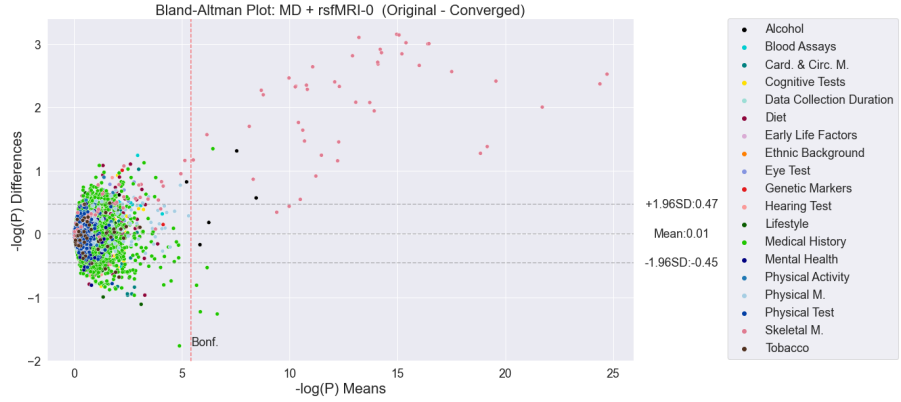
Figure 4.17: Manhattan plots relating brain age deltas to UK Biobank nIDPs for original and converged ElasticNet experiments for a subset of clusters, each dot representing the statistical significance of the correlation. The False Discovery Rate (FDR) and Bonferroni thresholds are also plotted.



(a) ElasticNet Cluster 1 original/converged Bland-Altman plot



(b) ElasticNet Cluster 2 original/converged Bland-Altman plot



(c) ElasticNet Cluster 3 original/converged Bland-Altman plot

Figure 4.18: Bland-Altman plots comparing $-log(p)$ for original and converged ElasticNet results for a subset of clusters. The $-log(p)$ correspond to the statistical significance of an nIDP-brain age delta association. The Bonferroni threshold is displayed as a red dotted line. The mean of the difference and the 95% limits of agreement are displayed as grey lines.

Table 4.6: $-log(p)$ correlations between original and converged ElasticNet ensembles. These were calculated for all nIDPs, as well as for only those nIDPs passing the False Discovery Rate (FDR) thresholds calculated for the original and converged datasets.

Ensemble Components	-log(p) Original-Converged Correlations (r)		
	All	Passing Original FDR Threshold	Passing Converged FDR Threshold
Cluster 1	0.928	0.881	0.914
Cluster 2	0.957	0.978	0.976
Cluster 3	0.976	0.989	0.989
Cluster 4	0.954	0.974	0.973
Cluster 5	0.988	0.995	0.995
Cluster 6	0.996	0.999	0.999
Cluster 7	0.980	0.969	0.968

Perfect convergence may remain elusive due to the inherent stochastic nature of neural network training, or may require a computationally prohibitive number of identical runs. In the context of this work, a combination of factors, such as random weight initialisation or the Adam optimiser, could be contributing to this fundamental limitation. Nonetheless, the work conducted in this section led to a key observation: ensembling 9 identical networks likely leads to predictions converging in terms of their accuracy, thereby becoming independent of the number of identical networks in an ensemble. This not only provides a favourable cost-benefit ratio in terms of predictive accuracy and computational efficiency, but it also leads to increased result coherence between runs in terms of delta variation.

While ensembles of 3 networks already exhibit better coherence amongst themselves than single networks, expanding the size of ensembles to 6 yields substantial improvements in the correlations between brain age deltas and several other metrics. Combined with the optimal number of 9 identical networks, this lends strong confidence in the stability and reliability of the observed results. Achieving stable and trustworthy results is crucial, especially when considering the future expansion of brain age prediction models.

Another important finding is that there is good agreement between the converged results and those obtained in the previous chapter. This agreement not only reinforces the robustness of the chosen methodology, but it also confirms that the results from Chapter 3 withstand stricter convergence criteria. However, comparing

the two sets of results underlines the necessity for rigorous statistical validations to safeguard against the possibility of spurious correlations. As highlighted by the case of the T2 FLAIR Nonlinear analysis, the presence of noise or non-Gaussian distributions can potentially yield misleading associations. These insights underscore the importance of vigilance in interpreting results and the need to implement robust statistical controls to ascertain their validity.

4.7 Comparing Fusion to Post-Training Linear Ensembling

Having addressed the questions surrounding the complex interplay between ensemble size, convergence, and prediction reliability, the final section in this chapter returns to the comparison between the fusion networks previously described and the equivalent ElasticNet ensembles. The methodology for achieving this is similar to that defined earlier in the chapter, except that 9 identical fusion networks are trained for each map cluster, and the results are compared against the equivalent converged ElasticNets. All 7 clusters defined earlier in this chapter were investigated. Firstly, the 4 progressively complex clusters were tested using all 6 fusion networks. Following this, the top 3 fusion networks, which appeared to perform best, were tested with the remaining 3 clusters, selected based on their association with a particular nIDP of interest. This decision was made due to the high computational demands. For these later clusters, in addition to evaluating their summary statistics, their association with the nIDP used to guide their selection was also evaluated and compared against the original results presented in Chapter 3.

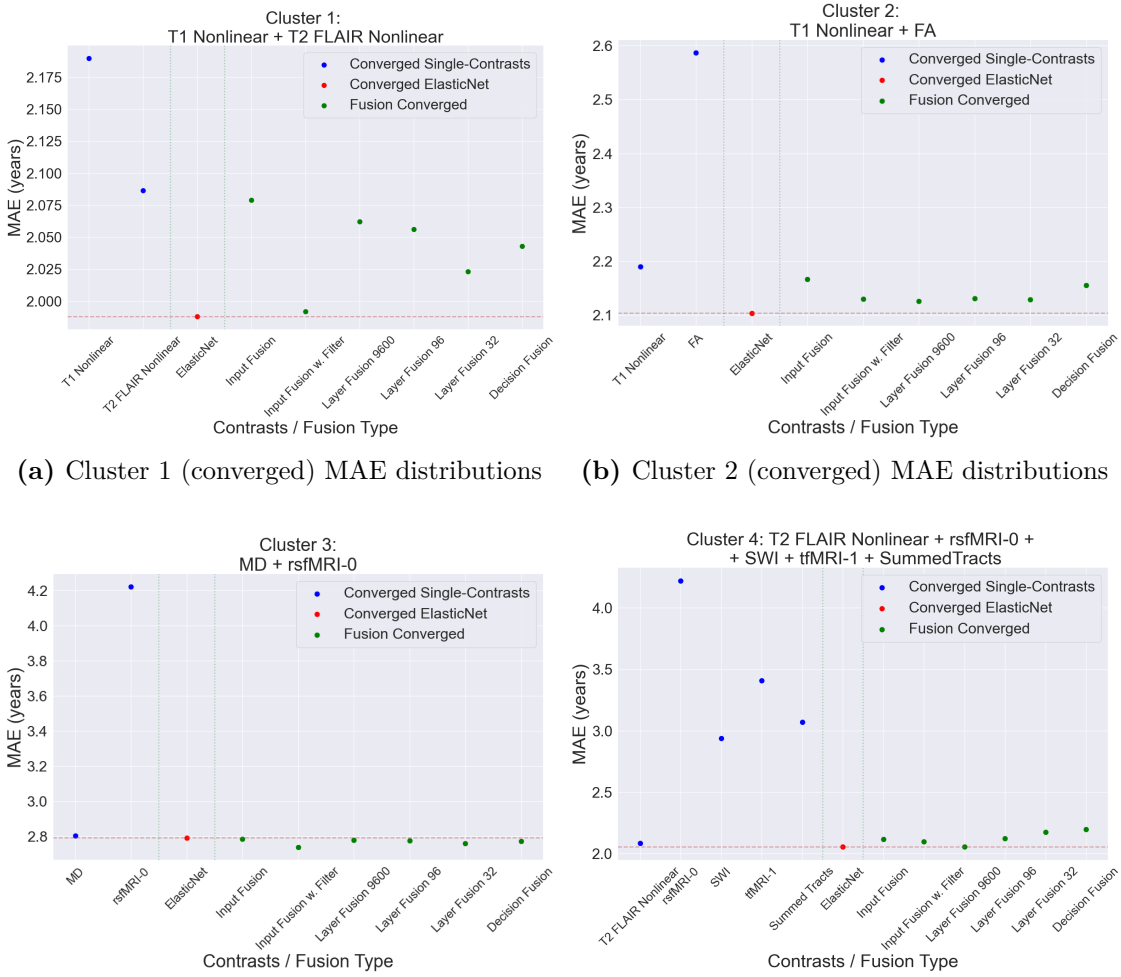
4.7.1 Results

The results obtained from the converged fusion networks are presented in Tables 4.7-4.10, and Figures 4.19-4.20 provide a more straightforward comparison between the converged single-map results, ElasticNet, and fusion networks in terms of their MAEs.

As depicted in Figures 4.19-4.20 and Table 4.7, the ElasticNet and fusion generally yield similar prediction accuracies. ElasticNet, however, outperforms the fusion networks for all clusters, with the sole exception of Cluster 3: MD + rsfMRI-0, which performs slightly better than its equivalent ElasticNet. Interestingly, in some cases, such as for Cluster 6 (Figure 4.20b) and Cluster 7 (Figure 4.20c), it can be observed that some or all of the fusion networks perform worse in terms of prediction accuracy than even the single-map components of that cluster. Similar results are obtained when linearly debiasing the brain age deltas (Table 4.8). However, it can be observed that when the deltas are debiased, a few more fusion architectures perform marginally better than the equivalent ElasticNet. Overall, while it is challenging to definitively determine which fusion network configuration performs best, the observed results suggest that the Input Fusion with Filter and Layer Fusion 9600 appear to be the top-2 performing architectures.

Further interesting observations arise when examining the correlations between the brain age deltas and debiased brain age deltas (Tables 4.9-4.10). Similar to the observations from the previous sections, high correlations can be seen here. These strong similarities are also evident in the associations with UK Biobank nIDPs, as displayed in Figures 4.21-4.22. Interestingly, for the majority of the considered clusters, the ElasticNet results appear to produce higher $-log(p)$ values (Figure 4.23), suggesting higher levels of statistical significance compared to the fusion networks. Moreover, when visually inspecting which nIDP associations pass the Bonferroni threshold for the two method classes, more associations were found for the ElasticNets across all clusters. In the few instances where one of the fusion networks has a statistically significant association to an nIDP which the equivalent ElasticNet did not, the nIDP in question was typically a binary variable, and the $-log(p)$ value of that association was close to the Bonferroni threshold, suggesting a potentially spurious correlation.

In evaluating the signal amplification potential of the fusion networks (Table 4.11), all correlations between the guide-nIDP and the debiased ensemble brain age deltas were found to be larger than the correlations to any of the individual



(a) Cluster 1 (converged) MAE distributions

(b) Cluster 2 (converged) MAE distributions

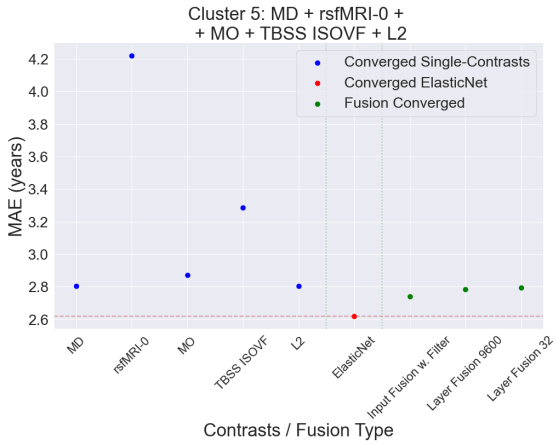
(c) Cluster 3 (converged) MAE distributions

(d) Cluster 4 (converged) MAE distributions

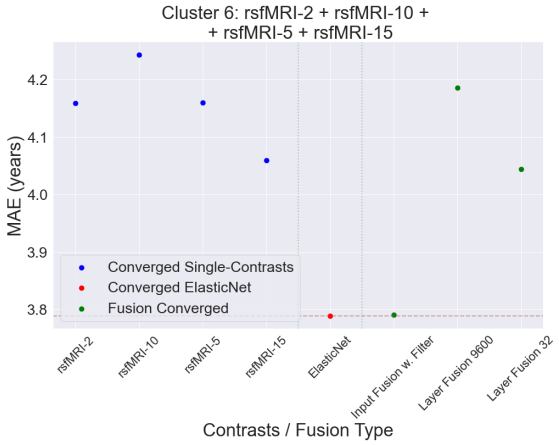
Figure 4.19: Fusion results (1/2) MAE distributions. The plots show the converged MAEs of the single-map HGL networks for each component, the converged MAE of the ElasticNet post-training linear ensemble utilising the cluster map components, and the converged results of the various deep fusion algorithms.

Table 4.7: Converged fusion results (MAE) for the various fusion networks, and the equivalent ElasticNet. The results are presented for the various map clusters, ordered from 1 to 7. The best results for each cluster are marked in bold letters. For clusters 5-7, experiments were performed only with the top performing fusion networks from clusters 1-4.

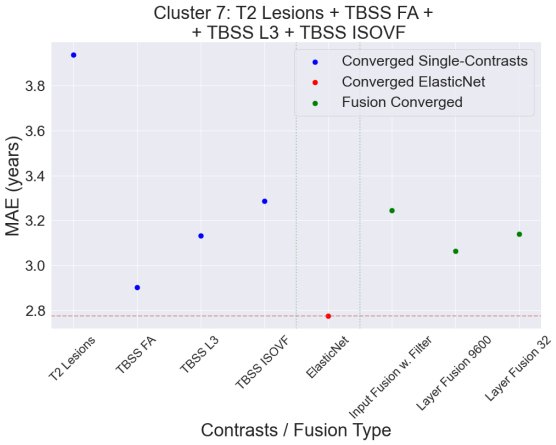
Ensemble Components (Clusters)	Converged ElasticNet	Converged Fusion					
		Input Fusion	Input Fusion w. Filter	Layer Fusion 9600	Layer Fusion 96	Layer Fusion 32	Decision Fusion
Cluster 1	**1.987**	2.079	1.992	2.026	2.056	2.023	2.043
Cluster 2	**2.103**	2.167	2.130	2.126	2.131	2.129	2.155
Cluster 3	2.792	2.787	**2.740**	2.779	2.776	2.762	2.772
Cluster 4	**2.055**	2.118	2.097	2.056	2.123	2.175	2.198
Cluster 5	**2.620**	-	2.739	2.785	-	2.794	-
Cluster 6	**3.789**	-	3.791	4.185	-	4.044	-
Cluster 7	**2.776**	-	3.244	3.063	-	3.140	-



(a) Cluster 5 (converged) MAE distributions



(b) Cluster 6 (converged) MAE distributions



(c) Cluster 7 (converged) MAE distributions

Figure 4.20: Fusion results (2/2) MAE distributions. The plots show the converged MAEs of the single-map HGL networks for each component, the converged MAE of the ElasticNet post-training linear ensemble utilising the cluster map components, and the converged results of the various deep fusion algorithms.

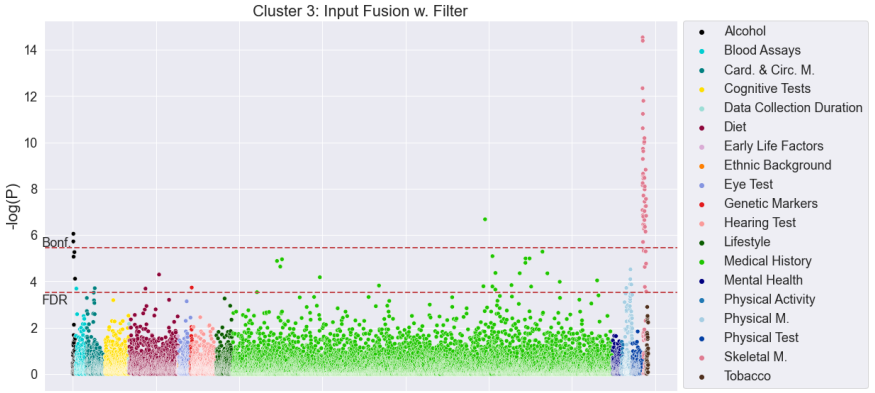
Table 4.8: Converged fusion results (MAE Debiased) for the various fusion networks, and the equivalent ElasticNet. The results are presented for the various map clusters, ordered from 1 to 7. The best results for each cluster are marked in bold letters. For clusters 5-7, experiments were performed only with the top performing fusion networks from clusters 1-4.

Ensemble Components (Clusters)	Converged ElasticNet	Converged Fusion					
		Input Fusion	Input Fusion w. Filter	Layer Fusion 9600	Layer Fusion 96	Layer Fusion 32	Decision Fusion
Cluster 1	2.124	2.222	**2.112**	2.206	2.201	2.158	2.179
Cluster 2	**2.275**	2.343	2.294	2.302	2.308	2.307	2.333
Cluster 3	3.117	3.157	**3.082**	3.150	3.153	3.136	3.132
Cluster 4	2.203	2.274	2.232	**2.197**	2.276	2.341	2.366
Cluster 5	**2.931**	-	3.098	3.153	-	3.186	-
Cluster 6	5.060	-	**4.955**	5.818	-	5.526	-
Cluster 7	**3.189**	-	3.935	3.632	-	3.745	-

(a) ElasticNet Cluster 1 correlations

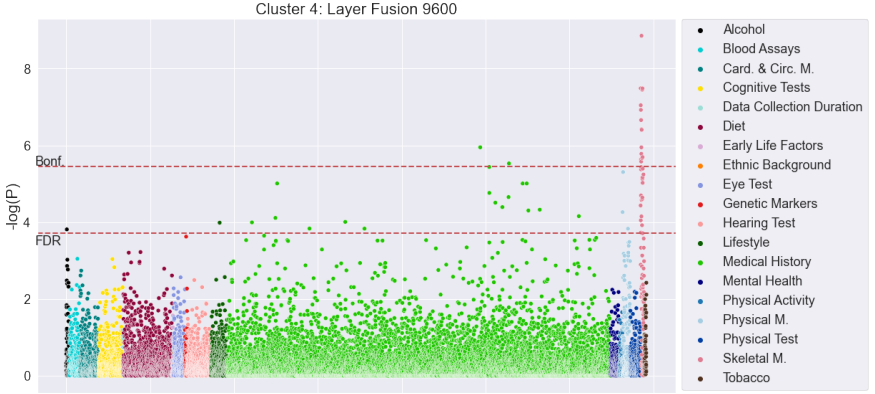
(b) Input Fusion with Filter Cluster 1 correlations

(c) ElasticNet Cluster 3 correlations



(d) Input Fusion with Filter Cluster 3 correlations

(e) ElasticNet Cluster 4 correlations



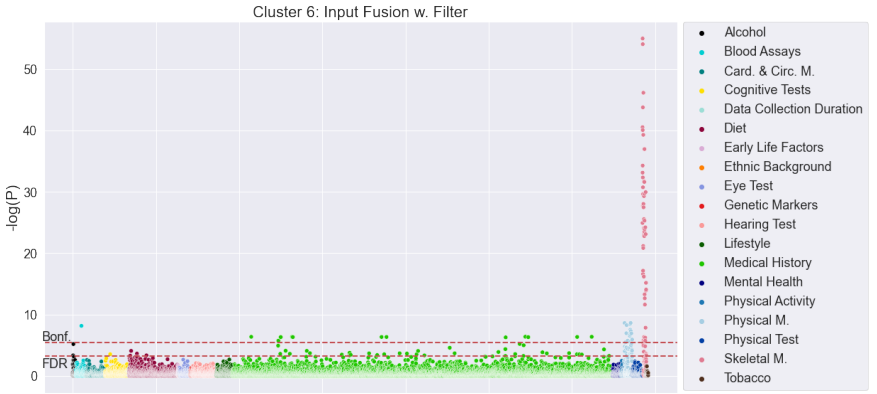
(f) Layer Fusion 9600 Cluster 4 correlations

Figure 4.21: Manhattan plots relating brain age deltas to UK Biobank nIDPs for the best fusion network per cluster and equivalent ElasticNet (1/2) for a subset of clusters, each dot representing the statistical significance of the correlation. The False Discovery Rate (FDR) and Bonferroni thresholds are also plotted.

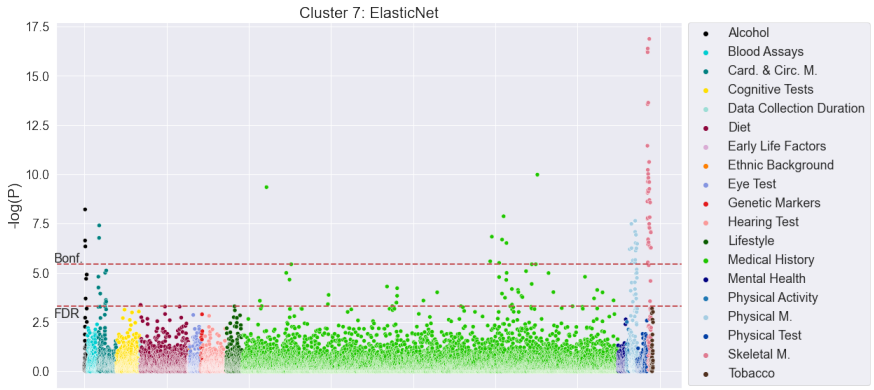
(a) ElasticNet Cluster 5 correlations

(b) Input Fusion with Filter Cluster 5 correlations

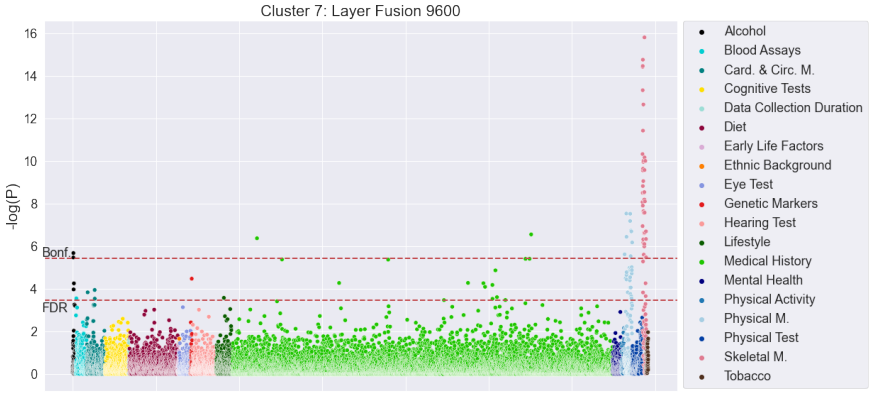
(c) ElasticNet Cluster 6 correlations



(d) Input Fusion with Filter Cluster 6 correlations



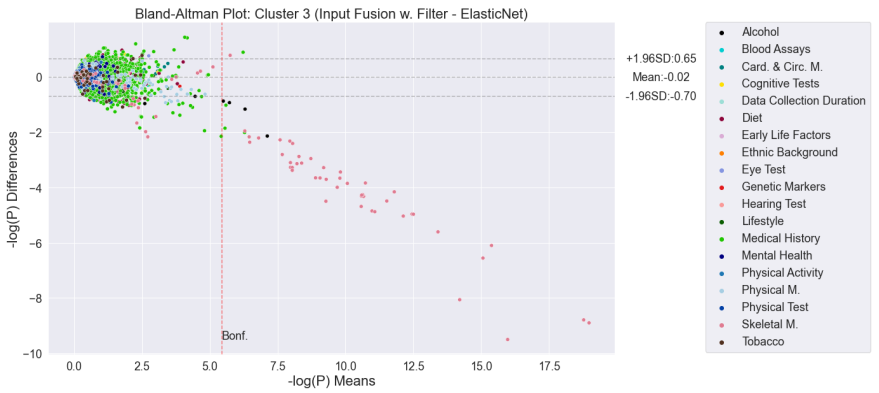
(e) ElasticNet Cluster 7 correlations



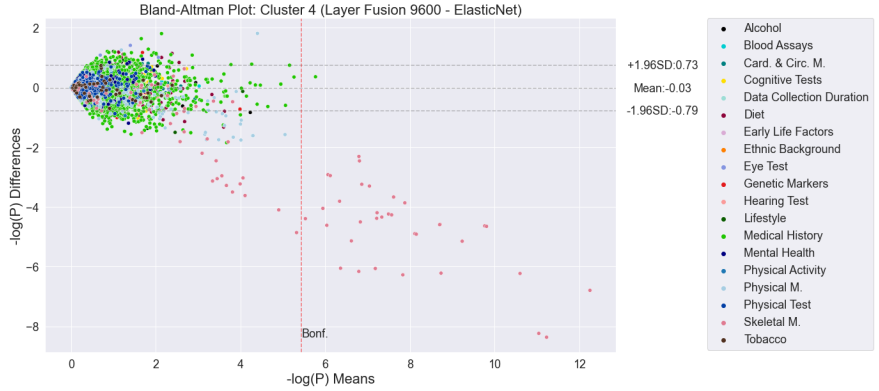
(f) Layer Fusion 9600 Cluster 7 correlations

Figure 4.22: Manhattan plots relating brain age deltas to UK Biobank nIDPs for the best fusion network per cluster and equivalent ElasticNet (2/2) for a subset of clusters, each dot representing the statistical significance of the correlation. The False Discovery Rate (FDR) and Bonferroni thresholds are also plotted.

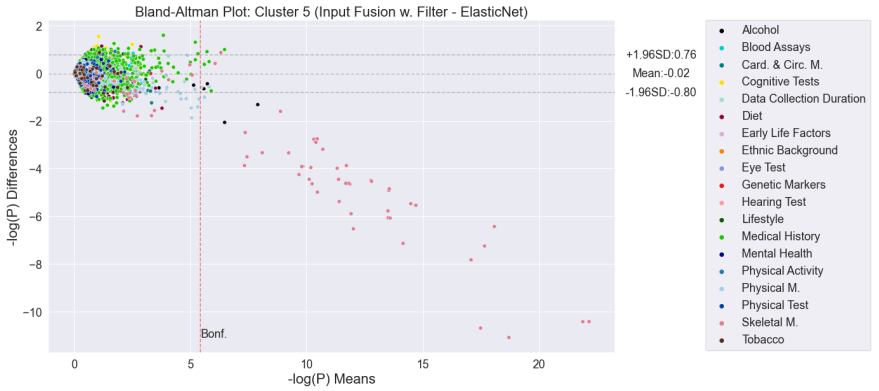
(a) Input Fusion with Filter vs ElasticNet results for Cluster 1



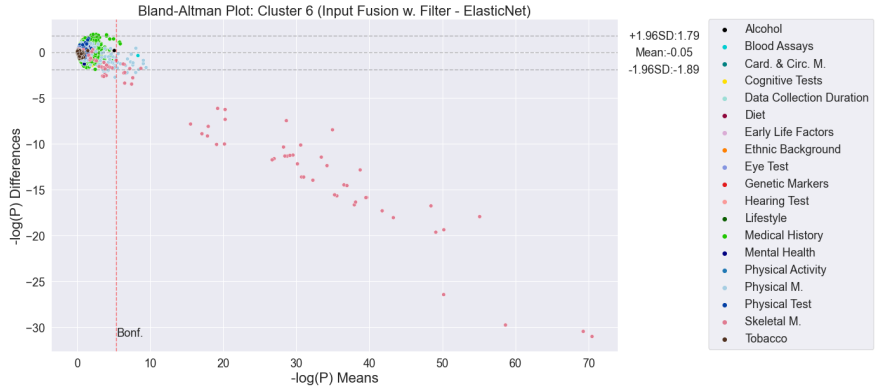
(b) Input Fusion with Filter vs ElasticNet results for Cluster 3



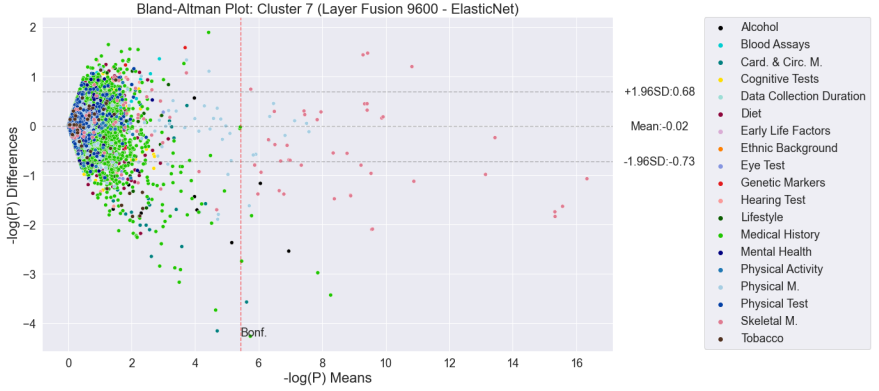
(c) Layer Fusion 9600 vs ElasticNet results for Cluster 4



(d) Input Fusion with Filter vs ElasticNet results for Cluster 5



(e) Input Fusion with Filter vs ElasticNet results for Cluster 6



(f) Layer Fusion 9600 vs ElasticNet results for Cluster 7

Figure 4.23: Bland-Altman plots comparing $-log(p)$ for the best fusion network per cluster and equivalent ElasticNet for a subset of clusters. The $-log(p)$ corresponds to the statistical significance of an nIDP-brain age delta association. The Bonferroni threshold is displayed as a red dotted line. The mean of the difference and the 95% limits of agreement are displayed as grey lines.

Table 4.9: Converged fusion results (brain age delta correlations) for the various fusion networks, and the equivalent ElasticNet. The results are presented for the various map clusters, ordered from 1 to 7. For clusters 5-7, experiments were performed only with the top performing fusion networks from clusters 1-4.

Ensemble Components (Clusters)	ElasticNet-Fusion Correlation (r)					
	Input Fusion	Input Fusion w. Filter	Layer Fusion 9600	Layer Fusion 96	Layer Fusion 32	Decision Fusion
Cluster 1	0.899	0.939	0.920	0.931	0.937	0.931
Cluster 2	0.926	0.940	0.924	0.929	0.933	0.935
Cluster 3	0.935	0.939	0.939	0.942	0.944	0.947
Cluster 4	0.894	0.937	0.897	0.915	0.908	0.911
Cluster 5	-	0.924	0.918	-	0.924	-
Cluster 6	-	0.870	0.808	-	0.876	-
Cluster 7	-	0.826	0.861	-	0.858	-

Table 4.10: Converged fusion results (brain age delta debiased correlations) for the various fusion networks, and the equivalent ElasticNet. The results are presented for the various map clusters, ordered from 1 to 7. For clusters 5-7, experiments were performed only with the top-performing fusion networks from clusters 1-4.

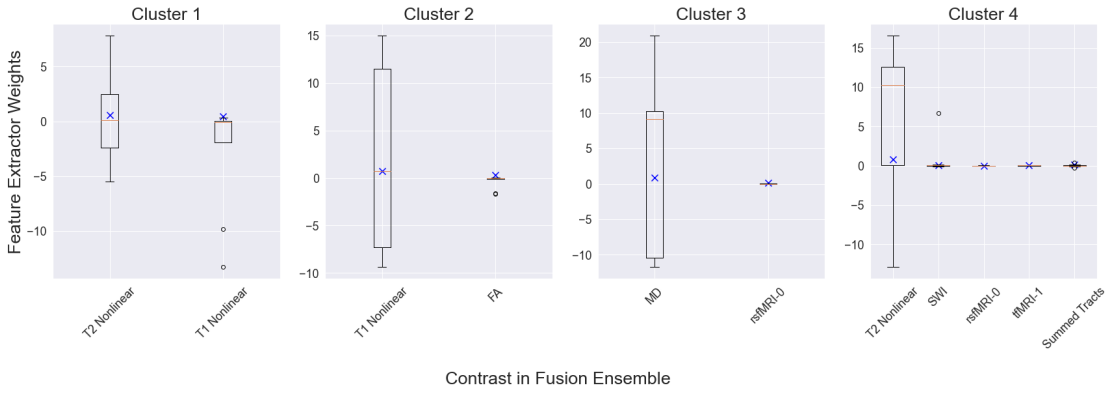
Ensemble Components (Clusters)	ElasticNet-Fusion Correlation (r)					
	Input Fusion	Input Fusion w. Filter	Layer Fusion 9600	Layer Fusion 96	Layer Fusion 32	Decision Fusion
Cluster 1	0.899	0.939	0.920	0.931	0.937	0.931
Cluster 2	0.926	0.940	0.924	0.929	0.933	0.935
Cluster 2	0.935	0.939	0.939	0.942	0.944	0.947
Cluster 4	0.894	0.937	0.897	0.915	0.908	0.911
Cluster 5	-	0.924	0.918	-	0.924	-
Cluster 6	-	0.870	0.808	-	0.876	-
Cluster 7	-	0.826	0.861	-	0.858	-

ensemble component maps. Overall, very small differences were observed between the original ElasticNet and the converged results, with the ElasticNets achieving higher correlations for Clusters 5 and 7, and the Layer Fusion 32 outperforming other methods for Cluster 6.

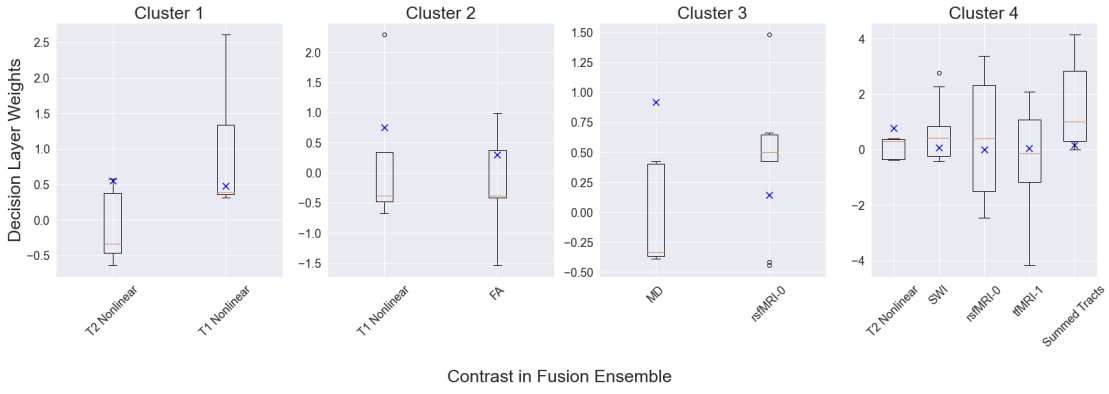
Table 4.11: UK Biobank based fusion ensembles, where groups of maps were selected and ensembled using the original and converged ElasticNet, as well as the various fusion networks, based on their high correlations with a specific UK Biobank nIDP. Groups of maps were selected for each major nIDP group by taking the top 3-5 most highly correlated maps with the particular nIDP of interest. For all groups, the resulting absolute ensemble correlation was higher than that of the components, except in the case of the final map cluster relating to a diagnosis of hypertension, where the information encoded in the T2 Lesions ends up dominating the information from the other maps for all ensembling techniques.

Measure Group	**Measure Name**	**Component Maps**	**Individual Correlations (r)**	**Original**	**Converged**			
				ElasticNet Ensemble Correlation (r)	**ElasticNet Ensemble Correlation (r)**	**Input Fusion w. Filter Correlation (r)**	**Layer Fusion 9600 Correlation (r)**	**Layer Fusion 32 Correlation (r)**
Alcohol	Frequency of drinking alcohol (0.0)	Cluster 5	0.129 0.129 0.127 0.127 0.127	**0.149**	0.140	0.128	0.132	0.134
Blood Assays	C-reactive protein (0.0)	Cluster 6	-0.108 -0.108 -0.107 -0.101	-0.130	-0.122	-0.119	-0.110	**-0.136**
Medical History	Diagnoses - secondary ICD10 (I10-I10 Essential (primary) hypertension)	Cluster 7	0.133 0.113 0.103 0.100	**0.128**	0.128	0.113	0.102	0.100

Lastly, in observing the high similarity between the ElasticNet and Fusion networks, the learned linear regression coefficients were compared against the learned weights of the Input Fusion with Filter and Decision Fusion networks for Clusters 1-4 (Figure 4.24). Although in some cases the linear coefficients (blue Xs) appear to coincide or be close to the weights distributions, this observation cannot be generalised. Moreover, particularly for Clusters 2, 3, and 4, and the Input Fusion with Filter network (Figure 4.24a), it can be seen that one map often dominates the cluster, while the other clusters are assigned weights close to 0. For the Input Fusion with Filter network, this implies that the network generally learns to prefer the best-performing single-map, which dominates the fusion thereafter. This pattern does not appear to hold for the Decision Fusion networks (Figure 4.24b).



(a) Input Fusion with Filter



(b) Decision Fusion

Figure 4.24: Learned fusion and regression weight distributions for map clusters 1-4. **(a)** displays the results for Input Fusion with Filter networks, whilst **(b)** shows those for the Decision Fusion architectures. The learned fusion weights are presented as box plots, whilst the ElasticNet coefficients are depicted as blue Xs.

4.7.2 Discussion

Upon considering the findings of this final section, it is apparent that, in most cases, modality-fusion does not outperform post-training ensembling. In fact, both the prediction accuracy, as measured by MAE, and the associations with nIDPs are stronger for the linear ElasticNet approach. Moreover, it is not generally the case that modality-fusion performs better than any single-modality component, and no single fusion approach can be singled out that consistently surpasses the others.

These shortcomings of deep fusion networks could partly be attributed to their substantial number of trainable parameters. This overparametrisation could lead to overfitting to noise, or, alternatively, contribute to noisy predictions due to random weight initialisation. This might explain why ElasticNet (trained with multiple single-map predictions), with a fewer number of trainable parameters, provides a more stable and reliable approach.

These observations culminate in a crucial insight: the higher computational costs of training bespoke deep fusion networks, coupled with the versatility and efficiency of ElasticNet models, potentially render deep fusion networks unattractive for multimodality neuroimaging brain age predictions.

Several pressing questions stem from these findings. Firstly, the need for explainable networks is underscored by the limited understanding derived from the fusion networks. While there are indications that the fusion networks do learn different filters from the equivalent single-map networks, potentially suggesting complementary features (Section 4.5.2.1), the exact nature of these features remains elusive. This lack of transparency calls for a more profound exploration, not only to understand the complementary features between maps but also to unravel the differences between various fusion architectures.

Secondly, the consistent difficulty in achieving brain age prediction accuracies with MAEs lower than $1.95 - 2.00$ years highlights a significant challenge. This could indeed be a reflection of reaching a point of maximum noise elimination from the brain age predictions, leaving only structural biological errors. However, it is crucial to acknowledge the inherent variability in brain age among individuals.

Even in a perfectly healthy population, individuals' true brain ages may naturally differ from their chronological ages due to a variety of factors such as genetics, lifestyle, and other health conditions. This natural variability poses a limit to the accuracy of brain age predictions. While further refinement of the CNN models and exploration of other datasets are essential, it is important to recognise this inherent limitation in brain age prediction accuracy.

4.8 Conclusion

This chapter has carried out an exploration of deep fusion networks in the context of brain age prediction, contrasting them with equivalent linear ElasticNet ensembles obtained using single-map networks. Through a sequence of methodological evaluations, experimental observations, and critical assessments, several insights and challenges were identified.

On the whole, the comparison between deep fusion networks, trained using multiple maps simultaneously, and ElasticNet, fed with the brain age predictions obtained from single-map networks, has underscored the relative efficiency and performance of the latter. Several augmentation techniques were attempted, such as the use of transfer learning, to enhance deep fusion networks, but without much success. In the case of transfer learning, its use exposed an unexpected degradation in performance, contradicting the conventional wisdom surrounding the efficacy of transfer learning in similar contexts. The insights obtained from the comparison between deep fusion networks and linear ensembling methods, such as the potential overparametrisation and unattractiveness of deep fusion networks for multimodality neuroimaging brain age predictions, have raised significant questions about their utility in this domain.

The other primary finding of this chapter relates to the exploration of network prediction convergence, leading to crucial insights that ensure the stability and reliability of predictions. The determination that ensembling ≈ 9 identical networks is likely to lead to convergence provides vital guidance for future work. Furthermore, the robustness of previously established methodologies has been confirmed.

