Imperial College London



Deep Learning for Interpretable Brain Age Estimation

Author:

Kyriaki Margarita Bintsi

Supervisors:

Prof. Daniel Rueckert

Prof. Alexander Hammers

A thesis submitted in fulfilment of the requirements

for the degree of Doctor of Philosophy in the

Biomedical Image Analysis Group, Department of Computing,

Imperial College London

April 2024

Biomedical Image Analysis Group,
Department of Computing,
Huxley Building,
Imperial College London,
180 Queens Gate,
South Kensington,
SW7 2AZ

Kyriaki Margarita Bintsi © 2024

In the journey of understanding the brain,
every image is a poem, every neuron a verse.

Dedicated to my family.

Acknowledgements

Firstly, I would like to express my heartfelt gratitude to my supervisor, Prof. Daniel Rueckert, for his constant guidance and support. Despite his busy schedule, he was always available and responsive when I needed him, teaching me what a great manager looks like, with his selflessness and big heart. He has been an inspiration and a role model, and I am immensely grateful for the privilege of working under his guidance.

I also extend my sincere thanks to my second supervisor, Prof. Alexander Hammers, for his kindness, support, and invaluable clinical insights, which were crucial for the completion of this thesis.

I was lucky to be part of a CDT, which provided an additional layer of support throughout my PhD journey. I never felt alone thanks to all the amazing community within the CDT. A special mention to Alina and Melissa, who were my buddies in this journey from day one, my go-to for sharing the challenges we faced. I would also like to thank Valeria, our CDT manager, for her invaluable assistance over the years.

Being part of the BioMedIA lab was a blessing, allowing me to meet and collaborate with remarkable people. I would like to thank each and every one that passed from BioMedIA all this years. A special mention to Harvey, for the "question of the day" that helped me immensely move forward when stuck; Gavin, for all the fun, laughs, and making the lab as enjoyable as it can be; Carolina, the person that can relate with me and understand me the most, and Michael, my conference/retreat travel buddy in the post-pandemic years. I also want to acknowledge Prof. Ben Glocker, and Prof. Bernhard Kainz, for their help and support.

Imperial College has been an exceptional place to study, filled with amazing people and huge support. A special thanks to Amani, for being the best point of contact possible, and for

providing help the moment I needed it. Also, to Rolandos, who was there every time I needed research chats and immensely helped me when needed.

I had the luck to collaborate with people from the AIM lab at TUM. Special mention to Tamara and Sophie, who shared similar challenges. Our meetings and chats helped all of us move forward, together.

The internships I did during my PhD, and the people I met, helped me grow both personally and professionally as a researcher. Thanks to Jannes and Pedro, as well as Anna, Benito, Andres, Robert, and Mira, with whom I built strong friendships.

I am also grateful to my friends outside the university, both in London, who made it a wonderful place to live, and in Greece. Special thanks to Olga and Lemonia, my ever-present pillars, and Petrina, who prepared me for my years abroad.

A special thanks to Vassilis, whose support and belief in me were stronger than my own at times.

Finally, my deepest appreciation goes to my family, for their endless love and support. To my little niece, Alexandra, whom I hope to inspire to chase her dreams. To my parents and my sister, who have been my constant source of encouragement and strength.

Declaration

I, Kyriaki Margarita Bintsi, hereby declare this work original unless otherwise stated.

The work detailed in this thesis was funded by the King's College London & Imperial College London EPSRC Centre for Doctoral Training in Medical Imaging (EP/L015226/1).

Copyright

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

Brain ageing, and more specifically the difference between the chronological and the biological age of a person, may be a promising biomarker for identifying neurodegenerative diseases, such as Alzheimer's Disease (AD). In healthcare, predictive tools must not only be accurate but also easily understood and trusted by doctors. In this thesis, we leverage deep learning for the estimation of brain age from Magnetic Resonance (MR) images, while also enhancing the interpretability of the employed architectures.

Firstly, we explore how Convolutional Neural Networks (CNNs) perform on the task of brain age regression. We leverage 3D patches of the brain to develop a localised brain age estimator.

Subsequently, we delve deeper into interpretability, constructing a pipeline that generates importance maps for brain age regression using noise masks. These maps discover the parts of the brain that are the most important for brain age.

A way to seamlessly integrate multi-modal information and capture the relationships among individuals is using population graphs along with Graph Neural Networks (GNNs). Here, we highlight the importance of a meaningful graph construction, which is a non-trivial task, and experiment with different static population-graph construction methods and their effect on GNN performance for brain age estimation. The findings indicate that static graph construction approaches are potentially insufficient for the task of brain age estimation.

Thus, we explore adaptive graph learning and we propose framework that learns a population graph structure optimised for the downstream task. An attention mechanism assigns weights to a set of imaging and non-imaging features (phenotypes), which are then used for edge extraction. Additionally, by visualising the attention weights that were the most important for the graph construction, we increase the interpretability of the graph.

Overall, this dissertation focuses on the application of deep learning techniques, particularly CNNs and GNNs, for brain age estimation, while concurrently increasing interpretability.

Contents

Acknowledgements **8**

Declaration **10**

Copyright **12**

Abstract **14**

List of Figures **24**

List of Tables **27**

List of Abbreviations **29**

1 Introduction **30**

1.1 Motivation 30

1.2 Thesis Objectives and Contributions 32

1.3 Thesis Outline 33

1.4 Publications 34

1.5 Statement of Originality 37

2 Clinical Background **38**

2.1 Alzheimer's Disease 38

2.1.1 Introduction to Alzheimer's Disease 38

2.1.2 Pathophysiology of Alzheimer's Disease 40

2.1.3 Clinical Presentation . . . 42
2.1.4 Diagnostic Approaches . . . 44
2.2 Brain Age Estimation as a biomarker for Alzheimer's Disease . . . 45
2.3 Brain MR Datasets . . . 47
2.4 Summary . . . 49

3 Technical Background and Related work **50**
3.1 Machine Learning . . . 50
3.1.1 Artificial Neural Networks (ANNs) . . . 51
3.1.2 Optimisation . . . 53
3.2 Convolutional Neural Networks (CNNs) . . . 56
3.3 Graph Neural Networks (GNNs) . . . 64
3.4 Related Works on Brain Age Estimation . . . 68
3.4.1 Pre-Machine Learning Approaches for Brain Age Estimation . . . 68
3.4.2 Traditional Machine Learning for Brain Age Estimation . . . 70
3.4.3 CNNs for Brain Age Estimation . . . 71
3.4.4 GNNs for medical tasks . . . 72
3.5 Bias Correction . . . 73
3.6 Interpretability for Deep Learning . . . 74
3.7 Summary . . . 76

4 Patch-based Brain Age Estimation from MR Images **78**
4.1 Introduction . . . 79
4.2 Materials & Methods . . . 80
4.2.1 Dataset & Preprocessing . . . 80
4.2.2 Convolutional Autoencoder . . . 81
4.2.3 ResNet18 Network . . . 83
4.3 Results . . . 86
4.3.1 Metrics . . . 86
4.3.2 Implementation details . . . 87
4.3.3 Experiments . . . 87

4.4 Discussion . . . 94
4.5 Summary . . . 98

5 Voxel-level Importance Maps for Interpretable Brain Age Estimation 100
5.1 Introduction . . . 101
5.2 Materials & Methods . . . 104
5.2.1 Dataset & Preprocessing . . . 104
5.2.2 Brain Age Estimation . . . 104
5.2.3 Localisation . . . 104
5.3 Results . . . 106
5.3.1 Age Estimation . . . 107
5.3.2 Population-based importance maps . . . 107
5.3.3 Ablation tests . . . 108
5.3.4 Comparison with other interpretability methods . . . 109
5.4 Discussion . . . 112
5.5 Summary . . . 113

6 A comparative study for population-based brain age estimation using Graph Neural Networks 116
6.1 Introduction . . . 117
6.2 Methods . . . 119
6.2.1 No edges . . . 119
6.2.2 Random Graph . . . 119
6.2.3 Clinical similarity score (Stankevičiūtė et al.) . . . 119
6.2.4 Similarity score (Parisot et al.) . . . 120
6.2.5 kNN graph . . . 121
6.3 Experiments . . . 121
6.3.1 Dataset . . . 121
6.3.2 Implementation Details . . . 122
6.3.3 Results . . . 122
6.3.4 Visualisations . . . 123

6.3.5 Phenotypes used for the population graph 124
6.4 Discussion and Summary 126

7 Multimodal brain age estimation using interpretable adaptive population-graph learning 132
7.1 Introduction 133
7.2 Methods 135
7.3 Experiments 137
7.3.1 Dataset 137
7.3.2 Baselines 138
7.3.3 Implementation Details 138
7.4 Results 138
7.4.1 Regression 138
7.4.2 Classification 139
7.4.3 Ablation studies 140
7.5 Interpretability 143
7.6 Summary 144

8 Conclusions and Future Directions 146
8.1 Summary 146
8.2 Limitations and Future Directions 148
8.3 Conclusion 150

Bibliography 152

List of Figures

2.1 Main parts of the human brain. (Image from Johns Hopkins Medicine) 40

2.2 Human brain of a healthy subject (left) and a patient of Alzheimer's disease (right). Alterations in the cerebral cortex, the hippocampus, and the ventricles can be observed. (by BrightFocus Foundation) 41

2.3 Left: Changes in the brain depicted in T1-weight MR images of an Alzheimer's patient that was asymptomatic at first. Symptoms appeared between scans 4 and 5 (arrow in right picture). Right: Plots of brain volume loss per scan in the case of (A) Alzheimer's patient, (B) normal ageing, and (C) very healthy subjects (average age of 70 years old) [78] 43

3.1 A biological neuron (a) and the perceptron (b) [109]. 51

3.2 Example of an Multi-Layer Perceptron (MLP) with an input layer, two hidden layers, and an output layer. 53

3.3 Visualisation of a convolutional layer's input and output. 58

3.4 The architecture of LeNet-5 [106]. 60

3.5 The GoogleNet architecture [178]. 61

3.6 The architecture of a residual block containing a skip connection as proposed in [67]. 62

3.7 Example CNN architectures as presented in [67]. Left: VGGNet model [170] (as reference). Middle: a simple model with 34 layers. Right: The architecture of Residual Network (ResNet)-34. 63

3.8 The architecture of U-Net as introduced in [156]. 64

3.9 Message-passing update with a GNN on a graph with five nodes. Each node has a node feature vector x_i (left), and an updated node feature vector x'_i after the aggregation from the neighbouring node representations. 65

3.10 Examples of graph with high homophily (left), and low homophily (right). The three different colours of the nodes represent three different classes. 68

4.1 Age distribution of the study population. 81

4.2 The architecture of the network. The encoder part, consisted of hidden layers, results to the bottleneck layer (latent code layer). The latent code is used for the reconstruction, by using a decoder, as well an age prediction task achieved by an extra branch. An optional sex classification task was added for some experiments as well. 82

4.3 Overview of the proposed method. N 3D patches are being extracted from the center-cropped image. Every patch is used as input in a ResNet model that is trained to predict the subject's brain age, hence N different age predictions y_i are produced. The individual predictions are then combined using an ensemble method in order to achieve an improved estimation y. 84

4.4 Average of saliency maps of the subjects that belong in the test set in the axial (a), coronal (b), and sagittal (c), planes for the baseline model. 89

4.5 (a) Uncorrected brain age delta vs Chronological age for the baseline model, (b) Corrected brain age delta vs Chronological age for the baseline model, (c) The plot of the predicted over the real ages of the subjects before bias correction on the test set for the baseline model, (d) The plot of the predicted over the real ages of the subjects after bias correction on the test set for the baseline model. 90

4.6 Slices of a brain MR image with three coloured patches in each case. The colours of the patches indicate the performance of this single-patch model in terms of Mean Absolute Error (MAE). Specifically: red patches indicate a MAE of less than 3 years, orange patches between 3 and 3.5 years, green patches between 3.5 and 4 years, and blue patches greater than 4 years. The patches that perform best include the ventricles (a), and the hippocampus (b). 91

4.7 The predicted over the real age of the subjects when the predicted ages result from a linear regression of all the patches before bias correction (a), and after bias correction (b). The black dashed line shows the line of identity ($x = y$), while the orange solid one corresponds to the fitted model. 92
4.8 Slice of a brain MR 64x64x64 patch with the 32x32x32 patch that performed the best coloured orange. The patch that performed the best includes part of the ventricles and reached a MAE of 3.84 years. 92
4.9 The predicted over the real age of the subjects when the predicted ages result from the sum of the predictions of the different patches extracted for every subject before bias correction (a), and after bias correction (b). The black dashed line shows the line of identity ($x = y$), while the orange solid one corresponds to the fitted model. 93
4.10 Average of saliency maps of the subjects that belong in the test set in the axial (a), coronal (b), and sagittal (c), planes for the model that was trained on random patches. 94
4.11 Average of the MAE of the predictions of the different 64x64x64 patches for the case of every pixel of the test set. Every fifth slice is presented in the axial (a), coronal (b), and sagittal (c), views. 95
4.12 Average of the MAE of the predictions of the different 32x32x32 patches for the case of every pixel of the test set. Every fifth slice is presented in the axial (a), coronal (b), and sagittal (c), views. 96

5.1 Overview of the proposed idea. A noise model is trained with the purpose of adding as much noise as possible to the input. The output is a noise mask, in which noise sampled from the standard normal distribution is added. The result is then added to the input image and is used as an input to a pretrained prediction model with frozen parameters. The aim is to create a noisy image that will maximise the noise level while also not harming the performance of the prediction model. 102

5.2 Six different slices of the population-based importance map on top of the normalised average of the brain MR Images of the test set. The most important parts of the brain for the model's prediction are highlighted in red. The most significant regions for the task of brain ageing are mesial temporal structures including the hippocampus, brainstem, periventricular and central areas. The results are in agreement with the relevant literature and previous studies. 107
5.3 Importance maps after keeping different percentage of important voxels. 108
5.4 Importance maps for the cases of different depths of the Noise model, Small (depth=2), Medium (depth=3), Large (depth=4). For the Large model, the case of the Pretrained Large model is also presented. 110
5.5 Six different slices of the population-based saliency maps using gradient backpropagation. We note that here, only the saliency maps are presented, and there is no overlay of the average brain MR Images of the test set. 111
5.6 Six different slices of the population-based saliency maps using Grad-CAM on top of the normalised average of the brain MR Images of the test set. 112

6.1 Visualisations of the population graphs extracted from the different graph construction methods. The nodes are coloured based on the age of the subject. Colder colours (i.e. blue) indicate subjects with an older age, while warmer colours (i.e. red) correspond to younger subjects. 125

7.1 Proposed methodology: The non-imaging features $\mathbf{q}_i$ and a subset of the imaging features $\mathbf{s}_i \subseteq \mathbf{x}_i$ per subject i are used as input to a Multi-Layer Perceptron (MLP), which produces attention weights for each one of these features. Based on these, the edges of the graph are stochastically sampled using the Gumbel-Top-k trick. The constructed graph, which uses the imaging features $\mathbf{X}$ as node features is used to train a Graph Convolutional Network (GCN) for brain age estimation tasks. 134
7.2 Attention weights for the top 5 imaging (blue) and top 5 non-imaging phenotypes (pink) averaged across 10 runs of the pipeline. 139

7.3 Population-graph visualisations. Left: graph constructed based on the cosine similarity of the phenotypes. Right: graph constructed based on the cosine similarity of the weighted phenotypes, where the attention weights are the ones extracted from the proposed pipeline. The colour of the nodes corresponds to the subject's age. It is evident that using the attention weights increases the homophily of the graph. 141

7.4 Attention weights extracted from the pipeline per phenotype. The mean and std of the attention weights from 10 runs with different initialisations are shown. The phenotypes are indexed 1-35. Full list of the phenotypes and the attention scores can be found in Table 7.6. The non-imaging phenotypes are coloured in pink, and the neuroimaging phenotypes are coloured in blue. The highest weights are given to two cognitive tasks, namely trail making tasks numeric (15) and alphanumeric (16), and measurements extracted from FA skeleton (28-32). 143

List of Tables

4.1 Comparison of the performance of the 2D and 3D baseline experiments in terms of Mean Absolute Error (MAE) and R^2 score. 87

4.2 Comparison of the performance of the baseline model when using images with different registration as input in terms of Mean Absolute Error (MAE) R^2 score. Here, the baseline model refers to the 3D ResNet18 network. 88

4.3 Performance comparison of the most important experiments for the 64x64x64 patches in terms of Mean Absolute Error (MAE) and R^2 score. Here, the baseline model refers to the 3D ResNet18 network. 91

6.1 MAE (in years) of the different graph construction methods along with various GNNs on the test set. The baseline Multi-Layer Perceptron (MLP) (*No edges*) has a performance of MAE=3.73 years. 123

6.2 R^2 score of the different graph construction methods along with various GNNs on the test set. The baseline Multi-Layer Perceptron (MLP) (*No edges*) has a performance of R^2 score of 0.56. 123

6.3 Homophily ratios for graphs extracted by various graph construction methods. . 124

6.4 Mean Absolute Error (MAE) (in years) of Graph Convolutional Network (GCN) and GraphSage models for a kNN graph that uses different combinations of imaging and non-imaging phenotypes for the nodes and the edges. 126

6.5 Neuroimaging phenotypes 129

6.6 Non-imaging phenotypes 130

7.1 Performance of the proposed approach in comparison with a traditional machine learning model, static graph baselines, and the state-of-the-art for age regression. 139
7.2 Performance of the proposed approach in comparison with a traditional machine learning model, static graph baselines, and the state-of-the-art for and 4-class age classification (right). 140
7.3 Ablation study: Effect of the choice of phenotypes used for the connection of the edges. 140
7.4 Ablation study: Performance of the proposed using different distances on the phenotypes to connect the edges. 141
7.5 Homophily ratios for the population graphs visualised in Figure 7.3. *Static (Phenotypes)* refers to the graph on the left side of the figure, while *Proposed* refers to the graph on the right side of figure. 141
7.6 List of phenotypes used for the connectivity of the graph along with the attention weights (mean & std after 10 runs with different initialisations) given from the pipeline. 142

Acronyms

AD Alzheimer's Disease

Adam adaptive moment estimation

ADNI Alzheimer's Disease Neuroimaging Initiative

ANN Artificial Neural Network

ASD Autism Spectrum Disorder

CNN Convolutional Neural Network

CSF Cerebrospinal Fluid

DL Deep Learning

DNN Deep Neural Network

GAT Graph Attention Network

GCN Graph Convolutional Network

GM Grey Matter

GNN Graph Neural Network

GPR Gaussian Process Regression

GPU Graphics Processing Unit

GraphSage Graph Sample and Aggregated

HCP Human Connectome Project

HMM Hidden Markov Models

ILSVRC ImageNet Large Scale Visual Recognition Challenge

Leaky ReLU Leaky Rectified Linear Unit

MAE Mean Absolute Error

MCI Mild Cognitive Impairment

ML Machine Learning

MLP Multi-Layer Perceptron

MNIST Modified National Institute of Standards and Technology database

MR Magnetic Resonance

MRA Magnetic Resonance Angiography

MRI Magnetic Resonance Imaging

MSE Mean Squared Error

NN Neural Network

OASIS Open Access Series of Imaging Studies

PCA Principal Component Analysis

PET Positron Emission Tomography

ReLU Rectified Linear Unit

ResNet Residual Network

SGD Stochastic Gradient Descent

UKBB UK Biobank

VBM Voxel-Based Morphometry

WM White Matter

Chapter 1

Introduction

1.1 Motivation

Alzheimer's Disease (AD) is a progressive neurodegenerative disease that poses a significant health challenge worldwide. It is estimated that approximately six million people are living with AD in the US [5, 55]. As the global population continues to age, the prevalence of this debilitating condition is expected to rise, with significant economical consequences if no effective interventions are developed. This underscores the urgent need for continued research, early detection, and innovative treatments to address the growing impact of AD on individuals and healthcare systems globally.

The distinction between a subject's chronological age and their brain age has emerged as a potential biomarker in the context of neurodegenerative diseases, such as Alzheimer's [34, 35]. While chronological age represents the number of years a person has lived, brain age seeks to account for the physiological and cognitive changes that occur in the brain over time. In the context of neurodegenerative diseases, this concept becomes particularly relevant as it offers insights into the degree of ageing an individual is experiencing at a cellular and cognitive level. Identifying disparities between chronological and brain age may help in early detection and monitoring of such conditions, potentially providing a valuable tool for assessing disease risk and progression [51].

Magnetic Resonance (MR) images are a valuable resource in the context of age estimation analysis, primarily because they offer crucial insights into the transformations taking place within the human brain over time [59]. These images provide a detailed and non-invasive means of observing structural, functional, and volumetric alterations that occur as individuals age. By capturing these intricate changes, MR images serve as a vital source of data for researchers and medical professionals seeking to understand and model the ageing process, as well as its relationship to various neurological conditions. Thus, MR images have been extensively used for the task of brain age estimation.

Recently, the availability of vast amounts of medical imaging data has paved the way for the application of machine learning in various medical imaging tasks, including brain age estimation. In the literature, a spectrum of approaches has been explored, ranging from traditional machine learning methods to Convolutional Neural Networks (CNNs) and more recently, Graph Neural Networks (GNNs). Traditional machine learning techniques [181, 64, 52], while interpretable, often struggle to deliver optimal performance in this context. CNNs [75, 147, 96], on the other hand, tend to yield superior results, but they frequently operate as black box models that may raise concerns among healthcare professionals. One limitation across these methodologies is the challenge of effectively incorporating multi-modal information and the relationships among individuals within the population. To address this, the use of population graphs, coupled with GNNs, has emerged as a promising avenue for medical imaging tasks [142, 87]. A key challenge lies in determining the most effective approach to construct a meaningful population graph for downstream medical tasks.

When it comes to healthcare, it's important that the tools used for prediction are not only accurate but also easy for doctors to understand and trust [57, 72]. Interpretability not only ensures that the outcomes of complex models are comprehensible to healthcare professionals, but also provides insight into how and why specific decisions are made. This transparency fosters a stronger bond of trust between advanced machine learning techniques and the clinicians who rely on them, ultimately enhancing the collaborative effort to deliver high-quality medical care.

1.2 Thesis Objectives and Contributions

Motivated by the above, this thesis aims to investigate brain age estimation using machine learning models, while also focusing on providing interpretable results that will be trusted by clinicians.

More specifically, the main research questions tackled by this thesis are summarised as follows:

- **Can the use of patches of the brain, instead of the entire brain scan, improve brain age prediction, while also providing more localised and interpretable results?**

  The first objective of this thesis is to investigate the potential of utilising brain patches in conjunction with deep learning models, and particularly CNNs, to yield a regional estimate of brain age for each patch. This approach aims to enhance both predictive accuracy and interpretability by generating localised visualisations that highlight the specific brain regions that significantly influence brain age estimation.
- **Can we increase interpretability by using population-based importance maps?**

  Moving on, we focus further into the interpretability of the predictions. We transition from using brain patches, which produce coarse attention maps, to generating saliency maps via a noise model. We introduce an interpretability technique for CNNs in regression tasks, and provide a population-based importance map for brain age regression.
- **Is the use of population graphs and GNNs promising for brain age estimation?**

  While CNNs demonstrate strong performance in brain age estimation, their emphasis is solely on an individual's imaging data, disregarding the inter-subject relationships. As the next step of the thesis, we delve into the exploration of the use of multi-modal information for brain age estimation. An inherent way to achieve that, is through the use of population graphs and GNNs, which have been often used in the literature for other medical imaging tasks.
- **How can we achieve the construction of a meaningful and interpretable population graph for brain age estimation?**

  Motivated by the previous findings, we highlight the significance of a meaningful population graph, a challenge that is not straightforward when dealing with brain age estimation. The

final contribution of this thesis, is the extraction of an optimal population graph structure through graph learning. Leveraging attention weights, we are able to visualise and pinpoint the most crucial phenotypes for the construction of the graph, thereby enhancing the interpretability of the graph and, consequently, brain age estimation.

1.3 Thesis Outline

In Chapter 2, we provide a brief overview of the clinical background that is relevant for this thesis. More specifically, we describe the statistics and principles of Alzheimer's disease. In addition, we introduce the task of brain age estimation, which is the primary focus of the thesis, and present how the task is important for the early diagnosis of neurodegenerative diseases, such as Alzheimer's disease.

In Chapter 3, we focus on the technical background related to this thesis. We introduce the necessary machine learning and deep learning concepts, with a focus on CNNs and GNNs, and how these models have been used for the task of brain age estimation.

In Chapter 4, we work with CNNs and we explore whether the use of patches of the brain, instead of the whole brain, can not only improve brain age estimation, but also provide more localised results regarding the regions of the brain that are the most relevant to brain age. Different ways of extracting patches and patch sizes are tested and the resulting attention maps are visualised.

In Chapter 5, we focus more on the interpretability aspect of the task. We keep exploring CNNs and provide a pipeline that extracts saliency maps for brain age regression with the use of noise masks. We compare our method with saliency maps extracted from other commonly used methods in the literature.

In Chapter 6, we examine how population graphs and GNNs perform on the task of brain age estimation. We combine multimodal information, both from imaging and non-imaging information, and implement static population graphs with multiple ways that have been previously used in the literature. We use the extracted graphs along with various GNN models and we investigate whether the use of static graphs can provide promising results for brain age regression.

In Chapter 7, we move to adaptive graph learning and present a framework that creates

an optimised population graph structure for brain age estimation. The graph is learnt through training, and the population graph and the GNN are optimised simultaneously. We compare the results of adaptive population graphs and static population graphs, and point out the interpretability aspect that can be added when using our pipeline for adaptive graph learning.

Finally, in Chapter 8 we provide a summary of the achievements of the thesis. Additionally, we discuss the limitations and we explore the potential future steps and new research directions that can be built on our findings.

1.4 Publications

Parts of this dissertation have been published as peer reviewed publications. First author publications related to this thesis include:

- Bintsi, K., Baltatzis, V., Kolbeinsson, A., Hammers, A., and Rueckert, D.: Patch-based brain age estimation from MR images. In: Kia, S.M., Mohy-ud-Din, H., Abdulkadir, A., et al. (eds.) Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology - Third International Workshop, MLCN 2020, and Second International Workshop, RNO-AI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4-8, 2020, Proceedings. Lecture Notes in Computer Science, vol. 12449, pp. 98–107. Springer (2020). doi: 10.1007/978-3-030-66843-3_10, https://doi.org/10.1007/978-3-030-66843-3_10
- Bintsi, K., Baltatzis, V., Hammers, A., and Rueckert, D.: Voxel-level importance maps for interpretable brain age estimation. In: Reyes, M., Abreu, P.H., Cardoso, J.S., et al. (eds.) Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data - 4th International Workshop, iMIMIC 2021, and 1st International Workshop, TDA4MedicalData 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings. Lecture Notes in Computer Science, vol. 12929, pp. 65–74. Springer (2021). doi: 10.1007/978-3-030-87444-5_7, https://doi.org/10.1007/978-3-030-87444-5_7
- Bintsi, K., Mueller, T.T., Starck, S., Baltatzis, V., Hammers, A., and Rueckert, D.: A comparative study of population-graph construction methods and graph neural networks

for brain age regression. In: Ahmadi, S. and Pereira, S. (eds.) Graphs in Biomedical Image Analysis, and Overlapped Cell on Tissue Dataset for Histopathology - 5th MICCAI Workshop, GRAIL 2023 and 1st MICCAI Challenge, OCELOT 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, September 23, and October 4, 2023, Proceedings. Lecture Notes in Computer Science, vol. 14373, pp. 64–73. Springer (2023). doi: 10.1007/978-3-031-55088-1_6, https://doi.org/10.1007/978-3-031-55088-1_6

- Bintsi, K., Baltatzis, V., Potamias, R.A., Hammers, A., and Rueckert, D.: Multimodal brain age estimation using interpretable adaptive population-graph learning. In: Greenspan, H., Madabhushi, A., Mousavi, P., et al. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2023 - 26th International Conference, Vancouver, BC, Canada, October 8-12, 2023, Proceedings, Part VIII. Lecture Notes in Computer Science, vol. 14227, pp. 195–204. Springer (2023). doi: 10.1007/978-3-031-43993-3_19, https://doi.org/10.1007/978-3-031-43993-3_19

Co-authored publications or first author publications related to research conducted during the PhD, but are not presented in this dissertation, are the following:

- Müller, T.T., Starck, S., Dima, A., et al.: A survey on graph construction for geometric deep learning in medicine: Methods and recommendations. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=sWlHhfijcS, survey Certification
- Mueller, T.T., Starck, S., Feiner, L.F., Bintsi, K., Rueckert, D., and Kaissis, G.: Extended graph assessment metrics for regression and weighted graphs. In: Ahmadi, S. and Pereira, S. (eds.) Graphs in Biomedical Image Analysis, and Overlapped Cell on Tissue Dataset for Histopathology - 5th MICCAI Workshop, GRAIL 2023 and 1st MICCAI Challenge, OCELOT 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, September 23, and October 4, 2023, Proceedings. Lecture Notes in Computer Science, vol. 14373, pp. 14–26. Springer (2023). doi: 10.1007/978-3-031-55088-1_2, https://doi.org/10.1007/978-3-031-55088-1_2
- Müller, T.T., Starck, S., Bintsi, K.M., et al.: Are population graphs really as powerful as believed? Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=TTRDCVnbjI

- Bintsi, K.M., Jarolim, R., Tremblay, B., et al.: Sunerf: Validation of a 3d global reconstruction of the solar corona using simulated euv images (2022)
- Jarolim, R., Tremblay, B., Muñoz-Jaramillo, A., et al.: Sunerf: 3d reconstruction of the solar euv corona using neural radiance fields. The Astrophysical Journal Letters **961**(2), L31 (2024)
- Tremblay, B., Jarolim, R., Jungbluth, A., et al.: Ai to enhance the capabilities of euv-observing satellites and estimate spectral irradiance. Tech. rep., Copernicus Meetings (2024)
- Mason, J.P., Bintsi, K.M., Gilly, C., et al.: Leveraging artificial intelligence to enhance the science return of 4π solar constellations. Bulletin of the American Astronomical Society **55**(3), 269 (2023)
- Potamias, R.A., Neofytou, A., Bintsi, K., and Zafeiriou, S.: Graphwalks: Efficient shape agnostic geodesic shortest path estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022. pp. 2967–2976. IEEE (2022). doi: 10.1109/CVPRW56347.2022.00335, https://doi.org/10.1109/CVPRW56347.2022.00335
- Baltatzis, V., Bintsi, K., Folgoc, L.L., et al.: The pitfalls of sample selection: A case study on lung nodule classification. In: Rekik, I., Adeli, E., Park, S.H., and Schnabel, J.A. (eds.) Predictive Intelligence in Medicine - 4th International Workshop, PRIME 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings. Lecture Notes in Computer Science, vol. 12928, pp. 201–211. Springer (2021). doi: 10.1007/978-3-030-87602-9_19, https://doi.org/10.1007/978-3-030-87602-9_19
- Baltatzis, V., Folgoc, L.L., Ellis, S., et al.: The effect of the loss on generalization: Empirical study on synthetic lung nodule data. In: Reyes, M., Abreu, P.H., Cardoso, J.S., et al. (eds.) Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data - 4th International Workshop, iMIMIC 2021, and 1st International Workshop, TDA4MedicalData 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings. Lecture Notes in Computer Science, vol. 12929, pp. 56–64. Springer (2021). doi: 10.1007/978-3-030-87444-5_6, https://doi.org/10.1007/978-3-030-87444-5_6

1.5 Statement of Originality

I declare that the work presented in this thesis is my own, unless specifically acknowledged.

Kyriaki Margarita Bintsi

Chapter 2

Clinical Background

2.1 Alzheimer's Disease

Alzheimer's Disease (AD) presents a significant challenge in the field of neurodegenerative diseases, significantly affecting individuals and their families. This section provides a foundational understanding of AD, covering its definition, prevalence, and its crucial role in the broader context of neurological research.

2.1.1 Introduction to Alzheimer's Disease

Definition and Overview

AD is a progressive neurodegenerative disease, which means that more and more parts of the brain are damaged over time [54, 98]. It is characterised by cognitive decline, memory loss, and behavioural changes. At its core, the disease stems from a complex interplay of genetic, environmental, and lifestyle factors, leading to the gradual degeneration of brain cells and disruptions in neural networks. As neurons are damaged, characteristic features of AD, such as amyloid-beta plaques and tau tangles, become prominent, laying the foundation for cognitive decline and functional impairment [54, 5, 55]. There are many risk factors that contribute to the development of AD [78]. Some of them, such as genetic inheritance, sex, and age cannot be altered, while others such as lifestyle and some kinds of health problems can be improved.

A nuanced understanding of Alzheimer's necessitates recognising the intricate cellular and molecular mechanisms involved, as well as appreciating the intricate interconnections within the neural circuitry of the brain. Given its multifaceted implications, addressing AD requires a comprehensive approach in both clinical management and scientific exploration.

Epidemiology

Alzheimer's disease poses a significant public health challenge on a global scale. According to the World Alzheimer Report [196], an estimated 50 million people worldwide were living with dementia in 2020, and this number is expected to nearly triple by 2050, reaching 152 million if no effective interventions are developed. This surge is particularly notable in high-income countries, where the prevalence of AD is higher.

In the United States alone, it is estimated that over 6 million people have Alzheimer's, making it the sixth leading cause of death [54, 5, 98]. The economic impact is equally substantial, with the annual cost of dementia care projected to exceed one trillion dollars globally.

These statistics emphasise the pressing need for addressing Alzheimer's disease from both a research and public health perspective. With an ageing population and increased life expectancy, the prevalence of AD has steadily risen, underscoring the urgent need for effective diagnostic and therapeutic strategies. The burden of AD extends beyond the individual, impacting families, caregivers, and healthcare systems.

Early diagnosis of the disease is therefore essential both for reducing the costs by providing the patients with the appropriate supportive therapies, but also for understanding the progression of the disease with the purpose of development of new treatments [140].

In the subsequent sections of this chapter, we will delve into the intricate pathophysiology of AD, explore the clinical presentation and staging, and examine diagnostic approaches. This comprehensive exploration sets the stage for the integration of brain age estimation as a novel biomarker in the context of Alzheimer's disease research.

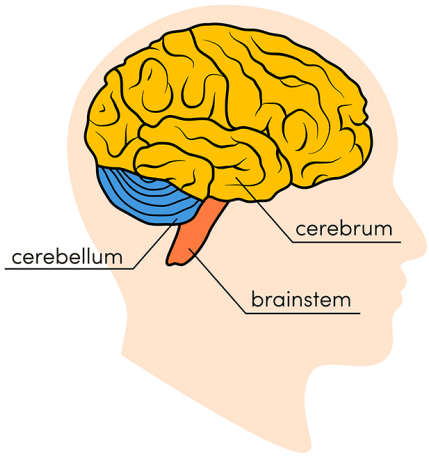


Figure 2.1: Main parts of the human brain. (Image from Johns Hopkins Medicine)

2.1.2 Pathophysiology of Alzheimer's Disease

Human Brain and Alzheimer's Disease

Human brains are made up of the cerebrum, the brainstem, and the cerebellum (Figure 2.1). The biggest part of the brain is consisted of the cerebrum, the surface of which is named cerebral cortex. The latter is an external layer of Grey Matter (GM), that encompasses the core of White Matter (WM). The cerebellum is under the cerebrum while the brainstem is considered as the connection of the cerebrum and the cerebellum to the spinal cord. A component of the brain that is crucial in AD is the hippocampus, which is detected under the cerebral cortex and is in charge of encoding long-term memory. The brain contains some cavities, named ventricles which are filled with Cerebrospinal Fluid (CSF) and is protected from injury by the skull.

It has been demonstrated that AD results in the shrinking of brain tissue [161]. More specifically, atrophy firstly manifests itself in the hippocampus and it then expands through the cerebral cortex. Meanwhile, cerebrospinal fluid (CSF) spaces, i.e. the ventricles inside the brain and cisterns outside the brain, are enlarged. The alterations in the brain because of AD can be depicted in Figure 2.2.