Throughout this research, the prevailing assumption has been that CNNs represent the optimal class of deep learning methods for brain age prediction, due to their extensive prevalence in the medical imaging domain. However, CNNs harbour several limitations that need to be addressed:

- Firstly, CNNs tend to overlook fine-grained details, which could be relevant for understanding brain ageing processes. This occurs as they operate on large receptive fields that encompass the whole image, employing progressively complex geometrical features.

- Secondly, the outcomes of CNNs are generally challenging to interpret as they are not inherently explainable. This opacity further underscores the need for more transparent, interpretable models.

The above conclusions set the stage for the final part of this thesis. Seeking to address these issues and build upon the insights and questions raised in this chapter, the next chapter will explore the use of a novel class of deep learning networks - the visual transformer. This new direction holds the promise of offering a fresh perspective and potentially addressing the challenges uncovered in the exploration of CNNs and deep fusion networks for brain age prediction.

5

BA-SWIN: Investigating the Brain Age Gap utilising SWIN Vision Transformers

Contents

5.1 Overview

Following on from the previous two chapters, this final results chapter of the thesis addresses some of the potential limitations of Convolutional Neural Networks (CNNs) in accurately predicting brain age and proposes a novel approach using Shifted Window (SWIN) Transformers — a distinct class of Vision Transformers (ViTs). Despite the widespread use of CNNs in medical imaging, their inherent lack of explainability and potential difficulties in modelling image contextual information [248], which may be crucial for understanding brain ageing processes, warrant the exploration of alternative methods.

To address these limitations, in this chapter I introduce a new SWIN-based network tailored for brain age prediction, referred to as Brain Age SWIN (BA-SWIN). Unlike conventional CNNs, which employ fixed-size convolutional kernels to process images, SWIN networks partition images into multiple non-overlapping patches. Each patch is independently processed through a series of non-linear operations, conserving and utilising the fine-grained details within each patch. Most notably, SWIN networks deploy an attention mechanism that dynamically adjusts the processing path for different patches based on their content, allowing the model to focus on the more informative regions of the image. This attention flow mechanism not only enhances the network's performance but also offers a level of explainability not commonly found in CNNs.

Following the methodologies established in earlier chapters, the BA-SWIN architecture is tested against equivalent CNNs in brain-age prediction tasks for several maps. Although the BA-SWIN networks did not surpass the CNNs in terms of prediction accuracy (as gauged by Mean Absolute Error - MAE), achieving comparable results, they managed to extract novel, potentially microstructure-related associations to biomedical and lifestyle measures, referred to as non-imaging derived phenotypes (nIDPs), for the fractional anisotropy (FA) map. This result suggests that SWIN networks are potentially better at capturing some non-geometric traits, likely tied to neurodegenerative conditions. This, however, requires further validation.

In examining data robustness, SWIN networks displayed superior performance when dealing with corrupted data, particularly when pretrained on research quality datasets such as UK Biobank. This observation aligns with existing literature, indicating that the global context and attention mechanisms utilised by SWINs aid in identifying key features, even amidst noisy data.

This chapter delves further into the explainability of SWIN networks, which demonstrated the capability to produce informative attention activation images at various network levels. These maps shed light on the features the network prioritises when making predictions. However, generating a global attention map proved challenging, pointing to a potential avenue for future exploration.

This chapter provides insight into the utility of Transformers for brain age prediction investigations. The chapter also provides insight into Transformer and CNN model behaviours when faced with clinical quality data. These findings underscore the potential of Vision Transformers in pioneering robust, explainable deep learning techniques in neuroimaging and lay the groundwork for future avenues of research.

5.2 Introduction

Over the past decade, a multitude of studies have delved into the brain age gap phenomenon, seeking to gain insights about the ageing brain and the factors potentially influencing this process, either accelerating or inhibiting it. As highlighted in Appendix A, Tables A.1-A.7, a significant proportion of these past studies (approximately 75%) utilised MRI image derived phenotypes (IDPs) to train an array of regression models. These models encompass Linear Regression [35], Ridge [59, 301], LASSO [58, 81], Support Vector Regression (SVR) [62, 97], Relevance Vector Machine (RVM) [61, 321], XGBoost regression [109, 146], as well as several others [322–324]. However, as discussed in Chapter 3 Section 3.2, leveraging IDP inputs can lead to potentially subjective judgements about feature importance. This issue has been tackled by a recent growing body of research adopting minimally-processed MRI volumetric data. The majority of these studies

have predominantly opted for CNNs as their deep learning regression model of choice [53, 74, 84, 90, 114], due to CNNs' ability to discern relevant data patterns and relationships autonomously in grid-like topologies. Coupled with the rise of architectures such as the fully convolutional network (FCN) [325] and U-Net [326], CNNs have been ubiquitously adopted across the medical imaging domain, from disease classification and diagnostics [327, 328], to medical image segmentation [329, 330] and reconstruction [331].

While CNNs are popularly used in medical image analysis, they come with certain limitations. One key issue is that the actual area a CNN "sees", termed the effective receptive field, is often smaller than what it is theoretically designed to observe. This discrepancy arises mainly because early stages of CNNs focus more on nearby, or "local", image features [248, 332, 333]. To account for a broader image context while preserving local features, modern CNNs often employ methods such as pooling and skip connections. Pooling helps the network to recognise larger patterns, while skip connections help in retaining finer details by bypassing certain layers. Another method employed is simply making the networks deeper, yet this can lead to its own set of problems. For instance, as a CNN grows deeper, it may downplay the significance of early layer features [334]. Without additional modifications, the addition of more layers may hinder the network, causing it to overlook crucial small-scale information in MRI images when making predictions. This occurs because the convolutional layers prioritise the formation of larger patterns by integrating new features at the cost of fine details, thus diluting or overlooking finer, local details. Skip connections were introduced to counteract this by forwarding the early features to later stages of the network. However, skip connections mainly enhance local feature learning, and their addition can quickly increase the network's complexity and computational demands, without guaranteeing an increase in performance. In addition, CNNs are not inherently explainable, functioning as effective black boxes, making it difficult to understand which features the network uses for a prediction [21, 335, 336].

Numerous modifications or alternatives to CNNs have been proposed over the years, attempting to address some of these issues, ranging from graph neural networks (GNNs) designed to better model long-range connections within a dataset [65], to recurrent neural networks (RNNs), which model data dependencies across the temporal spectrum in longitudinal datasets [77], to modern adaptations of the CNN architecture such as ConvNeXt [250], which improves on the original design by adding modern components, such as GELU activations, displaying both performance and scalability improvements [248]. In terms of CNN explainability, various techniques have been developed and deployed, including attention gates, saliency maps and gradient-based methods [21, 337]. Although these techniques offer some insight into feature importance, their output often lacks detailed interpretability, especially for complex regression tasks such as brain age prediction.

Over the last few years, a new class of network architectures has emerged that promises to address the limitations of CNNs. Based on the Language Transformer model [245] which underpins modern large language models such as BERT [233] and GPT [338], Vision Transformers (ViTs) [332] offer several advantages over CNNs. These include the ability to model long-range information dependencies while also still considering local features [339], inherent explainability, and increased robustness to noisy or corrupted data [340].

In the following paragraphs, a broad, top-level overview of ViTs is provided. For the underlying mathematics of these methods, the reader is directed to Chapter 2 Section 2.2.3. For an in-depth discussion of ViTs in the context of medical image analysis, the work by Li Chen et al. [248] is recommended.

The main innovation introduced by the ViT architecture comes from how the model handles input data, focusing on the most important parts of the input to enhance result accuracy and model efficiency. Unlike other models, which process the entire image input at once, ViTs use an attention mechanism to dynamically prioritise the most important parts of the input data. This is achieved by means of a weighted aggregation function, which applies varying levels of importance to specific sections of an input. This approach provides the model with a holistic understanding

of an input image [341, 342]. ViTs utilise a version of this mechanism known as Self-Attention (SA). While traditional attention functions allocate significance based on an external context, such as a query or information queue, SA functions by capturing the internal correlation in the input data or latent features [341]. This is achieved by mapping an input into query, $\mathbf{Q}$, key, $\mathbf{K}$, and value, $\mathbf{V}$, matrices using learnable parameters. The attention matrix, $\mathbf{A}$, is determined by applying the softmax transformation to the normalised product of the $\mathbf{Q}$ and $\mathbf{K}$ arrays; the former acts as a look-up array, while the latter serves as an embedding matrix. The SA block's output is then derived by multiplying matrices $\mathbf{A}$ and $\mathbf{V}$. This operation allows the SA block to assess and assign different weights to different parts of an input, based on the relationships and interactions within the input data itself, enabling it to find complex relationships and dependencies. To allow ViTs to simultaneously process different aspects of the input data, multiple self-attention blocks are usually used in parallel. These operations are termed multi-head self-attention (MSAs) blocks, and allow the network to focus on various features simultaneously, such as textures, shapes and colour within an image, leading to a more comprehensive and nuanced understanding of the input. Each MSA captures different aspects, or subspaces of the input data, with the final output being the concatenation of outputs from all heads. While it might intuitively seem that more MSAs would enhance performance, by allowing the model to capture a wider range of features and relationships, this is not always the case. Aspects such as overparametrisation and the coordination overheads among different blocks can lead to potential problems, such as overfitting and high computational costs [343]. To leverage the attention mechanism, ViTs typically begin by dividing an input image, or volume, into fixed-size non-overlapping patches and linearly projecting them into tokens. This approach might appear counterintuitive given ViTs' ability to model long-range image dependencies, as segmenting an input into patches disrupts its internal structure. However, by adopting this method, each patch essentially becomes an independent data point. This structure allows the MSA blocks to process information from each patch independently, making the

learning of long-range dependencies more efficient [344]. Breaking an image into patches and processing them independently can cause the model to lose spatial context. To compensate for this, positional embeddings are added to retain the spatial relationships among patches. This technique not only preserves spatial or sequential context but also adds inductive bias, which gives the model a set of guiding assumptions which help it generalise better to new, unseen data [248]. After processing the tokenised patches through the MSAs, the latent features are then presented to Multi-Layer Perceptrons (MLPs). These act as feedforward neural networks within Transformer blocks, introducing non-linearity and further inductive bias. They are essential for capturing patterns and relationships in the data, thereby enhancing the model's capability to make accurate predictions.

Given their architecture, ViTs represent a paradigm shift from traditional CNNs for visual tasks. This leads to several key qualities that make ViTs interesting models for medical imaging applications, such as brain age predictions. Firstly, as discussed above, ViTs are better than CNNs in modelling long-range dependencies and exploiting contextual information. This is due to their multi-head self-attention mechanism, which allows them to model interactions between any distant section of an input image, creating a global context at each layer of the network. In map, CNNs rely on a series of local convolution operations and pooling layers to gradually expand their receptive field, meaning that at each layer of the network, each element is only connected to a small, localised region of the input. This allows ViTs to exploit contextual information, or the interrelations and dependencies between parts of an image that contribute to its overall understanding, resulting in a significantly larger effective receptive field, similar to that of GNNs [345]. In addition, by embedding local image patches and progressively refining them at the same scale, ViTs are also more capable than CNNs at capturing and retaining detailed local information in deeper layers [346]. These two mechanisms should also make Transformers more robust to noise or other data perturbations [347]. And, finally, ViTs have the advantage of being inherently explainable. This is a characteristic of the SA mechanism. Utilising techniques such as attention rollout [348], the individual

attention activation maps can be utilised to create a global attention map, which provides information on features a Transformer is utilising in making a prediction.

On the other hand, compared to CNNs, ViTs have their own set of shortcomings. A notable challenge with ViTs is their intense computational and data demands during training. CNNs, due to their consistent weight-sharing across the image, come with inherent assumptions, known as inductive biases, that help them generalise effectively on unseen data, exhibiting qualities like locality and scale invariance [349]. ViTs, on the other hand, operate with fewer built-in assumptions, relying mainly on positional embeddings to derive any inductive bias [332]. While this makes them more flexible, it does imply that ViTs could potentially require more data and training resources to achieve similar generalisations. As such, training them, especially with smaller datasets, or in the initial stages for the first epochs, becomes notably challenging [332].

Across the medical image analysis field, in the past several years, Transformers have been utilised in hundreds of applications, including medical image segmentation, recognition and classification, object detection, registration, reconstruction, and enhancement [248]. Owing to the differences in properties between CNNs and ViTs [332], numerous past studies have sought to combine CNNs and Transformers using several approaches, including Conv-Transformer hybrids. These hybrid networks utilise, for instance, convolutional operations as a backbone to reduce the size of the input data by projecting it into a lower-dimensional space, and then the SA mechanism to extract long-range relationships [350]. A similar approach to this is being utilised by He et al [68], which, as far as it could be determined at the time of writing, represents the only published application of Transformers for the task of brain age prediction. Other CNN-ViT hybrid methods exist besides this approach, such as those which use the two networks together in parallel, fusing the information from the two branches at a deeper point in the network [351], or those that alternate convolutions and SA blocks sequentially [349]. However, such hybrid models can also have certain disadvantages, particularly in terms of any explainable

insights which can be extracted from a network. Given this, the work presented in this chapter will focus solely on pure Transformer architectures.

Based on all the observations discussed thus far, in this chapter the following questions will be addressed:

1. Can Vision Transformers (ViTs), or a version of them, be utilised to predict brain age?

2. What would be an optimal Transformer architecture for brain age predictions?

3. How do Transformers compare to CNNs in terms of their prediction accuracies, training times, and associations with nIDPs?

4. Can the attention mechanism be utilised to extract insights regarding what features of an image are most relevant for brain age predictions?

To address these questions, in this chapter I first started by exploring which pure Transformer architectures would be best suited for making brain age predictions using large 3D neuroimaging inputs. After deciding on a model, I performed several experiments aimed at better understanding its performance and fine-tuning its hyperparameters. Following this, I tested the proposed Transformer architecture against the HGL CNN model discussed in Chapter 3 for 18 different 3D maps derived from 5 core MRI modalities. These are the same maps used in Chapter 4, their use being justified by both the fact that they cover a wide spectrum of map types and imaging modalities, but also by the fact that sufficient networks had previously been trained to facilitate a comparison between MAE converged CNNs and Transformers for single-map brain age prediction tasks. The two model types were also compared in terms of their ability to deal with corrupted or lower-quality clinical data. Finally, the chapter concludes with a qualitative investigation of explainability for the chosen Transformer model.

5.3 Common Methods

In this section, the common methods used throughout this chapter are introduced. Many of these are similar to the general methods introduced in Chapters 3 and 4. For this reason, the interested reader is invited to revisit the relevant previous sections, such as Sections 3.3, 4.6.1.1 and 4.6.3, introducing aspects such as the utilised datasets and employed processing steps, the HGL architecture, how the brain age and brain age deltas are obtained and associated to nIDPs.

Below, a discussion is carried out regarding Transformer architectures, their limitations when dealing with large 3D volumetric inputs, and the chosen model for this chapter. Similar to the Introduction section above, this discussion will be top-level and aim to provide intuition. For the mathematical underpinning of the chosen transformer models, the reader is invited to visit Chapter 2 Section 2.2.3. Finally, the section concludes with a description of the experimental setup employed throughout this chapter.

5.3.1 The BA-SWIN Architecture

While Transformers have numerous advantages over CNNs, they are also very computationally and memory demanding to train. This can prove to be extremely problematic when working with 3D volumetric data, which is the format of choice for most medical images, as well as the UK Biobank data utilised in this work [248]. Over the past few years, several models have been proposed which address this problem [352–355]. The issue with most of these networks is that they were designed to perform segmentation tasks, eliminating the need for computationally intensive fully connected layers. For instance, in the case of a pure ViT network [355], assuming a patch size of 3 voxels, for $1mm$ isotropic resolution input cropped to $160 \times 192 \times 160$ voxels, this would result in $262k$ patches after padding, each patch containing 27 voxels. The information is then embedded internally, using an embedding dimension of 1024 and passed through a series of MSA blocks. The issue emerges at the output of the Transformer blocks, where, if flattened, the output would be of size $NumberPatches \times EmbeddingDimension = 268M$ parameters,

making it computationally infeasible for passing to a fully connected layer. To circumvent this issue, one could either increase the patch size, reduce the embedding dimensionality or downsample the data to $2mm$ isotropic resolution. None of these solutions is particularly appealing, as a higher patch size could imply less information being passed to the Transformer, a lower internal dimensionality could lead to less informative internal operations, and a reduction in resolution would not allow for a direct comparison with CNN results.

Thus, the only solution to this issue, given the hardware constraints imposed by limited GPU memory, is to somehow reduce the dimensionality of the data, either prior to feeding it to the Transformer by using a CNN, or internally in the Transformer. The former solution is undesirable, as it would not allow for a direct comparison between pure CNN and Transformer architectures, leaving only the latter approach. This can be achieved using a variation of the ViT architecture, referred to as a Shifted Window Transformer, or SWIN [246], which was designed specifically for working with high-resolution imaging data. Figure 5.1 shows the 3D SWIN network utilised for this brain age prediction work, named BA-SWIN. This has been adapted from the original 2D SWIN Transformer [246] and the 3D SWIN-UNETR networks [247], using the SWIN-UNETR code provided by Project MONAI [356].

Compared to the original ViT architecture, SWIN introduces several modifications, as described by the original authors [246]. Firstly, rather than processing the entire input in a global manner by dividing it into a sequence of non-overlapping patches to which self-attention (SA) is applied simultaneously, SWIN divides the image into smaller local windows, each containing a number of patches, and applies SA to these windows. This division into smaller windows and localised processing significantly reduces the computational complexity. While typical self-attention scales quadratically with the number of patches, due to the need to calculate pairwise attention scores for all pairs of image patches, SA scales linearly, since it limits self-attention to fixed-size, non-overlapping windows within each image. As the image size increases, the number of windows increases linearly, ensuring the overall computational cost for self-attention also increases linearly. This is

further augmented by SWIN's hierarchical structure, where adjacent windows are progressively merged, which enables the network to capture multi-scale information. This approach can, however, lead to the loss of long-range information pathways. SWIN accounts for this by using a shifted window technique between consecutive SA layers, which facilitates information exchange between neighbouring windows, allowing the network to understand relationships between adjacent regions and maintain a large receptive field without added computational complexity. In addition, rather than using global positional embeddings, SWIN employs a relative position bias mechanism, allowing for custom positional embeddings to be calculated for each SA layer. This enables the model to understand the relative positions of different tokens within each window, which is essential for maintaining the spatial information lost due to the window partitioning.

When comparing the original HGL CNN network (Figure 3.2) to the proposed BA-SWIN, it can be observed that a similar structure was maintained between the two. Essentially, the four 3D convolutional blocks were replaced by four 3D SWIN Transformer blocks for the BA-SWIN network. The latter are composed of two modules, one containing a window MSA (W-MSA), and the other a shifted windows MSA (SW-MSA). The MSA operations are preceded by LayerNorm operations, and followed by a 2-layer MLP also preceded by LayerNorm. Residual connections were applied after each operation block, as shown in Figure 5.1. After the final SWIN Transformer block, similar to HGL, the resulting latent volume is flattened and passed through three Fully Connected layers of sizes 96×1, 32×1 and 1×1. The first two fully-connected layers are followed by ReLU nonlinearities, and the final by a linear activation. The other hyperparameters were kept similar to those described in the original SWIN and SWIN-UNETR papers. Thus, the patch size was set to $2 \times 2 \times 2$ voxels, and the original window size to $7 \times 7 \times 7$ patches. When shifted, the windows were displaced by $(\lfloor \frac{7}{2} \rfloor, \lfloor \frac{7}{2} \rfloor, \lfloor \frac{7}{2} \rfloor) = (3, 3, 3)$, the displacement being tied to the window size. The embedding feature dimensionality was set to 24, doubling each time the input is downsampled. Thus, to maintain the ratio of data points per head constant, the number of heads for each Transformer block was set

to $[3, 6, 12, 24]$. The interested reader can find a more detailed explanation behind these hyperparameters and the SWIN architecture in Chapter 2 Section 2.2.3.

Although the two proposed networks follow similar design principles, they are substantially different in terms of their number of trainable parameters. While HGL has $\approx 1.3M$ trainable parameters, the vanilla form of BA-SWIN based on SWIN-UNETR has $\approx 7.7M$ trainable parameters. This has the immediate effect that the maximum training batch size is reduced from 12 to 3 for BA-SWIN. Moreover, this could lead to potentially longer training times and raises the question of the network overfitting.

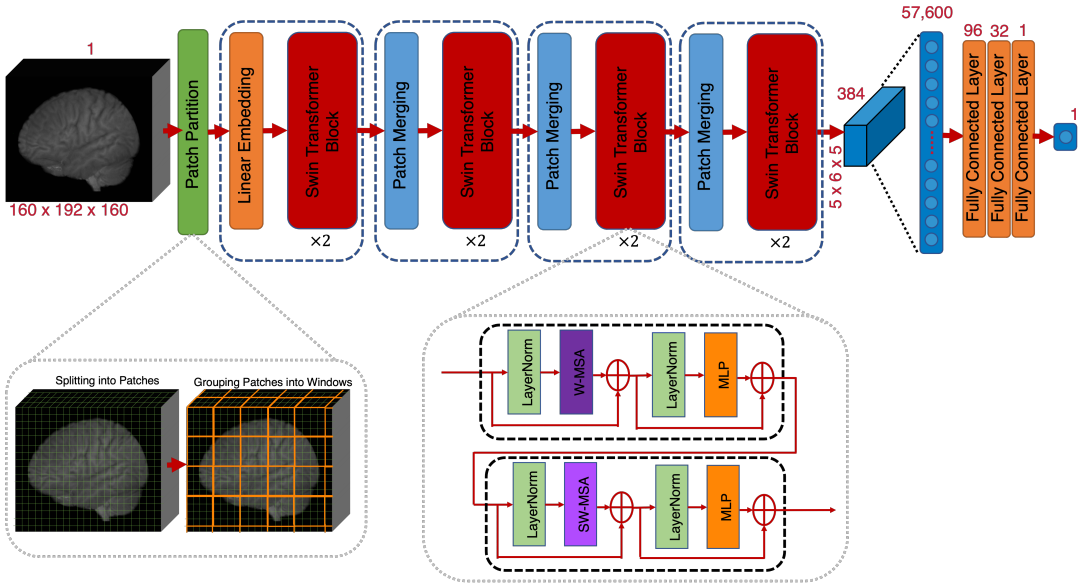


Figure 5.1: Vanilla BA-SWIN transformer architecture used throughout this chapter.

5.3.2 Experimental Setup

The experimental setup in this chapter follows the trends outlined in the previous chapters, to facilitate easier comparison between the HGL CNN and BA-SWIN Transformer networks. Thus, only the data for the female subjects was utilised, due to its larger volume as compared to the male dataset (15691 vs. 13640), and an anticipation of analogous results between the CNN and Transformer methods.

Also, 18 distinct 3D maps related to the core 5 MRI modalities were examined [41], their distribution being:

- 3 sMRI maps: T1 Nonlinear, T2 FLAIR Nonlinear, T2 Lesions;
- 1 swMRI map;
- 5 resting-state fMRI (rsfMRI) maps: rsfMRI-0, rsfMRI-2, rsfMRI-5, rsfMRI-10, rsfMRI-21;
- 1 task fMRI (tfMRI) map: tfMRI-1;
- 8 diffusion MRI (dMRI) maps: FA, L2, MD, MO, TBSS FA, TBSS ISOVF, TBSS L3, Summed Tracts;

For each map, a set of 9 identical networks were trained separately to convergence for both the HGL and BA-SWIN architectures. The exact number of identical networks required for the Transformer was confirmed independently, as described in the next section. For each group of identical networks, the reported results were calculated by averaging all predictions at the subject level.

Both network architectures were trained using the Mean Squared Error loss function. While the HGL networks were trained using the ADAM optimiser, the BA-SWINs utilised a slightly modified version named AdamW [243] in line with the original SWIN configuration [246]. This optimiser introduces a correction to the manner in which weight decay is calculated, decoupling it from the optimisation steps and adjusting the weights directly during the parameter update step. This is needed as in the vanilla ADAM algorithm, the weight decay term is integrated with the gradients, potentially leading to distortions in the weight updates due to the adaptive learning rates. AdamW mitigates this by applying weight decay directly to the weights, independent of the gradient calculations. HGL did not utilise any weight decay, but BA-SWIN employs a weight decay coefficient of $1e-5$. Overall, AdamW was found to result in better generalisation performance than ADAM when a non-zero weight decay coefficient is used [243].

BA-SWIN was trained for a maximum of 300 epochs, compared to HGL's 200, allowing for the potentially slower convergence associated with Transformers. To reduce overfitting, the training of both architectures was stopped if the validation loss did not improve for 40 epochs, with the network being saved at the epoch with the best validation loss. While HGL's initial learning rate was set to $1e-2$ and subsequently reduced by half whenever the validation loss did not improve for 15 epochs, BA-SWIN utilised an initial learning rate of $1e-4$ with 4 epochs of linear warm-up and a cosine annealing learning rate scheduler [356]. While some literature argues in favour of utilising pre-training steps when fitting Transformer networks [248], the original SWIN paper found that this would only bring about marginal gains for small datasets (ImageNet-1k) and that improvements rapidly diminish for larger datasets, such as ImageNet-22k [246]. Given this consideration, and the large size of the female dataset, it was decided not to utilise any pre-training steps, and train the BA-SWIN networks end-to-end.

To improve generalisability during training, voxels were randomly shifted by $0-2$ voxels along each axis [67] for both networks. The effectiveness of this augmentation for BA-SWIN was tested in the following section.

5.4 Finding an Adequate SWIN Architecture

In the rapidly evolving landscape of deep learning, Vision Transformers (ViTs) and their derivatives, like the SWIN architecture, have shown significant promise. Yet, their application in the domain of 3D medical imaging remains relatively nascent. Recognising the novelty of these architectures and the unique challenges posed by 3D medical imaging analysis, including brain age prediction, this section undertakes an exploration of hyperparameter optimisation for deploying SWIN transformers in brain age prediction. The approach to this optimisation is four-pronged. Initially, using the T1 Nonlinear map, several preliminary runs with the standard BA-SWIN architecture were made, gaining insight into the expected training duration for several dataset sizes. This foundational understanding was crucial as transformers, in general, are anticipated to have extended training

times compared to CNNs. With the overarching objective to comprehensively probe numerous hyperparameters, ensuring efficient training time was paramount. Following this first step, a phased hyperparameter search was initiated, with each phase building on the insights of its predecessor, progressively refining the network configurations. Once a robust configuration was found, its efficacy was validated across several maps. Following this systematic approach, a convergence study was undertaken to determine the required number of identical network runs needed to ensure BA-SWIN predictions accuracies converge.

5.4.1 Methods: Hyperparameter Search

Given the large number of hyperparameter variables that could be tuned, and the lack of intuition surrounding SWIN transformers due to their relative novelty, the hyperparameter search was conducted in four distinct phases:

1. **Phase 1 - Training Speed**. Given that Transformers are expected to take longer to train than CNNs, in this first phase three standard BA-SWIN models were trained using 3 progressively larger versions of the female subjects dataset, each using $1k$, $3k$, and $9k$ training subjects respectively. The $9k$ dataset is the same one utilised throughout this Thesis. The two smaller datasets were created using a similar methodology to the larger dataset, as described in Chapter 3 Section 3.3.1. The aim of this Phase was to provide intuition regarding the approximate training times needed for BA-SWIN, and find a good compromise between this and the dataset sizes. This is important as the subsequent hyperparameter search could be time and computationally demanding. As this is a preliminary step, only one network was trained for each condition.

2. **Phase 2 - Finding Adequate Hyperparameters**. In this Phase, different hyperparameters were tested to optimise the BA-SWIN network in terms of increasing accuracy, reducing training time and also reducing the number of trainable parameters. For each hyperparameter, 3 identical networks were

trained instead of 9. This was deemed a good compromise between noise attenuation and runtime, particularly in light of the numerous networks needing to be trained. The following hyperparameters and hypotheses were tested:

- The effectiveness of the random shift augmentation was tested: it was hypothesised that this augmentation might lead to performance degradation by interfering with the relative position bias added to the feature maps;
- The network depth: a shallower network with fewer trainable parameters could lead to faster training;
- Number of MSA heads: reducing the number of heads might allow more information to be distributed to each head, possibly enhancing the learning of more discriminative features. This could improve the model's ability to capture complex patterns and relationships, especially between shifted windows;
- Embedding feature dimensionality: having a smaller number of latent features could reduce the number of trainable parameters, improving speed and reducing possible overfitting;
- Number of warm-up epochs: adjusting this parameter will help ascertain the optimal number of warm-up epochs or determine if they are necessary for the training process;
- Weight decay penalty: varying this parameter can provide intuition on what is a good weight decay to improve the generalisation performance of the models;
- Cosine annealing number of iterations: while the original papers fixed this value at the maximum number of training epochs, it could be the case that a cyclical learning rate scheduler could help with convergence, particularly if the model gets stuck on an optimisation plateau, where it learns to predict only the population average age, or a local minimum;

- Patch size: increasing the patch size results in a reduced resolution of the input representation, aggregating more information into each individual patch. This larger context can aid early transformer blocks in capturing more extensive image features, which could be beneficial for tasks involving sparse maps. However, this aggregation could also lead to the loss of finer details. Conversely, reducing the patch size increases the resolution of the input representation, potentially capturing more detailed information at the cost of increasing the number of patches. While this approach might lead to an increase in the number of trainable parameters, potentially slowing down the training process, it allows the model to access more granular information from the outset;

3. **Phase 3 - Validating the Hyperparameters**. After finding a good set of hyperparameters, these were validated with several other maps. All experiments up to this point were carried out solely with the T1 Nonlinear map. To validate them, BA-SWIN networks were also trained for the FA, TBSS FA, rsfMRI-0, and Summed Tracts maps. These were chosen as they are representative of the various modalities within UK Biobank.

4. **Phase 4 - Convergence Study**. Once a good set of hyperparameters was found, a convergence study similar to the one in Chapter 4 was carried out for a standard BA-SWIN network and one with the chosen hyperparameters. The aim of this was to determine both the number of identical networks required for prediction convergence, but also to quantify any differences between the standard and optimised BA-SWIN networks.

5.4.2 Results

5.4.2.1 Training Speed Considerations

The results from the first set of tests conducted with BA-SWIN are presented in Table 5.1. Firstly, it can be observed that all BA-SWIN architectures are able to predict reasonable brain ages for the T1 Nonlinear map. Then, as expected, a positive

correlation can be observed between the number of subjects in a training dataset and the training times, both per epoch and total. Interestingly, when comparing the BA-SWIN to HGL for the large training dataset, it can be observed that, although achieving similar accuracies, the transformer architecture takes approximately $5\times$ longer than the CNN to train for each epoch. This observation justifies the need for a smaller SWIN architecture for the following hyperparameter search.

While Table 5.1 provides some information regarding the performance of the three BA-SWIN configurations, it is also important to consider their training performance. Figure 5.2 provides this information. Firstly, it can be observed that, for all three networks, a training plateau occurs in the first $20 - 30$ training epochs, where the network appears to be stuck either in a flat loss landscape or in a local minima. Evaluating the networks at epochs corresponding to these plateaus revealed they tend to make predictions close to the mean age of the population. This behaviour is confirmed by literature, which has found that the SA mechanism promotes flatter loss landscapes in transformers when compared to CNNs [357]. While useful in some cases, this behaviour can also be problematic if the networks are unable to escape this plateau.

Figure 5.2 also reveals that, although it facilitates the fastest training speeds per epoch, the small $1k$ subject dataset is unable to converge to an optimal value and overfits quickly after escaping the plateau. Meanwhile, it can be seen that the networks trained with $3k$ and $9k$ subjects display good convergence behaviour.

Based on these observations, the $3k$ subject dataset was chosen for the hyperparameter search in the following section, as it provides the best compromise between speed and accuracy. In addition, due to the long training times involved, the maximum number of training epochs for this baseline network was limited to 120, as Figure 5.2 indicates that at this point the network is mostly converged based on its validation curve.

Table 5.1: BA-SWIN transformer training speed for several dataset sizes and a single run per network. Results for an HGL CNN are added in the final row for comparison purposes. These were estimated based on the training runs from previous chapters.

Network Type	Number of Train Dataset Subjects	Total Train Epochs	Train Time (Days & Hours)	Approximate Time / Epoch	Single Run Test Dataset MAE (years)
BA-SWIN	1k	78	16.25 hours	8-10 minutes	4.05
	3k	223	5 days, 18 hours	15-30 minutes	2.85
	9.4k	115	8 days, 16 hours	1-2.5 hours	2.37
HGL CNN	9.4k	100	1-2 days	20-30 minutes	2.40

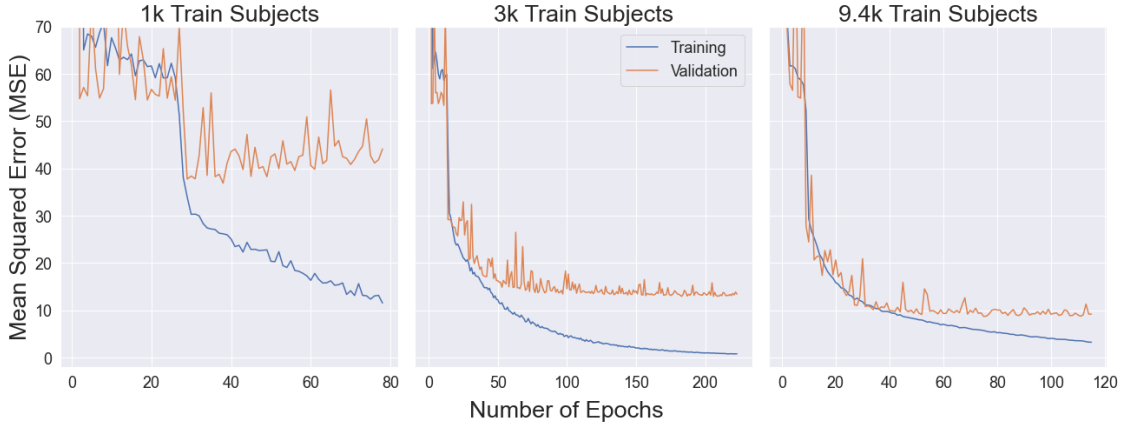


Figure 5.2: BA-SWIN training curves for three training dataset sizes. Training dataset curves are displayed in blue, while the results for the validation dataset are shown in orange.

5.4.2.2 Hyperparameter Search

Using the baseline configuration for BA-SWIN described in the previous subsection, the hyperparameter search investigation was then carried out. Tables 5.2-5.3 display the obtained results.

Firstly, it was observed that the random shift augmentation has a considerable positive impact on the prediction accuracies. Thus, for this reason, it was maintained for all other hyperparameter combinations.

Then, when considering the other experiments, they can be roughly split thematically into three categories, based on the performed augmentation: network architecture augmentations, network training augmentations and data decomposition augmentations. From Table 5.2, it can be seen that none of the hyperparameter changes versus the baseline from the former two categories leads to any major

improvements in any of the three metrics which the optimisation targets: accuracy, speed, number of trainable parameters. This changes when considering the latter category. It can be seen that by increasing the patch size, the training time per epoch decreases dramatically, being reduced by $70-73\%$ when increasing the patch size from 2 to 3, and up to $83-85\%$ when the patch size is set to 5. In addition, this major improvement in speed occurs without a large degradation in prediction accuracy, the $patch = 5$ network achieving an MAE comparable to the baseline.

Based on this observation, a further hyperparameter refinement was carried out using the $patch = 5$ network. Firstly, it was hypothesised that, given the larger amount of information in each patch, the network could benefit from a larger feature embedding dimension. However, as shown by the final three rows in Table 5.2, this was not the case. Moreover, these networks routinely became stuck on the plateaus observed in Figure 5.2 and discussed in the previous section. To address this, more experiments were run varying several parameters, including the initial learning rate, training patience, total number of training epochs, type of learning rate scheduler and the feature embedding dimensions (Table 5.3). Overall, it can be seen that replacing the linear warm-up and cosine annealing learning rate scheduler with a learning rate step scheduler, similar to that proposed for HGL, produces the best results, without impacting training performance in terms of prediction accuracy or speed. Moreover, when fixing the number of training epochs to 120, similar to the baseline in Table 5.2, only minor differences can be observed between the optimised networks and the baseline.

Based on these observations, a patch of 5 and a step learning rate scheduler were incorporated into the BA-SWIN architecture. Interestingly, despite a large increase in the number of training parameters, there are only small differences between the transformers trained with progressively larger embedding feature dimensions, making it difficult to decide on what the optimal value is for this hyperparameter. This could be due to the effectiveness of the attention mechanism to focus on only those features most relevant for brain age prediction, yet confirming this would require further testing. Regarding the other tested hyperparameters, as they did

Table 5.2: BA-SWIN hyperparameter search results (1/2) comparing an established Baseline model against several punctual configuration changes.

Tested Network Configuration	MAE	R^2	Predicted-Chronological Correlation (r)	Weighted MAE	Trainable Parameters (million)	Training Time/Epoch (minutes)
Baseline Configuration						
Baseline (3k Train Subjects)	2.828	0.711	0.878	0.076	7.7	30-60
Data Loader Augmentations						
No Random Shifting Augmentation	3.297	0.692	0.832	0.089	7.7	30-60
Network Architecture Augmentations						
Network Depth = (1,1,1,1)	2.818	0.774	0.880	0.076	7.0	20-40
Transformer Heads = (2, 4, 8, 16) ; Batch = 4	2.901	0.763	0.874	0.078	7.6	20-40
Embedding Feature Dimension = 3	2.829	0.771	0.878	0.076	0.9	25-50
Embedding Feature Dimension = 6	2.887	0.766	0.875	0.078	1.7	25-50
Embedding Feature Dimension = 12	2.842	0.771	0.878	0.077	3.5	25-50
Network Training Augmentations						
Number of Warmup Epochs = 0	63.420	0.000	0.003	1.708	7.7	30-60
Number of Warmup Epochs = 8	2.830	0.773	0.879	0.076	7.7	30-60
Number of Warmup Epochs = 12	2.866	0.770	0.877	0.077	7.7	30-60
Weight Decay Penalty = 1e-2	2.891	0.767	0.876	0.078	7.7	15-30
Weight Decay Penalty = 1e-3	3.435	0.764	0.874	0.092	7.7	17-34
Weight Decay Penalty = 1e-4	2.859	0.775	0.880	0.077	7.7	30-60
Weight Decay Penalty = 0	2.878	0.765	0.875	0.077	7.7	30-60
Cosine Annealing Iterations = 60	2.984	0.746	0.864	0.080	7.7	30-60
Cosine Annealing Iterations = 120	2.903	0.761	0.872	0.078	7.7	30-60
Cosine Annealing Iterations = 200	2.872	0.769	0.877	0.077	7.7	30-60
Data Decomposition Augmentations						
Patch Size = 3	2.938	0.758	0.871	0.079	4.5	8-16
Patch Size = 4	2.916	0.756	0.870	0.079	3.2	6-14
Patch Size = 5	2.847	0.772	0.879	0.077	2.6	4-10
Patch Size = 5, Embedding Features = 48	3.433	0.763	0.874	0.092	8.9	5-7
Patch Size = 5, Embedding Features = 72	4.684	0.740	0.860	0.126	19.2	5-9
Patch Size = 5, Embedding Features = 96	4.700	0.729	0.854	0.127	33.4	6-11

not produce any important improvements in performance over the baseline, they will retain their values as defined for the standard model.

5.4.2.3 Validating Hyperaprameter Findings with Several Maps

The refinement work carried out in the previous section managed to yield a BA-SWIN network configuration which substantially reduces training times without sacrificing accuracy. The following modifications have been made to the baseline network:

- The patch size was increased from 2 to 5 voxels;
- Given that each patch contains more information, the embedding feature dimension was increased from 24 to 48. Although this could be increased further, experiments indicate that no additional gains can be won by doing this;
- The learning rate scheduler was changed from cosine annealing with linear warmup to a step scheduler with a patience of 15 epochs;

Table 5.3: BA-SWIN hyperparameter search results (2/2) comparing an established Baseline model against several refined configuration changes.