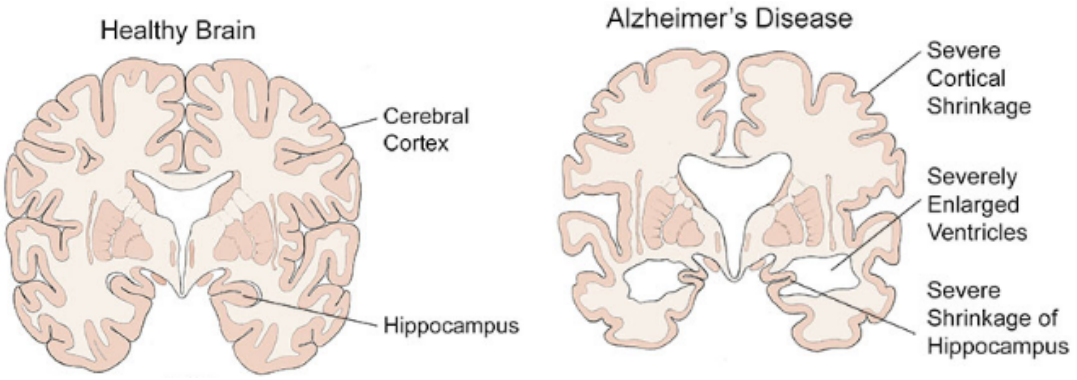


Figure 2.2: Human brain of a healthy subject (left) and a patient of Alzheimer's disease (right). Alterations in the cerebral cortex, the hippocampus, and the ventricles can be observed. (by BrightFocus Foundation)

Amyloid-Beta Plaques and Tau Tangles

A prominent characteristic of AD is the presence of abnormal protein accumulations within the brain. Amyloid-beta plaques and tau tangles play a central role in the progression of the disease [54, 5].

Amyloid-Beta Plaques. Amyloid-beta is a protein fragment that forms plaques between nerve cells [194]. These plaques disrupt communication and signaling between neurons, leading to cognitive dysfunction. The accumulation of amyloid-beta is believed to trigger a series of events resulting in damage and death of neurons. Researchers hypothesise that toxic forms of amyloid-beta, particularly soluble oligomers, may be more detrimental to synaptic function and contribute to the cognitive decline observed in AD [194, 153].

Tau Tangles. Tau is a protein that typically plays a role in maintaining the structural stability of neurons [194, 153]. In AD, tau undergoes abnormal modifications, resulting in the creation of neurofibrillary tangles within neurons. These tangles interfere with the transportation and cellular components, causing a decline in neuronal function and contributing to cell death. The progression of tau pathology in the brain follows a distinct pattern, correlating with the progression of cognitive decline.

Understanding the interplay between amyloid-beta plaques and tau tangles is crucial for unraveling the complex pathophysiology of AD. While the precise relationship between these

pathological features and the mechanisms underlying their toxic effects are areas of active research, their presence remains a hallmark of Alzheimer's.

Neurotransmitter Deficiency

Beyond the proteinaceous aggregates, alterations in neurotransmitters, particularly acetylcholine, significantly contribute to the progression of Alzheimer's disease [50].

Acetylcholine is a neurotransmitter crucial for various cognitive functions, including memory and learning [50]. In AD, there is a substantial reduction in acetylcholine levels due to the degeneration of neurons that produce this neurotransmitter, particularly in the basal forebrain. The decline in acetylcholine is associated with memory impairment and cognitive decline observed in individuals with AD.

Acetylcholine deficiency is targeted by therapeutic strategies that involve medications known as cholinesterase inhibitors [125, 160]. These medications aim to enhance cholinergic transmission by inhibiting the breakdown of acetylcholine, providing symptomatic relief in some individuals with Alzheimer's disease.

The intricate interplay between amyloid-beta pathology and neurotransmitter deficiencies highlights the intricate nature of Alzheimer's disease. Addressing both aspects in research and therapeutic interventions is essential for advancing our understanding and developing effective treatments for this complex neurodegenerative disorder.

2.1.3 Clinical Presentation

Early Symptoms

Memory Impairment. One of the earliest and most recognisable symptoms is memory impairment. Individuals may struggle to recall recent events, overlook important details, or face difficulty in forming new memories [32, 5].

Difficulty in Problem-Solving. As the disease progresses, individuals may encounter challenges in problem-solving and decision-making [32, 54]. Tasks that once seemed routine may become increasingly complex, leading to frustration and a sense of cognitive decline.

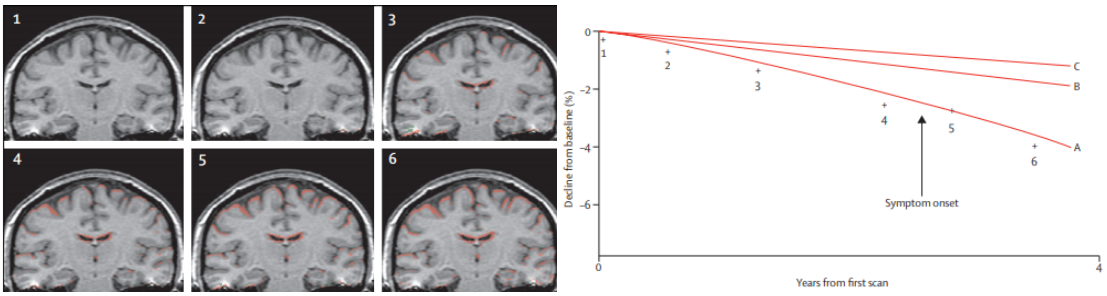


Figure 2.3: Left: Changes in the brain depicted in T1-weight MR images of an Alzheimer's patient that was asymptomatic at first. Symptoms appeared between scans 4 and 5 (arrow in right picture). Right: Plots of brain volume loss per scan in the case of (A) Alzheimer's patient, (B) normal ageing, and (C) very healthy subjects (average age of 70 years old) [78]

Confusion. Confusion is a common early symptom, with individuals experiencing disorientation regarding time, place, or the identity of familiar individuals. This cognitive confusion adds to a feeling of bewilderment and distress [32, 190].

However, it is important to note that it has been shown that AD starts spreading more than 20 years before the symptoms appear [5]. In Figure 2.3, T1-weighted Magnetic Resonance (MR) images from an asymptomatic at first patient, as well as plots of the percentage of brain volume lost over the years since the first scan in the cases of healthy and diseased subjects, are shown. It is clear that there is a delay in the appearance of symptoms.

Progression Stages

Alzheimer's disease is clinically categorised into distinct stages, each characterised by varying degrees of cognitive and functional decline [54, 5].

Mild Stage. In the mild stage, individuals may exhibit memory lapses, struggle with finding the right words, and encounter difficulty in organisational tasks. Despite these challenges, they often retain a degree of independence in daily activities.

Moderate Stage. As AD progresses to the moderate stage, cognitive decline becomes more pronounced. Language difficulties become apparent, affecting communication. Individuals may experience mood changes, such as increased irritability or withdrawal. Completing routine tasks, including personal care, becomes progressively challenging.

Severe Stage. The severe stage is marked by profound cognitive and functional decline. Individuals may lose the ability to communicate effectively, recognise loved ones, and perform basic activities of daily living. Dependence on caregivers for all aspects of care becomes necessary.

Understanding the nuanced progression of symptoms in Alzheimer's disease is crucial for timely diagnosis, intervention, and the provision of appropriate support for affected individuals and their families. The recognition of these stages facilitates a comprehensive approach to care and management throughout the course of the disease.

2.1.4 Diagnostic Approaches

Clinical Assessment

The diagnosis of Alzheimer's disease primarily relies on comprehensive clinical evaluations, which include cognitive tests and consultations with both patients and their caregivers [5, 32].

Cognitive Tests. A crucial component of the diagnostic process involves conducting cognitive tests to assess various aspects of mental function, such as memory, language, and problem-solving skills. These assessments contribute to a thorough understanding of the individual's cognitive status.

Patient and Caregiver Interviews. In-depth interviews with patients and their caregivers provide valuable insights into the onset and progression of symptoms. The subjective experiences shared during these interviews complement objective test results, enhancing the accuracy of the diagnostic evaluation.

Neuroimaging

In addition to clinical assessments, neuroimaging techniques play a crucial role in diagnosing Alzheimer's disease by providing visual insights into structural changes in the brain [163, 32, 8].

Magnetic Resonance Imaging (MRI). Detailed images of the brain's structure obtained through MRI scans aid in identifying any abnormalities or atrophy associated with Alzheimer's.

These images contribute to a comprehensive evaluation, facilitating early detection and confirmation of the disease.

Positron Emission Tomography (PET) Scans. PET scans serve to visualise metabolic activity in the brain. Through the detection of abnormal protein deposits, such as amyloid-beta, PET scans assist in identifying specific biomarkers linked to Alzheimer's. This enhances diagnostic precision and provides valuable information for developing targeted treatment plans.

The integration of clinical assessments with neuroimaging techniques ensures a comprehensive and multifaceted approach to diagnosing Alzheimer's disease. Early and accurate diagnosis is essential for implementing timely interventions and support strategies for individuals and their families as they navigate with the challenges posed by the disease [8].

2.2 Brain Age Estimation as a biomarker for Alzheimer's Disease

Biomarkers are measurable indicators of some biological state or condition, and in the context of AD, they play a critical role in diagnosis and disease monitoring [40]. Traditional biomarkers for Alzheimer's include CSF biomarkers, such as tau proteins and beta-amyloid levels, and neuroimaging techniques like MRI and PET scans that can visualise changes in brain structure and function [63]. These biomarkers have been instrumental in our understanding of Alzheimer's, providing insights into its pathophysiology and progression. The identification of reliable biomarkers is crucial for early detection, monitoring disease progression, and evaluating therapeutic interventions, thereby offering hope for improved patient outcomes [40].

Brain age estimation represents an innovative approach to understanding neurological health, positing that the "age" of one's brain may not necessarily match their chronological age. Healthy ageing results in the changing of brain structures following a specific pattern [30, 2]. A characteristic hallmark of neurodegenerative diseases such as AD is that the connections between neurons stop working and as a result many regions start to shrink. This process is called atrophy and leads to great loss of brain volume, especially in the final stages of the disease. Therefore, neurodegenerative diseases may result to accelerated ageing of the human brain [41, 65]. Hence,

a potential indicator used in AD is the estimation of the brain age of a subject which can then be compared with the real age of the subject [34, 51]. The motivation behind using brain age as a biomarker stems from its potential to offer a more nuanced understanding of brain health beyond traditional metrics. Unlike conventional biomarkers, which focus on specific pathological features, brain age estimation provides a holistic view of brain health, integrating multiple aspects of neurological function and structure.

Although the practicality of the use of this biomarker clinically is yet to be demonstrated, it seems to be a quite promising indicator of possible patients of Alzheimer's [34]. Implementing brain age estimation in a clinical setting could revolutionise AD management by enabling early detection and intervention. By identifying individuals whose brain age significantly exceeds their chronological age, clinicians could target early therapeutic strategies to potentially slow down or alter the disease trajectory. Moreover, brain age estimation can contribute to the creation of personalised management strategies by considering the biological condition of a person's brain, thereby customising treatments to ensure the best possible outcomes [128].

The integration of brain age estimation with traditional biomarkers holds the promise of enhancing diagnostic accuracy and understanding of AD. This synergistic approach allows for a comprehensive assessment of disease presence and progression, leveraging the strengths of each biomarker. For instance, while traditional biomarkers can indicate the presence of pathological changes, brain age estimation can provide a holistic view of the brain's health. This combination could lead to improved diagnostic precision and a deeper understanding of the disease's heterogeneity, facilitating the development of targeted therapies.

Brain Age Estimation for other applications

The concept of brain age estimation extends far beyond its application in AD, presenting itself as a promising biomarker for a wide range of neurological and psychiatric conditions [51, 174]. By comparing an individual's brain age to their chronological age, researchers and clinicians can gain insights into the health and ageing processes of the brain, potentially identifying early signs of disease or abnormal brain ageing before clinical symptoms manifest.

In conditions such as Parkinson's disease [47, 154], multiple sclerosis [36, 71], and even schizophrenia [97], brain age estimation has the potential to offer novel insights. For instance,

in Parkinson's disease, accelerated brain ageing may correlate with motor and cognitive decline, providing a useful tool for assessing disease severity and progression. Similarly, in multiple sclerosis, discrepancies between brain age and chronological age could reveal the extent of neurodegeneration, aiding in the prediction of disease course and response to therapy.

Furthermore, the use of brain age estimation in psychiatry holds promise for uncovering biomarkers of psychiatric conditions like major depressive disorder and bipolar disorder, where traditional biomarkers have been elusive [174, 37]. By identifying patterns of brain ageing associated with these conditions, clinicians could better understand the neurobiological underpinnings of psychiatric diseases, leading to improved diagnosis, prognosis, and personalised treatment strategies.

The potential for brain age estimation to serve as a biomarker across a spectrum of conditions underscores its value in neuroscientific research and clinical practice. It offers a unique lens through which the complex interplay between brain ageing and disease can be studied, opening up new avenues for early detection, monitoring, and intervention. As research in this area progresses, it is anticipated that brain age estimation will find broader applications, paving the way for more integrated and personalised approaches to neurological and psychiatric care [51].

2.3 Brain MR Datasets

MRI stands as a foundational element in neuroimaging, offering unmatched insights into the complex structures and functions of the human brain. This modality produces high-resolution images that are crucial for detailed examination of brain structures, playing a significant role in identifying age-related changes within the brain [59]. The ability of MRI to capture these intricate details makes it an invaluable resource in developing algorithms that can detect deviations from normal ageing processes, potentially indicating the early onset of neurodegenerative diseases. With the growing availability of large MR image datasets, deep learning methods are now at the forefront, using the power of artificial intelligence to improve how we diagnose and analyse brain health. As a result, MR data are mainly used in many advanced regression models, marking a major move towards using data-driven approaches to better understand and predict the ageing of the brain [10].

Several key brain MR datasets have been pivotal in advancing neuroimaging and brain age estimation research.

Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI (https://adni.loni.usc.edu/) is renowned for its focus on Alzheimer's disease, providing a rich collection of data that includes brain scans, biomarker profiles, and clinical assessments. It has been instrumental in understanding Alzheimer's progression and developing diagnostic tools. ADNI provides a large collection of MR images taken from elderly people who could potentially be AD patients and has been extensively used in the literature [148].

IXI. The IXI dataset (http://brain-development.org/ixi-dataset/), is a collection of about 600 MR brain images from healthy subjects aged 19-86 years old. Concerning the protocol of the image acquisition, T1, T2 as well as PD-weighted images are provided, along with some subjects having diffusion-weighted and Magnetic Resonance Angiography (MRA) images. Collected from three different hospitals in London, it offers a diverse sample that enriches the analysis and development of image processing techniques.

Human Connectome Project (HCP). The HCP (https://www.humanconnectome.org/) aims to map neural connections in the brain, offering high-resolution imaging data that sheds light on the brain's structural and functional network architecture. This dataset is crucial for studies on how these networks change with age or disease [58].

Open Access Series of Imaging Studies (OASIS). Focusing on the ageing brain, OASIS (https://www.oasis-brains.org/) datasets include cross-sectional and longitudinal MRI data from young, middle-aged, older adults, and individuals with AD, facilitating studies on normal ageing and neurodegeneration [124].

UK Biobank (UKBB). Last but not least, the UKBB (https://www.ukbiobank.ac.uk/) stands out for its scale, including imaging data from over 100,000 participants, aged between 40 and 69 at recruitment, which enables comprehensive studies on ageing and brain health across a diverse population. This wealth of data provides unprecedented opportunities to explore the

relationships between brain imaging findings, genetic predisposition, and lifestyle factors affecting brain health and disease. The UKBB was designed with the purpose of sharing data and includes a huge collection of brain MR images from over 60,000 subjects acquired with many different protocols such us T1-weighted, T2-weighted and Diffusion MRI [3] .

Regarding brain age prediction, the large sample size, combined with the depth and variety of data, provides a robust foundation for training algorithms that can accurately reflect the brain's ageing process, potentially identifying early signs of neurodegeneration. Although mostly T1-weighted images are used as input, it has been demonstrated that better prediction performance can be achieved by using a combination of different MR modalities [111, 31].

2.4 Summary

Alzheimer's disease is a serious and growing concern, with its prevalence projected to increase in the coming years, affecting more individuals and their families. Early diagnosis is essential in order to provide timely interventions and support. One intriguing avenue is the concept of "brain age," where the difference between a person's actual age and their brain age can serve as a potential biomarker for the disease. The estimation of brain age is becoming a meaningful task in the field of Alzheimer's research. With the abundance of MR images available at a large scale, deep learning methods have emerged as promising tools, harnessing the potential of artificial intelligence to advance our understanding and diagnostic capabilities.

In the following chapter, we will delve into the fundamental machine learning principles relevant to this thesis, and we will explore the promising machine learning techniques for brain age estimation as documented in the existing literature.

Chapter 3

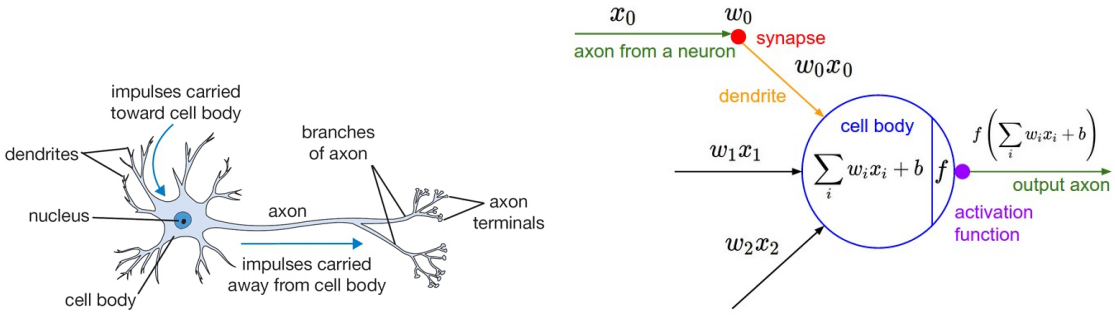
Technical Background and Related work

Within this chapter, we delve into the technical foundation that is relevant for this dissertation. We explore the broader field of machine learning (Section 3.1), delving deeper into the specifics of Convolutional Neural Networks (CNNs) (Section 3.2) and Graph Neural Networks (GNNs) (Section 3.3). These discussions set the stage for their application on brain age estimation, a focal point of this thesis, in Sections 3.4.2, 3.4.3, and 3.4.4 respectively. In Section 3.5, we address the significant bias problem in brain age regression tasks and we elaborate on corrective measures. Finally, in Section 3.6, we emphasise the critical role of interpretability, especially in healthcare-related tasks, underscoring its importance in ensuring meaningful and trustworthy outcomes.

3.1 Machine Learning

Machine Learning (ML) is a field dedicated to the development algorithms capable of extracting insights and patterns from data, ultimately facilitating informed predictions and decisions. It encompasses three main subcategories: supervised, unsupervised, and reinforcement learning.