Tested Network Configuration	MAE	R^2	Predicted-Chronological Correlation (r)	Weighted MAE	Trainable Parameters (million)	Training Time/Epoch (minutes)
Baseline Configuration						
Baseline (3k Train Subjects)	2.828	0.771	0.878	0.076	7.7	30-60
Patch Size = 5, Initial Learning Rate = 5e-3, Training Patience = 30, Training Epochs = 200						
Embedding Features = 48	6.248	0.000	-	0.168	8.9	5-7
Embedding Features = 72	6.249	0.009	0.094	0.168	19.2	5-9
Embedding Features = 96	6.254	0.000	-	0.168	33.4	6-11
Patch Size = 5, Training Patience = 40, Training Epochs = 200						
Embedding Features = 48	2.874	0.767	0.876	0.077	8.9	5-7
Embedding Features = 72	2.889	0.767	0.876	0.078	19.2	5-9
Embedding Features = 96	3.444	0.779	0.882	0.093	33.4	6-11
Patch Size = 5, Learning Patience = 40, Learning Rate Step Scheduler, Training Epochs = 200						
Embedding Features = 48	2.743	0.783	0.885	0.074	8.9	5-7
Embedding Features = 72	2.924	0.767	0.876	0.079	19.2	5-9
Embedding Features = 96	2.712	0.788	0.888	0.073	33.4	6-11
Patch Size = 5, Learning Patience = 40, Learning Rate Step Scheduler, Training Epochs = 120						
Embedding Features = 48	2.884	0.766	0.875	0.078	8.9	5-7
Embedding Features = 72	3.017	0.755	0.869	0.081	19.2	5-9
Embedding Features = 96	2.878	0.763	0.874	0.077	33.4	6-11

- Just for the hyperparameter search baseline, the maximum number of training epochs was increased from 120 to 200, with a training cessation patience of 40 epochs, similar to HGL. For the final version of BA-SWIN discussed in subsequent sections, the maximum number of training epochs will be set to 300, in line with the discussion presented in the methods.

Despite these modifications, two questions remain unanswered. Firstly, how will the networks optimised for T1 Nonlinear perform when encountering other maps? And, secondly, can the number of trainable parameters be further reduced so that the risks associated with overparameterisation might be mitigated?

Based on the observations from the previous section, the latter question can be addressed by either reducing the embedding feature dimension back to 24, or by reducing the depth of the network from 4 SWIN Transformer blocks to 3. Meanwhile, the former question can be addressed by performing experiments with several other modalities and maps. As the optimisation now tries to control for two variables, experiments were run addressing both factors, the results being presented in Table 5.4. From this, it can be seen that it is difficult to determine which configuration performs overall better than the others. A heuristic which can help is to assign

a score from 1-to-4 to each experiment for a given map, 4 corresponding to the best experiment in terms of MAE and 1 to the worst. Then, for each architecture configuration, add those scores across maps. Using this approach, it was observed that the network with a depth of $(2, 2, 2)$ and embedding feature dimension 48 performed slightly better than the other configurations. It also manages to reduce the number of trainable parameters from $8.7M$ to $5.0M$ and does not show any changes in the training time per epoch. Thus, this configuration was selected for the remaining experiments discussed in this chapter.

5.4.2.4 Prediction Convergence

The final part of the hyperparameter refinement work was concerned with finding an optimal number of identical runs at which the MAE prediction accuracy converges. Table 5.5 presents the results of the convergence study for standard and optimised BA-SWIN networks. The methodology utilised for this work is that described in Chapter 4 Section 4.6.1.1. It can be observed that both networks achieve $\approx 95\%$ convergence after 9 runs, which is similar to the HGL CNN architecture. It was also observed that while the standard model took between 1 and 2.75 hours to train per epoch, depending on the GPU architecture, the previously observed $\approx 90\%$ training time reduction was retained for the optimised architecture, with training durations varying between 9 and 15 minutes. It should be noted that this is also $\approx 50\%$ faster than the typical HGL CNN. In addition, despite the proposed modification, the two versions of the BA-SWIN architecture achieve similar MAE prediction accuracies, with 2.294 years for the standard model and 2.297 for the optimised architecture. Moreover, a high correlation 0.894 was observed between the brain age deltas of the two networks, suggesting a high degree of alignment between the two sets of results despite the introduction of several modifications.

Furthermore, in the light of results obtained with the SFCN network [67] in Chapter 3, the KDE distributions and density scatter plots of the standard (vanilla) and optimised (referred to from now on as simply BA-SWIN) architectures were observed (Figure 5.3). As no visible anomalies were detected, and given

Table 5.4: BA-SWIN multi-map hyperparameter refinement search results comparing several refined configurations across multiple maps. The best results per map for the MAE, R^2 and Pearson-r columns are highlighted in bold.

Tested Network Configuration (Patch Size = 5)	MAE	R^2	Predicted-Chronological Correlation (r)	Weighted MAE	Trainable Parameters (million)	Training Time/Epoch (minutes)
T1 Nonlinear						
Embedding Features = 24 Network Depth = (2,2,2,2)	2.815	0.776	**0.881**	0.076	2.6	5-7
Embedding Features = 48 Network Depth = (2,2,2,2)	2.769	**0.786**	0.886	0.075	8.9	5-7
Embedding Features = 24 Network Depth = (2,2,2)	**2.757**	0.779	0.883	0.074	2.1	5-7
Embedding Features = 48 Network Depth = (2,2,2)	2.791	0.778	0.882	0.075	5.0	5-7
FA						
Embedding Features = 24 Network Depth = (2,2,2,2)	**3.072**	**0.730**	0.854	0.083	2.6	5-7
Embedding Features = 48 Network Depth = (2,2,2,2)	3.113	0.721	0.849	0.084	8.9	5-7
Embedding Features = 24 Network Depth = (2,2,2)	3.153	0.716	**0.846**	0.085	2.1	5-7
Embedding Features = 48 Network Depth = (2,2,2)	3.076	0.727	0.853	0.083	5.0	5-7
TBSS FA						
Embedding Features = 24 Network Depth = (2,2,2,2)	3.475	0.656	0.810	0.094	2.6	5-7
Embedding Features = 48 Network Depth = (2,2,2,2)	**3.474**	0.656	0.810	0.094	8.9	5-7
Embedding Features = 24 Network Depth = (2,2,2)	3.500	**0.659**	0.812	0.094	2.1	5-7
Embedding Features = 48 Network Depth = (2,2,2)	3.506	0.650	**0.806**	0.094	5.0	5-7
rsfMRI-0						
Embedding Features = 24 Network Depth = (2,2,2,2)	5.401	0.225	**0.474**	0.145	2.6	5-7
Embedding Features = 48 Network Depth = (2,2,2,2)	5.395	0.230	0.480	0.145	8.9	5-7
Embedding Features = 24 Network Depth = (2,2,2)	5.317	0.234	0.484	0.143	2.1	5-7
Embedding Features = 48 Network Depth = (2,2,2)	**5.305**	**0.252**	0.502	0.143	5.0	5-7
Summed Tracts						
Embedding Features = 24 Network Depth = (2,2,2,2)	3.712	0.611	0.781	0.100	2.6	5-7
Embedding Features = 48 Network Depth = (2,2,2,2)	3.683	0.627	0.792	0.099	8.9	5-7
Embedding Features = 24 Network Depth = (2,2,2)	3.629	0.633	0.795	0.098	2.1	5-7
Embedding Features = 48 Network Depth = (2,2,2)	**3.569**	**0.643**	**0.802**	0.096	5.0	5-7

Table 5.5: BA-SWIN transformer independence study numerical results. Both transformer architectures were tested: the standard, or vanilla architecture, originally described at the start of this work, and the optimised architecture resulted from the hyperparameter search. Both the MAE and Convergence (Conv.) results are presented. The Convergence results represent, in percentages, how close an MAE value is to the best MAE achieved by that experiment. The point where each convergence metric reaches a convergence of $\approx 95\%$ is displayed in bold.

Number of Networks in Ensemble / Ensemble Type	Standard BA-SWIN						Optimised BA-SWIN)					
	Naive		Combinatorial		Bootstraped		Naive		Combinatorial		Bootstraped	
	MAE (years)	Conv. (%)	MAE (years)	Conv. (%)	MAE (years)	Conv. (%)	MAE (years)	Conv. (%)	MAE (years)	Conv. (%)	MAE (years)	Conv. (%)
1	2.397	0.0	2.449	0.0	2.448	0.0	2.456	0.0	2.461	0.0	2.461	0.0
2	2.333	0.591	2.362	0.541	2.369	0.542	2.374	0.496	2.37	0.536	2.377	0.541
3	2.331	0.611	2.333	0.726	2.343	0.722	2.349	0.649	2.339	0.719	2.349	0.722
4	2.323	0.682	2.318	0.818	2.33	0.812	2.328	0.772	2.323	0.81	2.334	0.816
5	2.327	0.643	2.309	0.873	2.321	0.871	2.308	0.897	2.314	0.866	2.327	0.865
6	2.313	0.779	2.303	0.91	2.316	0.906	2.312	0.874	2.307	0.903	2.32	0.907
7	2.304	0.858	2.299	0.935	2.312	0.936	2.301	0.941	2.303	0.93	2.316	0.934
8	2.298	0.917	**2.296**	**0.955**	**2.309**	**0.953**	2.292	0.995	**2.299**	**0.951**	**2.313**	**0.952**
9	**2.294**	**0.954**	2.294	0.97	2.306	0.972	**2.297**	**0.962**	2.296	0.967	2.311	0.968
10	2.286	1.025	2.292	0.982	2.305	0.983	2.292	0.991	2.294	0.981	2.309	0.981
11	2.287	1.018	2.29	0.992	2.303	0.993	2.291	0.997	2.292	0.992	2.307	0.993
12	2.289	1.0	2.289	1.0	2.302	1.0	2.291	1.0	2.291	1.0	2.306	1.0

that the predicted distributions resemble in shape the ground truth distribution, the BA-SWIN architecture was considered suitable for the subsequent brain age prediction experiments.

5.4.3 Discussion

The overarching objective of this section was to explore a broad range of hyperparameters and find an optimal SWIN architecture for predicting brain age. An evolutionary approach was adopted, starting from a foundational, or "vanilla", architecture based on the SWIN [246] and SWIN-UNETR [247] models, and moving systematically towards a more refined configuration.

The first important finding of this section is that the proposed vanilla SWIN transformer architecture is able to predict brain age without relying on pretraining, underscoring its inherent robustness and ability to adapt to domain-specific tasks [248]. However, this capability does come with caveats. Specifically, the vanilla architecture requires approximately 8 times more trainable parameters than the HGL CNN, which also contributes to the longer training times. Given this challenge, optimisation work was focused towards achieving three distinct yet

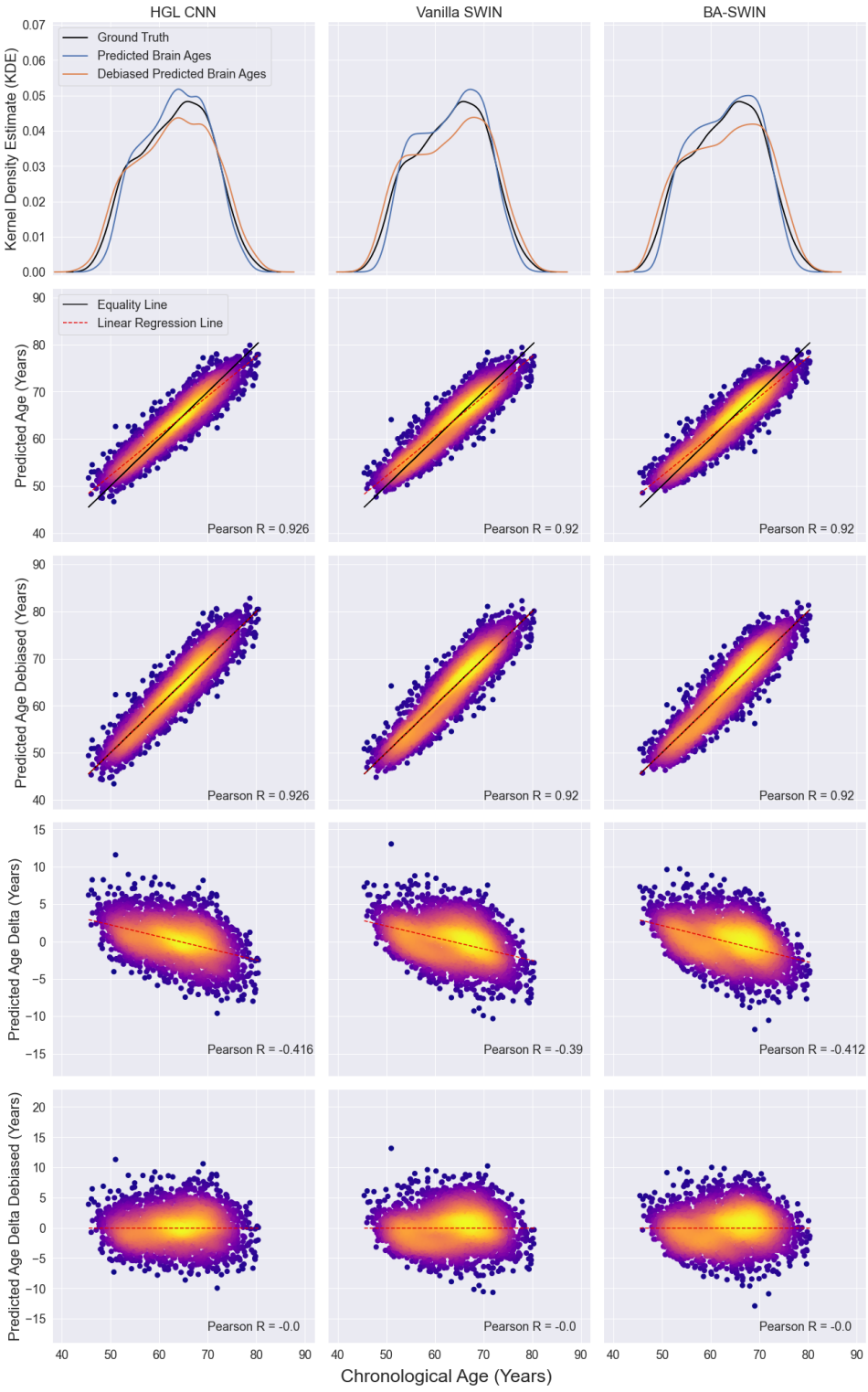


Figure 5.3: Kernel density estimates (KDE) and density plots (DP) of BA-SWIN predicted ages for the female dataset. The first column of the figure corresponds to a typical HGL CNN result, while the later two columns to the standard/vanilla BA-SWIN and the optimised BA-SWIN. Each row corresponds to a plot with respect to chronological age: KDE of predicted and debiased predicted ages distributions for the left-out subjects, DPs of predicted age and predicted age linearly debiased, and DPs of predicted age deltas and predicted age deltas debiased.

interrelated objectives: enhancing training efficiency, improving predictive accuracy, and reducing the number of trainable parameters.

The most important finding of this effort is concerned with the impact that the input patch size has on training efficiency. By adjusting the patch size from 2 to 5 voxels, a large $85-90\%$ reduction in computational time was achieved. Crucially, this increase was not accompanied by a degradation in prediction accuracy, indicating that the network still managed to effectively extract relevant information. This suggests that larger patch sizes may inherently encapsulate sufficient contextual cues to drive the transformer's attention mechanism without necessitating higher granularity. Furthermore, it was found that a shallower network architecture could facilitate a reduction in the number of training parameters without impacting the overall network performance.

5.5 SWIN vs. CNN: Single Map Predictions

Building on the hyperparameter refinement work presented above, which resulted in an optimised BA-SWIN architecture, in this section a comparison is carried out between this method and the HGL CNN architecture employed throughout this thesis. Experiments were carried out for the same 18 maps described in Chapter 4. These were selected not only because they are representative of the data contained in UK Biobank, but also because results for the HGL network were already available for them. These CNN-baseline results were compared against those obtained with the BA-SWIN transformer in terms of their summary statistics and associations to UK Biobank nIDPs.

5.5.1 Results

The numerical results for the comparison between HGL and BA-SWIN are presented in Table 5.6. Figures 5.4-5.5 enable a qualitative comparison between the nIDP associations obtained with the two methods for a group of maps, while Figure 5.6 and Table 5.7 facilitate a quantitative comparison using Bland-Altman plots and correlations between the calculated $-log(p)$ vectors respectively.

Generally, HGL and BA-SWIN can be seen to produce very similar results, with BA-SWIN generally slightly underperforming HGL. The only instances where this is not the case is for the L2 and MD maps (Table 5.6), where BA-SWIN performed marginally better than HGL. In addition, in the case of the maps derived from rsfMRI and tfMRI, it can be observed that BA-SWIN performed visibly worse than the equivalent CNN. This could be caused by the fact that these maps, given that they are z-statistic maps, are generally more noisy than those derived from other modalities. However, further testing is required to confirm this hypothesis.

Table 5.6 also reveals high correlations between the predicted brain age deltas from the two methods. However, these are not as high as those seen in Chapter 4 Sections 4.7- 4.6, in Tables 4.3, 4.5, 4.9 and 4.10. This could potentially indicate that, while the CNNs and Transformers converge to similar brain age prediction accuracies, they might be capturing different aspects of brain ageing.

While an initial examination of Figures 5.4-5.5 appears to indicate similar nIDP associations between the HGL and BA-SWIN, a more detailed investigation reveals certain subtle differences. For instance, across several maps, such as FA (Figures 5.4c-5.4d), rsfMRI-5 (Figures 5.5a-5.5b) and dMRI Summed Tracts (Figures 5.5e-5.5f), it can be seen that the CNN methods have multiple significant associations from the Physical Measurements category passing the Bonferroni threshold, while this does not seem to be the case for the Transformer. BA-SWIN also presents several differences vs. HGL. For instance, for the T2 FLAIR Nonlinear map (Figures 5.4a-5.4b), it shows a significant association to a Lifestyle measurement, while for the FA map (Figures 5.4c-5.4d) it shows many more significant associations to Medical History nIDPs than the equivalent CNN results. The observed discrepancies also manifest in the correlations between the $-log(p)$ values seen in Table 5.7. For example, despite the prevalent high correlations between the two methods across various maps, notably smaller correlations were recorded for the T2 FLAIR Nonlinear and FA maps.

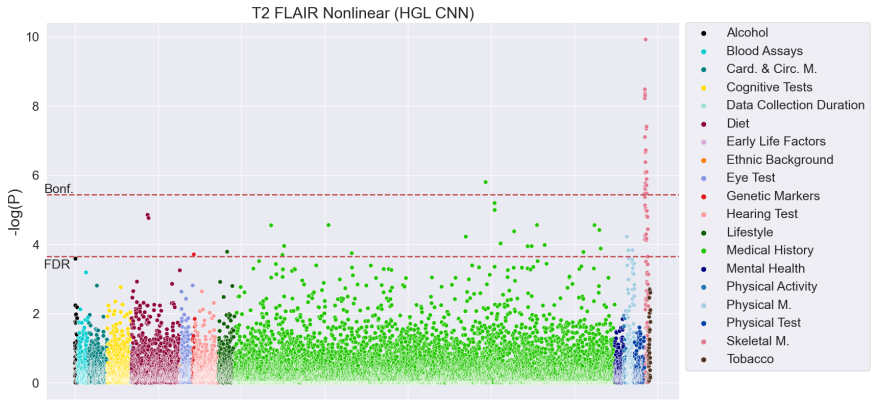
Tables 5.8 and 5.9 present all nIDP associations which passed the Bonferroni threshold for the BA-SWIN architecture but not for HGL, and vice versa, respectively.

Table 5.6: BA-SWIN vs. HGL results split by the network architecture utilised for each map. The final two columns indicate the correlations between the Original and Converged brain age deltas in their raw and debiassed forms. The Weighted MAE, which allows for easier comparison between this and other studies, was calculated using the method proposed by Cole et al [8]. The bolded results indicate metrics where BA-SWIN outperformed HGL.

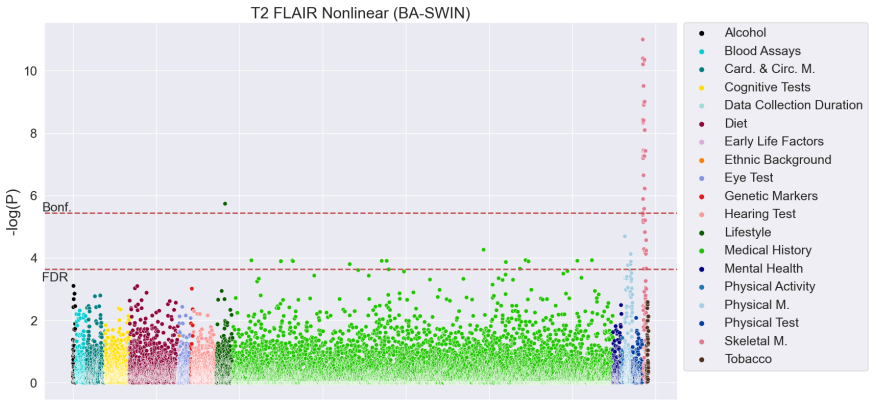
Map	BA-SWIN				HGL				BA-SWIN v. HGL CNN	
	MAE	R^2	Predicted-Chronological Correlation (r)	Weighted MAE	MAE	R^2	Predicted-Chronological Correlation (r)	Weighted MAE	Delta Correlation (r)	Delta (Linear Debias) Correlation (r)
sMRI Maps										
T1 Nonlinear	2.366	0.836	0.914	0.064	2.190	0.862	0.926	0.059	0.839	0.811
T2 FLAIR Nonlinear	2.222	0.855	0.924	0.060	2.086	0.867	0.933	0.056	0.844	0.820
T2 Lesions	4.076	0.507	0.712	0.110	3.936	0.540	0.733	0.106	0.921	0.856
SWI	3.034	0.724	0.851	0.082	2.939	0.749	0.862	0.079	0.872	0.827
rsfMRI Maps										
rsfMRI-0	4.846	0.332	0.576	0.130	4.220	0.518	0.700	0.114	0.872	0.710
rsfMRI-2	4.856	0.331	0.575	0.131	4.159	0.507	0.705	0.112	0.876	0.717
rsfMRI-5	4.876	0.322	0.568	0.131	4.160	0.502	0.700	0.112	0.859	0.679
rsfMRI-10	4.832	0.320	0.566	0.130	4.243	0.487	0.690	0.114	0.875	0.678
rsfMRI-21	4.862	0.323	0.568	0.131	4.060	0.520	0.716	0.109	0.861	0.679
tfMRI Maps										
tfMRI-1	3.812	0.575	0.758	0.103	3.407	0.647	0.812	0.092	0.848	0.762
dMRI Maps										
Summed Tracts	3.236	0.684	0.827	0.087	3.070	0.742	0.847	0.083	0.844	0.787
TBSS FA	3.120	0.714	0.845	0.084	2.903	0.752	0.866	0.078	0.889	0.847
TBSS ISOVF	3.384	0.658	0.811	0.091	3.286	0.700	0.828	0.088	0.890	0.838
TBSS L3	3.254	0.689	0.830	0.088	3.132	0.724	0.841	0.084	0.878	0.826
FA	2.680	0.786	0.887	0.072	2.587	0.801	0.895	0.070	0.844	0.798
L2	**2.787**	0.768	0.876	0.075	2.806	0.782	0.877	0.076	0.875	0.840
MD	**2.767**	0.773	0.879	**0.074**	2.804	0.780	0.876	0.075	0.885	0.851
MO	3.022	0.726	0.852	0.081	2.873	0.801	0.869	0.077	0.861	0.811

Although no straightforward conclusion can be drawn, these tables support the observations made using the Manhattan plots, in the sense that the Transformers appear to capture more nIDPs related to factors such as Diet, Lifestyle, and Medical History, while the CNNs generally capture more details relating to factors which are predominantly intrinsic, such as Physical Measurements.

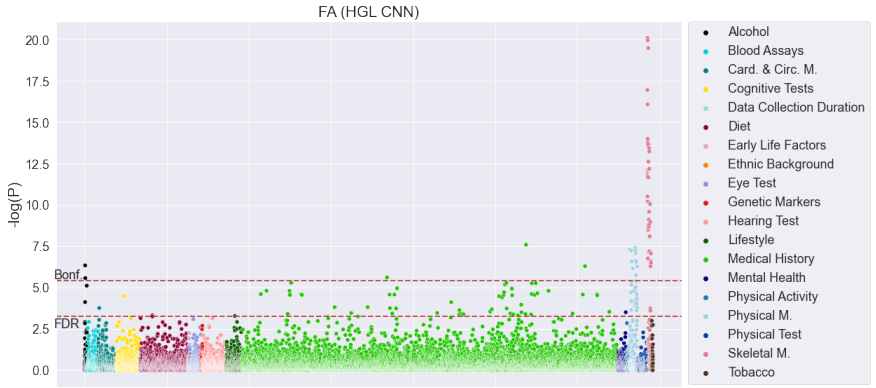
To complete the observations regarding nIDP associations, when examining the Bland-Altman plots in Figure 5.6, it can be seen that, generally, the CNNs produced more significant correlations, represented by higher $-log(p)$ values. However, this is not always the case, as seen for the T2 Lesions (Figure 5.6c) and dMRI Summed Tracts (Figure 5.6f). In the case of the T2 Lesions, this observation could be explained by the limited amount of information encoded in each of the binary masks, leading to both BA-SWIN and HGL learning identical brain age prediction features. For the dMRI Summed Tracts, the better performance of the Transformer model could be explained by its larger effective receptive field, which enables it to be better at capturing long-range information than the equivalent CNN.



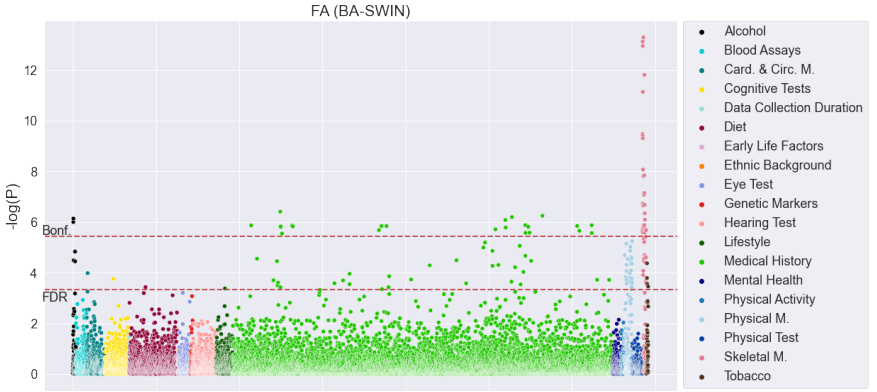
(a) HGL: T2 FLAIR Nonlinear



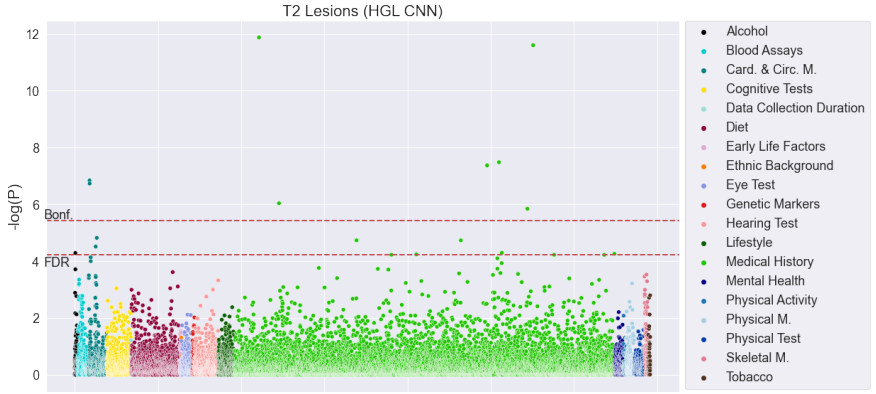
(b) BA-SWIN: T2 FLAIR Nonlinear



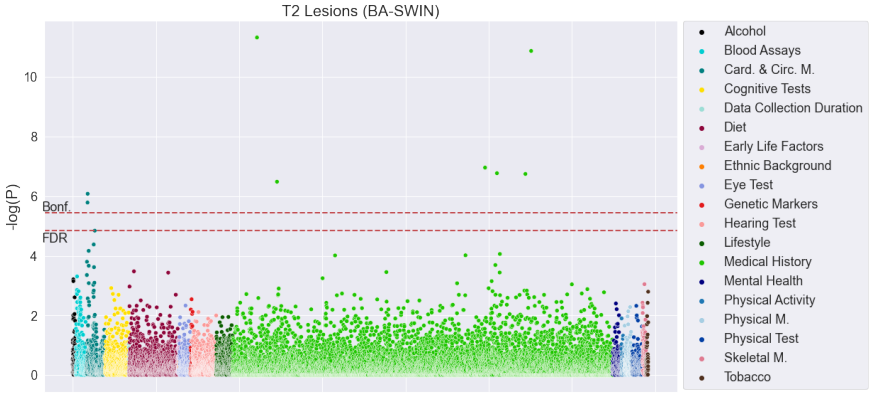
(c) HGL: FA



(d) BA-SWIN: FA

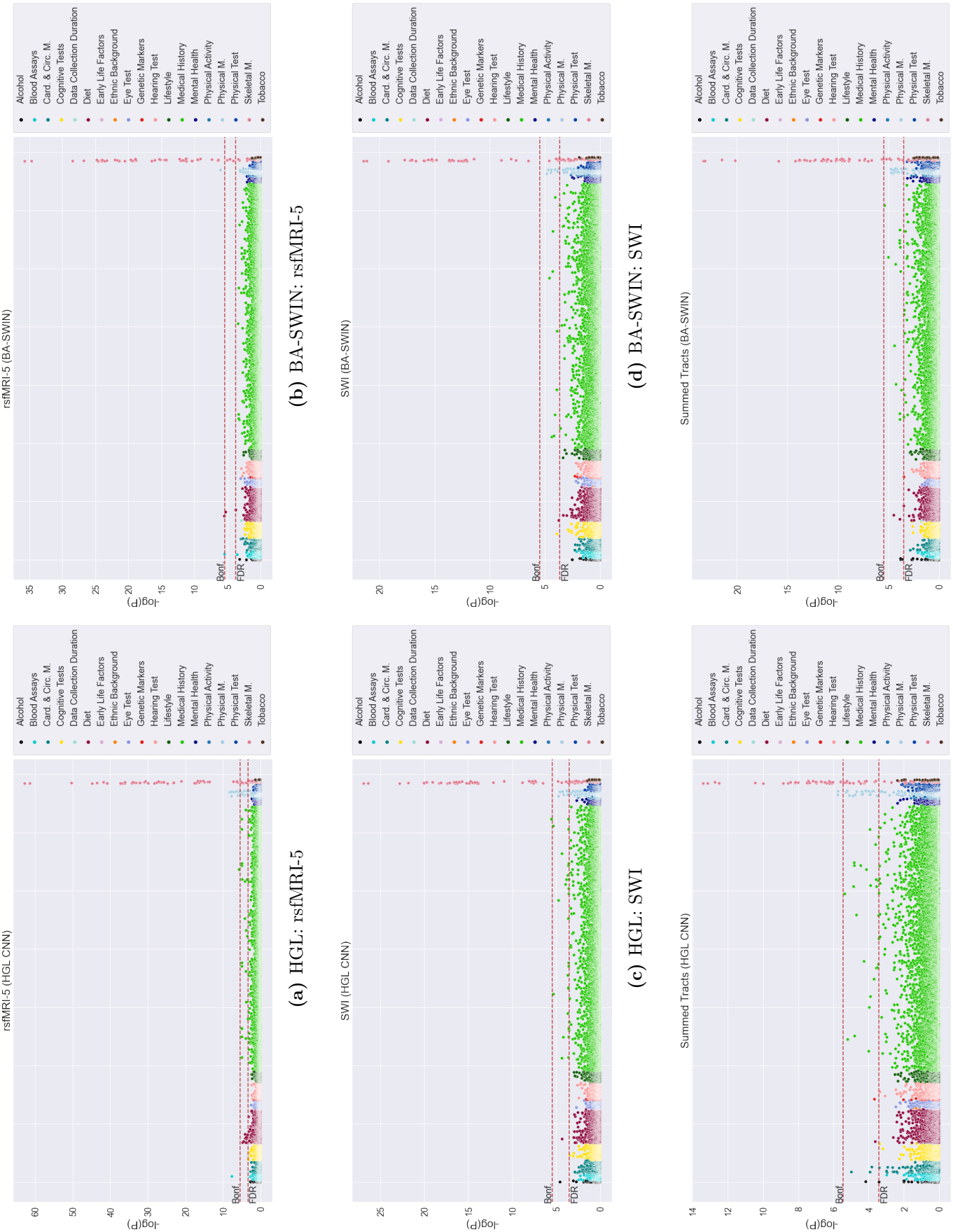


(e) HGL: T2 Lesions



(f) BA-SWIN: T2 Lesions

Figure 5.4: Manhattan Plots Relating Brain Age Deltas to UK Biobank nIDPs for BA-SWIN and HGL (1/2) for a subset of clusters, each dot representing the statistical significance of the correlation. The False Discovery Rate (FDR) and Bonferroni thresholds are also plotted.

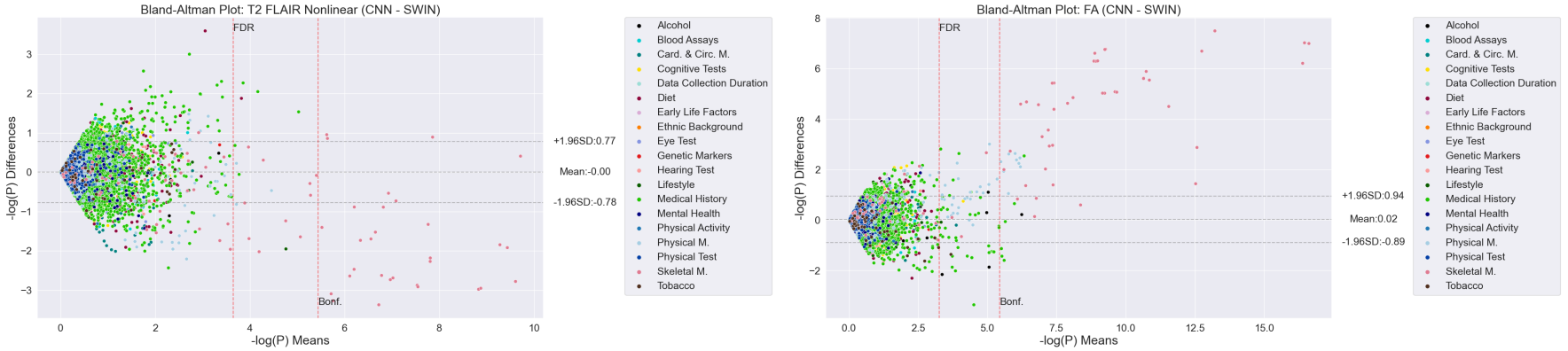


(a) HGL: rsfMRI-5 (b) BA-SWIN: rsfMRI-5

(c) HGL: SWI (d) BA-SWIN: SWI

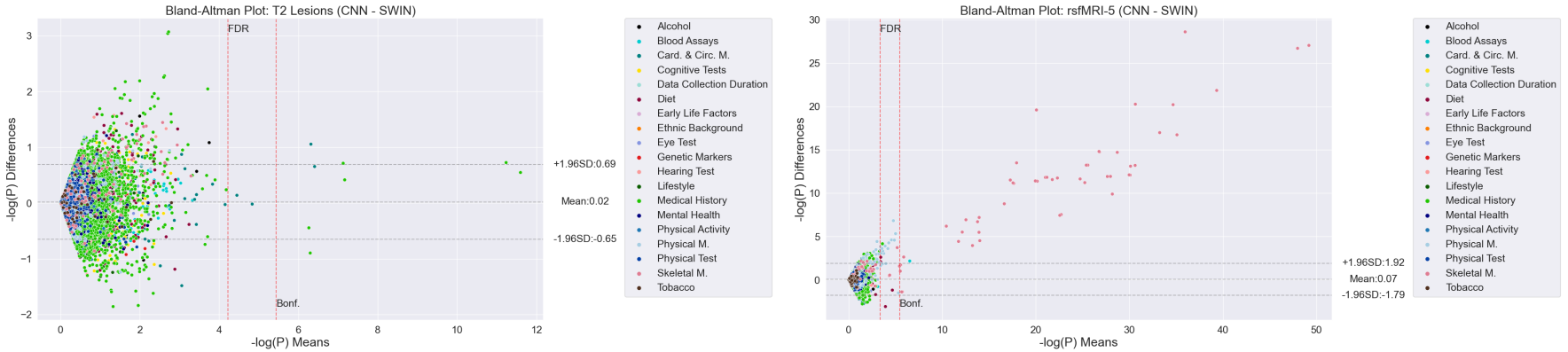
(e) HGL: Summed Tracts (f) BA-SWIN: Summed Tracts

Figure 5.5: Manhattan Plots Relating Brain Age Deltas to UK Biobank nIDPs for BA-SWIN and HGL (2/2) for a subset of clusters, each dot representing the statistical significance of the correlation. The False Discovery Rate (FDR) and Bonferroni thresholds are also plotted.



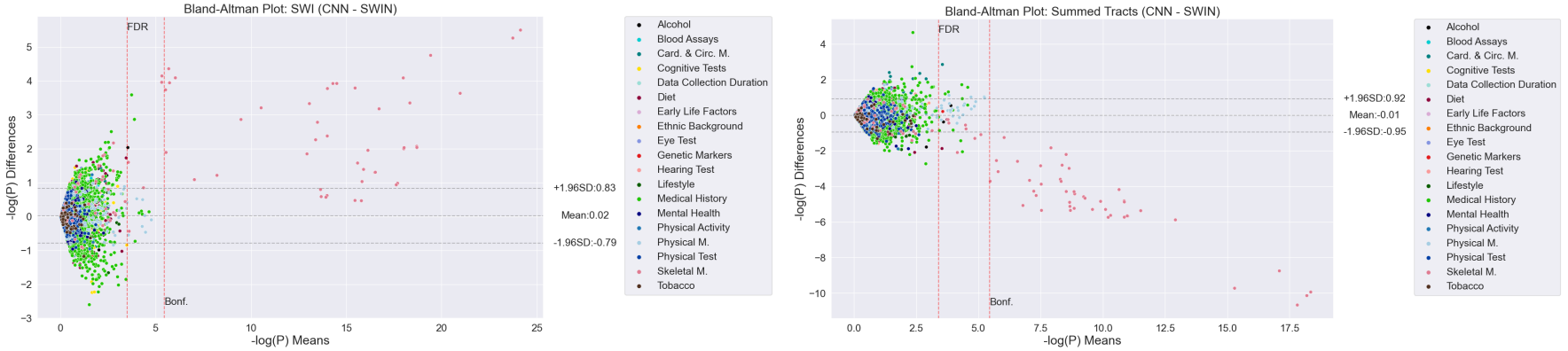
(a) T2 FLAIR Nonlinear

(b) FA



(c) T2 Lesions

(d) rsfMRI-5



(e) SWI

(f) Summed Tracts

Figure 5.6: Bland-Altman plots comparing $-log(p)$ for the BA-SWIN and HGL networks for a subset of maps. The $-log(p)$ marks the statistical significance of an nIDP-brain age delta association. The Bonferroni and FDR thresholds appear as red dotted lines. Though different, the FDRs are approximately equal. The mean of the difference and the 95% limits of agreement appear as grey lines.

Table 5.7: Correlations between the $-log(p)$ values obtained with BA-SWIN and HGL. These were calculated for all nIDPs, as well as for only those nIDPs passing the False Discovery Rate (FDR) thresholds calculated for the original and converged datasets.

Map	-log(p) BA-SWIN v. HGL CNN Correlations (r)		
	All	Passing BA-SWIN FDR Threshold	Passing HGL CNN FDR Threshold
sMRI Maps			
T1 Nonlinear	0.862	0.964	0.963
T2 FLAIR Nonlinear	0.806	0.809	0.802
T2 Lessions	0.801	0.897	0.896
SWI	0.918	0.980	0.981
rsfMRI Maps			
rsfMRI-0	0.898	0.970	0.971
rsfMRI-2	0.963	0.989	0.981
rsfMRI-5	0.928	0.979	0.978
rsfMRI-10	0.893	0.988	0.988
rsfMRI-21	0.881	0.973	0.973
tfMRI Maps			
tfMRI-1	0.917	0.979	0.980
dMRI Maps			
Summed Tracts	0.849	0.939	0.938
TBSS FA	0.881	0.908	0.908
TBSS ISOVF	0.845	0.874	0.873
TBSS L3	0.836	0.914	0.914
FA	0.852	0.810	0.809
L2	0.878	0.953	0.952
MD	0.896	0.951	0.950
MO	0.911	0.981	0.981

Table 5.8: Associations between UK Biobank nIDPs and brain age deltas appearing in BA-SWIN but not HGL, with *positive correlations* suggesting accelerated brain ageing. For brevity, the following abbreviations have been used: M. for Measurements, and Card. & Circ. for Cardiac & Circulatory.

Map	-log(p)	Correlation (Pearson r)	Variable Category	Variable Description	Variable ID
rsfMRI-5	5.47	0.178	Diet	*Drinking water intake (2.0)*	100150-2.0
T2 FLAIR Nonlinear	5.736	-0.137	Lifestyle	Mother's age at death (0.0)	3526-0.0
FA	6.196	0.099	Medical History	*Treatment/medication code (1140884600 - metformin)*	20003-0.1140884600
FA	6.08	0.098	Medical History	*Treatment/medication code (1140872198 - sodium valproate)*	20003-0.1140872198
FA	5.878	0.096	Medical History	*Diagnoses - secondary ICD10 (G403 - G40.3 Generalised idiopathic epilepsy and epileptic syndromes)*	41204-0.3303
FA	5.873	0.096	Medical History	*Diagnoses - main ICD10 (E230 - E23.0 Hypopituitarism)*	41202-0.2428
FA	5.866	0.096	Medical History	*Operative procedures - secondary OPCS4 (Z584 - Z58.4 Tibialis anterior)*	41210-0.10584
FA	5.862	0.096	Medical History	*Operative procedures - secondary OPCS4 (A843 - A84.3 Nerve conduction studies)*	41210-0.533
FA	5.847	0.096	Medical History	*Operative procedures - main OPCS4 (T811 - T81.1 Percutaneous biopsy of muscle)*	41200-0.7301
FA	5.841	0.096	Medical History	*Operative procedures - OPCS4 (T811 - T81.1 Percutaneous biopsy of muscle)*	41272-0.7301
FA	5.841	0.096	Medical History	*Diagnoses - ICD10 (M7969 - M79.69 Pain in limb (Site unspecified))*	41270-0.8949
FA	5.827	0.096	Medical History	*Diagnoses - ICD10 (E274 - E27.4 Other and unspecified adrenocortical insufficiency)*	41270-0.2456
FA	5.826	0.096	Medical History	*Operative procedures - secondary OPCS4 (A842 - A84.2 Electromyography)*	41210-0.532
FA	5.815	0.096	Medical History	*Diagnoses - secondary ICD10 (E230 - E23.0 Hypopituitarism)*	41204-0.2428
FA	5.798	0.096	Medical History	*Diagnoses - ICD10 (E230 - E23.0 Hypopituitarism)*	41270-0.2428
FA	5.783	0.095	Medical History	*Treatment/medication code (1140871732 - buprenorphine)*	20003-0.1140871732
FA	5.577	0.094	Medical History	*Diagnoses - secondary ICD10 (F009 - F00.9 Dementia in Alzheimer's disease, unspecified)*	41204-0.2713
FA	5.575	0.094	Medical History	*Diagnoses - secondary ICD10 (G309 - G30.9 Alzheimer's disease, unspecified)*	41204-0.3274
FA	5.545	0.093	Medical History	*Diagnoses - ICD10 (F009 - F00.9 Dementia in Alzheimer's disease, unspecified)*	41270-0.2713
rsfMRI-2	5.546	-0.253	Physical M.	Android bone mass (2.0)	23244-2.0
rsfMRI-2	6.107	-0.167	Skeletal M.	Heel quantitative ultrasound index (QUI), direct entry (right) (0.0)	4123-0.0
rsfMRI-2	5.511	-0.157	Skeletal M.	Heel quantitative ultrasound index (QUI), direct entry (left) (0.0)	4104-0.0
Summed Tracts	5.509	-0.158	Skeletal M.	Heel broadband ultrasound attenuation (right) (0.0)	4120-0.0
rsfMRI-2	5.496	-0.157	Skeletal M.	Heel broadband ultrasound attenuation (left) (0.0)	4101-0.0

Table 5.9: Associations between UK Biobank nIDPs and brain age deltas appearing in HGL but not BA-SWIN, with *positive correlations* suggesting accelerated brain ageing. For brevity, the following abbreviations have been used: M. for Measurements, and Card. & Circ. for Cardiac & Circulatory.

Map	-log(p)	Correlation (Pearson r)	Variable Category	Variable Description	Variable ID
MD	5.824	0.119	Alcohol	*Amount of alcohol drunk on a typical drinking day (0.0)*	20403-0.0
FA	5.584	0.094	Alcohol	*Alcohol intake frequency. (0.0)*	1558-0.0
T1 Nonlinear	6.237	0.1	Medical History	*Non-cancer illness code, self-reported (1134 - oesophageal disorder)*	20002-0.1134
rsfMRI-21	5.619	0.094	Medical History	*Operative procedures - main OPCS4 (X369 - X36.9 Unspecified blood withdrawal)*	41200-0.9159
L2	5.597	0.094	Medical History	*Treatment/medication code (1140883066 - insulin product)*	20003-0.1140883066
T1 Nonlinear	5.454	0.092	Medical History	Operative procedures - OPCS4 (A842 - A84.2 Electromyography)	41272-0.532
FA	7.082	-0.113	Physical M.	Total mass (2.0)	23283-2.0
TBSS ISOVF	6.891	-0.107	Physical M.	Arm fat percentage (right) (2.0)	23119-2.0
rsfMRI-21	6.881	-0.111	Physical M.	Android total mass (2.0)	23248-2.0
FA	6.289	-0.102	Physical M.	Arm fat mass (right) (2.0)	23120-2.0
FA	6.164	-0.1	Physical M.	Hip circumference (2.0)	49-2.0
rsfMRI-21	6.001	-0.099	Physical M.	Arm fat percentage (left) (2.0)	23123-2.0
rsfMRI-21	5.944	-0.098	Physical M.	Body mass index (BMI) (2.0)	21001-2.0
rsfMRI-5	5.854	-0.102	Physical M.	Trunk total mass (2.0)	23287-2.0
TBSS ISOVF	5.822	-0.102	Physical M.	Android tissue fat percentage (2.0)	23247-2.0
rsfMRI-21	5.729	-0.096	Physical M.	Leg fat mass (right) (0.0)	23112-0.0
TBSS ISOVF	5.709	-0.097	Physical M.	Leg fat percentage (right) (2.0)	23111-2.0
TBSS ISOVF	5.697	-0.095	Physical M.	Trunk fat mass (0.0)	23128-0.0
TBSS ISOVF	5.67	-0.096	Physical M.	Leg fat percentage (left) (2.0)	23115-2.0
TBSS ISOVF	5.607	-0.094	Physical M.	Body fat percentage (0.0)	23099-0.0
tfMRI-1	5.599	-0.099	Physical M.	Legs total mass (2.0)	23277-2.0
TBSS ISOVF	5.573	-0.094	Physical M.	Whole body fat mass (0.0)	23100-0.0
TBSS ISOVF	5.529	-0.099	Physical M.	Arms tissue fat percentage (2.0)	23259-2.0
rsfMRI-21	5.526	-0.095	Physical M.	Leg fat-free mass (right) (2.0)	23113-2.0
tfMRI-1	5.446	-0.098	Physical M.	Gynoid total mass (2.0)	23265-2.0
rsfMRI-21	8.158	-0.266	Skeletal M.	Femur upper neck BMD (bone mineral density) T-score (right) (2.0)	23217-2.0
rsfMRI-21	5.991	-0.103	Skeletal M.	Spine bone area (2.0)	23311-2.0
rsfMRI-0	5.907	-0.234	Skeletal M.	Heel quantitative ultrasound index (QUI), direct entry (left) (1.0)	4104-1.0
rsfMRI-0	5.749	-0.231	Skeletal M.	Speed of sound through heel (left) (1.0)	4103-1.0
rsfMRI-21	5.69	-0.1	Skeletal M.	Trunk bone area (2.0)	23304-2.0

5.5.2 Discussion

The core assumption which underpins the "brain age paradigm" is that observed errors are a combination of random Gaussian noise and structured biologically relevant errors. The goal of any regression method, trained with a loss function which encodes a Gaussian prior in the distribution of noise, such as the mean squared error (MSE), is to filter out the noise component, leaving only the biologically relevant error. Thus, in an ideal scenario, any two or more methodologically different models trained on the same population would produce identical results.

When looking at the comparison between the Transformer-based BA-SWIN and the CNN-based HGL models, it can be seen that, given sufficient training samples and iterations, they converge to very similar results when trained and evaluated on identical subject populations, despite the fundamental architectural differences between the two networks. Yet, while being very similar, the evaluation results are not identical. Thus, for a good comparison between the two models, a good understanding is required of the factors that promote these discrepancies, but also

of those which promote similarities between the methods.

The first factor that could be responsible for the observed discrepancies is the presence of residual Gaussian noise in the results. Across all experiments, both networks were trained to $\approx 95\%$ prediction accuracy convergence. This was considered sufficient to ensure result stability while preventing computational resource over-usage. However, this does mean that a residual amount of noise, originating from aspects such as weight initialisation or the stochastic behaviour of the ADAM and AdamW optimisers, could remain in the observed brain age deltas.

In addition, the degree to which the two models can overfit or otherwise be impacted by noise can also be influenced by their number of trainable parameters. While HGL has $1.3M$ trainable parameters, BA-SWIN is $3.8\times$ larger, employing $5.0M$. Generally, a larger number of parameters can sometimes help models fit the training data better and avoid getting stuck in local minima, as they have more degrees of freedom [358, 359]. This can, however, also be detrimental, as a larger model may be more prone to overfitting to the data, which could lead to reduced performance in terms of generalisation and higher variance in the test result [360]. The attention mechanism could help mitigate this, particularly at inference [347], however, the results presented in Table 5.6 suggest that the performance of BA-SWIN degrades more for inherently noisy maps, such as those derived from functional MRI modalities. While additional experiments are required to test this hypothesis, these considerations regarding the impact of noise on results could also be supported by the observation that, even for less inherently noisy maps such as those derived from sMRI, the nIDP associations obtained with HGL are more statistically significant than those obtained with BA-SWIN, as indicated by the Bland-Altman plots in Figure 5.6.

This brings the discussion towards the nature of data encoded by the various maps, and how this impacts the performance of the two networks. It can be seen that, primarily for maps encoding geometric information, HGL and BA-SWIN converge to very similar results both in terms of their prediction accuracy and nIDP associations. This is most clear for the T2 Lesions map, where very high

correlations for both the brain age deltas and $-log(p)$ values are presented (Tables 5.6 and 5.7). This, together with the other high correlations achieved for these metrics across the other maps, supports the idea that similar factors contributing to brain ageing are captured by the two methods, and that these factors potentially lie primarily in the larger geometrical features of the maps.

However, BA-SWIN's architecture might give it an advantage in recognising and utilising information from smaller geometrical features. Unlike CNNs, which have a limited local receptive field and tend to reduce the importance of local information as they become deeper [332, 334, 340, 352], BA-SWIN's attention mechanism and window-based processing can dynamically allocate importance across the image, enabling a more flexible and comprehensive analysis of both local and distributed spatial features. This can potentially lead to a richer and more nuanced understanding of the image, including the capturing and emphasising of essential microstructural information crucial for accurate brain age predictions.

An example where this might be the case comes from Table 5.8, where it can be observed that strong negative associations exist between the brain age deltas predicted by BA-SWIN for the FA map and several nIDPs belonging to the Medical History category, which are not otherwise found in the results obtained with HGL. The FA map, derived from the dMRI modality, contains information on how directional the diffusion of water molecules is in each voxel. FA can be influenced by various factors, such as the density, orientation, coherence, and diameter of axons, the degree of myelination, and the presence of crossing fibres [361]. Therefore, FA can reflect some aspects of the brain's microstructure, such as the integrity, organisation, and complexity of white matter tracts and grey matter regions. Neurological and neurodegenerative conditions have been found to impact these microstructural properties of the brain.

When considering the significant associations obtained for the FA map with BA-SWIN, several stand out. For instance, significant associations were observed in several nIDPs related to neurological conditions, such as Alzheimer's disease, dementia, and epilepsy. When considering the links between these conditions and

the brain's microstructure, past studies have found that subjects with Alzheimer's disease exhibit changes such as the presence of amyloid plaques and neurofibrillary tangles, granular vascular degradation, synaptic loss [362], calcium homeostasis, oxidative stresses, inflammation, and neuronal loss through autophagy [363], as well as the presence of small lesions in the anterior cingulate cortex, the prefrontal cortex, and the frontal-limbic circuit [364]. Microstructural changes, including white matter degradation, reduced axonal density, and dysmyelination, were also found in patients suffering from epilepsy, particularly in the corpus callosum and temporal lobe, as well as several other brain regions [365, 366]. These changes could also explain the strong associations found between FA and sodium valproate, an anticonvulsant drug used in the treatment of epilepsy.

Additional significant associations were found between certain hormonal imbalances and their associated conditions, such as adrenocortical insufficiency and hypopituitarism, and BA-SWIN results for FA. These also have correspondences in medical literature, with traumatic brain injury being linked to adrenocorticotropic hormone (ACTH) deficiency [367] and believed to potentially cause hypopituitarism [368, 369].

Particularly interesting is the significant negative association observed in Table 5.8 between the T2 FLAIR Nonlinear map and the Lifestyle nIDP named "Mother's age at death". Several past studies support this finding, having found links between higher maternal age at death and larger grey matter volumes in regions involved in memory, emotion, and executive functions, as well as better performance on tests of verbal memory, processing speed, and executive function [370, 371]. The authors of these papers suggest that a higher maternal age at death might reflect potential genetic advantages which promote resilience to brain ageing in children, as well as to neurological conditions.

Despite the findings presented in Tables 5.8-5.9, not all associations can be intuitively understood. For instance, discerning the underlying factors for the heightened sensitivity of CNN models to Physical Measurements proves challenging. While this situation highlights the potential utility of exploring explainable deep learning models or techniques in future work, the current study does not delve deeply

into this aspect. Nevertheless, even a preliminary exploration into the distinctive feature selection between different architectures could offer insights. Understanding these differences in feature selection could provide clues as to why certain models exhibit heightened sensitivity to specific biomedical parameters over others.

These findings contribute to two main ideas. Firstly, as mentioned in the discussion carried out in Chapter 3 Section 3.6, at no point should causal links be inferred based on the observed correlations between predicted brain age deltas and nIDPs, each of these requiring more in-depth, and sometimes longitudinal, investigations for determining causality. Secondly, these findings support the idea that, although generally the same factors influencing ageing can be found by different methods, careful consideration of the employed method, and how it relates to the map under investigation, is still warranted. This is because, as can be seen from these results, using the wrong method could lead to a loss of sensitivity to factors of interest.

5.6 Transfomers vs. CNNs: Dealing with Perturbed Data

In the preceding section, it could be observed that while BA-SWIN networks generally lagged behind HGL CNN models across all maps, their performance appeared especially compromised in the face of maps containing a higher degree of noise, such as those arising from functional MRI modalities. This trend raised pertinent questions regarding the susceptibility of these models to noise and their capacity to handle various data perturbations. In light of these observations, this section delves deeper into comparing the resilience of the two architectures against various perturbation techniques. Initially, the performances of the models are examined as progressively more noise is added to a series of input volumes for several maps. Subsequently, potential artefacts that could emerge in clinical contexts were simulated, evaluating the models against random anisotropy, random intensity bias field, and random rotations.

5.6.1 Methods

5.6.1.1 General Methdology

For every perturbation variant discussed in this chapter, two distinct experimental sets were carried out: one involved the use of networks pre-trained on clean, research quality data from UK Biobank, evaluated without any additional fine-tuning on perturbed data, whilst the other employed a freshly initialised network trained exclusively on perturbed data.

For both methodologies, every perturbation operation was configured to be governed by a singular hyperparameter. This hyperparameter was modulated, spanning from its minimum, null value, indicative of a scenario where no perturbation is applied, to an extreme value where the resulting image distortion made it challenging for a human observer to discern individual features. This expansive range was chosen intentionally to push the boundaries of the architectures, ensuring a rigorous examination of their robustness.

Also, for both experimental methodologies, neither BA-SWIN nor HGL underwent network ensembling. The rationale behind this was rooted in the presumption that if any pronounced disparities were present between the results obtained with the two architectures, they would overshadow the marginal denoising advantages conferred by ensembling.

When network retraining was carried out, measures were instituted to control for the stochastic differences emerging between identical yet randomly initialised networks. Specifically, the *seed()* method, responsible for initiating the random number generator utilised in the random weight initialisation, was fixed. In addition, during training, the seeds dictating both training data shuffling and perturbation processes were also fixed. Later, at the inference stage, it should be noted that for both sets of experiments, the seeds were not fixed.

For consistency, all perturbations were applied using the TorchIO library [372]. The following subsections give a top-level presentation of the various applied perturbations, but the interested reader is invited to consult the original paper and online documentation.

5.6.1.2 Random Noise Addition

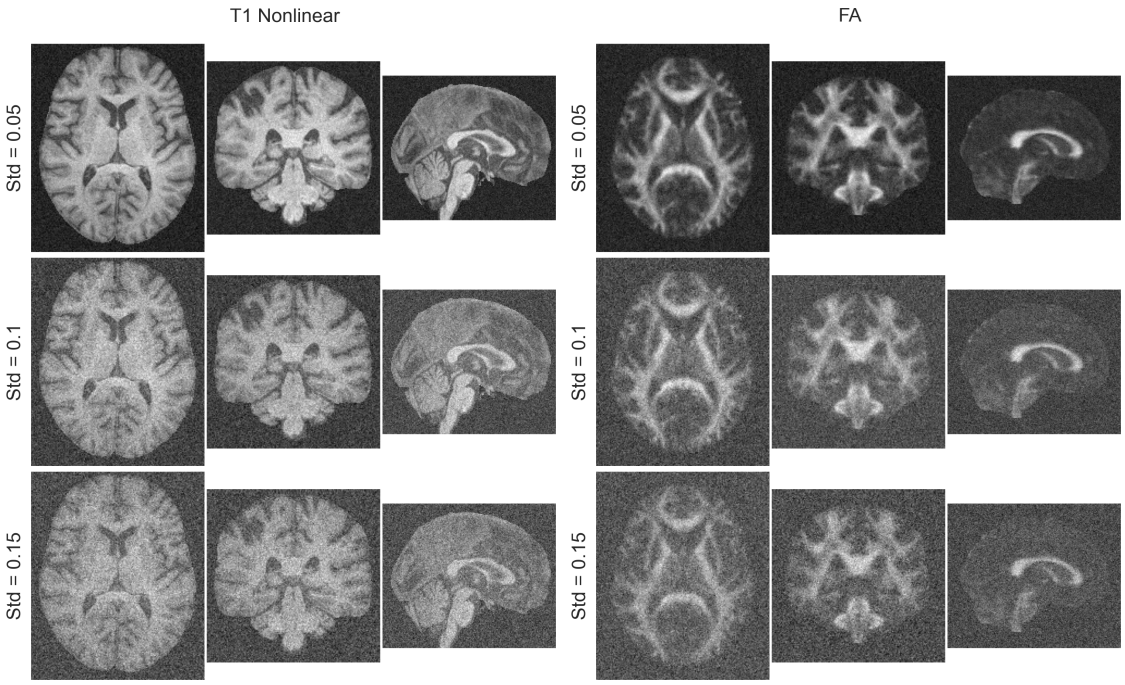
This perturbation added random Gaussian noise to the inputs. The noise was sampled from a normal distribution, with a fixed mean of 0 and a variable standard deviation, in the range $[0, 0.2]$. This range was selected based on visual observations of the impact this perturbation has on the T1 Nonlinear map (Figure 5.7a).

For the experiments using pre-trained networks, this perturbation was applied to several maps, including T1 Nonlinear, SWI, FA, rsfMRI-0, tfMRI-1, Summed Tracts, and T2 Lesions. These were selected to represent each of the core modalities. Examples of the impact of noise addition on some of these maps can be seen in Figure 5.7, where it can be observed that for certain inputs, such as the Summed Tracts or rsfMRI-0, even small amounts of noise can lead to indiscernible features.

While the noise addition perturbation was applied to several maps, all the following perturbations were tested only on T1 Nonlinear inputs. This is partially due to the fact that a similar behaviour was observed across the maps, as seen in Figure 5.11, but also due to the fact that all maps represent summary measures derived from core modalities. Thus, the impact of many artefacts would be corrected or at least accounted for while the maps are synthesised. For instance, dMRI UK Biobank data undergoes steps for motion-induced distortion and bias field correction [36, 173] aiming to reduce imaging artefacts prior to the image being further processed to produce different maps. This means that such artefacts would not be expected in the FA and Summed Tracts maps.

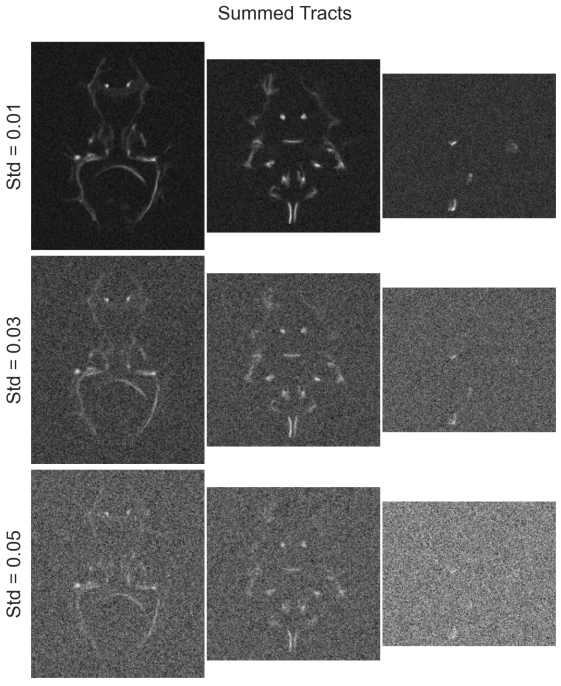
5.6.1.3 Random Anisotropy

The purpose of the Random Anisotropy perturbation is to generate images akin to those encountered in clinical settings using anisotropic voxel sizes. The perturbation downsamples an MRI volume along a random set of axes using nearest neighbour interpolation, after which it upsamples the image back into the initial space. Generally, there is an established axis about which the downsampling is carried out, as in a clinical setting the anisotropic spacing depends on the interest of the clinician. However, for the purposes of this work, the downsampling axis is

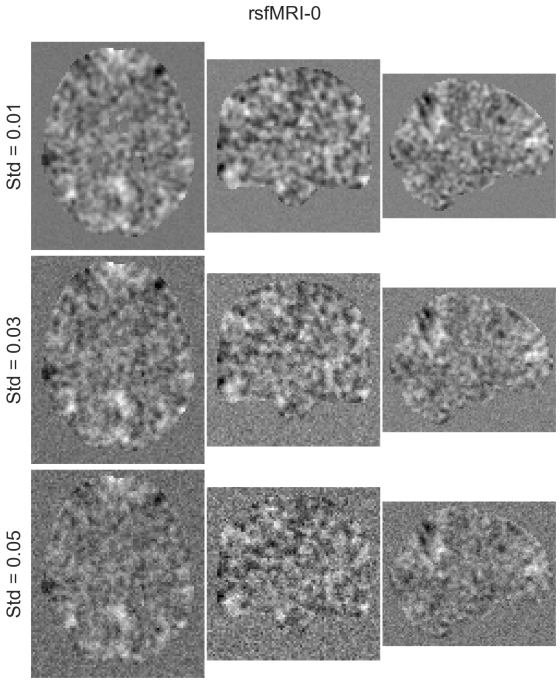


(a) Random noise addition to T1 Nonlinear

(b) Random noise addition to FA



(c) Random noise addition to Summed Tracts



(d) Random noise addition to rsfMRI-0

Figure 5.7: Example random noise addition for several maps, including **(a)** T1 Nonlinear, **(b)** FA, **(c)** Summed Tracts, and **(d)** rsfMRI-0.

a random parameter. Although several interpolation methods are available for upsampling, linear interpolation was selected for this work due to its speed. The hyperparameter controlling this perturbation was chosen to be the downsampling factor, which was varied in the range $[1, 10]$ (Figure 5.8).

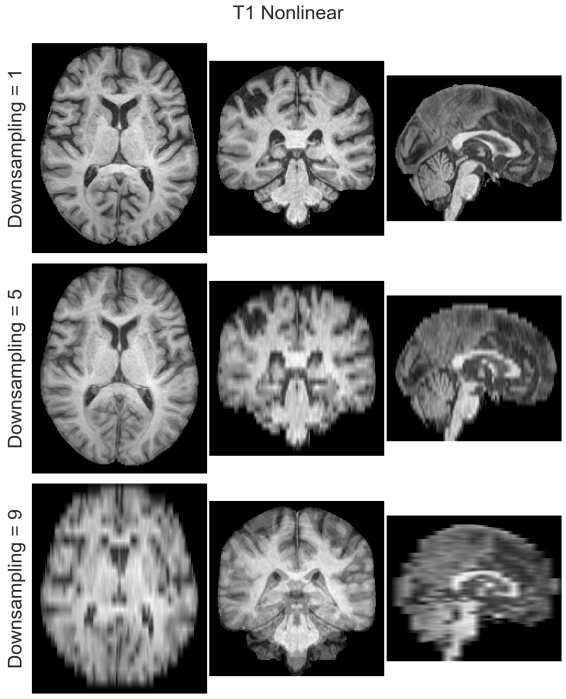


Figure 5.8: Random anisotropy applied to T1 Nonlinear for several downsampling values.

5.6.1.4 Random Bias Field

Clinical quality MRI scans sometimes exhibit intensity bias field artefacts, which can be randomly added to inputs using this perturbation technique. The perturbation works by modelling the intensity distortion created by MRI magnetic field inhomogeneities using a linear combination of polynomial basis functions [373, 374]. The hyperparameter chosen to control this perturbation is the coefficients factor, which represents the maximum magnitude of the polynomial coefficients. For this work, this was selected to be in the range $[0, 2]$, while the order of the basis polynomial function was kept at a default value of 3 (Figure 5.9).

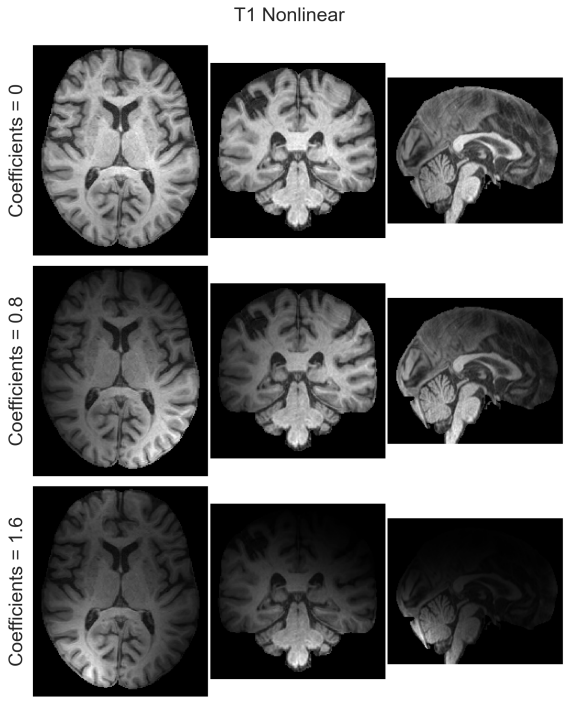


Figure 5.9: Random bias field applied to T1 Nonlinear for several values of the maximum magnitude of the polynomial coefficients.

5.6.1.5 Random Affine Rotation

The final tested perturbation is designed to simulate a random rotation of the subject in the scanner. It achieved this by applying a random affine transformation to the image, followed by a resampling. The controlling parameter for these experiments was chosen to be the degrees parameter, which defines the rotation of the image in degrees. The value was set in the range $[0, 10]$ degrees, with the rotation angle about each of the three axes being randomly sampled from a uniform distribution $U\left(-degree, +degree\right)$ (Figure 5.10).

5.6.2 Results

Figures 5.11-5.15 present the results for all four perturbation techniques employed in this section. Firstly, the noise addition perturbation was added to 7 different maps. Figure 5.11 reveals a similar trend for all maps, with the results for the HGL CNN degrading much faster than those obtained with the BA-SWIN Transformer.

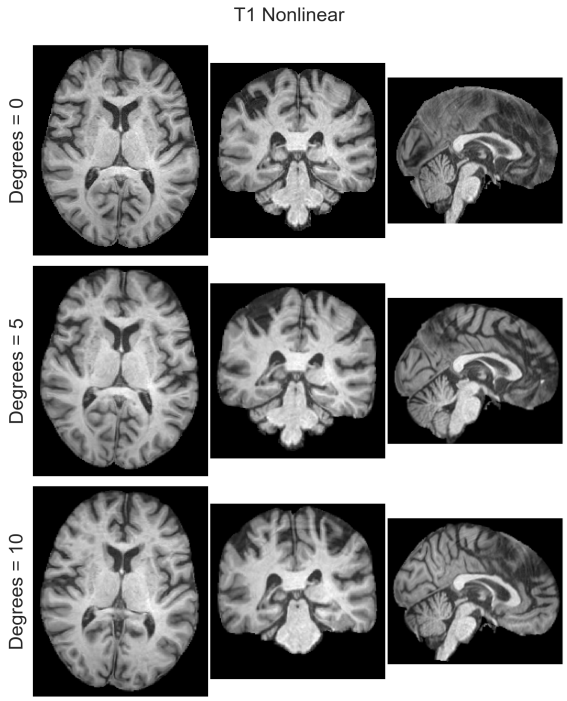


Figure 5.10: Random affine rotations applied to T1 Nonlinear for several rotation angle values in degrees.

Following these observations, both experimental setups were tested, the results being presented for Random Noise Addition in Figure 5.12, for Random Anisotropy in Figure 5.13, for Random Bias Field in Figure 5.14 and for Random Affine Rotation in Figure 5.15.

For the Random Noise Addition, it can be observed that when new networks are trained exclusively on noisy data, the previously observed trend reverses, with the CNN network now outperforming the Transformer.

In the case of the perturbations aimed at reproducing clinical quality images, when utilising pre-trained networks, the trend observed for the Random Noise Addition is generally maintained, with the BA-SWIN networks performing better than the equivalent HGLs as the perturbations become progressively more aggressive. The one case where this is different is for the Random Affine Rotation, where for small angles the networks show similar performances, after which the metrics for the BA-SWIN network rapidly degrade. When considering the experimental approach where newly initialised networks are trained using perturbed data, for all clinical

quality perturbations the performances of BA-SWIN and HGL are comparable, generally degrading at a similar pace with more aggressive perturbation.

5.6.3 Discussion

During this comparison of the performance of the BA-SWIN Transformer and the HGL CNN networks when faced with different types of data perturbation techniques, several important observations were made.

In the instances where networks pre-trained using research high-quality data were tested with progressively perturbed test datasets, it was observed that the BA-SWIN networks appeared to have an advantage over their HGL counterparts, demonstrating resilience to all perturbations bar one: the Random Affine Rotation. However, when subjected to training from scratch using random perturbations, the HGL networks performed better than BA-SWIN for Random Noise Addition, while both networks exhibited analogous performances in the case of all the other perturbations, especially within disturbance ranges expected in practical scenarios.

To understand what causes this behaviour, the mechanics underpinning the two models must be considered. Starting off with the pre-trained model experiments, it was observed that the HGL CNN networks encountered challenges across most perturbations. CNNs usually make use of high-frequency features when learning a given task [376]. In the presence of perturbations, the patterns and distributions previously learned by the networks are disrupted. This causes a shift in distribution between the training and testing datasets, undermining the discriminative strength of CNN filters [377]. In map, BA-SWIN Transformers operate on the entire image, using global information rather than just smaller, local pieces of information like CNNs. This means they do not rely as heavily on the high-frequency features of an image. Thus, when there are disruptions in these high-frequency features, Transformer performance is not severely hampered [378].

In this context, the self-attention mechanism in the BA-SWIN networks can potentially offer robustness against certain perturbations. By dynamically weighing features based on their relevance, the self-attention mechanisms in these networks

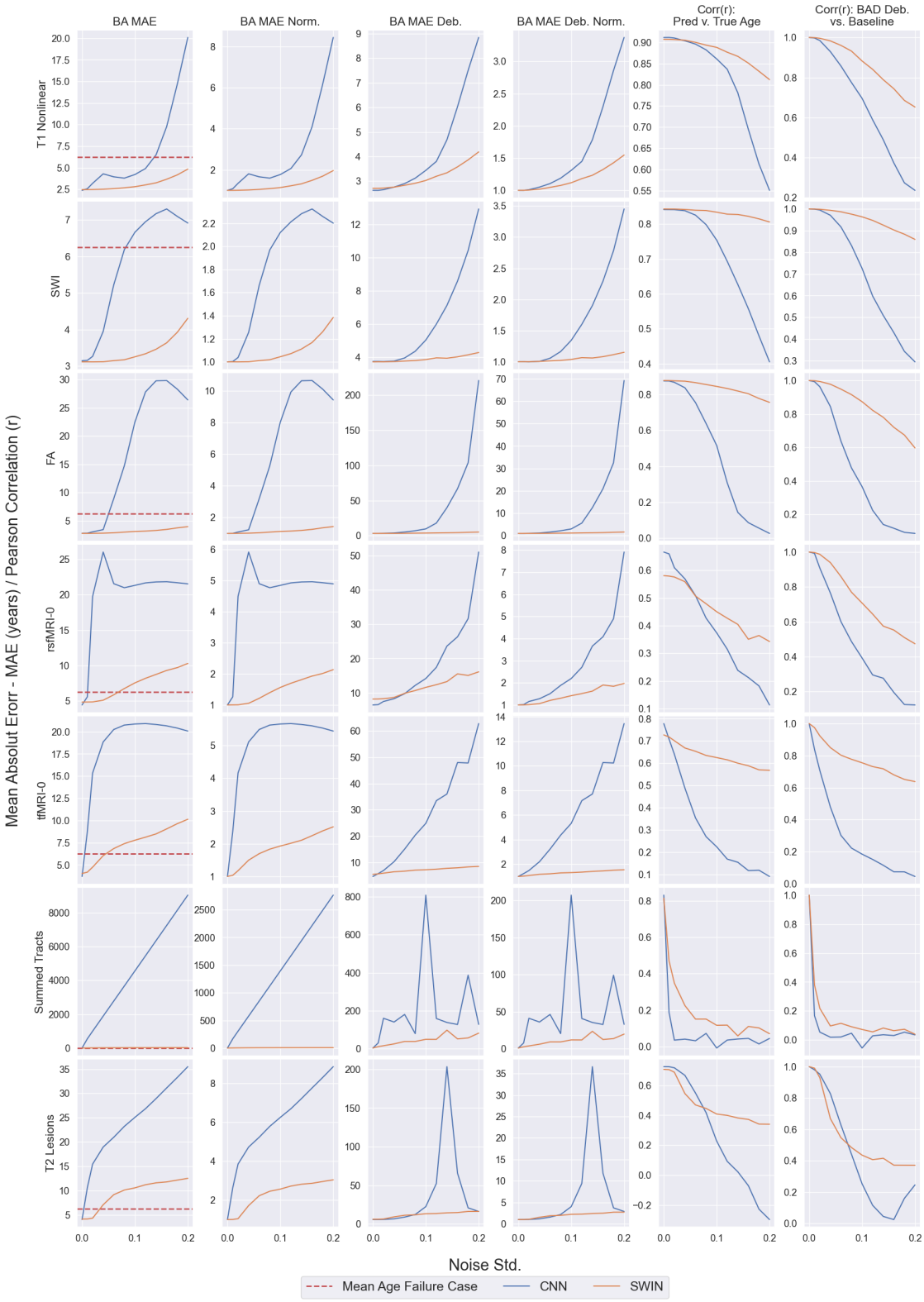


Figure 5.11: Random noise addition results with pre-trained networks for several maps. For brevity, the following abbreviations have been used to identify each column: BA MAE (Brain Age MAE), BA MAE Norm. (Brain Age MAE normalised with respect to the original unperturbed results), BA MAE Deb. (Brain Age MAE linearly debiased), BA MAE Deb. Norm (Brain Age MAE debiased and normalised as above), Corr(r): Pred. v. True Age (Correlation between the predicted and true chronological age), Corr(r): BAD Deb vs. Baseline (Correlation between the debiased brain age deltas and those obtained with original unperturbed data).