In this thesis, the focus is mainly on supervised learning models. Supervised learning focuses



(a) (b)

Figure 3.1: A biological neuron (a) and the perceptron (b) [109].

on labeled datasets, pairing inputs with corresponding outputs. Its primary objectives include classification and regression tasks. In contrast, unsupervised learning dives into unlabeled data, seeking intrinsic patterns without explicit guidance. Clustering and dimensionality reduction are common endeavors within this realm. Finally, reinforcement learning introduces an agent navigating a dynamic environment. The agent learns optimal decision-making strategies through a feedback loop of rewards and penalties, following a trial-and-error approach.

Key components of the ML process include feature extraction, model training, evaluation, and deployment. Deep Learning (DL), a subcategory of ML centered around Artificial Neural Networks (ANNs), does not rely on manually crafted features. Instead, DL directly processes raw data as input, representing a distinctive shift from conventional machine learning approaches.

3.1.1 Artificial Neural Networks (ANNs)

ANNs are Neural Networks (NNs) inspired from the structure and functioning of the human brain. Similarly to its biological counterpart (3.1a), a neuron (3.1b) receives input from other neurons, processes it, and then transmits the result to other neurons. ANNs consist of interconnected simple neurons, and it's these connections that determine how the entire network operates.

In 1958, Rosenblatt et al. [157] introduced the concept of the perceptron, representing a more refined version of the artificial neuron, as depicted in Figure 3.1. This evolutionary step marked a significant advancement in the field.

The formula of the perceptron is given by:

$$y = \phi(\sum_i w_i x_i + b), \tag{3.1}$$

where x, and y are the inputs and the outputs of the neuron respectively, b is the bias term, w_i is the weight of the neuron, and $\phi(\cdot)$ is the activation function.

The activation function plays a pivotal role in neural networks by introducing non-linearity, a crucial element for capturing intricate patterns in data. This non-linearity is essential for the network's capacity to learn complex relationships. There are several activation functions that have been proposed in the literature [138]. These include the hyperbolic tangent (tanh) function (Equation (3.2)), Rectified Linear Unit (ReLU) [53] (Equation (3.3)), and Leaky Rectified Linear Unit (Leaky ReLU) [123] (Equations (3.4) or (3.5)):

$$tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} = \frac{1 - e^{-2z}}{1 + e^{-2z}} \tag{3.2}$$

$$ReLU(z) = max(0, z) \tag{3.3}$$

$$LeakyReLU(z) = max(0, z) + negative_slope \cdot min(0, x) \tag{3.4}$$

or

$$LeakyReLU(z) = \begin{cases} z, & \text{if } z \geq 0 \\ negative_slope \cdot z, & otherwise \end{cases} \tag{3.5}$$

When multiple perceptrons are interconnected, they form a Multi-Layer Perceptron (MLP), a fundamental architecture in ANNs. An MLP typically consists of three main layers: the input layer, hidden layer(s), and the output layer. The input layer receives the initial data, and the hidden layer(s) process this information through weighted connections and activation functions, capturing complex patterns and relationships within the data. Finally, the output layer provides the network's prediction. An example of an MLP with two hidden layers can be found in Figure 3.2.

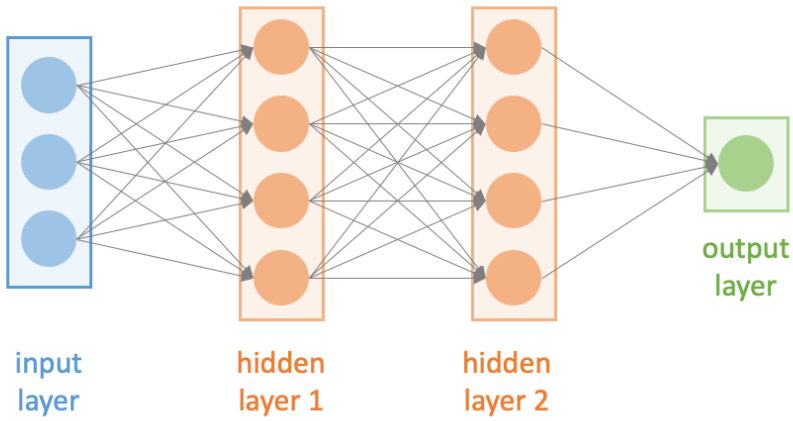


Figure 3.2: Example of an MLP with an input layer, two hidden layers, and an output layer.

Thanks to recent advancements in Graphics Processing Units (GPUs), we now have the capability to construct ANNs with numerous hidden layers, leading to what is commonly referred to as Deep Neural Networks (DNNs). The depth in DNNs arises from the incorporation of multiple hidden layers, allowing these networks to capture intricate patterns and representations in data. The increased depth contributes to the improved learning and abstraction capabilities of the model, enhancing its performance across various complex tasks.

3.1.2 Optimisation

Optimisation plays a pivotal role in machine learning, serving as a fundamental aspect in the training of models. Within the context of NNs, the optimisation process revolves around adjusting the weights to minimise a specified loss (or objective) function. This function quantifies the disparity between the predicted output of the NN and the desired output, essentially encapsulating how well the model is performing on a given task.

The mechanism through which the weights are updated is rooted in the concept of backpropagation, an algorithm introduced by Rumelhart et al. in 1986 [159]. Backpropagation involves the computation of gradients, which represent the rate of change of the loss function with respect to each weight in the network. These gradients guide the optimisation process by indicating the direction in which the weights should be adjusted to reduce the overall loss.

Gradient Descent

Gradient descent is the primary optimisation algorithm in ML [158]. The fundamental idea behind gradient descent is to minimise the loss function. The objective is to iteratively update the parameters of a model in the opposite direction of the gradients of the loss function with respect to those parameters. This iterative process allows the algorithm to navigate towards the minimum of the loss function, ultimately leading to optimised model parameters that enhance the model's performance on a given task.

In batch gradient descent, the gradients of the entire dataset are used to update the model parameters in each iteration. The formula is given by Equation (3.6):

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta) \tag{3.6}$$

In Equation (3.6), η represents the learning rate, determining the step size for each iteration.

Batch gradient descent typically converges smoothly towards the minimum of the loss function. However, when it comes to large datasets, it leads to high computational and memory costs, making it less suitable.

One possible solution involves adjusting the parameters after processing each individual sample, a technique known as Stochastic Gradient Descent (SGD). This is expressed by the formula:

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta; x_i; y_i) \tag{3.7}$$

This process introduces higher variance, resulting in fluctuations in the loss function. However, the advantage lies in its speed of execution.

A middle ground between gradient descent and SGD is represented by mini-batch gradient descent, which strikes a balance by dividing the dataset into smaller batches, and it is defined as:

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta; x_{(i:i+n)}; y_{(i:i+n)}) \tag{3.8}$$

Various alternative optimisation methods have been proposed to enhance traditional gradient descent [152, 137, 46]. One widely employed approach is adaptive moment estimation (Adam),

which individually updates each parameter. The update rule is as follows:

$$\theta = \theta - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon}\hat{m}_t \tag{3.9}$$

where η is the learning rate, m^t is the biased estimate of the first moment, v^t is the biased estimate of the second moment, and ϵ is a small constant added for numerical stability.

Loss Functions

The loss function, also known as objective or cost function, is a scoring function, which quantifies the difference between the predicted and the desired outputs of a model. The goal during training is to minimise the loss function. In this thesis, we focus on supervised ML methods, so here we present the most commonly used loss functions for classification and regression.

In classification scenarios, where the model's output is categorical (discrete), cross-entropy is widely utilised [136]. For binary classification tasks, the binary cross-entropy function is expressed as:

$$\mathcal{L}_{BCE} = -(y \log(p) + (1 - y) \log(1 - p)), \tag{3.10}$$

y denotes the true class label (0 or 1), and p represents the predicted probability of belonging to class 1.

For more general multi-class classification tasks, the cross-entropy loss function takes the form:

$$\mathcal{L}_{CE} = -\sum_{c=1}^{m} y_c \log(p_c), \tag{3.11}$$

where y_c signifies the true class label for class c, and p_c represents the predicted probability for class c.

When it comes to regression problems, where the output is a numerical value, the loss functions that are the most commonly used are the Mean Absolute Error (MAE), the Mean Squared Error (MSE), and the Huber Loss.

The MAE or L1 Loss, estimates the error as the distance between predicted and real values,

and is given by:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{3.12}$$

with n is the number of samples, and y_i and $\hat{y}_i$ are the real and predicted values for the i-th sample respectively.

The MSE or L2 Loss, calculates the average of the squared difference between predicted and real values. The formula is given by:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.13}$$

where n is the number of samples, and y_i and $\hat{y}_i$ are the real and predicted values for the i-th sample respectively.

The MSE tends to heavily penalise large errors, making it less robust to outliers. On the other hand, the MAE treats small and large errors with equal weight due to a consistent gradient loss. To strike a balance between the robustness of MAE and the sensitivity to small errors of MSE, the Huber Loss is employed. The formula for Huber Loss is provided in Equation (3.14).

$$\text{Huber Loss} = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & \text{if } |y - \hat{y}| \leq \delta \\ \delta(|y - \hat{y}| - \frac{1}{2}\delta) & \text{otherwise} \end{cases} \tag{3.14}$$

Here, y_i and $\hat{y}_i$ are the real and predicted values for the i-th sample respectively, the loss is quadratic for absolute differences less than or equal to δ, the loss is linear for absolute differences greater than δ.

3.2 Convolutional Neural Networks (CNNs)

A promising deep learning technique are CNNs [105] which usually require unprocessed images as input with the purpose of many different tasks, such as lesion classification. In this case, if the appropriate volume of data is available, the network by itself is able to find patterns among the raw data and transform them in a format suitable for the prediction by using many different layers of convolutions [35, 34].

CNNs can have inputs in the form of timeseries (1D grid), images (2D grid), or volumes (3D grid). Their exceptional suitability for image data has led to their widespread use in fields such as computer vision and medical imaging. Through a hierarchical learning process, CNNs automatically extract features, with lower layers capturing simple elements and higher layers discerning more intricate patterns.

CNN layers possess three dimensions: height, width, and depth/channels. Typical CNN structure comprises of a series of convolutional, pooling, normalisation layers, and, for classification or regression models, one or more fully connected layers. This architecture enables the network to learn and represent features at various levels of complexity.

Convolutional layer

The convolutional layer stands as a foundational element in CNNs, playing a pivotal role in the extraction of relevant features and patterns from input data. Its purpose is to systematically apply convolutional operations, a process that involves sliding filters known as kernels across the input.

In this layer, a local receptive field is employed, ensuring that a small region of the input is processed at a time. This localised approach allows the network to focus on capturing intricate local patterns, a crucial aspect for tasks such as image recognition.

One distinctive feature of the convolutional layer is the use of shared learnable parameters. This strategy offers two significant advantages. Firstly, it dramatically reduces the number of parameters, enhancing the efficiency of the network, particularly when compared to MLPs. Secondly, this parameter sharing endows the network with the capability to identify similar patterns throughout the entire input space.

The output size of a convolutional layer, denoted as $W_o \times H_o \times D_o$, which takes an input of dimensions $W_i \times H_i \times D_i$, can be determined using the following formula:

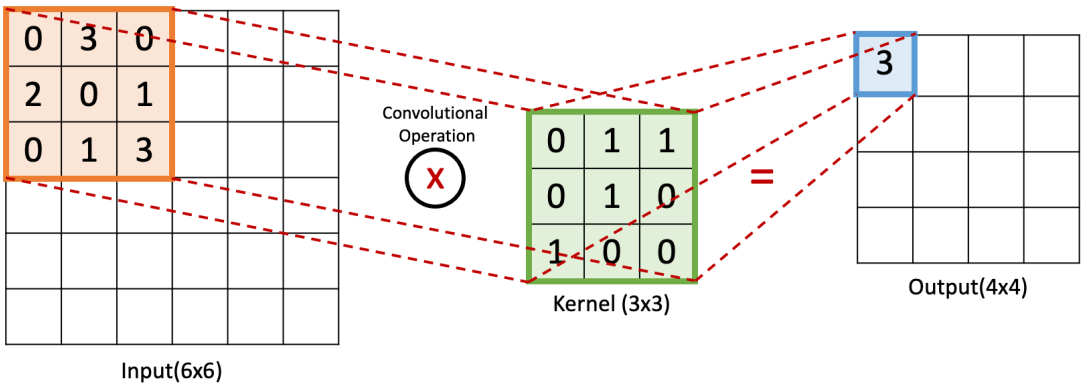


Figure 3.3: A convolutional layer with $K = 1$ filter of size $F = 3$, stride $S = 1$ and padding $P = 0$. Input dimensions are $W_1 = 6$, $H_1 = 6$ and $D_1 = 1$, and output dimensions are $W_2 = 4$, $H_2 = 4$ and $D_2 = 1$.

$$W_o = \frac{W_i - F + 2P}{S} + 1 \tag{3.15a}$$

$$H_o = \frac{H_i - F + 2P}{S} + 1 \tag{3.15b}$$

$$D_o = K \tag{3.15c}$$

Here, K represents the number of filters in the convolutional layer, F is the receptive field of each filter, S is the stride (i.e., the step size of the filter across the input), and P is the padding, often applied to maintain the input size at the output.

An illustration of the convolutional layer's application is presented in Figure 3.3. Consider an input with dimensions $W_i = 6$, $H_i = 6$, and $D_i = 1$. For this scenario, a single filter with $K = 1$, a size of $F = 3$, stride $S = 1$, and no padding ($P = 0$) is employed. Utilising equations (3.15a) through (3.15c), we can estimate the output's dimensions. The resulting width (W_o) is calculated as $\frac{6-3+0}{1} + 1 = 4$, the height (H_o) as $\frac{6-3+0}{1} + 1 = 4$, and the depth (D_o) remains 1.

Typically, CNNs consist of multiple convolutional layers, enabling the network to learn hierarchical representations. Lower layers specialise in capturing simpler features, while higher layers combine these features, allowing the network to discern and represent more complex patterns in the input data.

Pooling layer

The pooling layer is used to systematically reduce the spatial dimensions to decrease computational complexity, while retaining vital information. Two prevalent types of pooling are max pooling and average pooling.

Max pooling [201] is the one most commonly used. It retains the maximum value from a group of the neighbouring pixels, emphasising prominent features. Average pooling calculates the average value of the pixel group. Both pooling types are used after each convolutional layer to effectively reduce the spatial dimensions (width and height) of the network.

Global average pooling is applied along the channel dimension and is positioned after the last convolutional layer. It is frequently applied in networks where fully connected layers are not utilised as the final layers.

Normalisation layer

Normalisation layers are pivotal for enhancing the stability and convergence of CNNs. Various normalisation techniques, including Batch Normalisation [79], Layer Normalisation [117], Instance Normalisation [183], and Group Normalisation [192], contribute to the efficiency of network training. The selection among these methods depends on the unique characteristics of the problem at hand and the specific architecture of the network.

Fully connected layer

Fully connected layers are a key component in CNNs and are often positioned as the final layers of the network. Every node in a given layer is connected to every node of the previous layer, similarly to the structure found in MLPs. While convolutional and pooling layers specialise in processing local input regions, fully connected layers focus on the integration of the high-level features acquired by the preceding convolutional layers. Their objective is to generate the final classification/regression.

CNN architectures

In this section, we briefly describe some of the most influential CNN architectures.

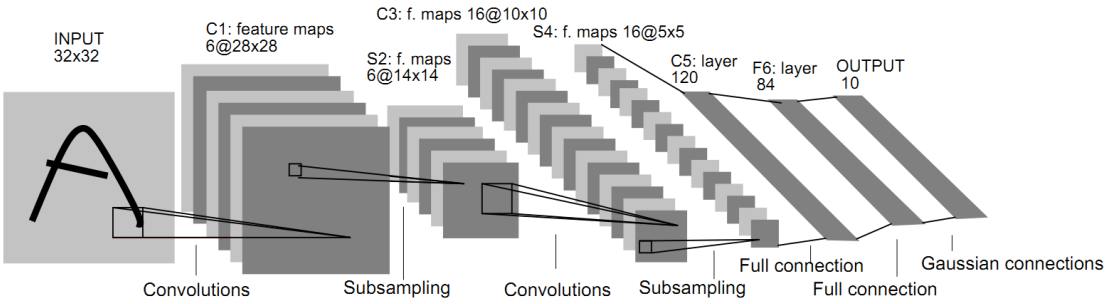


Figure 3.4: The architecture of LeNet-5 [106].

Among the pioneers is LeNet-5 [105, 106], proposed by LeCun et al. in 1998. This seminal architecture, depicted in Figure 3.4, was devised for handwritten digit classification, particularly on the Modified National Institute of Standards and Technology database (MNIST) dataset. LeNet-5, with its novel convolutional and pooling layers, laid the foundation for subsequent advancements in the field.

A milestone in the evolution of CNNs is AlexNet [100] , a deeper architecture introduced by Krizhevsky et al. in 2012. This groundbreaking model not only facilitated the learning of more complex features but also emerged as the champion of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [44] (an image database with more than 14 million images of 1000 classes) in 2012, outperforming its competitors by a significant margin. The victory of AlexNet underscored the capacity of deep neural networks to excel in large-scale image classification tasks.

GoogleNet (Inception) [178] by Szegedy et al. won ILSVRC in 2014. Its novelty was that it introduced the revolutionary inception module. This module allowed the simultaneous utilisation of various filter sizes, effectively reducing the number of required parameters. The inception module's impact resonates in architectures to this day, emphasising the significance of parallelised processing for enhanced performance. A visualisation of GoogleNet can be found in Figure 3.5.

VGGNet [170], proposed in the same year as GoogleNet, claimed the runner-up spot in the ILSVRC. Crafted by Simonyan et al., VGGNet stands out for its simplicity and depth, proving to be highly influential. Its architecture, characterised by stacking multiple convolutional layers, paved the way for deeper networks, and demonstrated the advantages of pretraining for improved performance. A visualisation of the network is given in Figure 3.7 (Left).



Figure 3.5: The GoogleNet architecture [178].