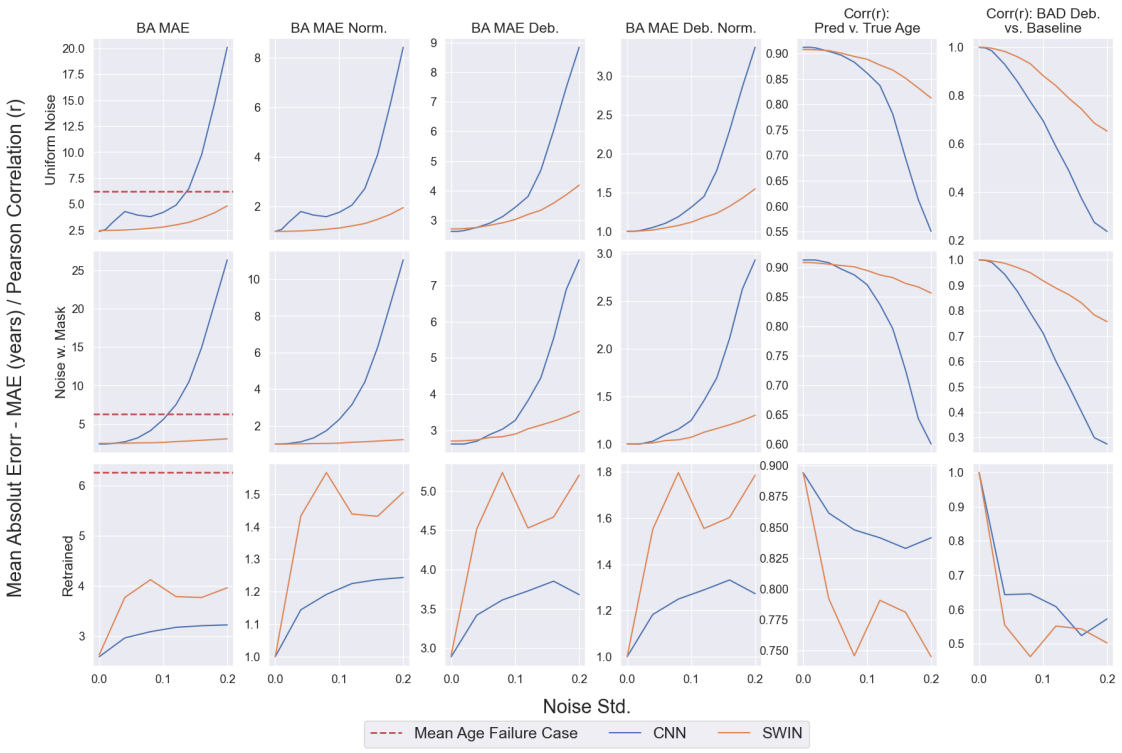


Figure 5.12: Random noise addition results for pre-trained and re-trained networks for T1 Nonlinear. The first row indicates the results obtained with the pre-trained networks. The second row also presents results obtained with pre-trained networks, but masked with the MNI152 T1 1mm brain mask diluted [375] so that noise is only added to the brain volume, not the background. The bottom row presents the data for the freshly initialised networks trained exclusively on perturbed data. For brevity, the following abbreviations have been used to identify each column: BA MAE (Brain Age MAE), BA MAE Norm. (Brain Age MAE normalised with respect to the original unperturbed results), BA MAE Deb. (Brain Age MAE linearly debiased), BA MAE Deb. Norm (Brain Age MAE debiased and normalised as above), Corr(r): Pred. v. True Age (Correlation between the predicted and true chronological age), Corr(r): BAD Deb vs. Baseline (Correlation between the debiased brain age deltas and those obtained with original unperturbed data).

might prioritise the more informative regions of an image over those affected by distortions such as noise, allowing the networks to focus on important features [378]. This global context processing might, in some instances, help the model discern true features from artefacts [379]. However, it is crucial to note that while CNNs are designed with a built-in translation invariance (i.e., treating all translations of a feature in the same way), they are not inherently rotationally invariant. Therefore, while CNNs might manage small translations effectively, their performance could

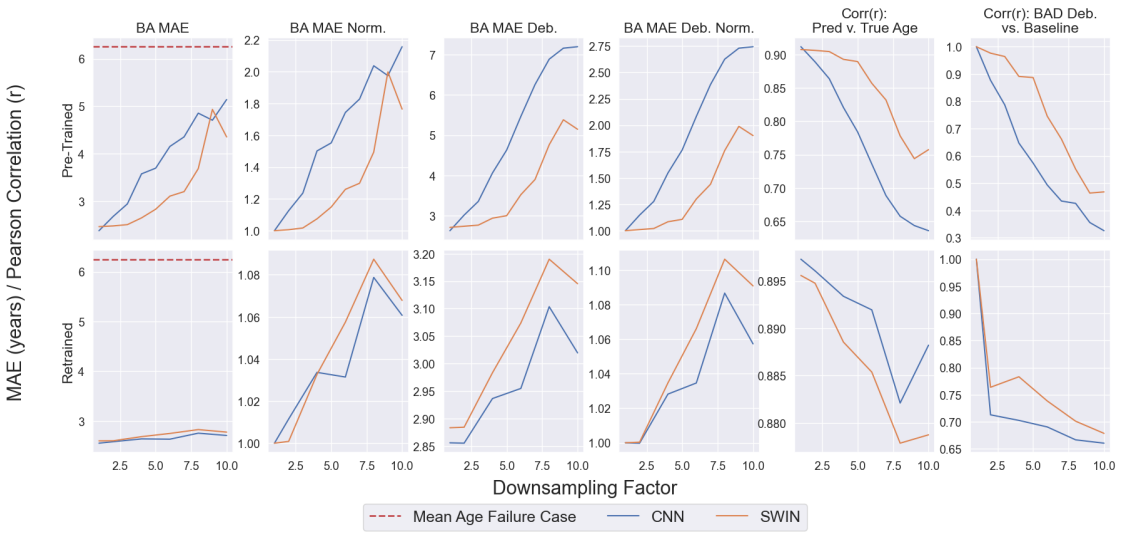


Figure 5.13: Random anisotropy perturbation results for pre-trained and re-trained networks for T1 Nonlinear. The top row indicates the results obtained with the pre-trained networks. The bottom row presents the data for the freshly initialised networks trained exclusively on perturbed data. For brevity, the following abbreviations have been used to identify each column: BA MAE (Brain Age MAE), BA MAE Norm. (Brain Age MAE normalised with respect to the original unperturbed results), BA MAE Deb. (Brain Age MAE linearly debiased), BA MAE Deb. Norm (Brain Age MAE debiased and normalised as above), Corr(r): Pred. v. True Age (Correlation between the predicted and true chronological age), Corr(r): BAD Deb vs. Baseline (Correlation between the debiased brain age deltas and those obtained with original unperturbed data).

vary with more pronounced perturbations. The key distinction lies in the way these architectures perceive and process spatial information, with Transformers capturing a more global context and CNNs focusing on local receptive fields.

Considering the case of Random Affine Rotations, both CNNs and BA-SWIN Transformers face challenges, albeit differently. CNNs, although not inherently rotationally invariant, have localised filters that may partially adapt to slight rotations within their receptive field. The issue arises when the rotation goes beyond the receptive field, causing misalignment and recognition issues [380]. On the other hand, BA-SWIN Transformers rely on global context and positional encodings. While this approach offers a more comprehensive understanding of the image, it assumes a consistent orientation across images, which is the case when images are always affine registered, such as in UK Biobank. During Random Affine Rotations, the global positional information gets misaligned, causing a significant

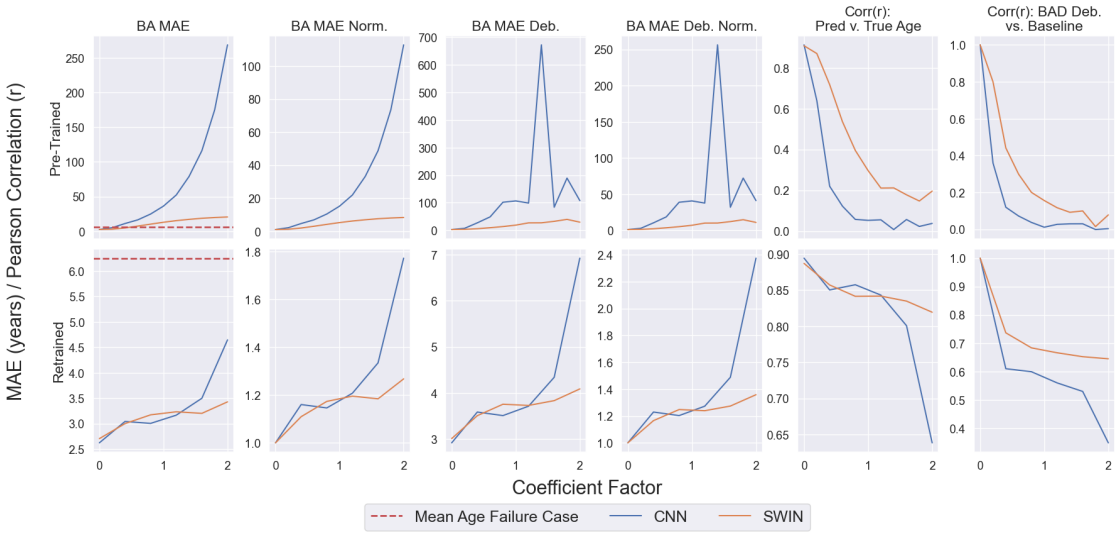


Figure 5.14: Random bias field perturbation results for pre-trained and re-trained networks for T1 Nonlinear. The top row indicates the results obtained with the pre-trained networks. The bottom row presents the data for the freshly initialised networks trained exclusively on perturbed data. For brevity, the following abbreviations have been used to identify each column: BA MAE (Brain Age MAE), BA MAE Norm. (Brain Age MAE normalised with respect to the original unperturbed results), BA MAE Deb. (Brain Age MAE linearly debiased), BA MAE Deb. Norm (Brain Age MAE debiased and normalised as above), Corr(r): Pred. v. True Age (Correlation between the predicted and true chronological age), Corr(r): BAD Deb vs. Baseline (Correlation between the debiased brain age deltas and those obtained with original unperturbed data).

degradation in the performance of BA-SWIN Transformers. The rotational data disrupts the global context, which BA-SWIN heavily relies upon, leading to a more pronounced impact on its performance compared to CNNs.

When models are retrained, however, these observations change. Now, the HGL CNNs, with their local receptive fields and shared weights, demonstrate enhanced robustness to noise. Weight sharing in CNNs contributes to translational invariance, allowing the model to generalise learned features across different parts of the image, providing a buffer against noise and other data variations, and aiding in stable training. Conversely, the absence of weight sharing in BA-SWIN, paired with its global context awareness, can pose challenges during training in environments exhibiting data variations [381]. An additional factor to consider is the number of trainable parameters for each model. With $5.0M$ trainable parameters compared to HGL's $1.2M$ trainable parameters, it is possible that BA-SWIN is more prone to

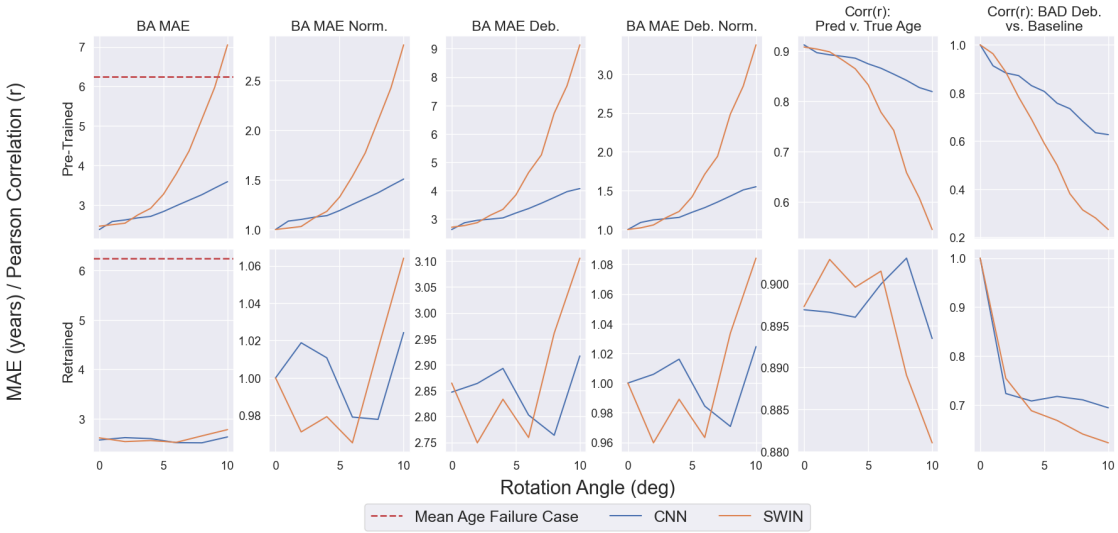


Figure 5.15: Random affine rotation perturbation results for pre-trained and re-trained networks for T1 Nonlinear. The top row indicates the results obtained with the pre-trained networks. The bottom row presents the data for the freshly initialised networks trained exclusively on perturbed data. For brevity, the following abbreviations have been used to identify each column: BA MAE (Brain Age MAE), BA MAE Norm. (Brain Age MAE normalised with respect to the original unperturbed results), BA MAE Deb. (Brain Age MAE linearly debiased), BA MAE Deb. Norm (Brain Age MAE debiased and normalised as above), Corr(r): Pred. v. True Age (Correlation between the predicted and true chronological age), Corr(r): BAD Deb vs. Baseline (Correlation between the debiased brain age deltas and those obtained with original unperturbed data).

overfitting to noise and other perturbations in the training data, due to its increased capacity to memorise complex patterns in the training data. This could lead the larger Transformer model to not generalise well at inference [379]. Furthermore, while transformers, initially crafted for natural language processing (NLP) tasks, efficiently handle disturbances like grammatical inconsistencies or semantic ambiguities using self-attention, the very nature of noise and other disturbances in visual tasks is distinct. This can mean that disturbances at the pixel or voxel level can be disruptive to the attention mechanism, as they affect the overall structure the model is trying to recognise. The majority of these considerations appear to apply to the Random Noise Addition perturbation. However, when considering the other perturbations it can be observed that the performance of both the HGL CNNs and BA-SWIN Transformers degrades at a similar pace, suggesting that in these cases both networks are equally good at extracting as much useful information as

possible from the lower quality data. This is also the case for the Random Affine Rotations perturbation, where now the BA-SWIN transformers are incentivised to learn rotationally invariant relationships through the self-attention mechanism.

5.7 Transformer Interpretation

Explainability in AI models is steadily gaining significance, particularly given the growth in model complexity. Understanding the inner workings of models is paramount, especially in critical fields such as medical imaging analysis, where model decisions can have direct clinical implications. This is even more crucial in brain age prediction studies, particularly when non-conventional maps are being utilised, beyond the standard sMRI derived ones. As the sophistication of models increases, it is important to not only produce stable and reliable results but also shed light on the features utilised for a prediction. Otherwise, opaque decision-making processes can lead to mistrust in and misuse of models, particularly in clinical settings.

The final section of this chapter delves into building intuition regarding which features the BA-SWIN networks utilise when making their predictions. For this purpose, the attention activation maps at various levels were extracted and projected back into 3D space for visualisation. Generally, for ViTs, a single global attention map is calculated using the activations in the entire network, using techniques such as Attention Rollout [348]. However, due to network architecture design choices, such as the path merging operation which builds hierarchical feature maps, and the shifted window approach which builds connections between windows, these techniques cannot be readily utilised with SWIN Transformers. Thus, in this section, feature visualisation was carried out for each individual MSA block individually. To account for the presence of multiple attention heads, the presented attention activation maps were created by averaging each head's attention activation [348] across the query dimension, as described in Appendix C. A visual analysis was carried out for 6 maps: T1 Nonlinear, FA, rsfMRI-0, Summed Tracts, T2 Lesions and TBSS FA. For each of these, attention activation maps were generated and visually inspected at different network depths.

Following the hyperparameter search carried out at the start of this chapter, the BA-SWIN networks utilised a patch size of 5 throughout this work. However, because of this lower resolution, it might be more difficult to interpret the observed attention maps. For this reason, for each of the considered maps, an additional SWIN network was trained, utilising a patch size of 2 and following the standard, vanilla, architecture defined in Section 5.3.1.

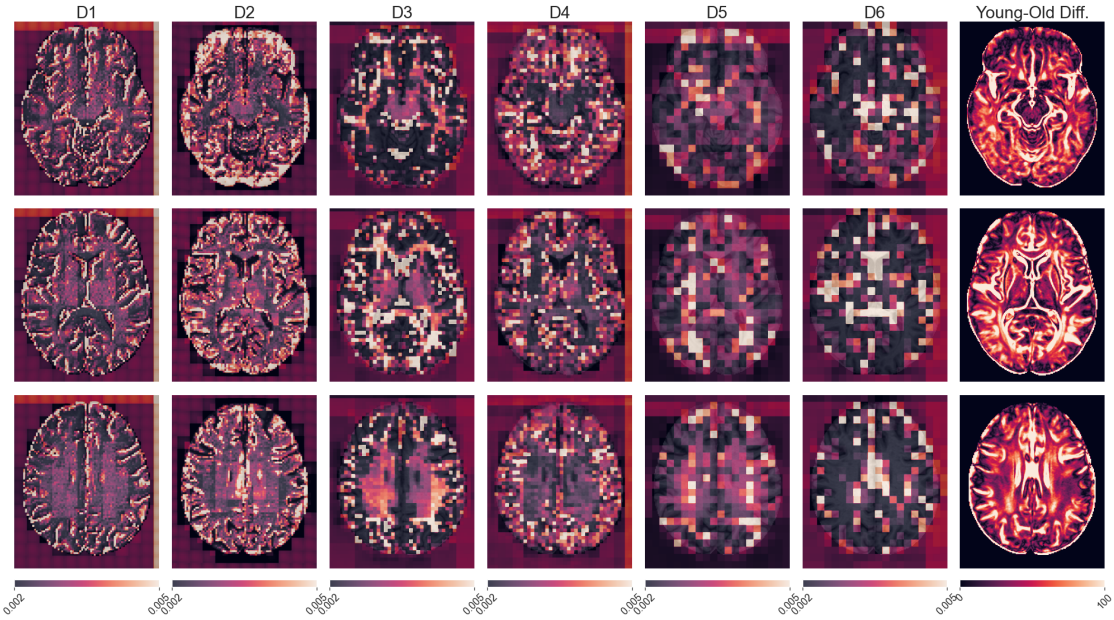
5.7.1 Results

The attention activation maps for the considered maps are presented in Figures 5.16-5.21. Each figure is composed of two subfigures, corresponding to the attention activation maps overlaid on a subject's original map data, obtained using high and low resolution models respectively. The columns in each of the subfigures show the attention activation maps at progressively larger depths in the network. Given the standard model has a depth of $(2, 2, 2, 2)$, while the BA-SWIN model is shallower, employing a $(2, 2, 2)$ depth structure, only the attention activation maps for the first 6 layers are shown. The final column in each of the subfigures is the absolute difference volume between the average maps obtained using the youngest and oldest predicted subject groups. This image acts as a control, showing the areas in each map which display the largest changes as individuals age. This image serves as a benchmark, the expectation being that the attention mechanism would primarily target these age-induced variations, along with any other features that hold predictive power regarding age progression. For each map, the control images were obtained by averaging all scans for those subjects predicted $45 - 52$ (Young) and $75 - 82$ (Old), and taking the voxel-wise difference between the two. The rows in each of the subfigures correspond to three different levels of the considered volume.

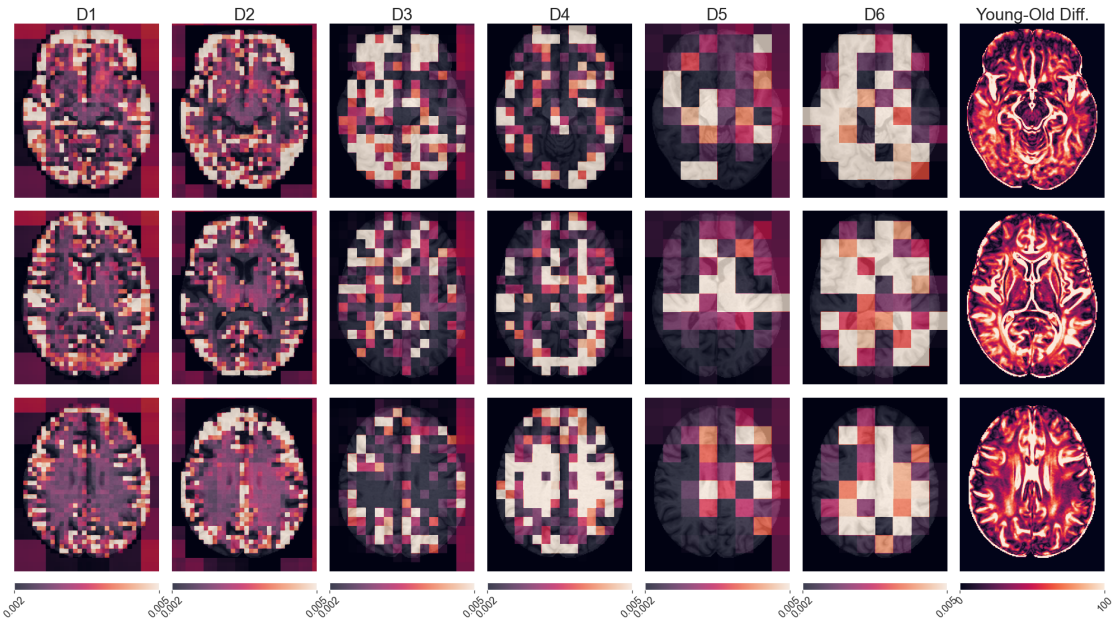
Several observations can be made from the results presented in this section. Firstly, due to the reduction in latent feature dimensionality, it is difficult to extract much useful information from the activation maps extracted from the deeper network layers. This is particularly true for the lower resolution BA-SWIN networks. However, when observing the attention activations which can be interpreted, it

can be seen that the Transformer generally focuses on areas where changes would be expected with ageing. For instance, in Figure 5.16, higher weights seem to be assigned to areas corresponding to the grey matter and ventricles for both networks. In addition, as seen for the other maps, both networks appear to generally focus on the same features, which are also the areas showing the highest values in the control Young-Old difference volumes shown in the rightmost columns. These observations build trust that the BA-SWIN models employed in this chapter are focusing on the brain ageing relevant areas for each map.

The one map which digresses from these general observations is the Summed Tracts derived from dMRI (Figure 5.18). While the lower resolution network appears to be quite adept at capturing the general shape of the subject's tractography data, it can be observed that this is not the case for the higher resolution $patch = 2$ model. In fact, the higher resolution $patch = 2$ model failed to converge, remaining stuck on a training plateau similar to those seen in Figure 5.2, predicting values close to the mean population age for all test subjects. The attention activation maps provide insight into what might have prevented this network from converging. Specifically, when looking at the earliest network layers (D1 and D2 in Figure 5.18a), one can observe that the smaller patches often encompass areas which appear devoid of any tract data, meaning they frequently cover regions that are essentially blank or uninformative. This highlights a potential drawback of the high-resolution SWIN transformer: when small patches are applied to a volumes which combine spatial sparsity with low intensity voxel values, like the Summed Tract volume, they often miss out on the vital tract details, impeding the network's ability to model the long-range dependencies effectively. However, as illustrated in Figure 5.18b, this limitation is not present in the network utilising larger patch sizes. This could be caused by the inherently larger receptive fields of larger patches, allowing them to capture sufficient contextual information and long-distance relationships. This is particularly useful in sparse datasets where relationships between distant parts of an image are vital for a correct global interpretation.

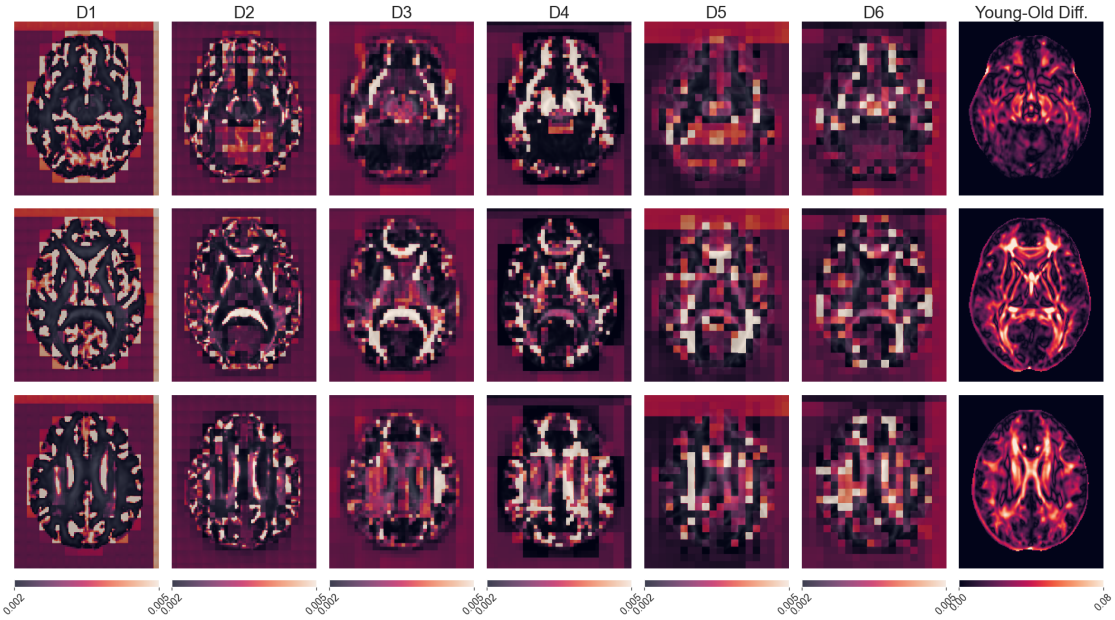


(a) Standard high resolution BA-SWIN

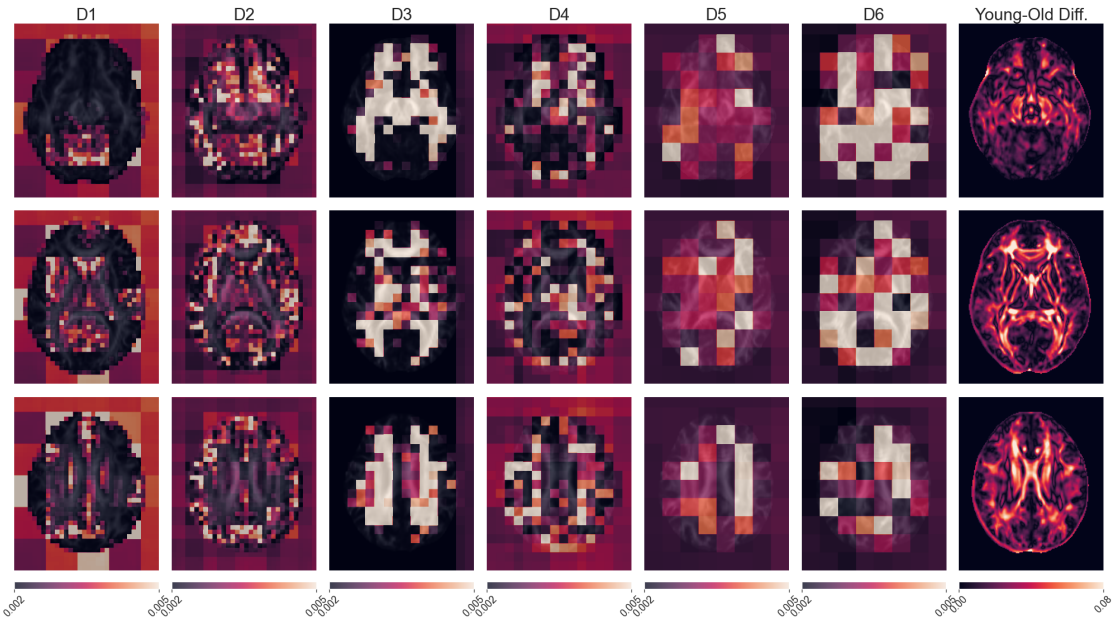


(b) Optimised low resolution BA-SWIN

Figure 5.16: T1 Nonlinear attention activation maps at several depths (D) in the network, for **(a)** a high resolution ($patch = 2$) BA-SWIN network and **(b)** the optimised low resolution ($patch = 5$) BA-SWIN network utilised in this chapter.

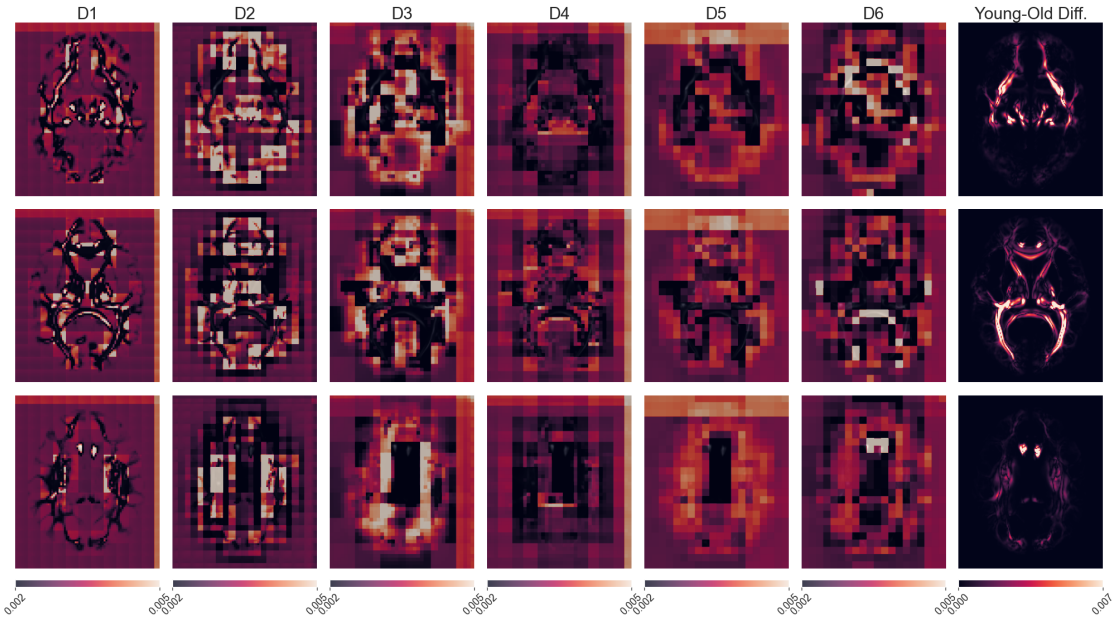


(a) Standard high resolution BA-SWIN

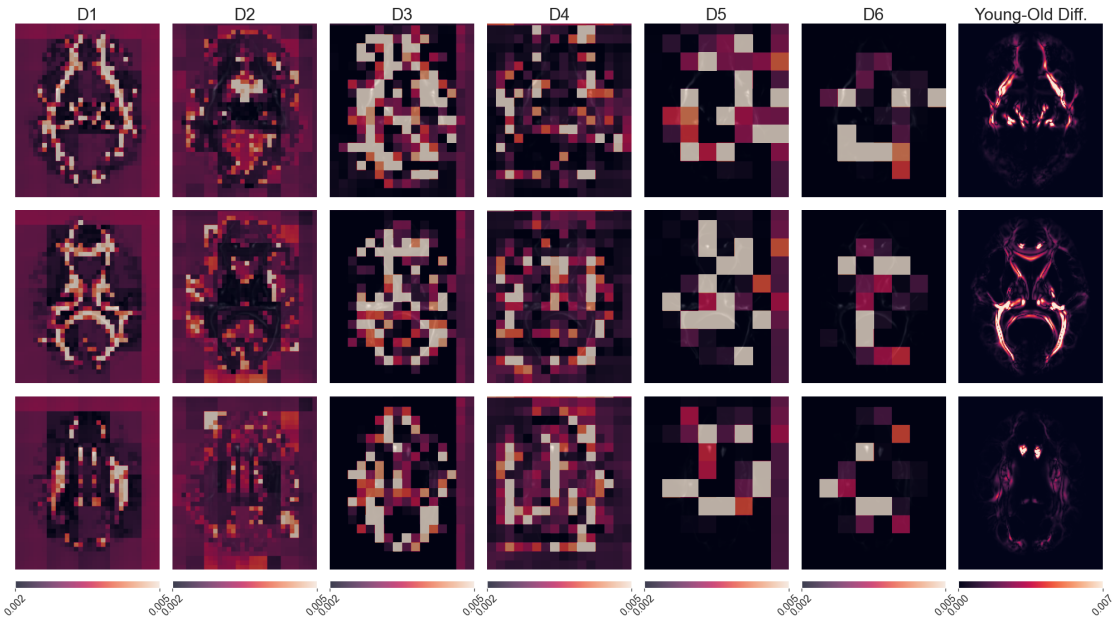


(b) Optimised low resolution BA-SWIN

Figure 5.17: FA attention activation maps at several depths (D) in the network, for **(a)** a high resolution ($patch = 2$) BA-SWIN network and **(b)** the optimised low resolution ($patch = 5$) BA-SWIN network utilised in this chapter.

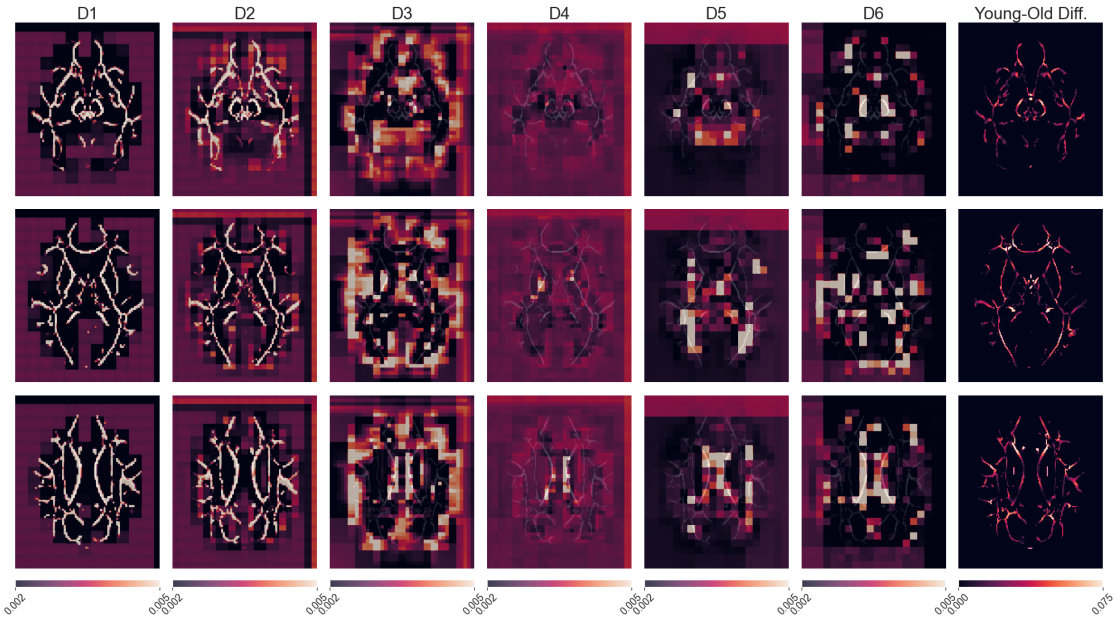


(a) Standard high resolution BA-SWIN

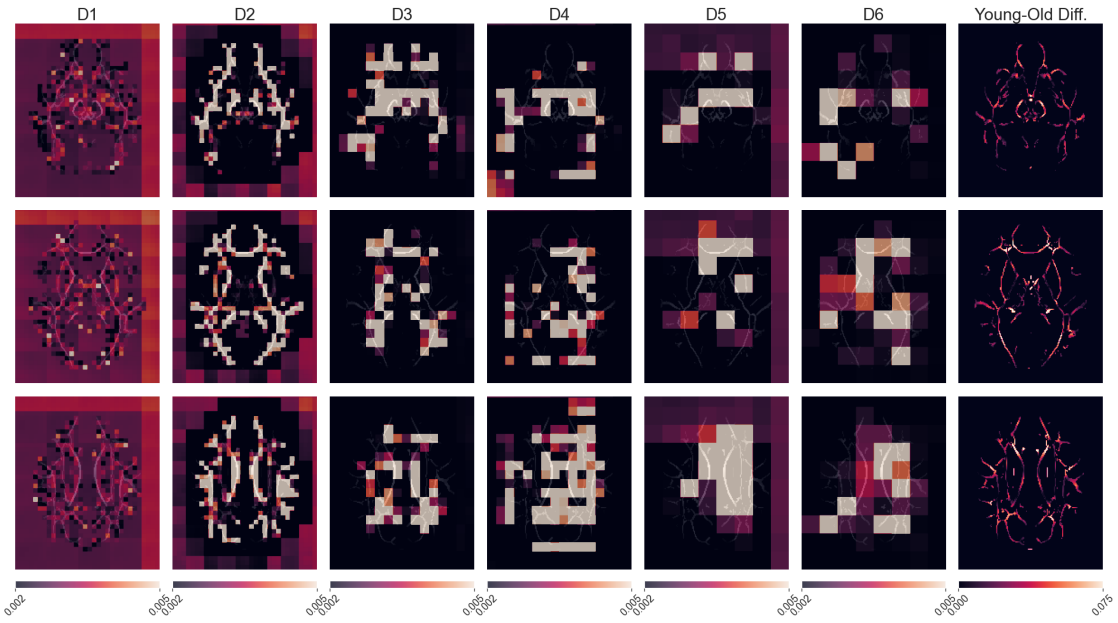


(b) Optimised low resolution BA-SWIN

Figure 5.18: Summed tracts attention activation maps at several depths (D) in the network, for **(a)** a high resolution ($patch = 2$) BA-SWIN network and **(b)** the optimised low resolution ($patch = 5$) BA-SWIN network utilised in this chapter.

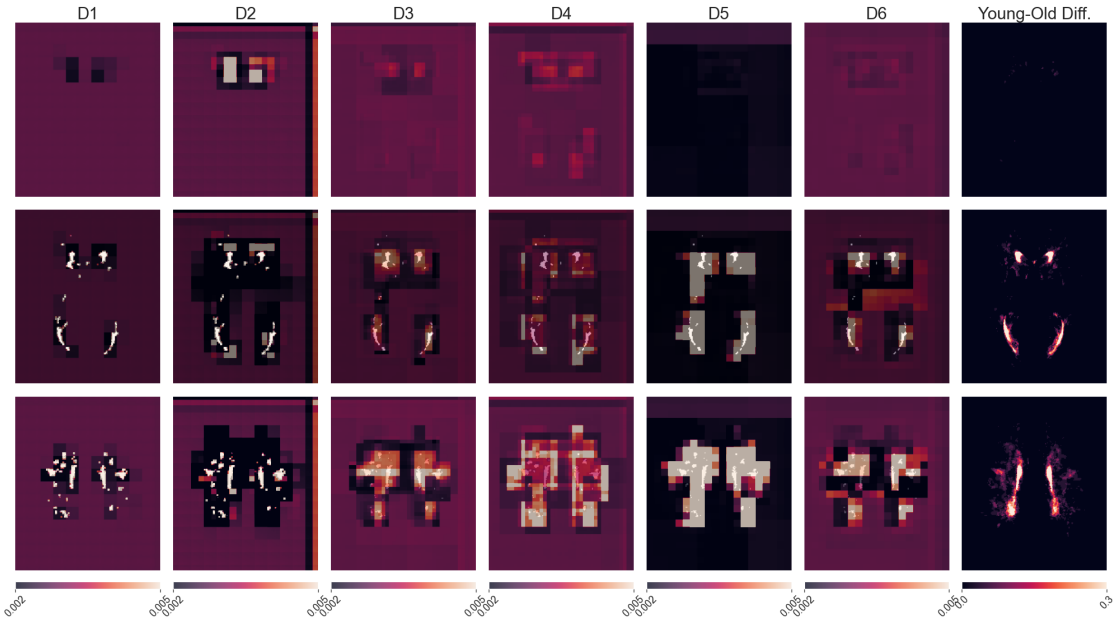


(a) Standard high resolution BA-SWIN

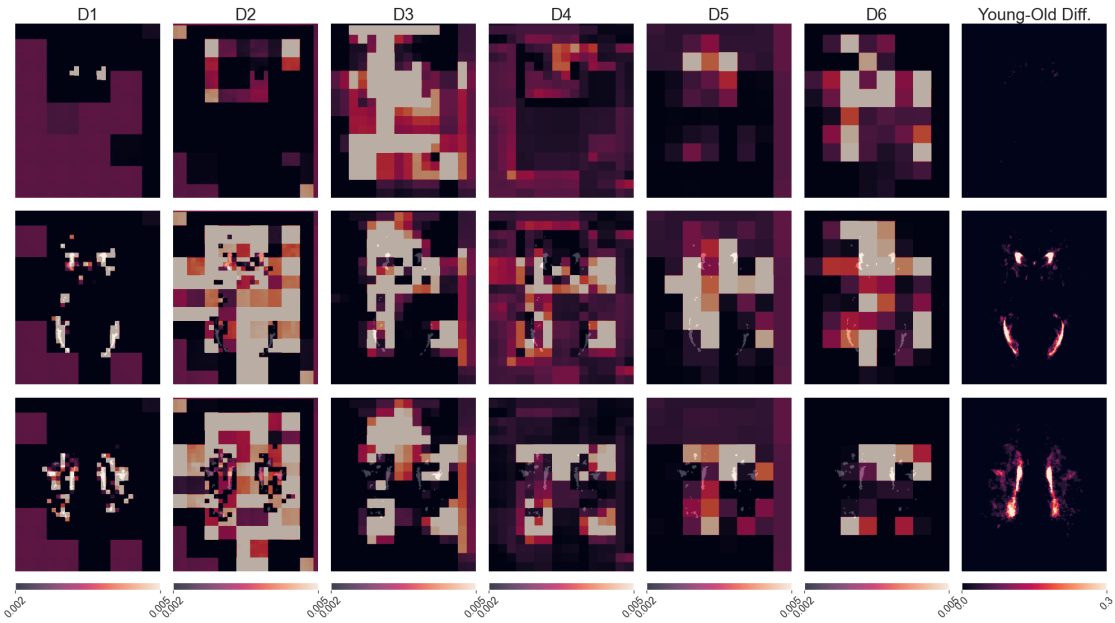


(b) Optimised low resolution BA-SWIN

Figure 5.19: TBSS FA attention activation maps at several depths (D) in the network, for **(a)** a high resolution ($patch = 2$) BA-SWIN network and **(b)** the optimised low resolution ($patch = 5$) BA-SWIN network utilised in this chapter.

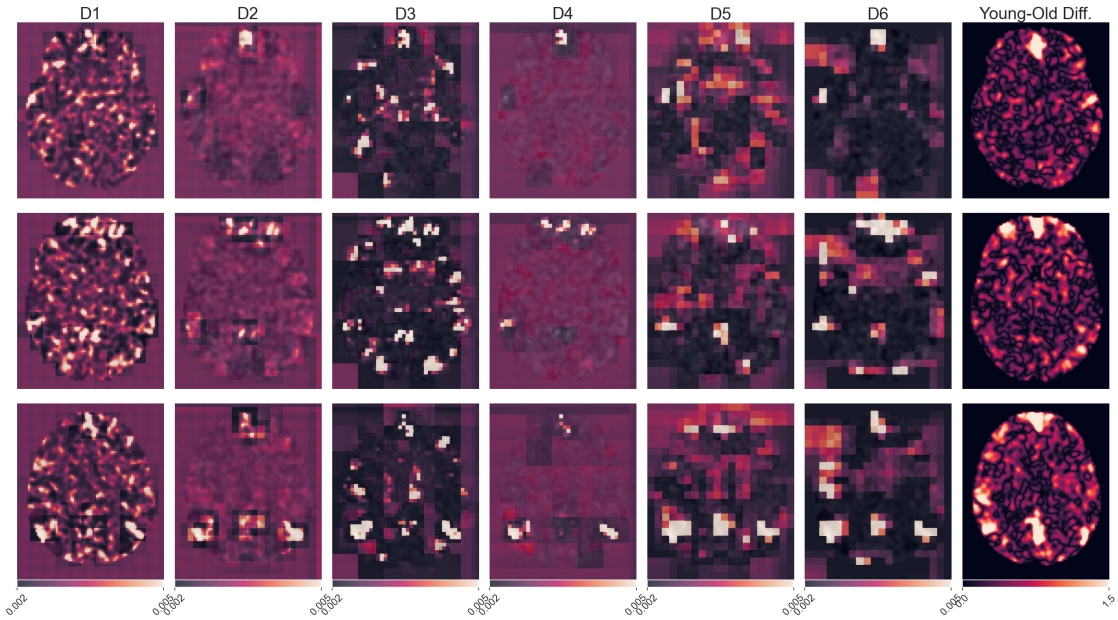


(a) Standard high resolution BA-SWIN

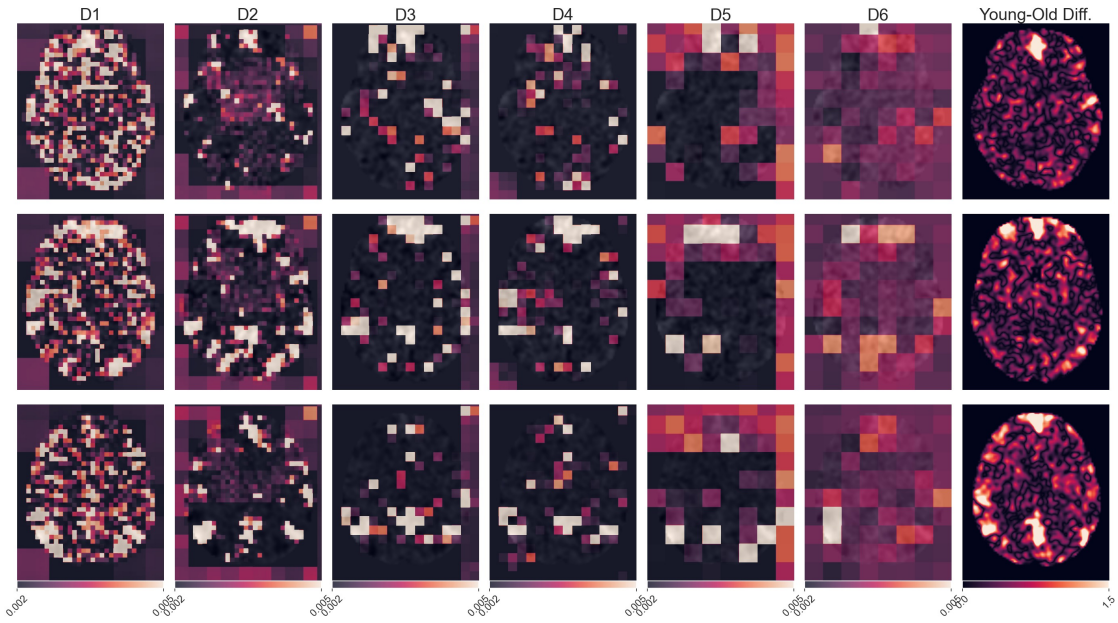


(b) Optimised low resolution BA-SWIN

Figure 5.20: T2 lesions attention activation maps at several depths (D) in the network, for **(a)** a high resolution ($patch = 2$) BA-SWIN network and **(b)** the optimised low resolution ($patch = 5$) BA-SWIN network utilised in this chapter.



(a) Standard high resolution BA-SWIN



(b) Optimised low resolution BA-SWIN

Figure 5.21: rsfMRI-0 attention activation maps at several depths (D) in the network, for **(a)** a high resolution ($patch = 2$) BA-SWIN network and **(b)** the optimised low resolution ($patch = 5$) BA-SWIN network utilised in this chapter.

5.7.2 Discussion

While not as informative as the other sections presented in this chapter, the work done herein nonetheless sheds light on the limitations surrounding explainable AI techniques applied to the BA-SWIN architecture. In particular, architectural choices such as the use of patch windows and window merging make the use of established explainable AI techniques, such as Attention Rollout, difficult. This means that a global attention map cannot be readily obtained, leaving observers only with individual attention activation maps. While providing insight into the focus of the network at each depth, these are still insufficient, as there is no straightforward way of determining how the information and network focus are transformed between Transformer blocks. By not being able to assess the attention flow from the start to the end of the network, this approach does not account for any learned interactions that might occur inside the network.

Thus, it appears that utilising the SWIN Transformer architecture introduces a compromise that users need to be aware of, as highlighted in this section. While enabling regression operations on much higher dimensional data than traditional ViTs, SWIN introduces the aforementioned limitations on the inherent explainable mechanisms of Transformers. At the time of writing (summer 2023), literature does not seem to provide a solution to this compromise. Only one paper claiming to have adapted the attention rollout mechanism to SWIN Transformers has been found, yet the authors neither explain how they have circumvented the inherent issues associated with SWIN, nor do they provide code which would enable their solution to be reverse engineered [351]. This does not mean, however, that solutions to this problem will not emerge in the future, either by modifying the SWIN architecture or the attention rollout mechanism. However, due to time limitations, attempting either of these was outside the scope of this work.

Some potential solutions to circumvent the described issues come from explainable AI literature. These could include already established explainable AI approaches, such as attention gates, saliency maps or gradient-based methods such as CAM [21, 337]. However, as seen in Chapter 1 Section 1.3.4, these can be

considered insufficient to provide good visualisation of features contributing to brain age prediction, as seen in the work carried out by Dinsdale et al [21], meaning that, ideally, methods utilising the inherent Transformer activation maps are still desirable.

5.8 Conclusion

Starting off from the limitations of CNNs, this chapter has proposed and then provided a detailed examination of a state-of-the-art Vision Transformer model, named BA-SWIN. This novel model was compared against HGL, the CNN-based network which constituted the backbone of the work discussed in this thesis. Through a series of rigorous analyses and comparisons, several key themes have emerged, each of which contributes to the current understanding of these models and highlights areas for future research.

After confirming that Transformers are capable of predicting brain age and optimising the architecture of the BA-SWIN Transformer, it was compared against HGL for several MRI maps derived from all 5 core modalities present in UK Biobank. This revealed different performances between BA-SWIN Transformers and HGL CNNs across various maps. While BA-SWIN networks generally lagged slightly behind HGL CNN models, and both methods demonstrated similar associations with nIDPs, the results obtained with BA-SWIN networks indicated that, due to their use of attention mechanisms and window-based processing, they are capable of dynamically allocating importance across an image, potentially allowing them to capture and accentuate the importance of information that would otherwise be potentially missed or discounted by HGL, such as microstructural information.

The BA-SWIN networks also showed remarkable resilience against certain types of noise and data perturbations. This resilience could be rooted in the global context operations and self-attention mechanism inherent to Transformers. When examining the behaviour of models trained from scratch using random perturbations, HGL CNNs demonstrated better performance in certain cases, such as Random Noise Addition. The BA-SWIN models, with their higher number of trainable parameters, and thus an increased capacity to memorise complex patterns in the data, were found

to not generalise as well at inference. This could be because larger models are more prone to overfitting to noise and other perturbations encountered during training, thus highlighting a trade-off between complexity and noise resilience. Yet, for other perturbations, both networks showed similar performance dynamics when trained on perturbed data. A detailed exploration of the architectural characteristics of both models unearthed insights into their robustness and susceptibility to different perturbations. While BA-SWIN Transformers could be less dependent on high-frequency features, they did exhibit vulnerability to large rotation angles. On the other hand, while not inherently rotationally invariant, as they focus on local receptive fields, CNNs were found to manage small translations effectively. This property however can cause them to underperform the Transformers in tasks involving distribution shifts.

The final section of this chapter addressed the critical aspect of explainability, particularly within the BA-SWIN architecture. The inherent limitations of SWIN Transformers regarding global attention map visualisation were underscored. Although potential workarounds were proposed, they were acknowledged as not ideal for providing good visualisation of features contributing to brain age prediction. It was also made clear that while the literature provides a rich foundation for understanding these models, the compromise between functionality and explainability within SWIN architecture requires further exploration. The absence of clear pathways to robust interpretation and visualisation marks an area for potential innovation and advancement.

To summarise, this chapter has illuminated the complexities, capabilities, and challenges inherent in the BA-SWIN Transformers and HGL CNN models within the context of brain age prediction. It has shed light on the trade-offs between performance, adaptability, noise resilience, and explainability. It also underscores the importance of continued research into new methods capable of tackling these multifaceted dimensions, as the integration of AI models into clinical and research landscapes continues to expand.

6

Conclusion and Further Work

Contents

6.1 Overview

In this section, I summarise the contributions made by the work presented in this thesis and discuss their implications for the wider field, but also what their practical applications could be. I then described the limitations of work done in this thesis, and starting from these, I present several future, potentially interesting and relevant research directions. This section, and the thesis, will then close with a few concluding remarks.

6.2 Summary of Contributions

The thesis began with an introduction to the concept of brain ageing and the brain age gap, and how deviations from normal ageing, signifying either accelerated ageing or resilience to ageing, are associated with both biomedical factors, such as cognitive decline, neurodegenerative and chronic diseases, but also to lifestyle factors, such as one's habits and diet. It also highlighted the wide-scale use of structural neuroimaging for predictions, discussing how this might lead to the overlooking of potentially informative brain ageing related functional and microstructural changes in the brain. This underscored the importance of understanding age-related structural and functional changes in the brain using maps and modalities which have not been considered so far for the task of brain age prediction.

The cornerstone of the first results chapter (Chapter 3) revolved around the power of neuroimaging as a predictive tool for brain ageing. By proposing a simple yet powerful CNN architecture, named HGL, and exploring various MRI imaging modalities, the thesis establishes that each modality offers unique and valuable insights, as different maps derived from these modalities encode different information about the ageing brain. Overall, 191 statistically significant associations between the brain age deltas and UK Biobank nIDPs were found, the majority of them coming from maps that have not been traditionally utilised for brain age prediction tasks. The first chapter results also showed that ensembling brain age predictions post-training leads to not only improved brain age predictions but also higher correlations to nIDPs of interest through signal amplification, when nIDPs are used to guide ensemble map selection. This maps the results obtained when ensembling was based purely on the criterion of improving prediction accuracies, without using nIDPs to inform map selection, which although led to improved brain age predictions over those of the ensemble components, also led to a reduction in the number of significant nIDP associations for the ensemble. The first chapter also addressed the limitations of SFCN, a widely employed deep learning model used in the study of the brain age gap, cautioning against the use of a singular metric for accuracy measurements (such as MAE) and highlighting the need for careful model

evaluation to ensure that architecture design choices do not impact the accuracy and underlying distributions of the resulting brain age predictions.

The second results chapter (Chapter 4) builds on the map ensembling findings of the first results chapter, introducing the concept of multi-modal deep fusion learning to bolster information extraction from neuroimaging data. This was motivated by the several limitations of linear models identified in the first chapter, particularly when dealing with maps which associate differently with specific nIDPs, which could be mediated by the use of more complex, non-linear methods. While multi-modal brain age predictions are not a new concept, a review of literature suggests this work represents the first ever use of end-to-end trained deep fusion networks for exploring the brain age gap. Three primary CNN-based fusion strategies are evaluated, resulting in the surprising revelation that these complex methods did not exceed the performance of the ElasticNet linear method. In fact, the presented data suggests that ElasticNet not only performed better than the multi-modal deep fusion models, but is also more versatile, allowing the user to combine any number of maps at a small fraction of the computational cost necessary for fusion models. Nevertheless, a useful insight emerged from the second chapter: ≈ 9 identical networks are necessary for achieving stable and reliable brain age predictions, which is important for any future deployment of brain age models in clinical settings.

Up until this point, all work was carried out using deep learning models based on the CNN architecture. For this reason, in the final analytical chapter (Chapter 5), the inherent CNN limitations were addressed, such as their inherent lack of explainability and potential difficulties in modelling image contextual information. For this, a new network was developed using a novel, cutting-edge method referred to as Vision Transformers (ViTs). In particular, a modified version of this architecture named Shifted Window Transformers (SWIN) was used, due to its proven ability to work with large, high-resolution volumetric inputs. Following extensive hyperparameter tuning, the resulting BA-SWIN network was compared against the established HGL CNN. As far as it could be determined, this represents the first use of a pure Transformer model for brain age prediction. Despite not outperforming CNNs in

prediction accuracy, the BA-SWIN models displayed very similar performances, while training $\approx 50\%$ faster, and showing remarkable robustness towards certain types of imaging artefacts and data perturbations, as well as having the capability to provide informative attention map images. The later aspect highlights an important advantage of SWINs over CNNs, which is their inherent explainability thanks to the self-attention. However, challenges in global attention map visualisation emerged for the BA-SWIN architecture, suggesting potential future avenues of exploration. The chapter concludes by emphasising the trade-offs among model efficiency, robustness, and interpretability, underscoring the continual need for research and innovation.

6.3 Potential Implications for the Field and Practical Applications

With the rising emphasis on brain ageing, neuroimaging is crucial for capturing age-related structural and functional changes. Understanding these changes could significantly benefit early diagnosis and interventions for neurodegenerative diseases. Thus, it is envisioned that one of the largest implications of this work is that it could be the start of future investigations into the causal links underpinning the observed brain age deltas-nIDP associations. There are numerous benefits for understanding which biomedical and lifestyle factors have a positive or negative impact on brain ageing. For brevity, these could be grouped based on their field of application into medical, pharmacological, and socio-political benefits.

In the medical field, a better understanding of the brain ageing gap can help with some of the following aspects:

- **Early neurological disease diagnosis** by capturing subtle age-related deviations from population norms, and using these together with lifestyle and biomedical factors to allow healthcare professionals to diagnose conditions like Alzheimer's disease or NMOSD at an earlier stage, allowing for better management and improved patient outcomes;

- **Continuous monitoring & Targeted interventions** for specific at-risk populations, such as groups of individuals suffering from chronic conditions, to prevent the emergence of neurological conditions;
- **Personalised treatment plans** for at-risk individuals from accelerated brain ageing, either to prevent further degeneration or to shield patients from medications which might inadvertently aggravate their conditions;

These advances would not be possible without further progress in the pharmacological field. Here, some benefits could include:

- **Drug development**, which could be aided by the identification of factors which inhibit accelerated brain ageing, leading to resilience;
- **Clinical trials** could benefit from the use of brain age models either for monitoring for off-target drug effects, or for recruiting subjects;
- **AI model development** for applied medical imaging, given that the work presented here suggests that, with sufficient training and identical models, the inherent stochasticity of deep learning models can be mediated, leading to stable, predictable results. Stable, predictable, and in the near future hopefully explainable results are crucial for increasing the confidence of both drug developers and clinicians in AI technologies and their use in hospitals or drug development pipelines;

Finally, over a longer time horizon, continued work in this area could also translate into several socio-political benefits:

- **Design of ageing-friendly environments**, ensuring that factors which might lead to accelerated ageing are mediated;
- **Establishing public services and programmes** which promote healthy brain ageing and overall better quality of life;
- **Creating educational campaigns** raising awareness about factors which could lead to accelerated brain ageing;

In addition to these, the results presented in this thesis also have several more key implications for the field of brain age research. These include:

- Presenting the community a well organised set of rules, acting as the selection and organisation criteria for the UK Biobank data utilised in this work, facilitating the reproduction of the presented results, and facilitate comparisons between current and future brain age prediction models;
- Adding to the body of work promoting the use of multiple imaging modalities simultaneously, either individually or as part of an ensemble, to boost prediction accuracies and the insight gained through various studies;
- Rethinking the use of linear models, such as ElasticNet, which despite their lightweight and simple nature have proven to be powerful when constructing multi-map ensembles;
- Highlighting that, although deep learning models have seen widespread use in the medical imaging analysis field, they should always be extensively validated for the task at hand and not assumed to guarantee superior performance;
- Raising awareness in regard to the importance of robust predictions in the face of inherent model stochastic noise, particularly for any networks or results that are aimed at clinical applications. In addition, consideration needs also be given to the performance of models when faced with potentially lower-quality datasets than those on which they were originally trained;
- Revealing that, although CNNs have dominated the medical imaging field for over a decade, new architectures are being constantly developed which address some of the inherent limitations of the CNN platform;
- Emphasising the need for more explainable models, particularly when utilising them in conjunction with previously unexplored maps, enabling developers and clinicians to gain a better understanding of which features are driving predictions;

6.4 Limitations and Future Work

Despite the contributions and potential applications of the work presented in this thesis, there are nonetheless several limitations that need to be considered. These can be broadly split into two main categories: data-induced and method-induced limitations.

6.4.1 Data Limitations

Regarding data-induced limitations, the first one that needs to be discussed is that all the data utilised in this study comes from the UK Biobank and was pre-processed using the UK Biobank pipeline. This has several implications. Firstly, it means that all models and results presented in this thesis are biased towards the demographics captured by the UK Biobank, consisting mostly of subjects with white European phenotypes, which in turn leads to reduced variability and potential problems for the models in generalising well for subjects from other ethnic and genetic backgrounds [382]. Additionally, even if the data were representative of a more diverse population, it still means that all models and data are still biased towards the UK Biobank ecosystem, including aspects such as data and pre-processing assumptions, and their inherent limitations. Given this, it is vital to validate the models on external datasets to truly assess their generalisability and robustness. Such validation is essential to ensure that the models perform well across different demographic and clinical settings. It is highly likely that this validation would result in requirements for further model training and refinement (such as the inclusion of harmonisation considerations between different scanners) before the successful deployment of any of the models outside of the UK Biobank ecosystem. Thus, the models in their current state are not ready to be deployed in a clinical setting without further extensive refinement and validation.

Another data-induced limitation comes from the fact that no convergence studies were carried out to measure the adequate sample sizes needed for training and testing. Consequently, the impact that those sample sizes have on the presented results has not been quantified. Peng et al. found that test MAE of their SFCN

network decreases by $0.3-to-0.4$ years each time the training dataset doubles, with a plateau beginning to emerge after $\approx 10,000$ subjects [67]. Given the sex-segregated assumption built in this thesis, it would have been difficult to utilise this many subjects for training while allowing for a sufficiently large left-out test population. Moreover, it is likely that convergence curves would be dependent on the map-sex pairings, making an exhaustive analysis impractical. Yet, there are several factors that might have mitigated the lack of this convergence analysis. Firstly, both the female and male datasets were constructed by randomly sampling subjects from across each population, ensuring the underlying age distributions for the test, train, and validation datasets matched those of the whole UK Biobank population. Then, as it was seen in Chapter 4, when a sufficiently large number of networks was trained, the result became independent of the number of trained networks. Moreover, the convergence curves also showed the benefits of refinement training with additional subjects also quickly diminish when this convergence occurs.

In regards to the observed nIDP associations, the reported results could also be improved by increasing the number of subjects utilised for determining the correlations. Currently, some differences exist between the associations reported in this thesis and those found in other studies which utilised similar methods and data from the UK Biobank. For instance, for female subjects, Dinsdale et al. [21] report significant associations between T1 Nonlinear map predicted brain age deltas and nIDP categories such as Cardiac and Circulatory Measurements, or, for male subjects, they report more Medical History nIDPs passing the Bonferroni Threshold than have been observed in this work. These differences can be associated with the stochastic nature by which the training and testing datasets were created, as different seeds were likely utilised. In addition, the inclusion of a larger pool of subjects in their work (2511 female and 2183 male test subjects, compared to 3654 and 3231 respectively for Dinsdale et al.) enables the establishment of both more numerous and more statistically significant associations, simply due to the higher statistical power and the larger variability inherent in a larger dataset. Thus, in time, as more subject data is added to the UK Biobank (and other datasets) more significant

associations will be identified. It should, however, be noted that Dinsdale et al. did not train the networks to convergence, only ensembling 3 identical networks. This in turn could be an additional source of stochasticity. Nevertheless, the incorporation of additional test subjects would also enable the carrying out of more complex forms of analysis, such as genome-wide association studies for the full genome, which could reveal genes and gene sequences responsible for the acceleration or deceleration of brain ageing, which can then be targeted by pharmacological research. Yet, as mentioned before, all the above results, as well as the others made available as part of this work, and through the extension of this work to other datasets, need to be carefully analysed in order to infer causality [288].

Another limitation of the work presented in this thesis comes from the cross-sectional nature of the data, with any longitudinal confounding and long-term consequences associated with accelerated or decelerated brain ageing being ignored. At the time of writing (summer 2023), the UK Biobank only has ≈ 4000 subjects that were scanned at two-time points 2 years apart, which were not included in the subject pool used herein. These subjects, however, represent an insufficiently large population for carrying out an adequate longitudinal study, given the large number of subjects required to train the models. Yet, this is set to change, as funding has been approved for the repeat-scanning of up to $60,000$ of the UK Biobank subjects at 2 and 4 year intervals. This will provide 3 data points for a large cohort of subjects, enabling future longitudinal work.

6.4.2 Methodological Limitations

A significant limitation in this thesis is the uniform application of HGL and BA-SWIN across all maps. Given that different maps encode distinct types of information, there is an opportunity to create map-specific deep learning models tailored to each. For instance, when processing maps with primarily geometric data, like those from sMRI, both HGL and BA-SWIN yielded similar results in terms of accuracy and nIDP associations. This suggests they are both well-suited for such data. However, when processing the FA map, BA-SWIN results showed significant

associations to a series of nIDPs which could be related to the brain's microstructure, suggesting that it might have a stronger sensitivity to microstructural tissue changes associated with ageing. Applying this observation more broadly, techniques such as data augmentation and network training can also be tailored to specific maps. Thus, the path forward lies in developing specialised deep learning architectures, fine-tuned for each map's unique data. However, as evidenced by SFCN, it is important to consider the impact on results of any network refinements and design decisions.

Another limitation of this work is concerned with how explainable the models are. In particular, it would be important to know which areas, or features, in each map contribute to the observed predictions. While the attention activation maps in BA-SWIN provided some insight into the features considered relevant by the network, the lack of a global attention map limits the utility of these observations. Explainability, or lack thereof, represents a large gap in the brain age prediction field as current explainable AI techniques, such as attention gates and saliency maps, are difficult to interpret even for traditional maps [21]. The development of explainable AI techniques, particularly for regression (most methods for interpretability developed thus far have been designed for classification tasks) is a technically complex effort and is outside of the scope of this work. In addition, validating these methods is also difficult, as it relies on features that can be discerned by human observers. This is easy for tasks such as identifying faces, features, or items, but not so much for brain age predictions, particularly those coming from maps such as those derived from fMRI, for instance. Moreover, brain ageing is a phenomenon that impacts the entire brain, with nonlinear temporal and spatial variations, which might make selecting a particular set of key regions or features difficult [9]. However, to enable the clinical adoption of these models and methods, a better understanding of the biological meaning of local, regional, and global brain factors that are driving the brain age predictions is imperative [53, 383].

Finally, another notable methodological limitation of this research, common to many deep learning projects, is the substantial computational demands of training these models. Coupled with hardware constraints, this often results in training

only a limited number of networks for each sex-map pairing. This study shows that training more identical networks enhances the denoising effect. Yet, achieving full convergence might be challenging due to the inherently stochastic nature of neural network training, and reaching it might be computationally impractical, especially as the benefits decrease with additional runs. Such limitations can impact various facets of a deep learning project, including the extent of hyperparameter tuning. In this research, while both HGL and BA-SWIN were optimised for several maps, the resulting hyperparameters were assumed suitable for other maps. This suggests potential room for improvement in the performance of various network-map-sex combinations with more fine-tuning. In addition to the above, the computational demands of training the models used herein also impose limitations on the reproducibility of this work.

6.5 Future Directions

Several future research directions have already been mentioned or hinted at in the previous sections. Here, a list of potential future research directions is presented:

- **Enhanced explainability**: Building on the advances made using Transformer networks such as BA-SWIN, work focusing on improving their explainability could set new benchmarks for AI transparency in medical imaging;
- **Longitudinal health data integration**: The addition of longitudinal data can provide a more holistic and dynamic understanding of brain ageing. By assessing changes over multiple time points, models can be better calibrated and can reveal patterns often missed in single time-point analyses. Given the UK Biobank's ongoing efforts, this direction can provide immediate and significant returns;
- **Genetic data integration and GWAS studies**: The continuous data addition efforts carried by the UK Biobank could also facilitate continued work into using genetic information to find which underlying genetic factors

influence brain ageing. Incorporating this alongside brain age deltas can potentially identify genetic markers associated with faster or slower ageing rates, thus providing preventive or intervention targets;

- **Generalisation across datasets**: More work is needed to ensure models are adaptable across diverse datasets, accounting for any built-in population biases, and robust enough to be deployed in clinical settings. This could include the use, or incorporation of, more geographically and demographically varied datasets, the use of augmentation techniques, and the inclusion of clinical-grade datasets, among other;

- **Causal studies on nIDPs**: Building on the results presented herein, future work transitioning from identifying associations to understanding causal relationships is required and can be extremely insightful;

- **Tailored architectures for specific applications**: As discussed above, utilising the same model, hyperparameters, training and data perturbation and augmentation techniques across all maps and modalities can be self-penalising. As such, the creation of task-bespoke models could be beneficial in extracting more information from the available datasets;

6.6 Concluding Remarks

As the final page of this chapter concludes, I find myself reflecting on the journey and pioneering spirit that has led me, under the wise guidance of my supervisors, to harness the power of deep learning and apply it to neuroimaging. What the work presented in this thesis shows is but a glimpse into the vast possibilities that the field of AI applied to neuroimaging, and in particular brain age prediction, has to offer. Every image, every voxel, every nIDP association, beckons us to delve deeper, to seek answers and extract meaning, and to venture into the unknown. For those reading this thesis at the start of their PhD journey, I can assure you that the confluence of neuroscience and AI holds promises far grander than you

might imagine at this stage. I am filled with optimism at the thought of all the benefits that continued work in this area can bring about, and I would like to thank the reader for reaching this point, and pass on the best well wishes to you in your own present and future research.

Appendices

A

Brief Summary of Prior Brain Ageing Research

Table A.1: sMRI brain age prediction studies (1/2). The Weighted MAE is calculated according to Cole et al. [8]. Abbreviations according to Dinsdale [73]: BG = Basal Ganglia; CNN = Convolutional Neural Network; CBR = Correlation-Based Regression; CSF = Cerebrospinal Fluid; DGP = Deep Gaussian Processes; GPR = Gaussian Processes Regression; GM = Grey matter; GNN = Graph Neural Network; HMM = Hidden Markov Model; MFN = Multi-Feature-Based Network; MVNL-R = Multivariate Non-Linear Regression; NN = Neural Network; RVR = Relevance Vector Regression; SVR = Support Vector Regression; VBM = Voxel Based Morphometry; WM = White Matter.

Study	Year	Training Data Type	Model	Age Range	MAE	Weighted MAE
Franke [52]	2010	T1 (GM)	RVR	20-86	4.61	0.07
Wang [69]	2011	T1 (GM + WM + CSF)	HMM	50-86	6.2	0.172
Brown [78]	2012	T1 (cortical thickens and area + subcortical structure volume)	MVNL-R	3-20	1.71	0.101
Brown [78]	2012	T2 (intensity measures with subcortical ROIs and white matter tracts)	MVNL-R	4-20	1.6	0.094
Franke [148]	2012	T1 (GM + WM)	RVR	5-19	1.2	0.086
Koutsouleris [93]	2013	T1 (GM)	SVR	18-65	4.6	0.098
Su [54]	2013	T1 (GM)	RVR	18-91	6.54	0.09
Wang [384]	2014	T1 (cortical thickness, area and curvatures)	RVR	20-82	4.57	0.074
Khundrakpam [276]	2015	T1 (cortical thickness)	ElasticNet	5-21	1.68	0.105
Kondo [385]	2015	T1 (GM + WM + CSF)	RVR	20-75	4.498	0.082
Cole [126]	2015	T1 (GM)	GPR	18-90	6.2	0.086
Cole [126]	2015	T1 (WM)	GPR	18-90	6.16	0.086
Erus [80]	2015	T1 (GM)	SVR	8-22	1.52	0.109
Erus [80]	2015	T1 (WM)	SVR	8-22	1.71	0.122
Erus [80]	2015	T1 (ventricle volume)	SVR	8-22	2.71	0.194
Fujimoto [85]	2016	T1 (GM + WM + CSF)	RVR	20-80	4.48	0.075
Cole [9]	2017	T1	GPR	18-90	11.81	0.164
Cole [9]	2017	T1 (GM)	GPR	18-90	4.66	0.065
Cole [9]	2017	T1 (WM)	GPR	18-90	5.88	0.082
Cole [9]	2017	T1 (GM + WM)	GPR	18-90	4.41	0.061
Cole [9]	2017	T1	CNN	18-90	4.65	0.065
Cole [9]	2017	T1 (GM)	CNN	18-90	4.16	0.058
Cole [9]	2017	T1 (WM)	CNN	18-90	5.14	0.071
Cole [9]	2017	T1 (GM + WM)	CNN	18-90	4.34	0.06
Valizadeh [63]	2017	T1 (cortical thickens, area and volumes)	NN	8-18	1.23	0.123
Valizadeh [63]	2017	T1 (cortical thickens, area and volumes)	NN	18-65	4.5	0.096
Valizadeh [63]	2017	T1 (cortical thickens, area and volumes)	NN	65-96	4.97	0.16
Huang [66]	2017	T1	CNN	20-80	4	0.067
Liem [75]	2017	T1 (cortical thickens and area + subcortical structure volume)	SVR	19-82	4.83	0.077
Guggenmos [59]	2017	VBM (parcelled volumes)	Ridge	21-65	6.9	0.157
Popescu [324]	2018	T1 (GM + WM)	DGP	18-88	4.44	0.063
Bagarinao [57]	2018	T1 (GM)	LASSO	21-86	7.2	0.111
Lancaster [386]	2018	T1 (GM)	SVR	16-90	5.08	0.069
Gutierrez Becker [70]	2018	T1 (VBM + cortical thickness and volume)	GPR	6-92	3.86	0.045
Pardoe [129]	2018	T1 (cortical thickness and surface)	RVR	8-69	7.2	0.118
Monte-Rubio [323]	2018	T1 (GM, WM, BG, multiple Jacobians, Divergence of Velocities, Scalar Momenta)	GPR	20-86	5	0.076
Eavani [113]	2018	T1 (GM)	SVR	50-96	4.41	0.096
Varikuti [58]	2018	VBM (parcelled volumes)	LASSO	55-75	3.4	0.17
Varikuti [58]	2018	VBM (parcelled volumes)	LASSO	18-81	4.9	0.078
Aycheh [387]	2018	T1 (cortical thickness)	Sparse Group LASSO	45-91	4.05	0.088
Bermudez [388]	2019	T1 + T1 volumetric features	CNN	4-96	4.08	0.044
Niu [260]	2019	T1 (cortical volume)	NN	8-21	1.649	0.127
Wang [122]	2019	T1 (GM)	CNN	45-96	4.45	0.087

Table A.2: sMRI brain age prediction studies (2/2). The Weighted MAE is calculated according to Cole et al. [8]. Abbreviations according to Dinsdale [73]: BG = Basal Ganglia; CNN = Convolutional Neural Network; CBR = Correlation-Based Regression; CSF = Cerebrospinal Fluid; DGP = Deep Gaussian Processes; GPR = Gaussian Processes Regression; GM = Grey matter; GNN = Graph Neural Network; HMM = Hidden Markov Model; MFN = Multi-Feature-Based Network; MVNL-R = Multivariate Non-Linear Regression; NN = Neural Network; RVR = Relevance Vector Regression; SVR = Support Vector Regression; VBM = Voxel Based Morphometry; WM = White Matter;

Study	Year	Training Data Type	Model	Age Range	MAE	Weighted MAE
Jonsson [160]	2019	T1	CNN	18-75	4.006	0.07
Jonsson [160]	2019	Jacobian	CNN	18-75	4.804	0.084
Jonsson [160]	2019	VBM	CNN	18-75	4.641	0.081
Jonsson [160]	2019	T1 (WM)	CNN	18-75	4.189	0.073
Jonsson [160]	2019	T1 + Jacobian + T1 (GM) + T1 (WM)	CNN	18-75	3.388	0.059
Jiang [389]	2020	T1 (GM parcellations)	CNN	18-90	5.55	0.077
Hong [390]	2020	T1	CNN	0-5	0.185	0.037
Rao [391]	2020	T1	CNN	12-30	1.96	0.109
Cole [81]	2020	T1 (145 phenotypes)	LASSO	45-80	4.14	0.118
Cole [81]	2020	T2-FLAIR (1 phenotype)	LASSO	45-80	5.653	0.162
Cole [81]	2020	T2* (14 phenotypes)	LASSO	45-80	5.78	0.165
Bashyam [133]	2020	T1	CNN	3-95	3.68	0.04
de Lange [138]	2020	T1(1118 GM features)	XGBoost Regression	60-84	3.6	0.15
de Lange [138]	2020	T1(246 WM features)	XGBoost Regression	60-84	3.51	0.146
Anatürk [392]	2021	T1 (GM and WM features)	XGBoost Regression	60-82	3.11	0.141
Mouches [82]	2021	T1	CNN	21-81	4.01	0.067
Baecker [321]	2021	T1 (cortical thickness, surface and curvature)	RVR	47-73	3.69	0.142
Hwang [393]	2021	T2	CNN	19-88	4.22	0.061
Hu [394]	2021	T1	CNN	6-18	1.01	0.084
Bellantuono [395]	2021	T1	NN (complex)	7-64	2.19	0.038
Peng [67]	2021	T1	CNN	44-80	2.14	0.059
Dinsdale [21]	2021	T1	CNN	44-80	2.71	0.075
Taylor [88]	2021	T1	CNN	18-89	6.546	0.092
Pina [65]	2021	T1 (GM parcellations)	GNN	45-82	4.27	0.115
Yang [271]	2021	T1	CNN	16-92	2.129	0.028
Mendes [396]	2021	T1	CNN	6-20	1.43	0.102
Lombardi [92]	2021	T1	CNN	6-48	2.7	0.064
Hepp [91]	2021	T1	CNN	20-72	3.2	0.062
Cheng [134]	2021	T1 (+ Subject Sex)	CNN	17-98	2.428	0.03
Cropley [97]	2021	T1 (111 features)	SVR	8-21	1.61	0.124
Leonardsen [53]	2022	T1	CNN	3-95	2.23	0.024
Wood [74]	2022	T2	CNN	18-95	2.97	0.039
Wood [74]	2022	T1	CNN	18-95	3.83	0.05
Wood [74]	2022	T1 + T2	CNN	18-95	3.35	0.044
Ren [84]	2022	T1	CNN	50-85	4.91	0.14
Sanford [90]	2022	T1	CNN	22-37	2.72	0.181
Poloni [114]	2022	T1 (hippocampal segmentation)	CNN	20-70	3.31	0.066
He [68]	2022	T1	CNN-Transformer Hybrid	0-97	2.38	0.025
Chen [94]	2022	T1 (GM + WM)	NN	16-62	3.99	0.087
Fu [250]	**2022**	**T1 (T1 + GM + WM)**	**ConvNeXt**	**3-97**	**2.097**	**0.022**
Dias [61]	2022	T1 (Deformation Fields)	RVR	20-86.3	7.1	0.107
Dias [61]	2022	T1 (GM)	RVR	20-86.3	7.96	0.12
Dias [61]	2022	T1 (WM)	RVR	20-86.3	9.99	0.151
Dias [61]	2022	T1 (CSF)	RVR	20-86.3	9.12	0.138
Dias [61]	2022	T1 (Deformation Fields + GM)	RVR	20-86.3	6.9	0.104
Linli [146]	2022	T1 (GM parcellations)	XGBoost Regression	44-81	4.22	0.114
Drobinin [109]	2022	T1 + T2 (121 cortical, subcortical and global features)	XGBoost Regression	9-19	1.49	0.149
Liu [322]	2022	T1 (cortical thicknes, area, curvature index, folding index, local gyrification index and sufrace area)	MFN	20-94	3.73	0.05
Ganaie [62]	2022	T1 (GM)	SVR	20-94	2.91	0.039

Table A.3: swMRI brain age prediction studies. The Weighted MAE is calculated according to Cole et al. [8]. Abbreviations according to Dinsdale [73]: CNN = Convolutional Neural Network

Study	Year	Training Data Type	Model	Age Range	MAE	Weighted MAE
Hofmann [83]	2022	SWI	CNN	18-82	5.74	0.09

Table A.4: dMRI brain age prediction studies. The Weighted MAE is calculated according to Cole et al. [8]. Abbreviations according to Dinsdale [73]: AD = Axial Diffusivity; ADC = Apparent diffusion coefficient; CBR = Case-based reasoning; CNN = Convolutional Neural Network; DWI = Diffusion Weighted Image; FA = Fractional Anisotropy; GPR = Gaussian Processes Regression; MD = Mean Diffusivity; MO = Mode of Anisotropy; NN = Neural Network; RD = Radial Diffusivity; RVR = Relevance Vector Regression; SVR = Support Vector Regression;

Study	Year	Training Data Type	Model	Age Range	MAE	Weighted MAE
Brown [78]	2012	dMRI (FA and ADC diffusivity measures in subcortical structures, and white matter tracts)	MVNL-R	4-20	1.74	0.102
Mwangi [251]	2013	dMRI FA	RVR	4-85	8.2	0.101
Mwangi [251]	2013	dMRI MD	RVR	4-85	7.1	0.088
Mwangi [251]	2013	dMRI RD	RVR	4-85	6.94	0.086
Mwangi [251]	2013	dMRI AD	RVR	4-85	7.161	0.088
Han [397]	2014	dMRI (structural connectivity)	CBR	4-85	7.656	0.095
Erus [80]	2015	dMRI FA	SVR	8-22	1.41	0.101
Erus [80]	2015	dMRI ADC	SVR	8-22	1.35	0.096
Lin [398]	2016	dMRI (structural connectivity)	NN	50-79	4.29	0.148
Niu [64]	2019	dMRI FA	GPR	8-21	1.965	0.151
Niu [64]	2019	dMRI MD	Ridge Regression	8-21	1.829	0.141
Cole [81]	2020	dMRI (675 phenotypes)	LASSO	45-80	3.897	0.111
Wood [74]	**2022**	**dMRI (DWI)**	**CNN**	**18-95**	**3.98**	**0.052**
Guo [399]	2022	dMRI FA	SVR	20-80	9.19	0.153
He [252]	2022	dMRI (FA and MD Features)	CNN	22-37	2.515	0.168
He [252]	2022	dMRI (FA)	CNN	22-37	2.768	0.185
Petersen [253]	2022	dMRI (FA and MD Features)	GPR	31-61	5.1	0.17

Table A.5: rsfMRI brain age prediction studies. The Weighted MAE is calculated according to Cole et al. [8]. Abbreviations according to Dinsdale [73]: CNN = Convolutional Neural Network; GPR = Gaussian Processes Regression; SVR = Support Vector Regression;