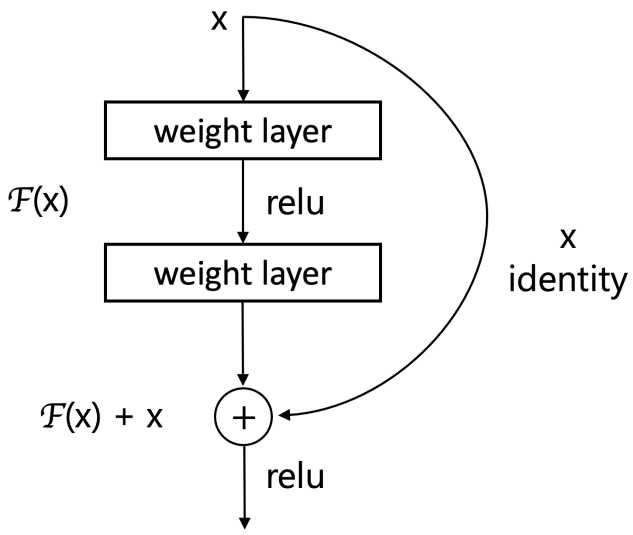


Figure 3.6: The architecture of a residual block containing a skip connection as proposed in [67].

Residual Network (ResNet) [67], introduced in 2015 by He et al., marked a paradigm shift with the incorporation of residual (or skip) connections. These connections addressed challenges related to vanishing gradients, enabling the creation of even deeper networks. The residual block (Figure 3.6), a cornerstone of ResNet's design, allowed for the successful training of networks with hundreds of layers. In Figure 3.7, a visualisation of ResNet-34 is provided.

Finally, for medical image segmentation tasks, U-Net [156], proposed by Ronneberger et al. in 2015, emerged as a pioneering architecture. Adopting an autoencoder structure with skip connections, U-Net facilitates the flow of high-resolution information from the encoder to the decoder, as illustrated in Figure 3.8. Renowned for its effectiveness, U-Net remains a cornerstone in computer vision applications.

Beyond these landmarks, other noteworthy architectures that have left their imprint on the landscape, such as EfficientNet [179], MobileNet [73], and Densenet [74], have each contributed unique innovations, further diversifying the toolbox of CNN architectures. These models continue to shape the trajectory of deep learning, illustrating the dynamic and ever-evolving nature of the field.

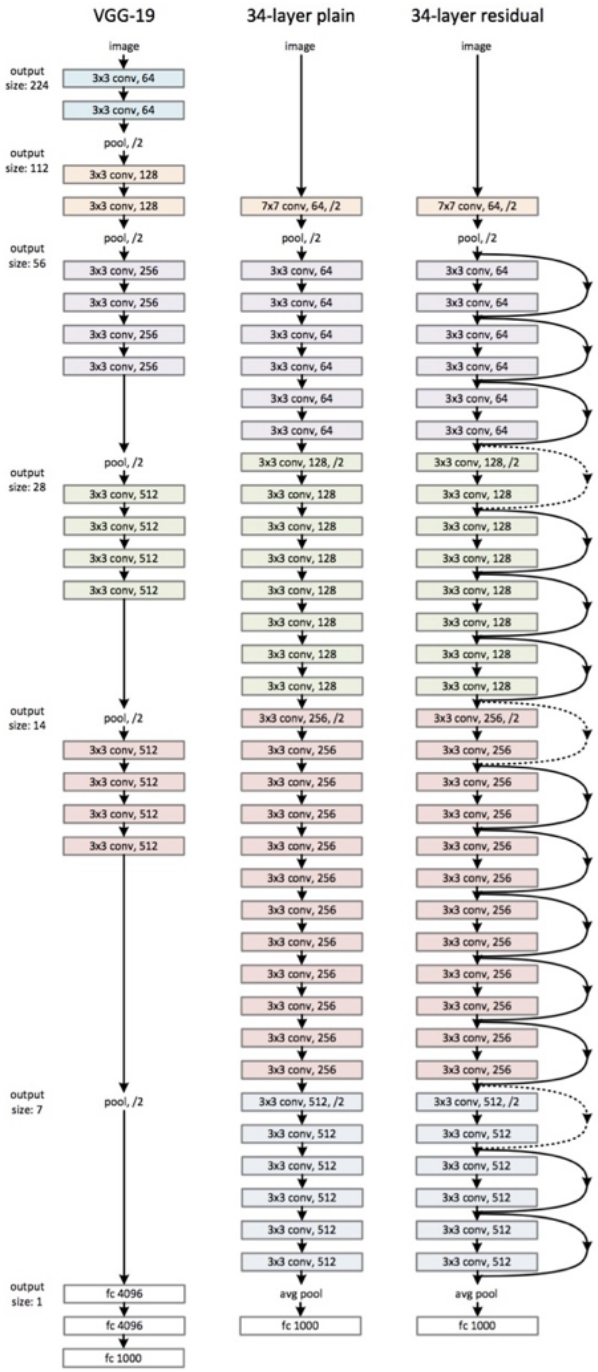


Figure 3.7: Example CNN architectures as presented in [67]. Left: VGGNet model [170] (as reference). Middle: a simple model with 34 layers. Right: The architecture of Residual Network (ResNet)-34.

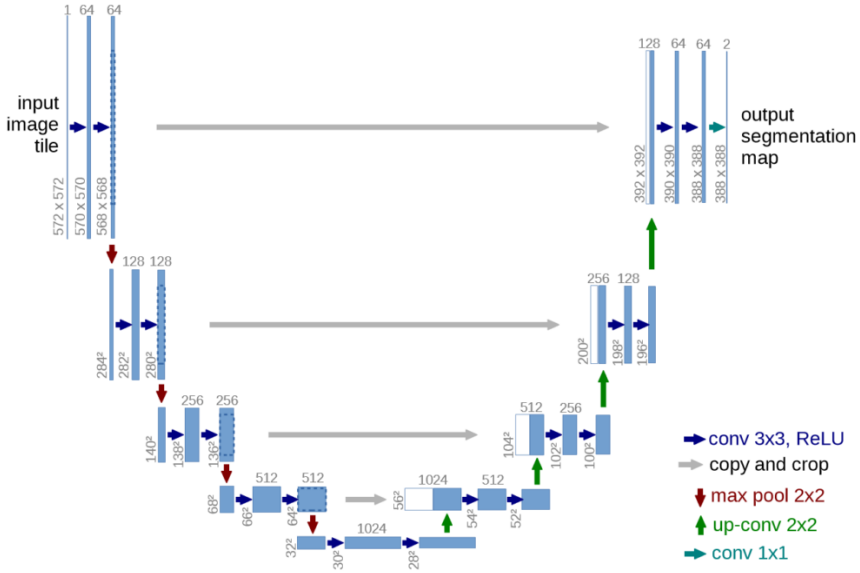


Figure 3.8: The architecture of U-Net as introduced in [156].

3.3 Graph Neural Networks (GNNs)

A graph is a data structure that is consisted of nodes (vertices) and edges, and it can represent multiple things, such as social networks, knowledge graphs [13], and 3D shapes [29].

There are three different categories of prediction tasks on graphs, node-level, edge-level, and graph-level tasks. For graph-level predictions, each sample is represented by a whole graph, e.g. a molecule, and a prediction is made for the entire graph. For edge-level tasks, predictions are made for the edges of the graph. Finally, regarding node-level tasks, we get a prediction for every node of the graph. When it comes to node-level predictions, the graph structure used is called a population graph. In a population graph, every node represents a subject and the associations among the nodes are represented by the edges.

More formally, given a collection of N entities with M features $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N] \in \mathbb{R}^{N \times M}$ and corresponding labels $\mathbf{y} \in \mathbb{R}^N$, a population graph is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. In this context, $\mathcal{V}$ represents a set of nodes, each corresponding to an individual entity, and $\mathcal{E}$ signifies a set of paired nodes, delineating the connections within the graph, i.e., the edges.

GNNs are the subcategory of NNs that use graphs as input by leveraging the concept of graph convolutions [60, 162]. GNNs work using a message-passing scheme, which was initially

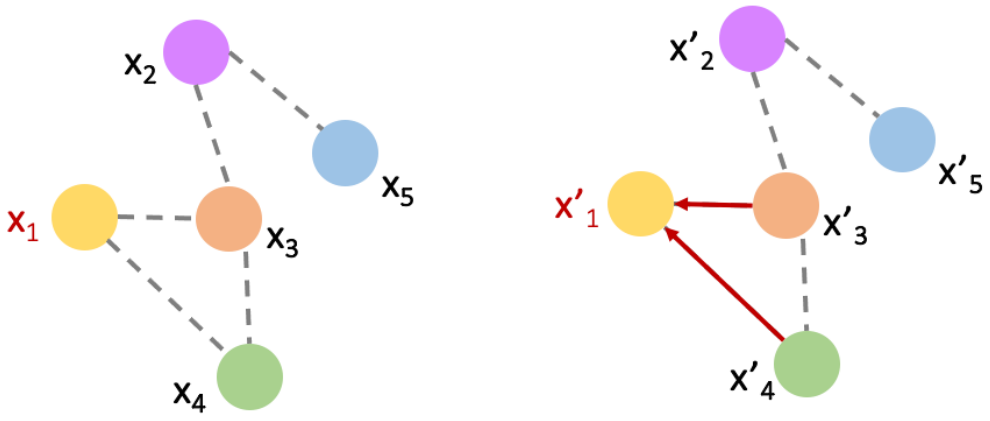


Figure 3.9: Message-passing update with a GNN on a graph with five nodes. Each node has a node feature vector x_i (left), and an updated node feature vector x'_i after the aggregation from the neighbouring node representations.

suggested by [56] and propagates the information of the nodes at their neighbourhood through the graph edges. An example of how message-passing works in a graph may be found in Figure 3.9.

GNN Architectures

There are multiple GNN architectures proposed in the literature [92, 66, 131, 193, 185]. We provide here the definitions of the GNNs that are the ones most commonly used.

Graph Convolutional Network (GCN) was one of the first GNNs introduced and it was suggested by [92]. The updated node feature emdeding follows the definition of Equation (3.16).

$$x'_i = \Theta^T \sum_{j \in \mathcal{N}_i \cup \{i\}} \frac{1}{\sqrt{\hat{d}_j \hat{d}_i}} x_j, \tag{3.16}$$

where $\hat{d}_i = 1 + \sum_{j \in \mathcal{N}_i} 1$, Θ is the weight matrix to be learnt during training, and $\mathcal{N}_i$ represents the neighbours of node i.

Graph Attention Network (GAT) was proposed by [185], and it incorporated the attention mechanism to the GCN in order to highlight the importance of different neighbours during the aggregation of information. More specifically, the node feature representation is provided by

3.17.

$$x'_i = \alpha_{ii}\Theta x_i + \sum_{j \in \mathcal{N}_i} \alpha_{ij}\Theta x_j, \tag{3.17}$$

where Θ is the weight matrix to be learnt during training, α_{ij} is the attention coefficient between two nodes i and j and is defined as follows:

$$\alpha_{ij} = \frac{\exp\left(\phi\left(a^T\left(\Theta x_i \parallel \Theta x_j\right)\right)\right)}{\sum_{k \in \mathcal{N}_i \cup i} \exp\left(\phi\left(a^T\left(\Theta x_i \parallel \Theta x_k\right)\right)\right)}, \tag{3.18}$$

where ϕ is commonly the LeakyReLU function and $\parallel$ indicates a concatenation of the values.

GraphSage was introduced by [66] and was designed to be able to work on large graphs. The updated node features are given by 3.19.

$$x'_i = \text{AGGREGATE}\left(\{x_j, \forall j \in \mathcal{N}(v_i)\}\right) \tag{3.19}$$

where x'_i is the updated feature vector for node v_i, $\mathcal{N}(v_i)$ is the set of sampled neighbours of node v_i, x_j is the feature vector of node v_j where v_j is a neighbour of v_i, and AGGREGATE is the aggregation function, which can be a mean, max, or LSTM-based aggregator, depending on the GraphSAGE variant.

Finally, Chebyshev [43] involves the use of Chebyshev polynomials to perform graph convolutions and updates the node features based on Equation (3.20)

$$x'_i = \sum_{k=0}^{K-1} \sum_{j \in \mathcal{N}_i} \frac{1}{\sqrt{d_i d_j}} x_j \Theta_k, \tag{3.20}$$

where K is the order of the Chebyshev approximation, x_j is the features of node j, Θ_k is the parameter matrix for the k-th Chebyshev polynomial, $\mathcal{N}_i$ represents the neighbours of node i, and d_i,d_j are the degrees of nodes i, j respectively.

Homophily on Graphs

The impact of graph structure on the efficiency of GNNs is a critical factor, as highlighted by various studies [120, 202]. The literature delves into assessing graph quality through metrics

like graph homophily/ heterophily [121], Cross-Class Neighbourhood Similarity (CCNS) [122], Normalised Total Variation (NTV), and Normalised Smoothness Value (NSV) [120]. These metrics play a pivotal role in understanding the dynamics of graph performance. Interestingly, a correlation emerges between these metrics and GNN performance, revealing that certain models, such as GCN and GAT, exhibit heightened sensitivity to graph structure, while others like GraphSage display a more resilient response.

The metric that is the most commonly used in the literature is homophily. In the context of population graphs, homophily refers to proportion of edges connecting nodes sharing the same labels to the edges connecting nodes with different labels. The underlying concept is that as GNNs transmit node features along edges, the lower the similarity among neighbours, the less likely it is for the GNN to acquire meaningful node feature embeddings for a given node. This, in turn, can have repercussions on the overall performance of the network.

There are three types of homophily: class homophily [112, 121], edge homophily [90], and node homophily [145]. They are similar with only small differences in the aspects of the graph structure that they highlight. Node homophily [145] is most commonly used, and it is defined as in Equation (3.21).

$$\text{homophily}(G, Y) := \frac{1}{|V|} \sum_{v \in V} \frac{|\{u | u \in \mathcal{N}_v, Y_u = Y_v\}|}{|\mathcal{N}_v|}, \tag{3.21}$$

where $G = (V, E)$ is the graph with labels $Y := \{y_u; u \in V\}$, $\mathcal{N}_v$ is the set of neighbouring nodes of v, and $|\cdot|$ the cardinality of a set.

A graph with homophily close to 1 is considered a *homophilic* or a graph of *high homophily*. On the other hand, when a graph has a homophily ratio close to 0, it is considered as a *heterophilic* or a graph of *low homophily*. In Figure 3.10, you may find examples of *homophilc* and *heterophilic* graphs.

When it comes to regression, there is little ongoing work on graph assessment metrics and how they could be extended to work for regression tasks, mostly by [132]. The homophily of the graphs for node regression, by estimating the distance among the labels of the nodes of a neighbourhood, can be found in Equation (3.22).

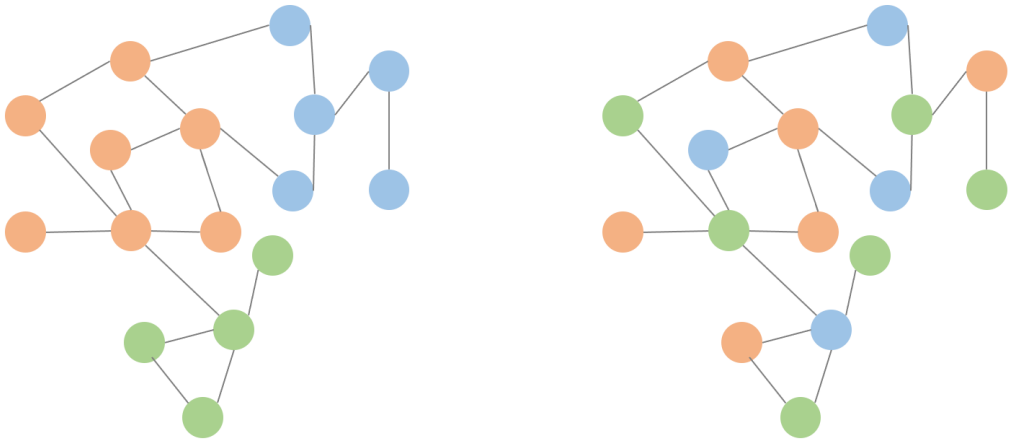


Figure 3.10: Examples of graph with high homophily (left), and low homophily (right). The three different colours of the nodes represent three different classes.

$$\text{homophily}_{\text{reg}}(G, Y) = 1 - \left(\frac{1}{|V|} \sum_{v \in V} \left(\frac{1}{|\mathcal{N}_v|} \sum_{n \in \mathcal{N}_v} \|y_v - y_n\|_1 \right) \right), \tag{3.22}$$

where $G := (V, E)$ is the graph with labels $Y := \{y_u; u \in V\}$, $\mathcal{N}_v$ is the set of neighbouring nodes of v, and $\|\cdot\|_1$ indicates the L_1 norm.

3.4 Related Works on Brain Age Estimation

3.4.1 Pre-Machine Learning Approaches for Brain Age Estimation

Before the era of ML, the field of neuroimaging relied heavily on traditional methods, which were predominantly manual and dependent on the specialised knowledge of clinicians and radiologists. These professionals would meticulously examine brain scans, searching for visible markers indicative of ageing or Alzheimer's Disease (AD), such as brain atrophy or the presence of lesions. This process was based on visual interpretation of images, employing a qualitative approach to assess the severity and progression of these conditions. While this method benefited from the acute observational skills of experienced practitioners, it was inherently subjective, with assessments varying between observers.

To complement these qualitative evaluations, researchers also developed basic statistical mod-

els that took a more quantitative approach. By collecting measurements from brain scans, such as the volumes of specific brain regions, and applying statistical analyses, this approach aimed to objectively correlate physical brain changes with ageing or disease progression [180, 116, 69]. These early quantitative methods marked a significant step towards understanding the structural manifestations of brain ageing and pathology.

However, these approaches were constrained by several limitations. The models used were relatively simple and lacked the ability to automate the analysis process, making it challenging to scale up studies or address the complex, multi-dimensional nature of brain imaging data effectively.

The advent of brain atlases marked a pivotal moment in the evolution of neuroimaging techniques [68, 4, 70, 191]. These atlases provided comprehensive maps of brain anatomy, becoming indispensable tools for the research community. They facilitated the standardised identification and analysis of brain regions across different individuals, enabling more systematic and objective analysis than the subjective methods that preceded them.

Voxel-Based Morphometry (VBM) further advanced the quantitative analysis of brain images [127, 59, 9]. By segmenting Magnetic Resonance (MR) images into distinct tissue types and assessing the volume of these tissues in a voxel-wise manner across the brain, VBM allowed researchers to identify subtle, diffuse patterns of change associated with ageing and disease, circumventing the need for manual region-specific measurements. This technique provided thorough insights into the process of regional brain shrinkage and revealed the relationship between structural changes in the brain and age-related cognitive decline.

The introduction of atlases and the development of VBM represented a significant leap forward in the pre-ML era of neuroimaging. These innovations bridged the gap between the subjective assessments of the past and the automated, data-driven approaches that machine learning would later enable. By laying the foundational techniques for standardised, quantitative analysis, they paved the way for the sophisticated, scalable methods that now underpin contemporary neuroimaging research.