Study	Year	Training Data Type	Model	Age Range	MAE	Weighted MAE
Wang [400]	2012	rsfMRI (functional connectivity)	Locally Adjusted SVR	8-79	7.5	0.106
Liem [75]	2017	rsfMRI (functional connectivity)	SVR	19-82	5.77	0.092
Liem [75]	**2017**	**rsfMRI (stacked volumes)**	**SVR**	**19-82**	**5.25**	**0.083**
Li [76]	2018	rsfMRI (3D Intrinsic Connectivity Networks)	CNN	8-22	2.15	0.154
Eavani [113]	2018	rsfMRI (functional connectivity)	SVR	50-96	5.54	0.12
Niu [64]	2019	rsfMRI (atlas based low-frequency fluctuations)	GPR	8-21	1.770	0.136
Niu [64]	2019	rsfMRI (atlas based regional homogeneity)	GPR	8-21	1.924	0.148
Cole [81]	2020	rsfMRI (210 phenotypes)	LASSO	45-80	5.261	0.15
de Lange [138]	2020	rsfMRI (3655 features)	XGBoost Regression	60-84	4.18	0.174
Gao [77]	2022	rsfMRI (atlas based time series)	RNN	10-80	6.507	0.093
Millar [115]	2022	rsfMRI (functional connectivity)	GPR	18-89	8.195	0.115
Luo [108]	2022	rsfMRI (functional connectivity)	Stacking Model	12-82	8.3055	0.119

Table A.6: tfMRI brain age prediction studies. The Weighted MAE is calculated according to Cole et al. [8]. Abbreviations according to Dinsdale [73]: CNN = Convolutional Neural Network

Study	Year	Training Data Type	Model	Age Range	MAE	Weighted MAE
Cole [81]	2020	tfMRI (14 phenotypes)	LASSO	45-80	5.929	0.169

Table A.7: Multi-modal brain age prediction studies. The Weighted MAE is calculated according to Cole et al. [8]. Abbreviations according to Dinsdale [73]: AD = Axial Diffusivity; ADC = Apparent diffusion coefficient; CNN = Convolutional Neural Network; CSF = Cerebrospinal Fluid; DTI = Diffusion Tensor Imaging; FA = Fractional Anisotropy; GM = Grey matter; GPR = Gaussian Processes Regression; ICA = Independent Component Analysis; MD = Mean Diffusivity; MVNL-R = Multivariate Non-Linear Regression; NN = Neural Network; PCA = Principal Component Analysis; RD = Radial Diffusivity; SVR = Support Vector Regression; WM = White Matter;

Study	Year	Training Data Type	Model	Age Range	MAE	Weighted MAE
Brown [78]	2012	T1 (cortical thickens and area + subcortical structure volume) + T2 (intensity measures with subcortical ROIs and white matter tracts) + dMRI (FA and ADC diffusivity measures in subcortical structures and white matter tracts)	MVNL-R	3-20	1.03	0.061
Erus [80]	2015	T1 (GM, WM, ventricle volume) + dMRI (FA and ADC)	SVR	8-22	1.22	0.087
Liem [75]	2017	T1 (cortical thickens and area + subcortical structure volume) + rsfMRI (stacked volumes)	SVR	19-82	4.29	0.068
Richard [299]	2018	T1 + DTI (MD, RD and AD)	XGBoost Regression	18-86	6.14	0.089
Niu [64]	2019	T1 (cortical volume) + dMRI (FA and MD)	GPR	8-21	1.395	0.107
Niu [64]	2019	T1 (cortical volume) + rsfMRI (atlas based low-frequency fluctuations and regional homogeneity)	GPR	8-21	1.463	0.113
Niu [64]	2019	dMRI (FA and MD) + rsfMRI (atlas based low-frequency fluctuations and regional homogeneity)	GPR	8-21	1.512	0.116
Niu [64]	2019	T1 (cortical volume) + dMRI (FA and MD) + rsfMRI (atlas based low-frequency fluctuations and regional homogeneity)	NN	8-21	1.381	0.106
Smith [35]	2019	2641 IDPs: T1 + T2+ dMRI + rsfMRI + tfMRI	Linear Regression	45-80	4	0.114
Cole [81]	2020	T1 (145 phenotypes) + T2-FLAIR (1 phenotype) + T2*(14 phenotypes) + dMRI (675 phenotypes) + tfMRI (14 phenotypes) + rsfMRI (210 phenotypes)	LASSO	45-80	3.515	0.1
Cole [81]	2020	T2-FLAIR (1 phenotype) + T2* (14 phenotypes) + dMRI (675 phenotypes) + tfMRI (14 phenotypes) + rsfMRI (210 phenotypes)	LASSO	45-80	3.752	0.107
Cole [81]	2020	T1 (145 phenotypes) + T2* (14 phenotypes) + dMRI (675 phenotypes) + tfMRI (14 phenotypes) + rsfMRI (210 phenotypes)	LASSO	45-80	3.572	0.102
Cole [81]	2020	T1 (145 phenotypes) + T2-FLAIR (1 phenotype) + dMRI(675 phenotypes) + tfMRI (14 phenotypes) + rsfMRI (210 phenotypes)	LASSO	45-80	3.598	0.103
Cole [81]	2020	T1 (145 phenotypes) + T2-FLAIR (1 phenotype) + T2*(14 phenotypes) + tfMRI (14 phenotypes) + rsfMRI (210 phenotypes)	LASSO	45-80	3.975	0.114
Cole [81]	2020	T1 (145 phenotypes) + T2-FLAIR (1 phenotype) + T2*(14 phenotypes) + dMRI (675 phenotypes) + rsfMRI (210 phenotypes)	LASSO	45-80	3.569	0.102
Cole [81]	2020	T1 (145 phenotypes) + T2-FLAIR (1 phenotype) + T2*(14 phenotypes) + dMRI (675 phenotypes) + tfMRI (14 phenotypes)	LASSO	45-80	3.522	0.101
Smith [56]	2020	3913 IDPs: T1 + T2+ dMRI + rsfMRI + tfMRI	PCA + ICA + Linear Regression	45-80	2.9	0.083
de Lange [138]	2020	5019 features: T1(1118 GM, 246 WM) + rsfMRI (3655)	XGBoost Regression	60-84	3.37	0.14
Wang [132]	2021	T1 (738 GM, WM and CSF features) + rsfMRI (functional connectivity)	Least Squares Regression	5-21.5	1.1	0.067
Mouches [82]	2022	T1 + Angiography	CNN	21-81	3.85	0.064
Wood [74]	**2022**	**T2 + dMRI**	**CNN**	**18-95**	**3.31**	**0.043**
Wood [74]	2022	T1 + T2	CNN	18-95	3.35	0.044
Hofmann [83]	2022	T1 + T2 FLAIR + SWI	CNN + Ridge	18-82	3.37	0.053

B
Magnetic Resonance Imaging

B.1 IDPs

Following the guidelines listed by Alfaro et al [173], the 3,921 UK Biobank IDPs can be thematically split into the following categories: Cortical Area (372), Cortical Grey-White Contrast (70), Cortical Thickness (306), Regional T2-star (14), Regional and Tissue Intensity (41), Regional and Tissue Volume (647), White Matter (WM) Tract Diffusivity (300), WM Tract FA (75), WM Tract ICVF (75), WM Tract ISOVF (75), WM Tract MO (75), WM Tract OD (75), White Matter Hyperintensity Volume (3), rfMRI Connectivity (1701), rfMRI Node Amplitude (76), tfMRI Activation (16) [173]. A comprehensive breakdown of all these IDPs is available in the UK Biobank resources [1] and the IDP spreadsheet [2].

B.2 nIDPs

While there is no definitive guideline for splitting the 17,526 nIDPs in UK Biobank, this thesis employs guidelines established by Alfaro et al [173] and Dinsdale et al [21, 173] to split them into into 19 thematically defined categories for interpretability as follows: Ethnic Background (3), Genetic Markers (63), Physical Activity (107),

[1] http://biobank.ctsu.ox.ac.uk/showcase/showcase/docs/brain$_m ri.pdf$
[2] https://www.fmrib.ox.ac.uk/ukbiobank/BrainAgingModes

Early Life Factors (32), Data Collection Duration (16), Lifestyle (868), Diet (1689), Alcohol (138), Tobacco (122), Physical Measurements (260), Skeletal Measurements (180), Cardiac & Circulatory Measurements (815), Hearing Test (755), Eye Test (434), Physical Test (333), Blood Assays (468), Cognitive Tests (1167), Medical History (9780), and Mental Health (297). Interested readers can explore these nIDPs by accessing the UK Biobank Showcase website [3].

B.3 Confounds

The 613 confounds listed in the UK Biobank were split into 6 categories, using the guidelines established by Alfaro et al [228], as follows: Subject-Specific Confounds (Age, Sex, Age-Sex, Headsize), Scanner Acquisition Protocol Processing Parameters (Site, Batch, Centre for Magnetic Resonance Research - CMRR - Software Version, Protocol, Service Pack, B0 field ramp-down/up events - SCAN RAMP, Cold Head replacement - Scan Cold Head, Head Coil replacements - Scan Head Coil, Scan Miscellaneous Events, Temporary unintended protocol changes in SWI imaging - Flipped SWI and Protocol, A confound variable describing if T2 FLAIR was used - FS T2, A confound variable describing heating-related effects on eddy currents - New Eddy, Global Intensity Scaling, and TE acquisition time), Head Motion Confounds (Structural Motion, DVARS, Head Motion, Head Motion over space and time - ST), Table-Position-related Confounds (Table Position, Eddy QC), Nonlinearities and Crossed Terms (Non Linearities, Cross terms), Date/Time Related Factors (Acquisition time, Acquisition Date).

[3] https://biobank.ndph.ox.ac.uk/showcase/

C
Transformers

C.1 Obtaining Attention Activation Maps

In this final subsection, an explanation is provided regarding the operations utilised to extract the attention activation maps **A** from Equation 2.2.3.2, used to determine what the network is paying attention to. This is done as the required operations are not entirely intuitive, and readers interested in reproducing these results would benefit from having this understanding.

As described above in Chapter 2 Section 2.2.3.1, a 3D volume input of size $H \times W \times D = 160 \times 192 \times 160$ passed to a vanilla 3D SWIN network first gets padded to $168 \times 196 \times 168$, then split into patches with patch size $n = 2$, which are then grouped into windows with $M = 7$ patches per window. This results in a window volume of size $12 \times 14 \times 12$.

For the following, it will be assumed that only the first layer of a SWIN is considered for simplicity. This has a feature embedding dimensionality $C = 24$. Thus, after embedding, the latent dimensionality of the input becomes $windows \times patches \times features = 2016 \times 343 \times 24$, equivalent to $(12 \times 14 \times 12) \times (7 \times 7 \times 7) \times 24$. When multiplying the first 3 terms by $8 = 2^3$ which is the dimension of a patch, the original input dimensionality can be recovered.

When passing through an MSA block, assuming $h = 3$ heads for the vanilla SWIN, $d = C/h = 8$, meaning that the latent dimensionality of the $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ matrices becomes $2016 \times 343 \times 3 \times 8$ each, which in turn results in a dimensionality of $2016 \times 3 \times 343 \times 343$ for $\mathbf{A}$, representing the [flattened patch window space x number of multi-heads x query patches x key patches].

As the interest is in retaining the attention corresponding to the keys, the query dimension can be averaged across and then squeezed, resulting in a latent dimensionality of $2016 \times 3 \times 343$, which can be reformed into a 3D volume for visualisation. To do this, first the latent representation is reshaped into $3 \times 2016 \times 343 = 3 \times 2016 \times (7 \times 7 \times 7) = 3 \times 2016 \times 7 \times 7 \times 7$. This assumes that the window size is kept constant at $M = 7$. The following step involves determining the number of windows from the flattened window space (i.e., 2016). This can be achieved using the same equations SWIN employs to reconstruct the latent representations between blocks, described in Equation C.1.0.1, where $n = 2$ is the patch size in the vanilla SWIN network, and the *depth* represents the order of a layer l in the network hierarchy $l \in [1, 4]$ for a 4-layer deep network.

$$n_{\text{windows x}} = \lceil \frac{\lceil \frac{H}{M \times n} \rceil}{2^{\text{depth}}} \rceil \tag{C.1.0.1}$$

$$n_{\text{windows y}} = \lceil \frac{\lceil \frac{W}{M \times n} \rceil}{2^{\text{depth}}} \rceil \tag{C.1.0.2}$$

$$n_{\text{windows z}} = \lceil \frac{\lceil \frac{D}{M \times n} \rceil}{2^{\text{depth}}} \rceil \tag{C.1.0.3}$$

Thus, the flattened window space can be reshaped from a vector to a 3D volume. After this, the original padded 3D volume can be recreated by rearranging the multi-dimensional feature representation. For instance, the $3 \times 2016 \times 7 \times 7 \times 7$ latent dimension can be rearranged to $3 \times 12 \times 14 \times 12 \times 7 \times 7 \times 7 = 3 \times 12 \times 7 \times 14 \times 7 \times 12 \times 7 = 3 \times 84 \times 98 \times 84$. At this point, nearest neighbour upsampling, with a scale factor of 2^{depth} can be used to obtain a 3D attention volume, which can be overlayed, after cropping, on the original input for further analysis. It should be mentioned that, in the case of SW-MSA, the shifted maps need to be rolled back into their original alignment.

References

[1] Carlos López-Otın et al. “The hallmarks of aging”. In: *Cell* 153.6 (2013), pp. 1194–1217.

[2] Thomas A Rando and Howard Y Chang. “Aging, rejuvenation, and epigenetic reprogramming: resetting the aging clock”. In: *Cell* 148.1-2 (2012), pp. 46–57.

[3] Sven Bocklandt et al. “Epigenetic predictor of age”. In: *PloS one* 6.6 (2011), e14821.

[4] Katja Franke and Christian Gaser. “Ten years of BrainAGE as a neuroimaging biomarker of brain aging: what insights have we gained?” In: *Frontiers in neurology* (2019), p. 789.

[5] Denise C Park and Patricia Reuter-Lorenz. “The adaptive brain: aging and neurocognitive scaffolding”. In: *Annual review of psychology* 60 (2009), pp. 173–196.

[6] Ian J Deary et al. “Age-associated cognitive decline”. In: *British medical bulletin* 92.1 (2009), pp. 135–152.

[7] Timothy A Salthouse. “Selective review of cognitive aging”. In: *Journal of the International neuropsychological Society* 16.5 (2010), pp. 754–760.

[8] James H Cole, Katja Franke, and Nicolas Cherbuin. “Quantification of the biological age of the brain using neuroimaging”. In: *Biomarkers of human aging*. Springer, 2019, pp. 293–328.

[9] James H Cole et al. “Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker”. In: *NeuroImage* 163 (2017), pp. 115–124.

[10] Anatole S Dekaban and Doris Sadowsky. “Changes in brain weights during the span of human life: relation of brain weights to body heights and body weights”. In: *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 4.4 (1978), pp. 345–356.

[11] Thomas L Kemper. “Neuroanatomical and neuropathological changes during aging and dementia.” In: (1994).

[12] Catriona D Good et al. “A voxel-based morphometric study of ageing in 465 normal adult human brains”. In: *Neuroimage* 14.1 (2001), pp. 21–36.

[13] BM Hubbard and JM Anderson. “Age, senile dementia and ventricular enlargement”. In: *Journal of Neurology, Neurosurgery & Psychiatry* 44.7 (1981), pp. 631–635.

[14] Mark P Mattson and Thiruma V Arumugam. “Hallmarks of brain aging: adaptive and pathological modification by metabolic states”. In: *Cell metabolism* 27.6 (2018), pp. 1176–1199.

[15] José Julio Rodrıguez, Harun N Noristani, and Alexei Verkhratsky. "The serotonergic system in ageing and Alzheimer's disease". In: *Progress in neurobiology* 99.1 (2012), pp. 15–41.

[16] Gregory Hannum et al. "Genome-wide methylation profiles reveal quantitative views of human aging rates". In: *Molecular cell* 49.2 (2013), pp. 359–367.

[17] Steve Horvath. "DNA methylation age of human tissues and cell types". In: *Genome biology* 14.10 (2013), pp. 1–20.

[18] Lilach Soreq et al. "Major shifts in glial regional identity are a transcriptional hallmark of human brain aging". In: *Cell reports* 18.2 (2017), pp. 557–570.

[19] Lars Bäckman et al. "Linking cognitive aging to alterations in dopamine neurotransmitter functioning: recent data and future avenues". In: *Neuroscience & Biobehavioral Reviews* 34.5 (2010), pp. 670–677.

[20] Joanna M Wardlaw et al. "Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration". In: *The Lancet Neurology* 12.8 (2013), pp. 822–838.

[21] Nicola K Dinsdale et al. "Learning patterns of the ageing brain in MRI using deep convolutional networks". In: *NeuroImage* 224 (2021), p. 117401.

[22] Monika Hollander et al. "Incidence, risk, and case fatality of first ever stroke in the elderly population. The Rotterdam Study". In: *Journal of Neurology, Neurosurgery & Psychiatry* 74.3 (2003), pp. 317–321.

[23] Amy Reeve, Eve Simcox, and Doug Turnbull. "Ageing and Parkinson's disease: why is advancing age the biggest risk factor?" In: *Ageing research reviews* 14 (2014), pp. 19–30.

[24] Chengxuan Qiu, Miia Kivipelto, and Eva Von Strauss. "Epidemiology of Alzheimer's disease: occurrence, determinants, and strategies toward intervention". In: *Dialogues in clinical neuroscience* (2022).

[25] Philip A Greiner, David A Snowdon, and Frederick A Schmitt. "The loss of independence in activities of daily living: the role of low normal cognitive function in elderly nuns." In: *American Journal of Public Health* 86.1 (1996), pp. 62–66.

[26] Ranmalee Eramudugolla et al. "Self-reported cognitive decline on the informant questionnaire on cognitive decline in the elderly is associated with dementia, instrumental activities of daily living and depression but not longitudinal cognitive change". In: *Dementia and geriatric cognitive disorders* 34.5-6 (2012), pp. 282–291.

[27] Maria M Johansson, Jan Marcusson, and Ewa Wressle. "Cognitive impairment and its consequences in everyday life: experiences of people with mild cognitive impairment or mild dementia and their relatives". In: *International psychogeriatrics* 27.6 (2015), pp. 949–958.

[28] Helena Chui et al. "Trajectories of depressive symptoms in old age: Integrating age-, pathology-, and mortality-related changes." In: *Psychology and aging* 30.4 (2015), p. 940.

[29] Hannah Ritchie and Max Roser. "Age structure". In: *Our World in Data* (2019).

[30] Rachael I Scahill et al. “A longitudinal study of brain volume changes in normal aging using serial registered magnetic resonance imaging”. In: *Archives of neurology* 60.7 (2003), pp. 989–994.

[31] Naftali Raz et al. “Trajectories of brain aging in middle-aged and older adults: regional and individual differences”. In: *Neuroimage* 51.2 (2010), pp. 501–511.

[32] Marco Lorenzi et al. “Disentangling normal aging from Alzheimer’s disease in structural magnetic resonance images”. In: *Neurobiology of aging* 36 (2015), S42–S52.

[33] Nhat Trung Doan et al. “Distinguishing early and late brain aging from the Alzheimer’s disease spectrum: consistent morphological patterns across independent samples”. In: *Neuroimage* 158 (2017), pp. 282–295.

[34] Jennifer H Barnett et al. “Early intervention in Alzheimer’s disease: a health economic study of the effects of diagnostic timing”. In: *BMC neurology* 14.1 (2014), pp. 1–9.

[35] Stephen M Smith et al. “Estimation of brain age delta from brain imaging”. In: *Neuroimage* 200 (2019), pp. 528–539.

[36] Cathie Sudlow et al. “UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age”. In: *PLoS medicine* 12.3 (2015), e1001779.

[37] Naftali Raz et al. “Regional brain changes in aging healthy adults: general trends, individual differences and modifiers”. In: *Cerebral cortex* 15.11 (2005), pp. 1676–1689.

[38] Anders M Fjell et al. “Critical ages in the life course of the adult brain: nonlinear subcortical aging”. In: *Neurobiology of aging* 34.10 (2013), pp. 2239–2247.

[39] Mohamad Habes et al. “White matter hyperintensities and imaging patterns of brain ageing in the general population”. In: *Brain* 139.4 (2016), pp. 1164–1179.

[40] Jeff Duyn. “MR susceptibility imaging”. In: *Journal of magnetic resonance* 229 (2013), pp. 198–207.

[41] Karla L Miller et al. “Multimodal population brain imaging in the UK Biobank prospective epidemiological study”. In: *Nature neuroscience* 19.11 (2016), pp. 1523–1536.

[42] Stamatios N Sotiropoulos and Andrew Zalesky. “Building connectomes using diffusion MRI: why, how and but”. In: *NMR in Biomedicine* 32.4 (2019), e3752.

[43] Simon R Cox et al. “Ageing and brain white matter structure in 3,513 UK Biobank participants”. In: *Nature communications* 7.1 (2016), pp. 1–13.

[44] Sheng-Kwei Song et al. “Demyelination increases radial diffusivity in corpus callosum of mouse brain”. In: *Neuroimage* 26.1 (2005), pp. 132–140.

[45] Nikos K Logothetis. “What we can do and what we cannot do with fMRI”. In: *Nature* 453.7197 (2008), pp. 869–878.

[46] Bharat Biswal et al. “Functional connectivity in the motor cortex of resting human brain using echo-planar MRI”. In: *Magnetic resonance in medicine* 34.4 (1995), pp. 537–541.

[47] Jessica S Damoiseaux et al. "Consistent resting-state networks across healthy subjects". In: *Proceedings of the national academy of sciences* 103.37 (2006), pp. 13848–13853.

[48] Zarrar Shehzad et al. "The resting brain: unconstrained yet reliable". In: *Cerebral cortex* 19.10 (2009), pp. 2209–2229.

[49] Li Deng. "The mnist database of handwritten digit images for machine learning research". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.

[50] Nicola K Dinsdale, Mark Jenkinson, and Ana IL Namburete. "Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal". In: *NeuroImage* 228 (2021), p. 117689.

[51] Nicola K Dinsdale et al. "Challenges for machine learning in clinical translation of big data imaging studies". In: *arXiv preprint arXiv:2107.05630* (2021).

[52] Katja Franke et al. "Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters". In: *Neuroimage* 50.3 (2010), pp. 883–892.

[53] Esten H Leonardsen et al. "Deep neural networks learn general and clinically relevant representations of the ageing brain". In: *NeuroImage* (2022), p. 119210.

[54] Longfei Su, Lubin Wang, and Dewen Hu. "Predicting the age of healthy adults from structural MRI by sparse representation". In: *International Conference on Intelligent Science and Intelligent Data Engineering*. Springer. 2012, pp. 271–279.

[55] Ender Konukoglu et al. "Neighbourhood approximation using randomized forests". In: *Medical image analysis* 17.7 (2013), pp. 790–804.

[56] Stephen M Smith et al. "Brain aging comprises many modes of structural and functional change with distinct genetic and biophysical associations". In: *Elife* 9 (2020), e52677.

[57] Epifanio Bagarinao et al. "An unbiased data-driven age-related structural brain parcellation for the identification of intrinsic brain volume changes over the adult lifespan". In: *Neuroimage* 169 (2018), pp. 134–144.

[58] Deepthi P Varikuti et al. "Evaluation of non-negative matrix factorization of grey matter in age prediction". In: *Neuroimage* 173 (2018), pp. 394–410.

[59] Matthias Guggenmos et al. "Quantitative neurobiological evidence for accelerated brain aging in alcohol dependence". In: *Translational psychiatry* 7.12 (2017), pp. 1–7.

[60] Budhachandra S Khundrakpam et al. "Prediction of brain maturity based on cortical thickness at different spatial resolutions". In: *Neuroimage* 111 (2015), pp. 350–359.

[61] Maria de Fátima Machado Dias et al. "Deformation fields: a new source of information to predict brain age". In: *Journal of Neural Engineering* 19.3 (2022), p. 036025.

[62] MA Ganaie, M Tanveer, and Iman Beheshti. "Brain age prediction with improved least squares twin SVR". In: *IEEE Journal of Biomedical and Health Informatics* (2022).

[63] SA Valizadeh et al. “Age prediction on the basis of brain anatomical measures”. In: *Human brain mapping* 38.2 (2017), pp. 997–1008.

[64] Xin Niu et al. “Improved prediction of brain age using multimodal neuroimaging data”. In: *Human brain mapping* 41.6 (2020), pp. 1626–1643.

[65] Oscar Pina et al. “Structural Networks for Brain Age Prediction”. In: *Medical Imaging with Deep Learning*. 2021.

[66] Tzu-Wei Huang et al. “Age estimation from brain MRI images using deep learning”. In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE. 2017, pp. 849–852.

[67] Han Peng et al. “Accurate brain age prediction with lightweight deep neural networks”. In: *Medical image analysis* 68 (2021), p. 101871.

[68] Sheng He, P Ellen Grant, and Yangming Ou. “Global-local transformer for brain age estimation”. In: *IEEE transactions on medical imaging* 41.1 (2021), pp. 213–224.

[69] Bing Wang and Tuan D Pham. “MRI-based age prediction using hidden Markov models”. In: *Journal of neuroscience methods* 199.1 (2011), pp. 140–145.

[70] Benjamin Gutierrez Becker et al. “Gaussian process uncertainty in age estimation as a measure of brain abnormality”. In: *NeuroImage* 175 (2018), pp. 246–258.

[71] Christopher R Madan and Elizabeth A Kensinger. “Predicting age from cortical structure across the lifespan”. In: *European Journal of Neuroscience* 47.5 (2018), pp. 399–416.

[72] James H Cole and Katja Franke. “Predicting age using neuroimaging: innovative brain ageing biomarkers”. In: *Trends in neurosciences* 40.12 (2017), pp. 681–690.

[73] Nicola K Dinsdale. “Optimising Convolutional Neural Networks for Large-Scale Neuroimaging Studies.” PhD thesis. University of Oxford, 2021.

[74] David A Wood et al. “Accurate brain-age models for routine clinical MRI examinations”. In: *NeuroImage* 249 (2022), p. 118871.

[75] Franziskus Liem et al. “Predicting brain-age from multimodal imaging data captures cognitive impairment”. In: *Neuroimage* 148 (2017), pp. 179–188.

[76] Hongming Li, Theodore D Satterthwaite, and Yong Fan. “Brain age prediction based on resting-state functional connectivity patterns using convolutional neural networks”. In: *2018 ieee 15th international symposium on biomedical imaging (isbi 2018)*. IEEE. 2018, pp. 101–104.

[77] Yunfei Gao and Albert No. “Age Estimation from fMRI Data Using Recurrent Neural Network”. In: *Applied Sciences* 12.2 (2022), p. 749.

[78] Timothy T Brown et al. “Neuroanatomical assessment of biological maturity”. In: *Current biology* 22.18 (2012), pp. 1693–1698.

[79] Adrian R Groves et al. “Benefits of multi-modal fusion analysis on a large-scale dataset: life-span patterns of inter-subject variability in cortical morphometry and white matter microstructure”. In: *Neuroimage* 63.1 (2012), pp. 365–380.

[80] Guray Erus et al. “Imaging patterns of brain development and their relationship to cognition”. In: *Cerebral cortex* 25.6 (2015), pp. 1676–1684.

[81] James H Cole. "Multimodality neuroimaging brain-age in UK biobank: relationship to biomedical, lifestyle, and cognitive factors". In: *Neurobiology of aging* 92 (2020), pp. 34–42.

[82] Pauline Mouches et al. "Multimodal biological brain age prediction using magnetic resonance imaging and angiography with the identification of predictive regions". In: *Human brain mapping* 43.8 (2022), pp. 2554–2566.

[83] Simon M Hofmann et al. "Towards the interpretability of deep learning models for multi-modal neuroimaging: Finding structural changes of the ageing brain". In: *NeuroImage* 261 (2022), p. 119504.

[84] Bingyu Ren et al. "Deep transfer learning of structural magnetic resonance imaging fused with blood parameters improves brain age prediction". In: *Human brain mapping* 43.5 (2022), pp. 1640–1656.

[85] Ryuichi Fujimoto et al. "Age estimation using effective brain local features from T1-weighted images". In: *2016 38th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2016, pp. 5941–5944.

[86] Fei Wang et al. "Residual attention network for image classification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 3156–3164.

[87] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034* (2013).

[88] Daniel Taylor et al. "Brain Structural Saliency Over The Ages". In: *arXiv preprint arXiv:2202.11690* (2022).

[89] Sebastian G Popescu et al. "A U-net model of local brain-age". In: *bioRxiv* (2021).

[90] N Sanford et al. "Sex differences in predictors and regional patterns of brain-age-gap estimates". In: *bioRxiv* (2022).

[91] Tobias Hepp et al. "Uncertainty estimation and explainability in deep learning-based age estimation of the human brain: Results from the German National Cohort MRI study". In: *Computerized Medical Imaging and Graphics* 92 (2021), p. 101967.

[92] Angela Lombardi et al. "Explainable deep learning for personalized age prediction with brain morphology". In: *Frontiers in neuroscience* (2021), p. 578.

[93] Nikolaos Koutsouleris et al. "Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders". In: *Schizophrenia bulletin* 40.5 (2014), pp. 1140–1153.

[94] Chang-Le Chen et al. "Detection of advanced brain aging in schizophrenia and its structural underpinning by using normative brain age metrics". In: *NeuroImage: Clinical* 34 (2022), p. 103003.

[95] Weiqi Man et al. "Brain age gap as a potential biomarker for schizophrenia: A multi-site structural MRI study". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2021, pp. 4060–4063.

[96] Pedro L Ballester et al. "Brain age in mood and psychotic disorders: a systematic review and meta-analysis". In: *Acta Psychiatrica Scandinavica* 145.1 (2022), pp. 42–55.

[97] Vanessa L Cropley et al. "Brain-predicted age associates with psychopathology dimensions in youths". In: *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 6.4 (2021), pp. 410–419.

[98] Hugo G Schnack et al. "Accelerated brain aging in schizophrenia: a longitudinal pattern recognition study". In: *American Journal of Psychiatry* 173.6 (2016), pp. 607–616.

[99] Igor Nenadić et al. "BrainAGE score indicates accelerated brain aging in schizophrenia, but not bipolar disorder". In: *Psychiatry Research: Neuroimaging* 266 (2017), pp. 86–89.

[100] Marian Kolenic et al. "Obesity, dyslipidemia and brain age in first-episode psychosis". In: *Journal of psychiatric research* 99 (2018), pp. 151–158.

[101] Tomas Hajek et al. "Brain age in early stages of bipolar disorders or schizophrenia". In: *Schizophrenia bulletin* 45.1 (2019), pp. 190–198.

[102] Tobias Kaufmann et al. "Common brain disorders are associated with heritable patterns of apparent aging of the brain". In: *Nature neuroscience* 22.10 (2019), pp. 1617–1623.

[103] Jaroslav Rokicki et al. "Multimodal imaging improves brain age prediction and reveals distinct abnormalities in patients with psychiatric and neurological disorders". In: *Human brain mapping* 42.6 (2021), pp. 1714–1726.

[104] Constantinos Constantinides et al. "Brain ageing in schizophrenia: evidence from 26 international cohorts via the ENIGMA Schizophrenia consortium". In: *medRxiv* (2022).

[105] Jiayuan Huang et al. "Multimodal Magnetic Resonance Imaging Reveals Aberrant Brain Age Trajectory During Youth in Schizophrenia Patients". In: *Frontiers in aging neuroscience* 14 (2022).

[106] Laura KM Han et al. "Brain aging in major depressive disorder: results from the ENIGMA major depressive disorder working group". In: *Molecular psychiatry* 26.9 (2021), pp. 5124–5139.

[107] Cherise Chin Fatt et al. "Accelerated Brain Aging in Major Depressive Disorder Predicts Poorer Treatment Outcomes with Sertraline: Findings From the EMBARC Study". In: *Biological Psychiatry* 91.9 (2022), S4.

[108] Yunsong Luo et al. "Accelerated functional brain aging in major depressive disorder: evidence from a large scale fMRI analysis of Chinese participants". In: *arXiv preprint arXiv:2205.04871* (2022).

[109] Vladislav Drobinin et al. "The Developmental Brain Age Is Associated With Adversity, Depression, and Functional Outcomes Among Adolescents". In: *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 7.4 (2022), pp. 406–414.

[110] Katja Franke and Christian Gaser. "Longitudinal changes in individual BrainAGE in healthy aging, mild cognitive impairment, and Alzheimer's disease". In: *GeroPsych* (2012).

[111] Christian Gaser et al. “BrainAGE in mild cognitive impaired patients: predicting the conversion to Alzheimer’s disease”. In: *PloS one* 8.6 (2013), e67346.

[112] Luise Christine Löwe et al. “The effect of the APOE genotype on individual BrainAGE in normal aging, mild cognitive impairment, and Alzheimer’s disease”. In: *PloS one* 11.7 (2016), e0157514.

[113] Harini Eavani et al. “Heterogeneity of structural and functional imaging patterns of advanced brain aging revealed via machine learning methods”. In: *Neurobiology of aging* 71 (2018), pp. 41–50.

[114] Katia Maria Poloni, Ricardo José Ferrari, Alzheimer’s Disease Neuroimaging Initiative, et al. “A deep ensemble hippocampal CNN model for brain age estimation applied to Alzheimer’s diagnosis”. In: *Expert Systems with Applications* 195 (2022), p. 116622.

[115] Peter R Millar et al. “Predicting brain age from functional connectivity in symptomatic and preclinical Alzheimer disease”. In: *NeuroImage* 256 (2022), p. 119228.

[116] Panteleimon Giannakopoulos et al. “Alzheimer resemblance atrophy index, BrainAGE, and normal pressure hydrocephalus score in the prediction of subtle cognitive decline: added value compared to existing MR imaging markers”. In: *European Radiology* (2022), pp. 1–10.

[117] Shalaila Haas et al. “P582. Local and Global Brain Ageing in Cognitive Subgroups of Early Psychosis”. In: *Biological Psychiatry* 91.9 (2022), S324–S325.

[118] Gundula Seidel et al. “Accelerated brain ageing in sepsis survivors with cognitive long-term impairment”. In: *European Journal of Neuroscience* 52.10 (2020), pp. 4395–4402.

[119] Einar A Høgestøl et al. “Cross-sectional and longitudinal MRI brain scans reveal accelerated brain aging in multiple sclerosis”. In: *Frontiers in neurology* 10 (2019), p. 450.

[120] Chang-Le Chen et al. “Validation of neuroimaging-based brain age gap as a mediator between modifiable risk factors and cognition”. In: *Neurobiology of Aging* 114 (2022), pp. 61–72.

[121] Junhong Yu et al. “Differences between multimodal brain-age and chronological-age are linked to telomere shortening”. In: *Neurobiology of Aging* 115 (2022), pp. 60–69.

[122] Johnny Wang et al. “Gray matter age prediction as a biomarker for risk of dementia”. In: *Proceedings of the National Academy of Sciences* 116.42 (2019), pp. 21213–21218.

[123] Maxwell L Elliott et al. “Brain-age in midlife is associated with accelerated biological aging and cognitive decline in a longitudinal birth cohort”. In: *Molecular psychiatry* 26.8 (2021), pp. 3829–3838.

[124] Mohammad SE Sendi, David H Salat, and Vince D Calhoun. “Brain age gap difference between healthy and mild dementia subjects: Functional network connectivity analysis”. In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2021, pp. 1636–1639.

[125] Yinan Zheng et al. "Mid-life epigenetic age, neuroimaging brain age, and cognitive function: coronary artery risk development in young adults (CARDIA) study". In: *Aging (Albany NY)* 14.4 (2022), p. 1691.

[126] James H Cole et al. "Prediction of brain age suggests accelerated atrophy after traumatic brain injury". In: *Annals of neurology* 77.4 (2015), pp. 571–581.

[127] Gershon Spitz et al. "Brain age in chronic traumatic brain injury". In: *NeuroImage: Clinical* (2022), p. 103039.

[128] Emily L Dennis et al. "Advanced brain age in deployment-related traumatic brain injury: A LIMBIC-CENC neuroimaging study". In: *Brain injury* (2022), pp. 1–11.

[129] Heath R Pardoe et al. "Structural brain changes in medically refractory focal epilepsy resemble premature brain aging". In: *Epilepsy research* 133 (2017), pp. 28–32.

[130] Daichi Sone et al. "Neuroimaging-based brain-age prediction in diverse forms of epilepsy: a signature of psychosis and beyond". In: *Molecular psychiatry* 26.3 (2021), pp. 825–834.

[131] Christophe E de Bézenac et al. "Association of Epilepsy Surgery With Changes in Imaging-Defined Brain Age". In: *Neurology* 97.6 (2021), e554–e563.

[132] Qi Wang et al. "Predicting brain age during typical and atypical development based on structural and functional neuroimaging". In: *Human brain mapping* 42.18 (2021), pp. 5943–5955.

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

[134] Jian Cheng et al. "Brain age estimation from MRI using cascade networks with ranking loss". In: *IEEE Transactions on Medical Imaging* 40.12 (2021), pp. 3400–3412.

[135] Trang T Le et al. "Effect of ibuprofen on BrainAGE: a randomized, placebo-controlled, dose-response exploratory study". In: *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 3.10 (2018), pp. 836–843.

[136] Holly Van Gestel et al. "Brain age in bipolar disorders: Effects of lithium treatment". In: *Australian & New Zealand Journal of Psychiatry* 53.12 (2019), pp. 1179–1188.

[137] Katja Franke et al. "Premature brain aging in humans exposed to maternal nutrient restriction during early gestation". In: *Neuroimage* 173 (2018), pp. 460–471.

[138] Ann-Marie G De Lange et al. "Multimodal brain-age prediction and cardiovascular risk: The Whitehall II MRI sub-study". In: *NeuroImage* 222 (2020), p. 117292.

[139] Katja Franke et al. "Gender-specific impact of personal health parameters on individual brain aging in cognitively unimpaired elderly subjects". In: *Frontiers in aging neuroscience* 6 (2014), p. 94.

[140] James H Cole et al. "Brain age predicts mortality". In: *Molecular psychiatry* 23.5 (2018), pp. 1385–1392.

[141] Katja Franke et al. “Changes of individual BrainAGE during the course of the menstrual cycle”. In: *Neuroimage* 115 (2015), pp. 1–6.

[142] Eileen Luders et al. “Potential brain age reversal after pregnancy: younger brains at 4–6 weeks postpartum”. In: *Neuroscience* 386 (2018), pp. 309–314.

[143] Katja Franke et al. “Advanced BrainAGE in older adults with type 2 diabetes mellitus”. In: *Frontiers in aging neuroscience* 5 (2013), p. 90.

[144] Jonathan C Ipser et al. “Limited evidence for a moderating effect of HIV status on brain age in heavy episodic drinkers”. In: *Journal of NeuroVirology* (2022), pp. 1–9.

[145] Alexanndra Angebrandt et al. “Dose-dependent relationship between social drinking and brain aging”. In: *Neurobiology of Aging* 111 (2022), pp. 71–81.

[146] Zeqiang Linli et al. “Associations between smoking and accelerated brain ageing”. In: *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 113 (2022), p. 110471.

[147] Nora Bittner et al. “When your brain looks older than expected: combined lifestyle risk and BrainAGE”. In: *Brain Structure and Function* 226.3 (2021), pp. 621–645.

[148] Katja Franke et al. “Brain maturation: predicting individual BrainAGE in children and adolescents using structural MRI”. In: *Neuroimage* 63.3 (2012), pp. 1305–1312.

[149] Geneviève Richard et al. “Brain age prediction in stroke patients: Highly reliable but limited sensitivity to cognitive performance and response to cognitive training”. In: *NeuroImage: Clinical* 25 (2020), p. 102159.

[150] Lisa Ronan et al. “Obesity associated with increased brain age from midlife”. In: *Neurobiology of aging* 47 (2016), pp. 63–70.

[151] Nicole M Ashpole et al. “Growth hormone, insulin-like growth factor-1 and the aging brain”. In: *Experimental gerontology* 68 (2015), pp. 76–81.

[152] Dennis M Hedderich et al. “Increased brain age gap estimate (brainage) in young adults after premature birth”. In: *Frontiers in aging neuroscience* (2021), p. 158.