3.4.2 Traditional Machine Learning for Brain Age Estimation

Machine learning has been extensively used for the prediction of the chronological age from MR data [181, 64, 52]. Machine learning algorithms are trained to find out correlations and patterns in the MR data in order to predict the brain age of the subject. These models require some 'hand-engineering', including the removal of non-brain tissue, image registration, smoothing and interpolation. The common steps adopted include the segmentation and normalisation of the MR data followed by feature extraction and feature selection or dimensionality reduction, such as Principal Component Analysis (PCA). These data are then used as input in a classifier with the purpose of age prediction.

The machine learning models are trained on a dataset of MR images of healthy subjects and are used to predict the estimated chronological age of the subjects. The prediction error is defined as the difference between the predicted age and the biological age. When working with the training set, the prediction error should be as small as possible while when working with the test set, which could include both healthy subjects and possible AD patients, a great difference is indicative of accelerated brain ageing, thus a possible AD patient. Apart from the prediction error, the authors in [64] proposed the use of Gaussian Process Regression (GPR) uncertainty in order to measure the alterations from the healthy ageing model as they suggested that accelerated ageing is not usually brain-wide.

A considerable amount of work has been done on brain age estimation from MR images using machine learning techniques. Many of the existing approaches take advantage of the segmentation of the T1-weighted MR images into Grey Matter (GM), White Matter (WM), and Cerebrospinal Fluid (CSF) regions and utilise one or more region masks as input. It has been demonstrated that the GM volume increases rapidly until reaching a peak at some point during the subject's childhood, after which it starts decreasing continuously [62, 168]. Concerning the WM volume, it expands until the age of 10-15 years old and it continues to grow but extremely slowly afterwards, while the CSF, remains stable until the age of 20 when, contrary to the GM volume, it starts rising monotonically until the age of 70 [81]. This approach [94] took advantage of the alterations of these volumes in order to come to a conclusion regarding the prediction of the chronological age of the subject. More specifically, local features were extracted from GM,

WM, and CSF volumes with the help of automated anatomical labelling (AAL) atlas, which were then used as an input to a Relevance Vector Machine (RVM) classifier. A MAE of 4.50 years was achieved. Another approach, which was presented in [186], utilised only the GM volume in order to predict brain age by using Hidden Markov Models (HMM).

In this study [52], the authors proposed a framework that included an affine registration of T1-weighted MR scans and dimensionality reduction via PCA followed by a RVM for regression. The method performed pretty well as a MAE of 5 years was noted as a measure of prediction accuracy. In [187], the authors outlined a cortical surface pattern (CSP) by associating the cortical thickness with curvature and predicted the chronological age with RVM regression reporting a MAE of 4.57 years.

Although many methods that use machine learning for brain age prediction provide quite accurate results, they require data preprocessing which is not only time-consuming and complex, but also prevents the algorithms from being able to generalise because of the hand-crafted features.

Deep learning has been greatly used in medical image analysis achieving high performance in many different tasks, such as brain segmentation and disease classification, outperforming classical machine learning approaches in most cases [113]. Contrary to machine learning, deep learning techniques require negligible or no image preprocessing at all.

3.4.3 CNNs for Brain Age Estimation

Recently, there have been some studies using deep learning, and more specifically CNNs, with the purpose of brain age prediction from MR images. CNN based architectures were applied with the purpose of an accurate brain age prediction while using the minimum amount of domain information [184, 177, 83]. In this work [75], the authors proposed a CNN that received the raw T1-weighted MR images as input in order to take into account all the available information that could possibly be omitted if some kind of segmentation was used. The method was both fast and accurate and state-of-the-art results were achieved with a MAE of 4 years. Another study [151] used a 3D CNN architecture in order to take advantage of the spatial information between the slices of the MR images. In addition, the authors extracted the high frequencies of the structures

of the brain and then extracted the features, taking into account that brain ageing results to morphological alterations of the brain. Additionally, they used a densely connect block method in order to enhance the fitting ability of their network. Their network performed very well with a MAE of 4.28 years. 3D CNNs were also used in [35], where both raw T1-weighted MR images and processed GM and WM images were used as input. The predictions were quite accurate in both cases, with a MAE of about 4.50 years, proving that sufficient features can be extracted from the raw 3D data without using the additional assumptions made during the preprocessing. It is worth mentioning that in this study a GPR approach was used for comparison and these methods performed equally well in terms of MAE. In addition, in a recent competition on brain age estimation held last year (https://www.photon-ai.com/pac2019) the top performing submission achieved results with less than 3 years in MAE. Last but not least, in [96], the authors implemented a 3D ResNet model, which is enhanced by a randomised tensor regression layer, achieving state-of-the-art results, and a MAE of approximately 2.6 years.

Deep learning techniques have demonstrated state-of-the-art performance in brain age estimation without the need for hand-crafted feature extraction [197, 113]. Avoiding the preprocessing steps not only eliminates possible biases and assumptions made, but also accelerates the pipeline thus providing the clinicians with a quicker prediction. If real time prediction is achieved, models will be available to help clinical decision making. However, in order to do so, the models should not be regarded as black boxes and their results would have to be interpretable in order to be trusted by the clinicians. Patches have been widely used for brain disease prediction [115, 110] as they can provide information about subtle changes at a local level that are probably disregarded when working globally. To our knowledge, although patches could provide more localised estimations, they have not been used successfully for age regression. It is believed that accurate localised predictions could provide interpretable insights [144] that would be similar to clinical ones [64, 33].

3.4.4 GNNs for medical tasks

GNNs have been used in the literature for a variety of medical tasks. They have been used on graphs that have an inherent structure, such as surface meshes for quantifying fatty tissue [133],

molecules for drug discovery [28], vessels for vessel segmentation [139], and brain connectivity graphs for the representation of the human connectome [20].

Nevertheless, in situations such as population graphs, a distinct and predefined structure may not be evident. As described before, in population graphs, each subject is typically represented as a node, with imaging features as node features, and edges are determined based on predefined criteria indicating the similarity between subjects. The construction of the graph can either be pre-decided and static [142], or learnt dynamically during training [85].

Within the medical literature, population graphs coupled with GNNs have been employed for various tasks [1], like AD classification, and Mild Cognitive Impairment (MCI) to AD conversion. There are methods that use static graphs [142, 87, 198], while others use adaptive graphs that are trained alongside the GNN [85, 199, 76]. There is also work on Autism Spectrum Disorder (ASD) classification [7, 142, 199], COVID-19 outcome prediction [88], and other applications [198, 146].

When it comes to brain age estimation, the primary focus of this thesis, there is little ongoing work on population graphs and GNNs. To our knowledge, only [176] has tackled the task of age regression by constructing a population graph with imaging information used for the node features, and non-imaging information determining the edges. However, the model's performance was modest, with a MSE of approximately 28. Additionally, a few studies have addressed the less challenging 4-class age classification problem using graph learning [85, 86, 38]. These works, where the population graph is learnt during training alongside the model, achieved accuracy ranging from 60% to 65%.

3.5 Bias Correction

A common issue observed when estimating brain age, is a bias, possibly because of regression dilution [104]. More specifically, there is an overestimation in younger subjects and an underestimation in older subjects, which arises because datasets usually follow a gaussian distribution with respect to age. Hence, the predictions, which a ML model makes, for cases on the lower and higher parts of the distribution tend to converge to the mean of the distribution. On the other hand, predicted age of subjects with chronological ages closer to the mean age tend to be

more accurate. Additionally, another hypothesis revolves around the idea that the relationship between real and biological age is non-linear. There is a lot of work [173, 18, 37, 42] going on towards the adjustment of the bias by applying a statistical bias correction method to the predicted ages. This recent commentary [102] compares some of the different existing approaches.

Recent works [42, 18] utilise the chronological age of the subject as a covariate with the purpose of predicting a bias-corrected age. Although this strategy cannot be helpful for predicting the bias-corrected age of a subject lacking the chronological age, it could prove important clinically for giving insights for the subject's brain age, as when people arrive at the hospital, they usually provide their demographics, including chronological age.

3.6 Interpretability for Deep Learning

Understanding the logic behind a model's decision is very important in assessing trust and therefore in the adoption of the model by the users [57]. For instance, the users should be certain that correct predictions are an outcome of proper problem representation and not of the mistaken use of artifacts. For this reason, some sort of interpretation may be essential in some cases. In the medical domain [72], the ability of a model to explain the reasoning behind its predictions is an even more important property of a model, as crucial decisions about someone's life need to be made based on its predictions.

Usually, there is a trade-off between accuracy and interpretability [27]. Although deep learning models have even better performance than humans in some cases, they are usually considered as black boxes [130]. On the other hand, classical machine learning techniques, such as decision trees and linear models, are more interpretable and this is the reason behind their extensive use in many applications even though they do not perform as well as other more complex models. Therefore, the need for tools that will not only be accurate, but also understandable by humans, is justified and the increasing interest in research related to explainable deep learning is reasonable.

A common technique to make a deep network interpretable is by using a proxy method in order to simplify its outputs and not interfere with the model's structure[129, 167, 16]. This method, termed network diagnosis by [197], focuses on understanding the model behavior rather than

the model function. According to [155], good explanation methods need to meet certain criteria, which involve being interpretable, locally faithful, and model-agnostic. Apart from that, they must provide a global perspective of the model. Keeping these characteristics in mind, the authors proposed their explanation technique, which can be used in combination with any classifier and not only provide explanations of the predictions by approximating locally an interpretable linear model around them, but also choose and present representative instances in order to provide the user with a general idea about how the classifier works. More recently, another study ([130]) suggested methods for making any kind of machine learning model, including nonlinear deep neural networks, interpretable by building prototypes and exploiting the use of Layer-wise Relevance Propagation.

Another approach to the explainability of deep learning is by understanding the function and the structure of the model. This approach is referred to as transparency by [129] and can be further distinguished in comprehension of the model, comprehension of the parameters and comprehension of a specific training algorithm [107]. [101] proposed modifying CNNs in order to make them interpretable rather than tackling the problem with network diagnosis. More specifically, each high-layer filter is trained to correspond to an object part of the image in order to create a network that can clarify which specific part is learnt by each convolutional filter.

A concept that could contribute to the explainability of deep learning is attention [6]. The idea behind it is the production of a context vector that assigns weights on the input. Attention Gates (AG) have been considerably used in computer vision and natural language processing domains [6, 82]. In the medical domain, [164] is a study worth mentioning as the authors proposed the application of Attention Gated (AG) models in medical images. The purpose of the study is to highlight salient features for classification or segmentation by incorporating AG networks into a standard CNN. It is believed that the focus of the model on specific regions could result in better understanding behind the model's decisions. Therefore, AG networks could be useful for explainable deep learning models.

A quite interesting study was reported in [19], where the authors tried to predict human survival in patients with a heart disease using deep learning, more specifically a Denoising Autoencoder (DAE) and a prediction branch for the survival. An important contribution of this work was the presentation of their results in an interpretable manner, with the use of Laplacian

eigenmaps and saliency maps, which facilitates the understanding of the logic behind the predictions of the model and enhances its trustworthiness. This approach could be applied for age estimation from brain images as well. In this way, not only a novel deep learning technique that is differentiated from the existing CNN models could be deployed, but also the results will be understandable by humans which could be the start for a whole new set of possibilities in the field of age estimation with the purpose of diagnosis of Alzheimer's disease.

3.7 Summary

Summarising all of the above, AD is a life-threatening disease and early diagnosis could contribute to the understanding of its progression. The estimation of the chronological age of a subject and the comparison with their biological age is considered as a promising biomarker. MR images have been extensively used in age prediction tools. Machine learning techniques have been developed for this purpose giving very good results with quite small MAE. However, these techniques include some preprocessing which is time consuming. In addition, this preprocessing that is tailored to the training dataset does not allow the trained model to generalise on unseen data. On the other hand, deep learning techniques have achieved the same, if not better, results, without any 'hand-engineering', making the potential of real time prediction and the involvement of deep learning models in clinical decisions feasible. However, the architectures that have been primarily developed with the purpose of age estimation are based on CNNs and other architectures, such as GNNs have not been particularly explored. Apart from this, the fact that deep learning models are regarded as black box models prevents their use by the clinicians since they are not trustworthy. Therefore, the development of deep learning models that are not only accurate, but also interpretable seems to be a territory that needs to be investigated.

Chapter 4

Patch-based Brain Age Estimation from MR Images

Chapter 3 highlighted the importance of a deep learning pipeline that can not only predict brain age as accurately as possible, but it is also interpretable and localised, so it will be trusted by clinicians. In this chapter, we explore the use of patches of the brain, instead of the whole brain, and obtain a regional estimate of brain age based on every patch. In this way, we allow the network to focus on the specific patch without being affected by the rest of the brain. Furthermore, we evaluate which patches include important regions for the age regression and obtain a visualisation on the areas that play the most vital role on brain age estimation, thus making the results more interpretable and providing the first step for even more localised predictions. We, then, combine the results of the different patches to produce a more accurate brain age estimation for each subject.

This chapter is an extension of the work that is published in [23]:

Bintsi, K., Baltatzis, V., Kolbeinsson, A., Hammers, A., and Rueckert, D.: Patch-based brain age estimation from MR images. In: Kia, S.M., Mohy-ud-Din, H., Abdulkadir, A., et al. (eds.) Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology - Third International Workshop, MLCN 2020, and Second International Workshop, RNO-AI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4-8, 2020, Proceedings. Lecture Notes in

Computer Science, vol. 12449, pp. 98–107. Springer (2020). doi: 10.1007/978-3-030-66843-3_10, https://doi.org/10.1007/978-3-030-66843-3_10

4.1 Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disease and the most common cause of dementia [5]. It has been demonstrated that AD results in atrophy of the brain[161] as well as changes in the morphology of brain tissues. Specifically, atrophy firstly manifests itself in the hippocampus and it then expands through the cerebral cortex. Meanwhile, Cerebrospinal Fluid (CSF) spaces, i.e. the ventricles inside the brain and cisterns outside the brain, are enlarged. Biomarkers are an important diagnostic tool for detection of the early stages of AD [40]. A characteristic hallmark of neurodegenerative diseases such as AD is that they may result in accelerated ageing of the human brain. Hence, a promising biomarker that has been used in AD and other neuroimaging studies is brain age [51, 34], which measures the difference between the biological brain age of a subject and their real (chronological) age.

Structural Magnetic Resonance (MR) images have been widely used for the measurement of brain changes related to age [59] and are often available in patients. For this reason, a considerable amount of work has been done on brain age estimation from MR images using Machine Learning (ML) techniques [17, 52, 181]. The segmentation of the T1-weighted MR images into Grey Matter (GM), White Matter (WM), and CSF regions, and region masks are usually used as input to a ML model [94, 52], achieving a Mean Absolute Error (MAE) of as low as 4.3 years [111]. Recently, various Convolutional Neural Network (CNN)-based approaches which provide accurate brain age estimation using raw T1-weighted MR images with MAEs ranging from 3.3 to 4 years have been proposed [83, 75, 34]. In a recent competition on brain age estimation held in 2019 (https://www.photon-ai.com/pac2019) the top performing submission achieved results with a MAE of less than 3 years. In [96], the authors implemented a 3D Residual Network (ResNet) model enhanced by a randomised tensor regression layer, achieving state-of-the-art results, and a MAE of 2.6 years.

A common bias observed when estimating brain age, possibly because of regression dilution [104], is the overestimation in younger subjects and the underestimation in older subjects. The

term bias correction used in age estimation should not be confused with the bias field correction that is used as a pre-processing step for MR images [84]. There is a substantial amount of work [173, 37, 102] going on towards the adjustment of the bias by applying a statistical bias correction method to the predicted ages. Recent works [103, 18] use the real age of the subject as a covariate to predict a bias-corrected age.

Patches have been widely used for brain disease prediction [110, 114] as they can provide information about subtle changes at a local level that are probably disregarded when working globally. To our knowledge, although patches could provide more localised estimations, they have not been used successfully for age regression. It is believed though that accurate localised predictions could provide interpretable insights [144] that would be similar to clinical ones [17, 33].

In this chapter, we propose the use of patches of the brain, instead of the whole brain, as input and train independent 3D ResNet18 networks for every patch in order to obtain a regional estimate of brain age. Thus, we allow the network to focus on the specific patch without being affected by the rest of the brain. Furthermore, we evaluate which patches include important regions for the age regression and obtain a visualisation of these areas, thus making the results more interpretable and providing the first step for even more localised predictions. We, then, combine the results of the different patches to produce a more accurate brain age estimation for each subject, achieving state-of-the-art results with a MAE of 2.13 years before and 1.96 years after bias correction.

4.2 Materials & Methods

4.2.1 Dataset & Preprocessing

We use the UK Biobank (UKBB) dataset [3] for the task of age regression. UKBB aims to scan nearly 100,000 subjects and is designed with the purpose of sharing data. It incorporates a large collection of Magnetic Resonance Imaging (MRI) scans of vital organs, including brain MR images acquired with many different protocols such us T1-weighted, T2-weighted and Diffusion MRI. The dataset used in this study, contains 3D T1-weighted brain images, sized 182x218x182, acquired from nearly 15,000 healthy subjects aged between 44 and 73 years old, 52.3% of which



Figure 4.1: Age distribution of the study population.

are women and 47.8% are men. The age distribution of the study population can be found in Figure 4.1. The images are skull-striped and are non-linearly registered to MNI152 standard space, while we also normalise each brain scan to zero mean and unit variance. From the 14,503 subjects provided, 753 did not provide their age and are removed from the study, thus leaving 13,750 3D brain images, 80% of which are used for the training set, 15% for the validation set and the remaining 5% for the test set.

4.2.2 Convolutional Autoencoder

We wanted to explore how convolutional autoencoders can perform for the task of brain age prediction. Therefore, a network based on [19] was implemented. The architecture of the network consisted of a convolutional autoencoder with an age prediction branch, as can be seen in Figure 4.2.