[153] Christos Pitsavos et al. “Association between low-grade systemic inflammation and type 2 diabetes mellitus among men and women from the ATTICA study”. In: *The review of diabetic studies: RDS* 4.2 (2007), p. 98.

[154] Mohammed S Ellulu et al. “Obesity and inflammation: the linking mechanism and the complications”. In: *Archives of medical science: AMS* 13.4 (2017), p. 851.

[155] Federica Zatterale et al. “Chronic adipose tissue inflammation linking obesity to insulin resistance and type 2 diabetes”. In: *Frontiers in physiology* 10 (2020), p. 1607.

[156] Eileen Luders, Nicolas Cherbuin, and Christian Gaser. “Estimating brain age using high-resolution pattern recognition: Younger brains in long-term meditation practitioners”. In: *Neuroimage* 134 (2016), pp. 508–513.

[157] Lars Rogenmoser et al. “Keeping brains young with making music”. In: (2017).

[158] Jason Steffener et al. “Differences between chronological and brain age are related to education and self-reported physical activity”. In: *Neurobiology of aging* 40 (2016), pp. 138–144.

[159] Elisa Scheller et al. “Brain aging and APOE $\varepsilon 4$ interact to reveal potential neuronal compensation in healthy older adults”. In: *Frontiers in aging neuroscience* 10 (2018), p. 74.

[160] Benedikt Atli Jónsson et al. “Brain age prediction using deep learning uncovers associated sequence variants”. In: *Nature communications* 10.1 (2019), pp. 1–10.

[161] Yanhui Cai et al. “TREK-1 pathway mediates isoflurane-induced memory impairment in middle-aged mice”. In: *Neurobiology of learning and memory* 145 (2017), pp. 199–204.

[162] Hreinn Stefansson et al. “A common inversion under selection in Europeans”. In: *Nature genetics* 37.2 (2005), pp. 129–137.

[163] R Rademakers, M Cruts, and C Van Broeckhoven. “The role of tau (MAPT) in frontotemporal dementia and related tauopathies”. In: *Human mutation* 24.4 (2004), pp. 277–295.

[164] David A Koolen et al. “A new chromosome 17q21. 31 microdeletion syndrome associated with a common inversion polymorphism”. In: *Nature genetics* 38.9 (2006), pp. 999–1001.

[165] Mats Nagel et al. “Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways”. In: *Nature genetics* 50.7 (2018), pp. 920–927.

[166] Anil Kuchinad et al. “Accelerated brain gray matter loss in fibromyalgia patients: premature aging of the brain?” In: *Journal of Neuroscience* 27.15 (2007), pp. 4004–4007.

[167] Roberta Riccelli et al. “Surface-based morphometry reveals the neuroanatomical basis of the five-factor model of personality”. In: *Social cognitive and affective neuroscience* 12.4 (2017), pp. 671–684.

[168] GJ Hervieu et al. “Distribution and expression of TREK-1, a two-pore-domain potassium channel, in the adult rat CNS”. In: *Neuroscience* 103.4 (2001), pp. 899–919.

[169] Stefan Bittner et al. “TREK-king the blood–brain-barrier”. In: *Journal of Neuroimmune Pharmacology* 9.3 (2014), pp. 293–301.

[170] James Cole. “Steps Towards Clinical Application of the Brain Age Paradigm”. In: *Biological Psychiatry* 91.9 (2022), S3–S4.

[171] Lea Baecker et al. “Machine learning for brain age prediction: Introduction to methods and clinical applications”. In: *EBioMedicine* 72 (2021), p. 103600.

[172] Mark Jenkinson et al. “Fsl”. In: *Neuroimage* 62.2 (2012), pp. 782–790.

[173] Fidel Alfaro-Almagro et al. “Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank”. In: *Neuroimage* 166 (2018), pp. 400–424.

[174] Stephen M Smith. “Fast robust automated brain extraction”. In: *Human brain mapping* 17.3 (2002), pp. 143–155.

[175] John C Mazziotta et al. “A probabilistic atlas of the human brain: theory and rationale for its development”. In: *Neuroimage* 2.2 (1995), pp. 89–101.

[176] John Mazziotta et al. “A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM)”. In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 356.1412 (2001), pp. 1293–1322.

[177] John Mazziotta et al. “A four-dimensional probabilistic atlas of the human brain”. In: *Journal of the American Medical Informatics Association* 8.5 (2001), pp. 401–430.

[178] Mark Jenkinson and Stephen Smith. “A global optimisation method for robust affine registration of brain images”. In: *Medical image analysis* 5.2 (2001), pp. 143–156.

[179] Xiangrui Li et al. “The first step for neuroimaging data analysis: DICOM to NIfTI conversion”. In: *Journal of neuroscience methods* 264 (2016), pp. 47–56.

[180] John P Mugler III and James R Brookeman. “Three-dimensional magnetization-prepared rapid gradient-echo imaging (3D MP RAGE)”. In: *Magnetic resonance in medicine* 15.1 (1990), pp. 152–157.

[181] Gwenaëlle Douaud et al. “Anatomically related grey and white matter abnormalities in adolescent-onset schizophrenia”. In: *Brain* 130.9 (2007), pp. 2375–2386.

[182] John P Mugler III. “Optimized three-dimensional fast-spin-echo MRI”. In: *Journal of magnetic resonance imaging* 39.4 (2014), pp. 745–767.

[183] Ludovica Griffanti et al. “BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities”. In: *Neuroimage* 141 (2016), pp. 191–205.

[184] Michela Simoni et al. “Age-and sex-specific rates of leukoaraiosis in TIA and stroke patients: population-based study”. In: *Neurology* 79.12 (2012), pp. 1215–1222.

[185] Linxin Li et al. “Population-based case–control study of white matter changes on brain imaging in transient ischemic attack and ischemic stroke”. In: *Stroke* 44.11 (2013), pp. 3063–3070.

[186] Laura Clarke et al. “Magnetic resonance imaging in neuromyelitis optica spectrum disorder”. In: *Clinical & Experimental Immunology* 206.3 (2021), pp. 251–265.

[187] Ferdinand Schweser et al. “Differentiation between diamagnetic and paramagnetic cerebral lesions based on magnetic susceptibility mapping”. In: *Medical physics* 37.10 (2010), pp. 5165–5178.

[188] Edith V Sullivan et al. “Relevance of iron deposition in deep gray matter brain structures to cognitive and motor performance in healthy elderly men and women: exploratory findings”. In: *Brain imaging and behavior* 3.2 (2009), pp. 167–175.

[189] Bang J Guo, Zhen L Yang, and Long J Zhang. “Gadolinium deposition in brain: current scientific evidence and future perspectives”. In: *Frontiers in molecular neuroscience* 11 (2018), p. 335.

[190] E Mark Haacke et al. “Susceptibility weighted imaging (SWI)”. In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 52.3 (2004), pp. 612–618.

[191] Karen A Tong et al. “Susceptibility-weighted MR imaging: a review of clinical applications in children”. In: *American Journal of Neuroradiology* 29.1 (2008), pp. 9–17.

[192] Sergi Martinez-Ramirez, Steven M Greenberg, and Anand Viswanathan. “Cerebral microbleeds: overview and implications in cognitive impairment”. In: *Alzheimer’s research & therapy* 6.3 (2014), pp. 1–7.

[193] Christian Beaulieu. “The basis of anisotropic water diffusion in the nervous system–a technical review”. In: *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo* 15.7-8 (2002), pp. 435–455.

[194] Stamatios N Sotiropoulos et al. “Effects of image reconstruction on fiber orientation mapping from multichannel diffusion MRI: reducing the noise floor using SENSE”. In: *Magnetic resonance in medicine* 70.6 (2013), pp. 1682–1689.

[195] Stamatios N Sotiropoulos et al. “Advances in diffusion MRI acquisition and processing in the Human Connectome Project”. In: *Neuroimage* 80 (2013), pp. 125–143.

[196] Junqian Xu et al. “Evaluation of slice accelerations using multiband echo planar imaging at 3 T”. In: *Neuroimage* 83 (2013), pp. 991–1001.

[197] Peter J Basser and Derek K Jones. “Diffusion-tensor MRI: theory, experimental design and data analysis–a technical review”. In: *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo* 15.7-8 (2002), pp. 456–467.

[198] Hui Zhang et al. “NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain”. In: *Neuroimage* 61.4 (2012), pp. 1000–1016.

[199] Yaniv Assaf and Ofer Pasternak. “Diffusion tensor imaging (DTI)-based white matter mapping in brain research: a review”. In: *Journal of molecular neuroscience* 34.1 (2008), pp. 51–61.

[200] Daniel B Ennis and Gordon Kindlmann. “Orthogonal tensor invariants and the analysis of diffusion tensor magnetic resonance images”. In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 55.1 (2006), pp. 136–146.

[201] Gordon Kindlmann et al. “Diffusion tensor analysis with invariant gradients and rotation tangents”. In: *IEEE Transactions on Medical Imaging* 26.11 (2007), pp. 1483–1499.

[202] Weikang Gong, Christian F Beckmann, and Stephen M Smith. “Phenotype discovery from population brain imaging”. In: *Medical image analysis* 71 (2021), p. 102050.

[203] Koji Kamagata et al. “Diffusion magnetic resonance imaging-based biomarkers for neurodegenerative diseases”. In: *International Journal of Molecular Sciences* 22.10 (2021), p. 5216.

[204] Stephen M Smith et al. “Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data”. In: *Neuroimage* 31.4 (2006), pp. 1487–1505.

[205] Jesper LR Andersson, Mark Jenkinson, Stephen Smith, et al. "Non-linear registration, aka Spatial normalisation FMRIB technical report TR07JA2". In: *FMRIB Analysis Group of the University of Oxford* 2.1 (2007), e21.

[206] Marius De Groot et al. "Improving alignment in tract-based spatial statistics: evaluation and optimization of image registration". In: *Neuroimage* 76 (2013), pp. 400–411.

[207] Timothy EJ Behrens et al. "Probabilistic diffusion tractography with multiple fibre orientations: What can we gain?" In: *neuroimage* 34.1 (2007), pp. 144–155.

[208] Timothy EJ Behrens et al. "Characterization and propagation of uncertainty in diffusion-weighted MR imaging". In: *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 50.5 (2003), pp. 1077–1088.

[209] Saad Jbabdi et al. "Model-based analysis of multishell diffusion MR data for tractography: How to get over fitting problems". In: *Magnetic resonance in medicine* 68.6 (2012), pp. 1846–1855.

[210] Moisés Hernández et al. "Accelerating fibre orientation estimation from diffusion weighted magnetic resonance imaging using GPUs". In: *PloS one* 8.4 (2013), e61892.

[211] Charles Smart Roy and Charles S Sherrington. "On the regulation of the blood-supply of the brain". In: *The Journal of physiology* 11.1-2 (1890), p. 85.

[212] Scott A Huettel, Allen W Song, Gregory McCarthy, et al. *Functional magnetic resonance imaging*. Vol. 1. Sinauer Associates Sunderland, 2004.

[213] BT Thomas Yeo et al. "The organization of the human cerebral cortex estimated by intrinsic functional connectivity". In: *Journal of neurophysiology* 106.3 (2011), pp. 1125–1165.

[214] Christopher J Honey, Jean-Philippe Thivierge, and Olaf Sporns. "Can structure predict function in the human brain?" In: *Neuroimage* 52.3 (2010), pp. 766–776.

[215] Jessica S Damoiseaux and Michael D Greicius. "Greater than the sum of its parts: a review of studies combining structural connectivity and resting-state functional connectivity". In: *Brain Structure and Function* 213.6 (2009), pp. 525–533.

[216] Evan M Gordon et al. "Precision functional mapping of individual human brains". In: *Neuron* 95.4 (2017), pp. 791–807.

[217] Stephanie Noble, Dustin Scheinost, and R Todd Constable. "A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis". In: *Neuroimage* 203 (2019), p. 116157.

[218] Pierre Bellec et al. "Multi-level bootstrap analysis of stable clusters in resting-state fMRI". In: *Neuroimage* 51.3 (2010), pp. 1126–1139.

[219] Aapo Hyvarinen. "Fast and robust fixed-point algorithms for independent component analysis". In: *IEEE transactions on Neural Networks* 10.3 (1999), pp. 626–634.

[220] Christian F Beckmann and Stephen M Smith. "Probabilistic independent component analysis for functional magnetic resonance imaging". In: *IEEE transactions on medical imaging* 23.2 (2004), pp. 137–152.

[221] Gholamreza Salimi-Khorshidi et al. "Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers". In: *Neuroimage* 90 (2014), pp. 449–468.

[222] Christian F Beckmann et al. "Group comparison of resting-state FMRI data using multi-subject ICA and dual regression". In: *Neuroimage* 47.Suppl 1 (2009), S148.

[223] Lisa D Nickerson et al. "Using dual regression to investigate network shape and amplitude in functional connectivity analyses". In: *Frontiers in neuroscience* 11 (2017), p. 115.

[224] Soojin Lee et al. "Amplitudes of resting-state functional networks–investigation into their correlates and biophysical properties". In: *NeuroImage* 265 (2023), p. 119779.

[225] Ahmad R Hariri et al. "The amygdala response to emotional stimuli: a comparison of faces and scenes". In: *Neuroimage* 17.1 (2002), pp. 317–323.

[226] Deanna M Barch et al. "Function in the human connectome: task-fMRI and individual differences in behavior". In: *Neuroimage* 80 (2013), pp. 169–189.

[227] Mark W Woolrich et al. "Multilevel linear modelling for FMRI group analysis using Bayesian inference". In: *Neuroimage* 21.4 (2004), pp. 1732–1747.

[228] Fidel Alfaro-Almagro et al. "Confound modelling in UK Biobank brain imaging". In: *NeuroImage* 224 (2021), p. 117002.

[229] Lloyd T Elliott et al. "Genome-wide association studies of brain imaging phenotypes in UK Biobank". In: *Nature* 562.7726 (2018), pp. 210–216.

[230] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[231] Kunihiko Fukushima. "Cognitron: A self-organizing multilayered neural network". In: *Biological cybernetics* 20.3-4 (1975), pp. 121–136.

[232] Dan Hendrycks and Kevin Gimpel. "Gaussian error linear units (gelus)". In: *arXiv preprint arXiv:1606.08415* (2016).

[233] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[234] Adam Paszke et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Ed. by H. Wallach et al. 2019. URL: `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

[235] Shizuo Kaji and Satoshi Kida. "Overview of image-to-image translation by use of deep neural networks: denoising, super-resolution, modality conversion, and reconstruction in medical imaging". In: *Radiological physics and technology* 12.3 (2019), pp. 235–248.

[236] Kaiming He et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[237] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International conference on machine learning*. PMLR. 2015, pp. 448–456.

[238] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. "Layer normalization". In: *arXiv preprint arXiv:1607.06450* (2016).

[239] Shun-ichi Amari. "Backpropagation and stochastic gradient descent method". In: *Neurocomputing* 5.4-5 (1993), pp. 185–196.

[240] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[241] Tijmen Tieleman and Geoffrey Hinton. "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude". In: *COURSERA: Neural networks for machine learning* 4.2 (2012), pp. 26–31.

[242] Robin M Schmidt, Frank Schneider, and Philipp Hennig. "Descending through a crowded valley-benchmarking deep learning optimizers". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 9367–9376.

[243] Ilya Loshchilov and Frank Hutter. "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101* (2017).

[244] Ilya Loshchilov and Frank Hutter. "Sgdr: Stochastic gradient descent with warm restarts". In: *arXiv preprint arXiv:1608.03983* (2016).

[245] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[246] Ze Liu et al. "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[247] Ali Hatamizadeh et al. "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images". In: *International MICCAI Brainlesion Workshop*. Springer. 2021, pp. 272–284.

[248] Jun Li et al. "Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives". In: *Medical image analysis* (2023), p. 102762.

[249] Aman Arora. *Swin Transformer - Hierarchical Vision Transformer using Shifted Windows: Swin Transformer Model Architecture explained with PyTorch implementation line-by-line.* July 2022. URL: `https://amaarora.github.io/posts/2022-07-04-swintransformerv1.html`.

[250] Yu Fu et al. "OTFPF: Optimal Transport-Based Feature Pyramid Fusion Network for Brain Age Estimation with 3D Overlapped ConvNeXt". In: *arXiv preprint arXiv:2205.04684* (2022).

[251] Benson Mwangi, Khader M Hasan, and Jair C Soares. "Prediction of individual subject's age across the human lifespan using diffusion tensor imaging: a machine learning approach". In: *Neuroimage* 75 (2013), pp. 58–67.

[252] Sheng He et al. "Deep Relation Learning for Regression and Its Application to Brain Age Estimation". In: *IEEE Transactions on Medical Imaging* (2022).

[253] Kalen J Petersen et al. "Machine learning quantifies accelerated white-matter aging in persons with HIV". In: *The Journal of Infectious Diseases* (2022).

[254] Neil J Tolentino et al. "Alcohol effects on cerebral blood flow in subjects with low and high responses to alcohol". In: *Alcoholism: Clinical and Experimental Research* 35.6 (2011), pp. 1034–1040.

[255] Frederik J Lange et al. "MMORF—FSL's MultiMOdal Registration Framework". In: *Imaging Neuroscience* (2024).

[256] Ana IL Namburete, Weidi Xie, and J Alison Noble. "Robust regression of brain maturation from 3D fetal neurosonography using CRNs". In: *Fetal, Infant and Ophthalmic Medical Image Analysis*. Springer, 2017, pp. 73–80.

[257] Nicola K Dinsdale, Mark Jenkinson, and Ana IL Namburete. "STAMP: Simultaneous Training and Model Pruning for Low Data Regimes in Medical Image Segmentation". In: *bioRxiv* (2021).

[258] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. URL: `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

[259] Ann-Marie G de Lange and James H Cole. "Commentary: Correction procedures in brain-age prediction". In: *NeuroImage: Clinical* 26 (2020).

[260] Xin Niu, Hualou Liang, and Fengqing Zhang. "Brain age prediction for post-traumatic stress disorder patients with convolutional neural networks: A multi-modal neuroimaging study". In: *Conference on Cognitive Computational Neuroscience, Philadelphia, PA*. 2018, p. 1121.

[261] Trang T Le et al. "A nonlinear simulation framework supports adjusting for age when analyzing BrainAGE". In: *Frontiers in aging neuroscience* 10 (2018), p. 317.

[262] Hualou Liang, Fengqing Zhang, and Xin Niu. *Investigating systematic bias in brain age estimation with application to post-traumatic stress disorders*. Tech. rep. Wiley Online Library, 2019.

[263] J Martin Bland and Douglas G Altman. "Statistic Notes: Regression towards the mean". In: *Bmj* 308.6942 (1994), p. 1499.

[264] Stephen M Stigler. "Regression towards the mean, historically considered". In: *Statistical methods in medical research* 6.2 (1997), pp. 103–114.

[265] Ellyn R Butler et al. *Pitfalls in brain age analyses*. Tech. rep. Wiley Online Library, 2021.

[266] Ann-Marie G de Lange et al. "Population-based neuroimaging reveals traces of childbirth in the maternal brain". In: *Proceedings of the National Academy of Sciences* 116.44 (2019), pp. 22341–22346.

[267] John W Tukey. "Tightening the clinical trial". In: *Controlled clinical trials* 14.4 (1993), pp. 266–285.

[268] Theo A Knijnenburg et al. "Fewer permutations, more accurate P-values". In: *Bioinformatics* 25.12 (2009), pp. i161–i168.

[269] Emma AM Stanley et al. "A fully convolutional neural network for explainable classification of attention deficit hyperactivity disorder". In: *Medical Imaging 2022: Computer-Aided Diagnosis*. Vol. 12033. SPIE. 2022, pp. 296–301.

[270] Jing Du et al. “White matter brain age as a biomarker of cerebrovascular burden in the ageing brain”. In: *medRxiv* (2022).

[271] Yanwu Yang et al. “Regularizing Brain Age Prediction via Gated Knowledge Distillation”. In: *Medical Imaging with Deep Learning*. 2021.

[272] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272.

[273] Geoffrey E Hinton et al. “Improving neural networks by preventing co-adaptation of feature detectors”. In: *arXiv preprint arXiv:1207.0580* (2012).

[274] Søren Rasmussen. *Pitfalls with Dropout and BatchNorm in regression problems*. Ed. by Medium.com. [Online; posted 19-November-2020]. Nov. 2020. URL: `https://towardsdatascience.com/pitfalls-with-dropout-and-batchnorm-in-regression-problems-39e02ce08e4d`.

[275] Laura E Suárez et al. “Linking structure and function in macroscale brain networks”. In: *Trends in Cognitive Sciences* (2020).

[276] Ankit N Khambhati et al. “Modeling and interpreting mesoscale network dynamics”. In: *NeuroImage* 180 (2018), pp. 337–349.

[277] KV Mardia, JT Kent, and JM Bibby. “Multivariate analysis”. In: (1979).

[278] Richard A Reinhardt et al. “Influence of estrogen and osteopenia/osteoporosis on clinical periodontitis in postmenopausal women”. In: *Journal of periodontology* 70.8 (1999), pp. 823–828.

[279] Robert Marcus. “Post-menopausal osteoporosis”. In: *Best practice & research Clinical obstetrics & gynaecology* 16.3 (2002), pp. 309–327.

[280] Meng-Xia Ji and Qi Yu. “Primary osteoporosis in postmenopausal women”. In: *Chronic diseases and translational medicine* 1.01 (2015), pp. 9–13.

[281] David Karasik et al. “Disentangling the genetic determinants of human aging: biological age as an alternative to the use of survival measures”. In: *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 60.5 (2005), pp. 574–587.

[282] Howard S Friedman. “Cardiovascular effects of alcohol with particular reference to the heart”. In: *Alcohol* 1.4 (1984), pp. 333–339.

[283] Joaquim Fernandez-Sola. “Cardiovascular risks and benefits of moderate and heavy alcohol consumption”. In: *Nature Reviews Cardiology* 12.10 (2015), pp. 576–587.

[284] Arthur L Klatsky. “Alcohol, cardiovascular diseases and diabetes mellitus”. In: *Pharmacological research* 55.3 (2007), pp. 237–247.

[285] Jo Wrigglesworth et al. “Factors associated with brain ageing-a systematic review”. In: *BMC neurology* 21.1 (2021), pp. 1–23.

[286] Terry W Du Clos. “Function of C-reactive protein”. In: *Annals of medicine* 32.4 (2000), pp. 274–278.

[287] Joseph Berkson. “Limitations of the application of fourfold table analysis to hospital data”. In: *Biometrics Bulletin* 2.3 (1946), pp. 47–53.

[288] Gail M Sullivan and Richard Feinn. "Using effect size—or why the P value is not enough". In: *Journal of graduate medical education* 4.3 (2012), pp. 279–282.

[289] Mellar P Davis and Declan Walsh. "Treatment of nausea and vomiting in advanced cancer". In: *Supportive care in cancer* 8 (2000), pp. 444–452.

[290] Martin R Tramèr et al. "Cannabinoids for control of chemotherapy induced nausea and vomiting: quantitative systematic review". In: *Bmj* 323.7303 (2001), p. 16.

[291] Robert G Twycross. "Palliative care formulary". In: *(No Title)* (1998).

[292] Dylan G Harris. "Nausea and vomiting in advanced cancer". In: *British medical bulletin* 96.1 (2010), pp. 175–185.

[293] Martin Hitier, Stephane Besnard, and Paul F Smith. "Vestibular pathways involved in cognition". In: *Frontiers in integrative neuroscience* 8 (2014), p. 59.

[294] Xuehao Zhang et al. "Vestibular dysfunction is an important contributor to the aging of visuospatial ability in older adults–Data from a computerized test system". In: *Frontiers in Neurology* 13 (2022), p. 1049806.

[295] Anna Jenul et al. "Rent—repeated elastic net technique for feature selection". In: *IEEE Access* 9 (2021), pp. 152333–152346.

[296] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[297] Alexei Botchkarev. "Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology". In: *arXiv preprint arXiv:1809.03006* (2018).

[298] Andrea Cherubini et al. "Importance of multimodal MRI in characterizing brain tissue and its potential application for individual age prediction". In: *IEEE journal of biomedical and health informatics* 20.5 (2016), pp. 1232–1239.

[299] Geneviève Richard et al. "Assessing distinct patterns of cognitive aging using tissue-specific brain age prediction based on diffusion tensor imaging and brain morphometry". In: *PeerJ* 6 (2018), e5908.

[300] Dong Nie et al. "3D deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer. 2016, pp. 212–220.

[301] Dong Nie et al. "Multi-channel 3D deep feature learning for survival time prediction of brain tumor patients using multi-modal neuroimages". In: *Scientific reports* 9.1 (2019), p. 1103.

[302] Tien Duong Vu et al. "Multimodal learning using convolution neural network and sparse autoencoder". In: *2017 IEEE international conference on big data and smart computing (BigComp)*. IEEE. 2017, pp. 309–312.

[303] Fan Zhang et al. "Multi-modal deep learning model for auxiliary diagnosis of Alzheimer's disease". In: *Neurocomputing* 361 (2019), pp. 185–195.

[304] Janani Venugopalan et al. "Multimodal deep learning models for early detection of Alzheimer's disease stage". In: *Scientific reports* 11.1 (2021), p. 3254.

[305] Zhe Guo et al. “Deep learning-based image segmentation on multimodal medical imaging”. In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 3.2 (2019), pp. 162–169.

[306] Nachwa Aboubakr, Mihaela Popova, and James L Crowley. “Color-based fusion of mri modalities for brain tumor segmentation”. In: *Proceedings of 2021 International Conference on Medical Imaging and Computer-Aided Diagnosis (MICAD 2021) Medical Imaging and Computer-Aided Diagnosis*. Springer. 2022, pp. 89–97.

[307] Muhammad Adeel Azam et al. “A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics”. In: *Computers in biology and medicine* 144 (2022), p. 105253.

[308] Zhe Guo et al. “Medical image segmentation based on multi-modal convolutional neural network: Study on image fusion schemes”. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 903–907.

[309] Tongxue Zhou, Su Ruan, and Stéphane Canu. “A review: Deep learning for medical image segmentation using multi-modality fusion”. In: *Array* 3 (2019), p. 100004.

[310] Said Yacine Boulahia et al. “Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition”. In: *Machine Vision and Applications* 32.6 (2021), p. 121.

[311] Guotai Wang et al. “Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks”. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3*. Springer. 2018, pp. 178–190.

[312] Chenhong Zhou et al. “One-pass multi-task convolutional neural networks for efficient brain tumor segmentation”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11*. Springer. 2018, pp. 637–645.

[313] Giovanna Maria Dimitri et al. “Multimodal and multicontrast image fusion via deep generative models”. In: *Information Fusion* 88 (2022), pp. 146–160.

[314] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. “Multimodal deep learning for biomedical data fusion: a review”. In: *Briefings in Bioinformatics* 23.2 (2022), bbab569.

[315] Lele Chen et al. “MRI tumor segmentation with densely connected 3D CNN”. In: *Medical Imaging 2018: Image Processing*. Vol. 10574. SPIE. 2018, pp. 357–364.

[316] Jose Dolz et al. “HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation”. In: *IEEE transactions on medical imaging* 38.5 (2018), pp. 1116–1126.

[317] Dong Nie et al. “Fully convolutional networks for multi-modality isointense infant brain image segmentation”. In: *2016 IEEE 13Th international symposium on biomedical imaging (ISBI)*. IEEE. 2016, pp. 1342–1345.

[318] Konstantinos Kamnitsas et al. “Ensembles of multiple models and architectures for robust brain tumour segmentation”. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3*. Springer. 2018, pp. 450–462.

[319] Mehmet Aygün, Yusuf Hüseyin Şahin, and Gözde Ünal. “Multi modal convolutional neural networks for brain tumor segmentation”. In: *arXiv preprint arXiv:1809.06191* (2018).

[320] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. “A survey of transfer learning”. In: *Journal of Big data* 3.1 (2016), pp. 1–40.

[321] Lea Baecker et al. “Brain age prediction: A comparison between machine learning models using region-and voxel-based morphometric data”. In: *Human brain mapping* 42.8 (2021), pp. 2332–2346.

[322] Xia Liu et al. “Brain age estimation using multi-feature-based networks”. In: *Computers in Biology and Medicine* 143 (2022), p. 105285.

[323] Gemma C Monté-Rubio et al. “A comparison of various MRI feature types for characterizing whole brain anatomical differences using linear pattern recognition methods”. In: *NeuroImage* 178 (2018), pp. 753–768.

[324] Sebastian G Popescu et al. “Deep Learning Methods for Estimating" Brain Age" from Structural MRI Scans”. In: (2018).

[325] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[326] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.

[327] Dan C Cireşan et al. “Mitosis detection in breast cancer histology images with deep neural networks”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16*. Springer. 2013, pp. 411–418.

[328] Christopher D Malon and Eric Cosatto. “Classification of mitotic figures with convolutional neural networks and seeded blob features”. In: *Journal of pathology informatics* 4.1 (2013), p. 9.

[329] Abhijit Guha Roy et al. “QuickNAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy”. In: *NeuroImage* 186 (2019), pp. 713–727.

[330] Fabian Isensee et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature methods* 18.2 (2021), pp. 203–211.

[331] Jianan Cui et al. “PET image denoising using unsupervised deep learning”. In: *European journal of nuclear medicine and molecular imaging* 46 (2019), pp. 2780–2789.

[332] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[333] Hong-Yu Zhou et al. “Convnets vs. transformers: Whose visual representations are more transferable?” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 2230–2238.

[334] Wenjie Luo et al. “Understanding the effective receptive field in deep convolutional neural networks”. In: *Advances in neural information processing systems* 29 (2016).

[335] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[336] Xinyang Feng et al. “Estimating brain age based on a uniform healthy population with deep learning and structural magnetic resonance imaging”. In: *Neurobiology of aging* 91 (2020), pp. 15–25.

[337] Gidon Levakov et al. “From a deep learning model back to the brain—Identifying regional predictors and their relation to aging”. In: *Human brain mapping* 41.12 (2020), pp. 3235–3252.

[338] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[339] Risheng Wang et al. “Medical image segmentation using deep learning: A survey”. In: *IET Image Processing* 16.5 (2022), pp. 1243–1267.

[340] Muhammad Muzammal Naseer et al. “Intriguing properties of vision transformers”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 23296–23308.

[341] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate”. In: *arXiv preprint arXiv:1409.0473* (2014).

[342] Kelvin Xu et al. “Show, attend and tell: Neural image caption generation with visual attention”. In: *International conference on machine learning*. PMLR. 2015, pp. 2048–2057.

[343] Elena Voita et al. “Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned”. In: *arXiv preprint arXiv:1905.09418* (2019).

[344] Kai Han et al. “Transformer in transformer”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 15908–15919.

[345] Chaitanya Joshi. “Transformers are graph neural networks”. In: *The Gradient* 7 (2020).

[346] Yanghao Li et al. “Exploring plain vision transformer backbones for object detection”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 280–296.

[347] Srinadh Bhojanapalli et al. "Understanding robustness of transformers for image classification". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10231–10241.

[348] Samira Abnar and Willem Zuidema. "Quantifying attention flow in transformers". In: *arXiv preprint arXiv:2005.00928* (2020).

[349] Stéphane d'Ascoli et al. "Convit: Improving vision transformers with soft convolutional inductive biases". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 2286–2296.

[350] Haiping Wu et al. "Cvt: Introducing convolutions to vision transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 22–31.

[351] Liang Huang et al. "Fine-grained ship classification by combining CNN and swin transformer". In: *Remote Sensing* 14.13 (2022), p. 3087.

[352] Hong-Yu Zhou et al. "nnformer: Interleaved transformer for volumetric segmentation". In: *arXiv preprint arXiv:2109.03201* (2021).

[353] Yutong Xie et al. "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*. Springer. 2021, pp. 171–180.

[354] Ali Hatamizadeh et al. "Unetr: Transformers for 3d medical image segmentation". In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2022, pp. 574–584.

[355] Davood Karimi, Haoran Dou, and Ali Gholipour. "Medical image segmentation using transformer networks". In: *IEEE Access* 10 (2022), pp. 29322–29332.

[356] M Jorge Cardoso et al. "Monai: An open-source framework for deep learning in healthcare". In: *arXiv preprint arXiv:2211.02701* (2022).

[357] Namuk Park and Songkuk Kim. "How do vision transformers work?" In: *arXiv preprint arXiv:2202.06709* (2022).

[358] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. "Learning and generalization in overparameterized neural networks, going beyond two layers". In: *Advances in neural information processing systems* 32 (2019).

[359] Hamed Hassani and Adel Javanmard. "The curse of overparametrization in adversarial training: Precise analysis of robust generalization for random features regression". In: *arXiv preprint arXiv:2201.05149* (2022).

[360] Sheng Liu et al. "Robust training under label noise by over-parameterization". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 14153–14172.

[361] Chase R Figley et al. "Potential pitfalls of using fractional anisotropy, axial diffusivity, and radial diffusivity as biomarkers of cerebral white matter microstructure". In: *Frontiers in Neuroscience* 15 (2022), p. 799576.

[362] Michael A DeTure and Dennis W Dickson. "The neuropathological diagnosis of Alzheimer's disease". In: *Molecular neurodegeneration* 14.1 (2019), pp. 1–18.

[363] Fei Yang et al. "Alzheimer's disease and epilepsy: An increasingly recognized comorbidity". In: *Frontiers in Aging Neuroscience* 14 (2022), p. 940515.

[364] Yaojing Chen, Mingxi Dang, and Zhanjun Zhang. "Brain mechanisms underlying neuropsychiatric symptoms in Alzheimer's disease: a systematic review of symptom-general and–specific lesion patterns". In: *Molecular Neurodegeneration* 16.1 (2021), p. 38.

[365] Ran Long et al. "Abnormalities of cerebral white matter microstructure in children with new-onset, untreated idiopathic-generalized epilepsy". In: *Frontiers in Neurology* 12 (2021), p. 744723.

[366] L Bonilha et al. "Altered microstructure in temporal lobe epilepsy: a diffusional kurtosis imaging study". In: *American Journal of Neuroradiology* 36.4 (2015), pp. 719–724.

[367] Sumati Bavisetty et al. "Chronic hypopituitarism after traumatic brain injury: risk assessment and relationship to outcome". In: *Neurosurgery* 62.5 (2008), pp. 1080–1094.

[368] Fatih Tanriverdi et al. "Brief communication: pituitary volume and function in competing and retired male boxers". In: *Annals of internal medicine* 148.11 (2008), pp. 827–831.

[369] Jeffrey C Ives, Mark Alderman, and Susan E Stred. "Hypopituitarism after multiple concussions: a retrospective case study in an adolescent male". In: *Journal of athletic training* 42.3 (2007), p. 431.

[370] Jurate Aleknaviciute et al. "Long-term association of pregnancy and maternal brain structure: the Rotterdam Study". In: *European Journal of Epidemiology* 37.3 (2022), pp. 271–281.

[371] Klara Mareckova et al. "Association of Maternal Depression During Pregnancy and Recent Stress With Brain Age Among Adult Offspring". In: *JAMA Network Open* 6.1 (2023), e2254581–e2254581.

[372] Fernando Pérez-Garcıa, Rachel Sparks, and Sébastien Ourselin. "TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning". In: *Computer Methods and Programs in Biomedicine* 208 (2021), p. 106236.

[373] Koen Van Leemput et al. "Automated model-based tissue classification of MR images of the brain". In: *IEEE transactions on medical imaging* 18.10 (1999), pp. 897–908.

[374] Carole H Sudre et al. "Longitudinal segmentation of age-related white matter hyperintensities". In: *Medical Image Analysis* 38 (2017), pp. 50–64.

[375] Alan C Evans et al. "3D statistical neuroanatomical models from 305 MRI volumes". In: *1993 IEEE conference record nuclear science symposium and medical imaging conference*. IEEE. 1993, pp. 1813–1817.

[376] Haohan Wang et al. "High-frequency component helps explain the generalization of convolutional neural networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 8684–8694.

[377] Jehan Ghafuri, Hongbo Du, and Sabah Jassim. "Sensitivity and stability of pretrained CNN filters". In: *Multimodal Image Exploitation and Learning 2021*. Vol. 11734. SPIE. 2021, pp. 79–89.

[378] Maithra Raghu et al. "Do vision transformers see like convolutional neural networks?" In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 12116–12128.

[379] Yutong Bai et al. "Are transformers more robust than cnns?" In: *Advances in neural information processing systems* 34 (2021), pp. 26831–26843.

[380] Alhassan Mumuni and Fuseini Mumuni. "CNN architectures for geometric transformation-invariant feature representation in computer vision: a review". In: *SN Computer Science* 2 (2021), pp. 1–23.

[381] Lin Xiang et al. "Computation of cnn's sensitivity to input perturbation". In: *Neural Processing Letters* 53 (2021), pp. 535–560.

[382] Paul M Thompson et al. "ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries". In: *Translational psychiatry* 10.1 (2020), pp. 1–28.

[383] Didac Vidal-Pineiro et al. "Individual variations in 'brain age'relate to early-life factors more than to longitudinal brain change". In: *Elife* 10 (2021), e69995.

[384] Jieqiong Wang et al. "Age estimation using cortical surface pattern combining thickness with curvatures". In: *Medical & biological engineering & computing* 52.4 (2014), pp. 331–341.

[385] Chihiro Kondo et al. "An age estimation method using brain local features for T1-weighted images". In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2015, pp. 666–669.

[386] Jenessa Lancaster et al. "Bayesian optimization for neuroimaging pre-processing in brain age classification and prediction". In: *Frontiers in aging neuroscience* 10 (2018), p. 28.

[387] Habtamu M Aycheh et al. "Biological brain age prediction using cortical thickness data: a large scale cohort study". In: *Frontiers in aging neuroscience* 10 (2018), p. 252.

[388] Camilo Bermudez et al. "Anatomical context improves deep learning on the brain age estimation task". In: *Magnetic Resonance Imaging* 62 (2019), pp. 70–77.

[389] Huiting Jiang et al. "Predicting brain age of healthy adults based on structural MRI parcellation using convolutional neural networks". In: *Frontiers in neurology* 10 (2020), p. 1346.

[390] Jin Hong et al. "Brain age prediction of children using routine brain MR images via deep learning". In: *Frontiers in Neurology* 11 (2020), p. 584682.

[391] Guangxiang Rao et al. "A High-Powered Brain Age Prediction Model Based on Convolutional Neural Network". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 1915–1919.

[392] Melis Anatürk et al. "Prediction of brain age and cognitive age: Quantifying brain and cognitive maintenance in aging". In: *Human brain mapping* 42.6 (2021), pp. 1626–1640.

[393] Inpyeong Hwang et al. “Prediction of brain age from routine T2-weighted spin-echo brain magnetic resonance images with a deep convolutional neural network”. In: *Neurobiology of Aging* 105 (2021), pp. 78–85.

[394] Guozhen Hu et al. “Accurate Brain Age Prediction Model for Healthy Children and Adolescents using 3D-CNN and Dimensional Attention”. In: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2021, pp. 800–806.

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

[396] Sergio Leonardo Mendes et al. “Estimating Gender and Age from Brain Structural MRI of Children and Adolescents: A 3D Convolutional Neural Network Multitask Learning Model”. In: *Computational intelligence and neuroscience* 2021 (2021).

[397] Cheol E Han et al. “Predicting age across human lifespan based on structural connectivity from diffusion tensor imaging”. In: *2014 IEEE Biomedical Circuits and Systems Conference (BioCAS) Proceedings*. IEEE. 2014, pp. 137–140.

[398] Lan Lin et al. “Predicting healthy older adult’s brain age based on structural connectivity networks using artificial neural networks”. In: *Computer methods and programs in biomedicine* 125 (2016), pp. 8–17.

[399] Yingying Guo et al. “A comparison between diffusion tensor imaging and generalized q-sampling imaging in the age prediction of healthy adults via machine learning approaches”. In: *Journal of Neural Engineering* 19.1 (2022), p. 016013.

[400] Lubin Wang et al. “Decoding lifespan changes of the human brain using resting-state functional connectivity MRI”. In: (2012).