The network utilises an autoencoder in order to create latent representations optimised for age prediction. The input vector is introduced to three consecutive blocks that are consisted of a convolutional, a batch normalisation and a ReLU activation layer. A convolutional, a ReLU and a max pooling layer follow and the output is fed to the bottleneck layer of the autoencoder that represents the latent code, which has a much lower dimensionality than the one of the input and can be regarded as a compressed representation. Since both 2D and 3D inputs were explored,

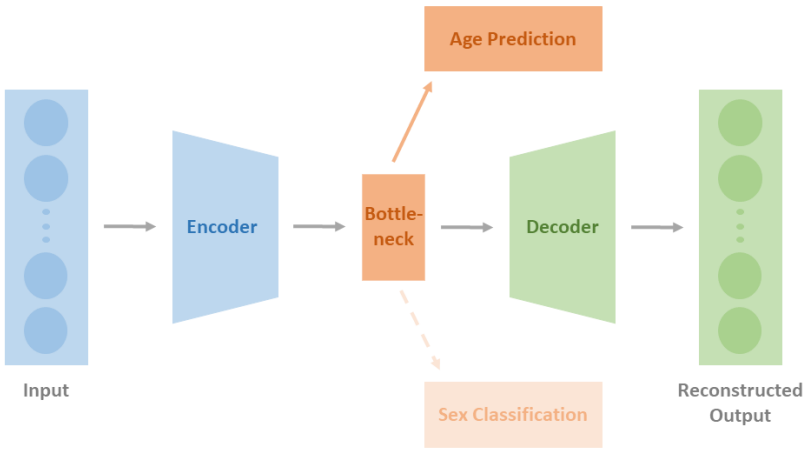


Figure 4.2: The architecture of the network. The encoder part, consisted of hidden layers, results to the bottleneck layer (latent code layer). The latent code is used for the reconstruction, by using a decoder, as well an age prediction task achieved by an extra branch. An optional sex classification task was added for some experiments as well.

two similar models were implemented using the 2D and 3D equivalent functions of the layers mentioned.

The latent code is then used as input in the second part of the autoencoder which is a multi-layer decoder network. The purpose of the decoder is to upsample the latent code, by applying a transposed convolution operator, in order to bring it back to the original dimensionality. The autoencoder has a symmetric architecture, meaning that the same number of hidden layers were used for both the encoder and the decoder parts.

The purpose of the autoencoder is to minimise the dissimilarity between the original input and the reconstructed output vectors of the autoencoder by minimising the loss function. For our network, Mean Squared Error (MSE) loss function $\boldsymbol{L_r}$ is chosen. In this way, the network learns to keep robust features in the latent code.

However, the aim of the study is age estimation. Therefore, we do not only need robust features for the reconstruction but also the most useful ones for the prediction. For this reason, apart from the decoder part, the bottleneck layer is also linked to a prediction branch, and more

specifically a linear regression, with the task of age estimation. The prediction branch is trained with the ages of the subjects as labels. In order to minimise the difference between the real and the predicted ages of the subjects, we use a mean squared error loss function $\boldsymbol{L_p}$ in this case as well.

The loss used for our network was a combination of $\boldsymbol{L_r}$ and $\boldsymbol{L_p}$. We define:

$$L_{network} = \alpha L_r + \beta L_p \tag{4.1}$$

where the terms α and β define the importance of each one of the losses to the global loss. For the training of the network, β was selected equal to $1 - \alpha$. The value of α chosen is 0.3.

4.2.3 ResNet18 Network

2D & 3D Baseline Models

We trained a 2D ResNet18 network using as input only the central transverse slice of the brain scans, as we wanted to check if only a slice of the brain would include enough information for the brain age estimation.

In addition, we implemented a 3D version of a ResNet18 model similar to the one used in [96] in order to use it as a baseline for the evaluation of our method. The architecture of the network follows the one of [67] but we use three-dimensional convolutions instead of 2D.

Patch-based Prediction

An overview of the proposed method can be found in Figure 4.3.

In brief, we utilise patches of the brain, instead of the whole brain scans, with the purpose of understanding the effect of different parts of the brain on the predictions, and hence acquiring more localised predictions. We center crop each brain scan from 182x218x182 to 140x176x140. Afterwards, we extract 3D patches sized 64x64x64 with a stride of 40x56x40 from each brain, which leads to 27 patches per brain scan. We then train 27 18-layer 3D ResNet models, similar to the baseline one described before. We train and test on the same patch, thus on a specific region of the brain. This allows the network to focus on the specific patch independently of the

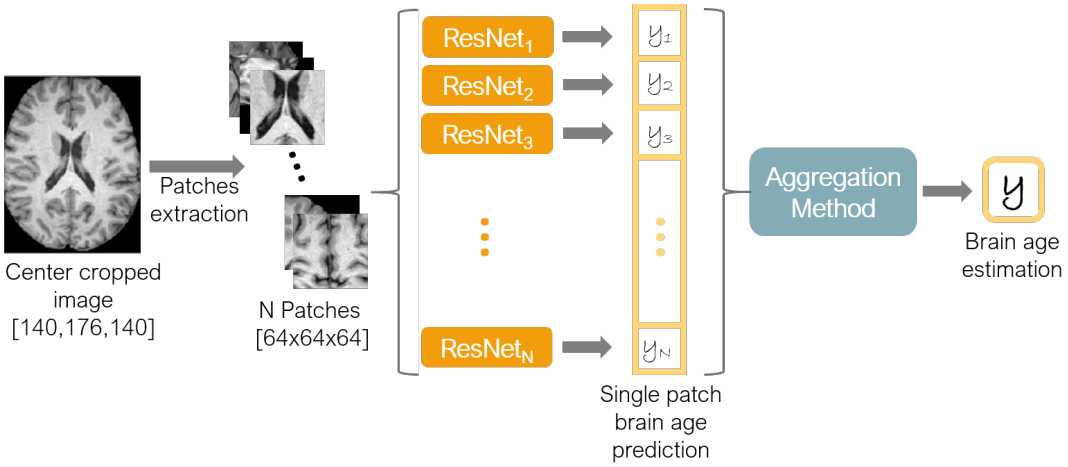


Figure 4.3: Overview of the proposed method. N 3D patches are being extracted from the center-cropped image. Every patch is used as input in a ResNet model that is trained to predict the subject's brain age, hence N different age predictions y_i are produced. The individual predictions are then combined using an ensemble method in order to achieve an improved estimation y.

rest of the image.

Ensemble Method We investigate whether the use of patches could lead to a more accurate brain age estimation. Therefore, we combine the predictions of the single-patch networks with an ensemble model. Every patch-based model provides a brain age estimation resulting in 27 independent predictions for each subject. We experiment with two different ways for the fusion of the patch-based models. The first involves taking the mean of the results of the multiple ResNets for every instance of the test set. So, if N is the total number of patch-based models, $P \leq N$ is a subset of these models and y_i is the prediction of every one of them, the prediction y produced by the ensemble model is given by Equation (4.2):

$$y = \frac{1}{P} \sum_{i=1}^{P} y_i \tag{4.2}$$

Moreover, since not all patches are equally important for age prediction, weighted averaging may produce better results. We, therefore, implement a linear regression model that predicts the brain age of each subject on the test set from y_i. The output prediction y in this case is given

by Equation (4.3):

$$y = w_0 + \sum_{i=1}^{P} w_i y_i \tag{4.3}$$

Here, w_0 is the intercept term and w_i are the learnable weights for each of the P individual models. In addition, based on the previous experiments, we make some conclusions on which patches contain important features for brain age and which not. For this reason, we choose the number of the selected patch-based models P so that each individual $MAE < threshold$ on the validation set. We choose that threshold heuristically. We investigate both averaging the predictions and using a meta-regressor similar to the one described above.

32x32x32 Patches Following up from the previous exploration, we extracted smaller patches out of a patch of the previous network in order to check if a more specialised localisation is feasible.

Therefore, we extracted patches of 32x32x32 out of a 64x64x64 patch used before and trained networks independently. The patch from which smaller patches were extracted can be found in Figure 4.8. A total of 8 patches were extracted from each brain and thus, 8 networks were trained.

Random Patches As the methods described before do not easily scale, we used random patches for the training of the network with the purpose of localised predictions. Therefore, a number for random patches was extracted from each brain scan. These could be overlapping and they were different for each scan.

Three different experiments were implemented, using different size for the random patches.

More specifically, for the first experiment, relatively big patches sized 64x64x64 were used, while for the next two, patches of 32x32x32 and 16x16x16 were used. A number of N patches were extracted from each brain volume and every patch was used independently as an input in the network. Therefore, there were $N * (Number of Subjects)$ inputs in total. The number of patches for the three experiments were 30, 50, and 70, respectively.

For the case of the 64x64x64 patches, we also average the predictions of all the patches extracted from the same subject with the purpose of getting one age estimation for every subject.

Bias Correction

The bias correction technique adopted in this work is the one proposed by [18]. We fit the relationship of the brain age delta and the chronological age using Equation (4.4):

$$\Delta = \alpha * Y_{chronological} + \beta \tag{4.4}$$

where $Y_{chronological}$ is the real age of the subject and Δ the brain age delta function. The α and β represent the slope and intercept respectively and are then used for the estimation of the corrected predicted age from Equation (4.5):

$$Y_{corrected} = Y_{predicted} - (\alpha * Y_{chronological} + \beta) \tag{4.5}$$

Saliency Maps

We utilised saliency maps [169] in order to get an insight into the reasoning behind the predictions provided by our models. Saliency maps are a visualisation tool that highlights the most important pixels of the image, meaning the ones that affect the most the decision made by the model. This can be done by computing the gradient of the loss function with respect to the pixels of the input image via backpropagation. The greatest the impact of a voxel on the prediction, the highest the pixel's intensity in the saliency map. Since the images are non-linearly registered to MNI152 standard space, this method can be used in cases where some voxels are present in more than one patches (e.g. for random patches or for the generation of a saliency map across subjects).

4.3 Results

4.3.1 Metrics

Numerous different experiments are implemented. For the evaluation of the age regression task in every model, MAE, and R^2 score are computed.

Table 4.1: Comparison of the performance of the 2D and 3D baseline experiments in terms of Mean Absolute Error (MAE) and R^2 score.

Experiment	MAE (years)	R^2 score
2D autoencoder	4.28	0.69
3D autoencoder	3.28	0.82
2D ResNet18	3.88	0.74
3D ResNet18	**2.64**	**0.87**

4.3.2 Implementation details

We train the network using backpropagation [159], with batch size 8 for 40 epochs. We use a mean squared error loss for the age regression. Furthermore, we use the adaptive moment estimation (Adam) optimiser [91] with an initial learning rate lr=0.0001, which is reduced by a factor of 10 every 15 epochs, and a weight decay of $\lambda = 0.0005$. The implementation of the network is done on an NVIDIA Titan RTX using Pytorch deep-learning library [143].

4.3.3 Experiments

Convolutional Autoencoder

For the 2D Convolutional Autoencoder we used the central transverse slice of the brain scans as input. The results were not promising as the MAE was around 4.28 years.

For the case of the 3D network, the whole brain scans were used as inputs. In this case, the results were improved compared to the case of the 2D slices of the brain as a MAE of 3.28 years and a R^2 score of 0.82 were observed. The results for the 2D and 3D convolutional autoencoders may be found in Table 4.1.

ResNet18

2D & 3D Baseline Models For the 2D network, we trained a ResNet18 network using as input only the central transverse slice of the brain scans in this case as well, as we wanted to check if only a slice of the brain would include enough information for the brain age estimation.

The network gave predictions with a MAE of around 3.88 years and a R^2 score of 0.74 (Table 4.1). Since the results were not that encouraging, we used the whole brain scans for the regression task.

Table 4.2: Comparison of the performance of the baseline model when using images with different registration as input in terms of Mean Absolute Error (MAE) R^2 score. Here, the baseline model refers to the 3D ResNet18 network.

Input images	MAE (years)	R^2 score
Non-rigid registration	**2.64**	**0.87**
Rigid registration	3.34	0.80
Affine registration	3.06	0.85

For the case of the 3D version of a ResNet18 model, the network that uses the whole brain scans as an input predicts accurate results with a MAE of around 2.64 years and a R^2 score of 0.87, as depicted in Figure 4.5 and Table 4.1. Comparing to [96], our model performs better to their simple 3D ResNet18 as we used a later release of the UKBB that includes a greater number of brain scans.

Apart from this, we also train the network with images after rigid, and affine registration to the MNI152 space used as inputs and compare the results with our baseline model. The results for the performance of the 3D ResNet18 model when using images with different registration as input can be found in Table 4.2.

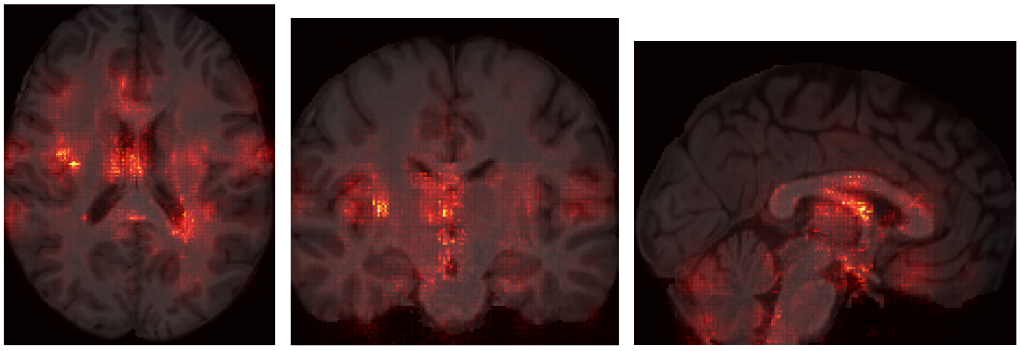
In both of the cases the network performs slightly worse, with a MAE of 3.34 years and a R^2 score of 0.80 for the rigid registration, and with a MAE of 3.06 years and a R^2 score of 0.85 for the affine registration one.

The model that was used as the baseline one was, therefore, the one that was trained on the non rigid to MNI152 space registered images as not only it performed better, but also is the one mostly used by literature since it is provided by UKBB.

For the case of the baseline model, bias correction was also applied. Bias correction led to the improvement of the results as a MAE of 2.43 years and a R^2 score of 0.8 were performed.

In addition, we computed the saliency maps in order to get a visualisation of the areas that played an important role in the prediction. In Figure 4.4, the central axial, coronal, and sagittal slices of the average of the saliency maps of the subjects of the test set can be seen.

Central Patch We wanted to check how the network would perform in case only the central part of the image was provided. For this reason, we trained two networks using only the 64x64x64 and 32x32x32 central patches of the brains, respectively.



(a) (b) (c)

Figure 4.4: Average of saliency maps of the subjects that belong in the test set in the axial (a), coronal (b), and sagittal (c), planes for the baseline model.

The results were interesting for the case of the 64x64x64 patch as the network performed similarly to when given the whole brain scan as input, and more specifically with a MAE of 2.74 and a R^2 score of 0.88.

For the 32x23x32 central patch, the results were not promising, as MAE was 3.64 years and R^2 score was 0.76.

Patch-based Prediction

When we use a 64x64x64 patch as input to a 3D ResNet18, the results are highly dependent on the area of the brain where the patch is extracted from. We train 27 ResNets for 27 patches, achieving MAE from around 2.5 to 4.2 years. The patches with the best performance are the ones that include the ventricles or the hippocampus, with MAE of less than 2.5 and 2.7 years, respectively, and R^2 scores of 0.80 and 0.76, respectively, as well as patches that include parts of these regions. On the contrary, patches that contain a lot of background perform much worse with a MAE of 3.5-4.2 years and a R^2 score of about 0.15-0.6. In Figure 4.6, numerous patches, coloured based on the performance of the corresponding models, are illustrated as part of a whole brain slice. The main results of the experiments can be found in Table 4.3.

Ensemble Method When we combine the single-patch models, we achieve an improved performance. Concerning the aggregation of the individual models, apart from the average of all the patches of the brain, we also calculate the average of the patches that perform better on the

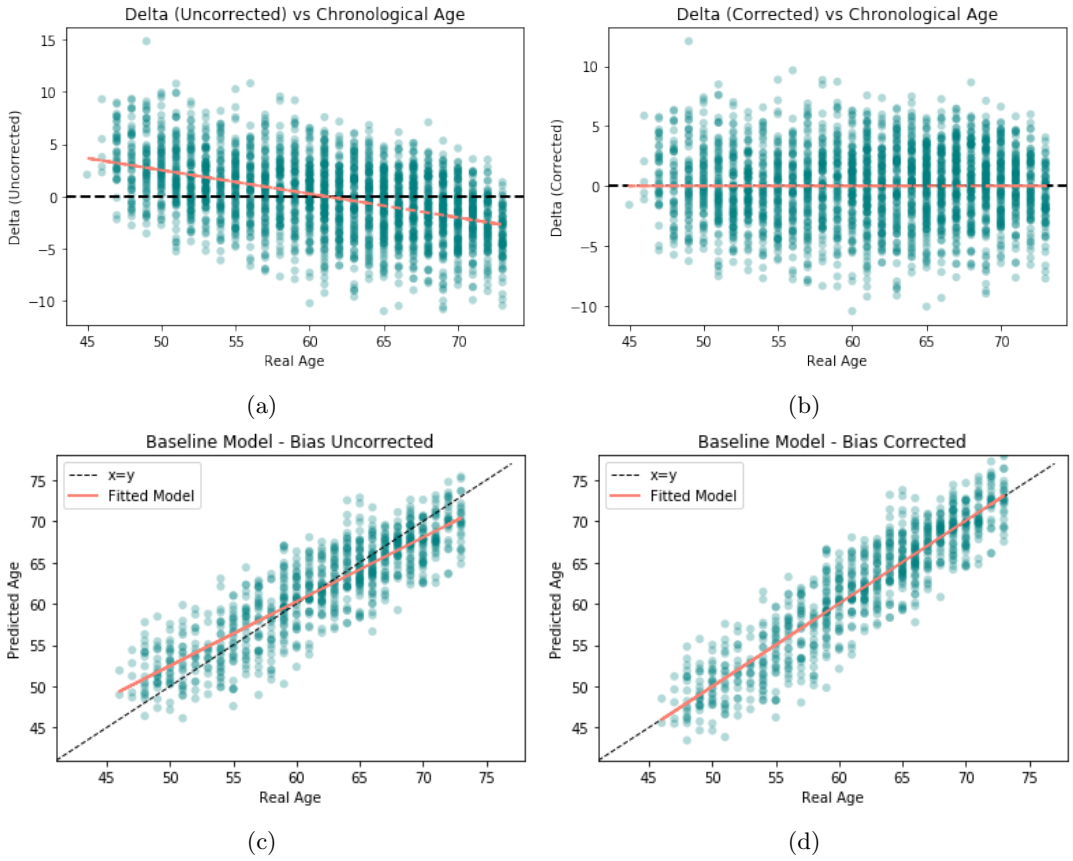
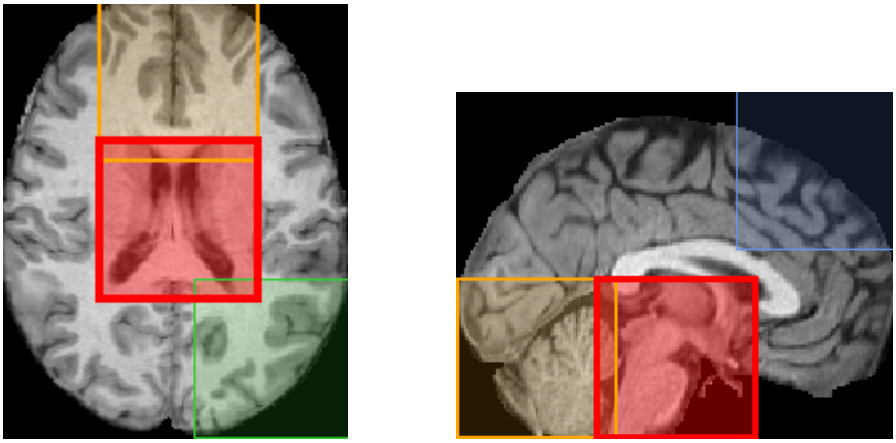


Figure 4.5: (a) Uncorrected brain age delta vs Chronological age for the baseline model, (b) Corrected brain age delta vs Chronological age for the baseline model, (c) The plot of the predicted over the real ages of the subjects before bias correction on the test set for the baseline model, (d) The plot of the predicted over the real ages of the subjects after bias correction on the test set for the baseline model.

task according to the performance on the validation set ($MAE < 3$ years), hence the ones that include important information on brain ageing. The MAE is 2.66 and 2.28 years, respectively, and the R^2 score 0.78 and 0.83, respectively.

The use of linear regression provides better results. The scatter plots of the real over predicted ages, before and after bias correction is applied, are shown in Figure 4.7. Specifically, we achieve a MAE of 2.13 years when all the single-patch models are used. Furthermore, when we use only the models of the patches with the most important information for the age of the brain, we achieve a MAE of 2.16 years. Regarding R^2 score, it is 0.85 in both cases. In this case, bias correction is applied as well and we achieve a MAE of 1.96 years, and an R^2 score of 0.87.



(a) (b)

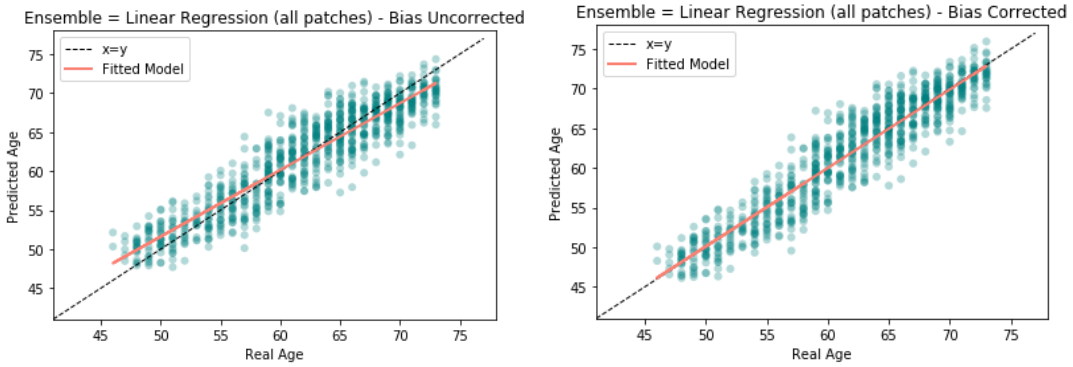
Figure 4.6: Slices of a brain MR image with three coloured patches in each case. The colours of the patches indicate the performance of this single-patch model in terms of Mean Absolute Error (MAE). Specifically: red patches indicate a MAE of less than 3 years, orange patches between 3 and 3.5 years, green patches between 3.5 and 4 years, and blue patches greater than 4 years. The patches that perform best include the ventricles (a), and the hippocampus (b).

Table 4.3: Performance comparison of the most important experiments for the 64x64x64 patches in terms of Mean Absolute Error (MAE) and R^2 score. Here, the baseline model refers to the 3D ResNet18 network.

Experiment	MAE (years)	R^2 score
Baseline model (uncorrected)	2.64	0.77
Best single-patch network	2.46	0.80
Worst single-patch network	4.19	0.14
Averaging (selected patches)	2.26	0.84
Linear Regression (all patches, uncorrected)	**2.13**	**0.85**
Baseline model (corrected)	2.43	0.80
Linear Regression (all patches, corrected)	**1.96**	**0.87**

32x32x32 Patches Regarding the results of the independent networks in the case of the 32x32x32 patches, the results varied with MAEs from 3.84 to 5.28 years. As it was expected, the patches that performed the worst were the ones including the background. The patch that performed the best was the one containing part of the ventricles. A slice of the 64x64x64 patch along with the smaller 32x32x32 patch that performed the best can be found in Figure 4.8.

When combining the results of the different patches with the ensemble method that uses linear regression described in the previous section, there was a slight improvement of the prediction with a MAE of 3.49 years and an R^2 score of 0.59.



(a) (b)

Figure 4.7: The predicted over the real age of the subjects when the predicted ages result from a linear regression of all the patches before bias correction (a), and after bias correction (b). The black dashed line shows the line of identity ($x = y$), while the orange solid one corresponds to the fitted model.

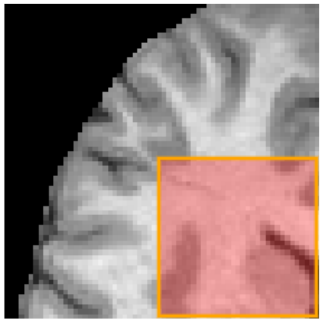
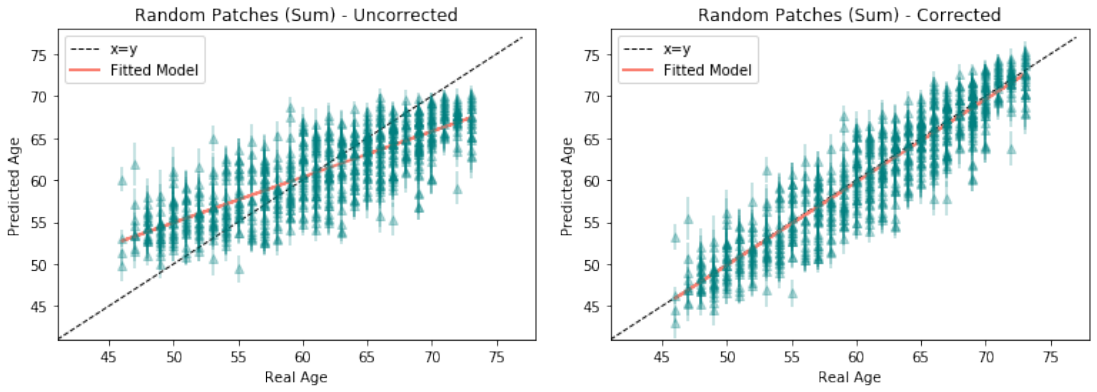


Figure 4.8: Slice of a brain MR 64x64x64 patch with the 32x32x32 patch that performed the best coloured orange. The patch that performed the best includes part of the ventricles and reached a MAE of 3.84 years.

Although the pipeline described here provides some kind of more detailed localisation, the procedure can not be easily scaled in order to examine the importance of the regions of the whole brain.

Random Patches For the case of the random 64x64x64 patches, the average of the predictions of all the patches extracted by the same subject was also calculated in order to examine whether the combination of different patches would improve the results.

Regarding the results, the performance decreased with the decrease of the patch size. The MAE for the large 64x64x64 patch was 3.70 years and the R^2 score was 0.58. When the predic-



(a) (b)

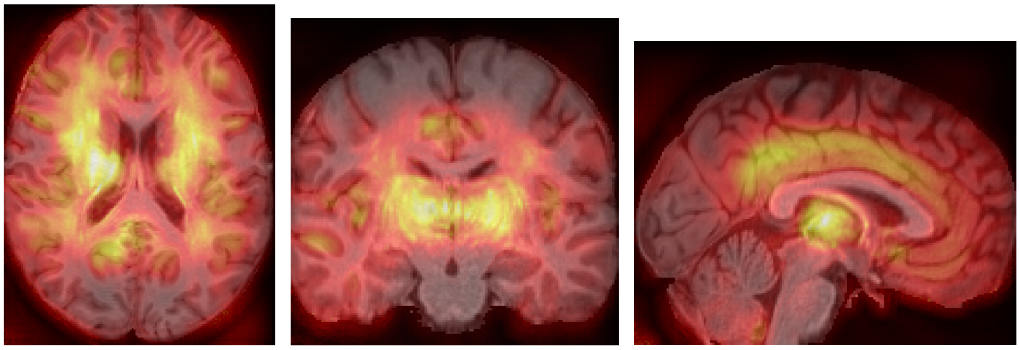
Figure 4.9: The predicted over the real age of the subjects when the predicted ages result from the sum of the predictions of the different patches extracted for every subject before bias correction (a), and after bias correction (b). The black dashed line shows the line of identity ($x = y$), while the orange solid one corresponds to the fitted model.

tions of the different patches were averaged, the results indeed improved as a MAE of 3.56 and 2.33 years and a R^2 score of 0.65 and 0.83, before and after bias correction respectively, were achieved. The MAE along with the standard deviation (std) of the different predictions of every subject can be seed in Figure 4.9, while in Figure 4.10 the average of the saliency maps for the patches of the test set can be found.

In addition, we estimate the average of the MAE of the different patches for every pixel. In other words, we calculate the number of patches that include a pixel, and we average their different predictions. We visualise the results in heatmaps for every fifth slice in the three planes in Figure 4.11.

For the 32x32x32 patches, the network performed worse with a MAE of 4.90 years and a R^2 score of 0.3, while for the small 16x16x16 patches the network did not manage to be trained at all.

In the case of 32x32x32 patches, we also estimate and create the heatmaps presented in Figure 4.12, following the same procedure mentioned for the larger patches.



(a) (b) (c)

Figure 4.10: Average of saliency maps of the subjects that belong in the test set in the axial (a), coronal (b), and sagittal (c), planes for the model that was trained on random patches.

4.4 Discussion

A plethora of deep learning methods based on CNNs which use the whole brain images as input, have been used for estimating brain age via regression, achieving good performance (MAE of around 2.6 years) [96].

Firstly, we decided to explore how autoencoders perform in this task and created a network that not only tries to find robust features for the reconstruction, but also predicts the age of the subject. We explored both the use of 2D and 3D inputs. The results were not comparable to the state-of-the-art and the training was slower than the ResNet18. Therefore, we utilised the model described in [96] as our baseline model.

For the experiments that used the ResNet18, different kind of registrations were used as input. The images that seemed to include the most useful information for the task, were the ones with the non rigid registration. These images are also the ones used for the relevant literature. Therefore, these images are the ones used for our baseline model.

We have also explored how patches of the brain can be used to perform this task. Firstly, we train a network for each patch that tries to find robust but patch-specific features for the age estimation. In addition, we combine the results of the different networks via an ensemble approach with the purpose of a more accurate estimation.

Based on the results, we conclude that the patches that include parts of the ventricles and the hippocampus tend to predict the age of the subject more accurately with a MAE of less than

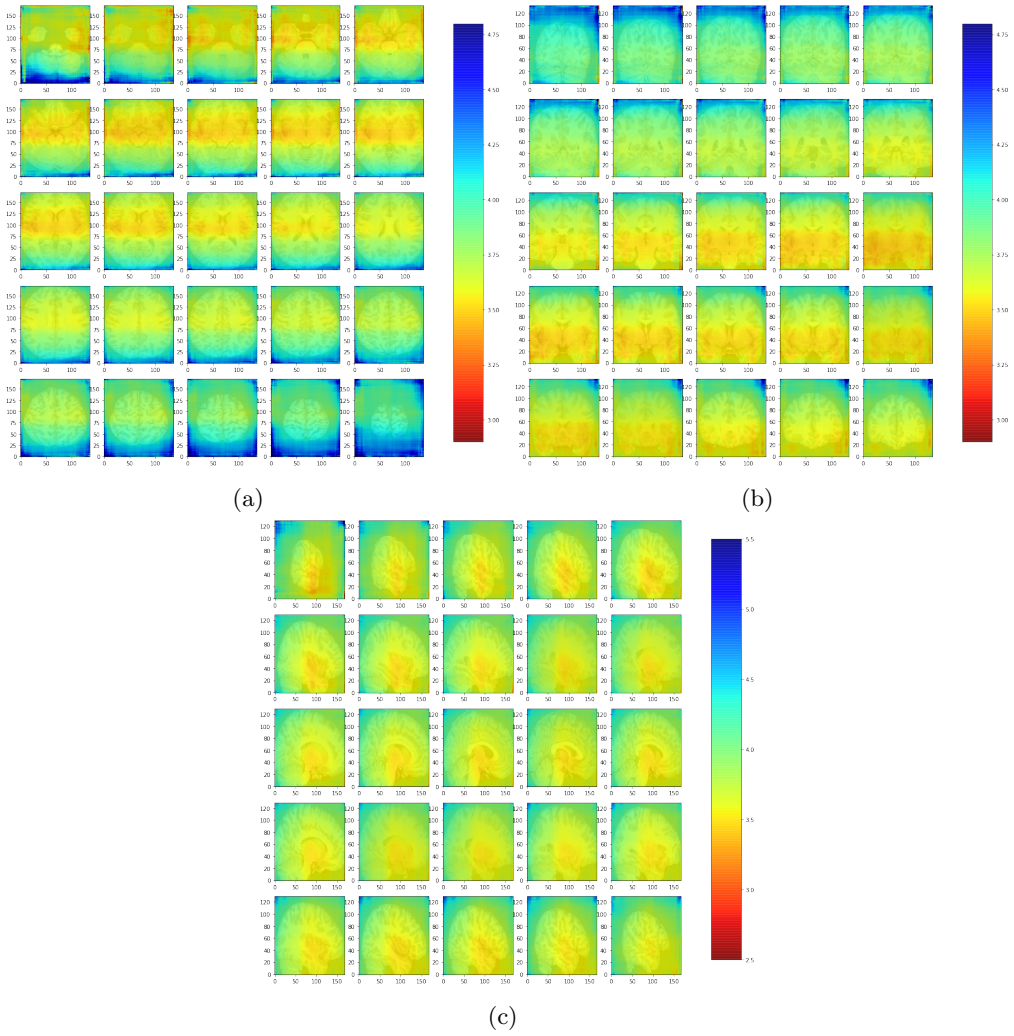


Figure 4.11: Average of the MAE of the predictions of the different 64x64x64 patches for the case of every pixel of the test set. Every fifth slice is presented in the axial (a), coronal (b), and sagittal (c), views.

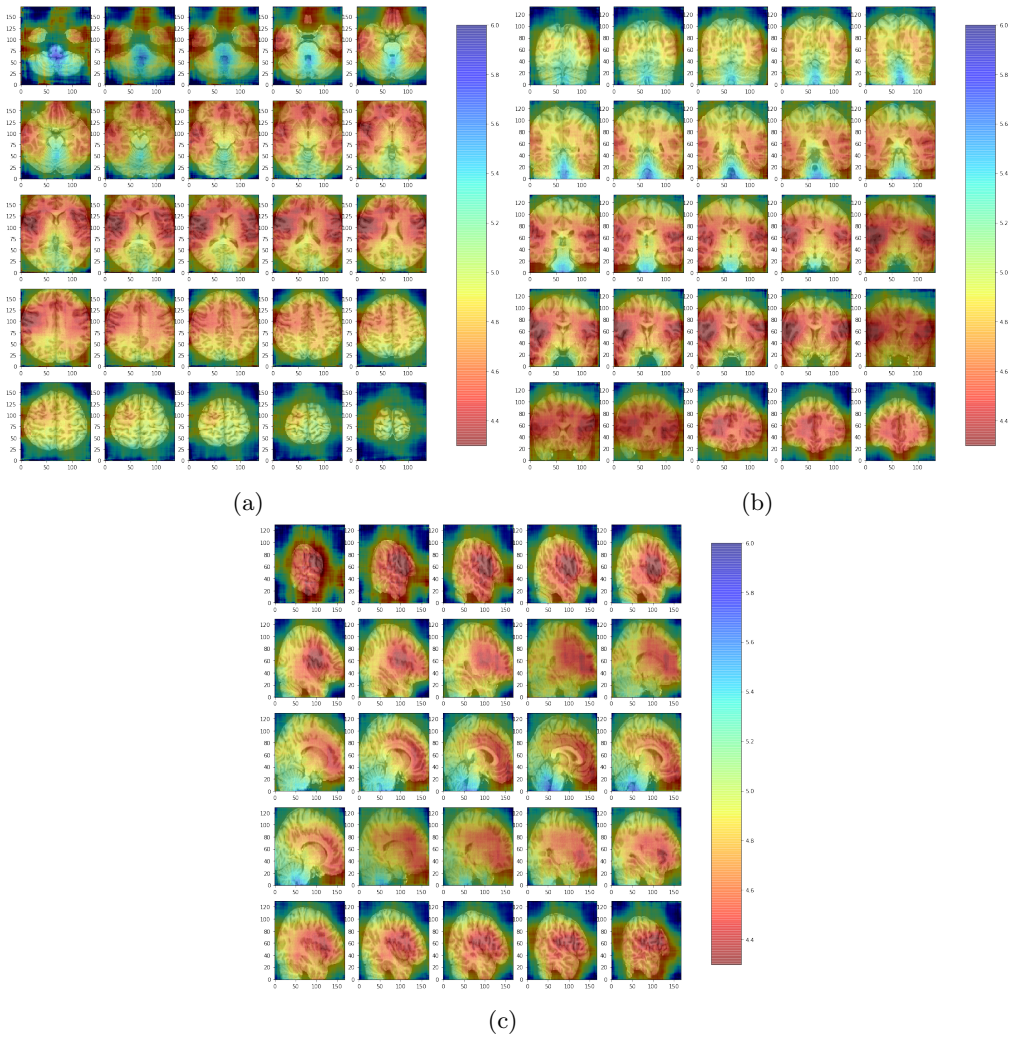


Figure 4.12: Average of the MAE of the predictions of the different 32x32x32 patches for the case of every pixel of the test set. Every fifth slice is presented in the axial (a), coronal (b), and sagittal (c), views.

3 years, thus our results are consistent with the existing literature [161]. The MAE in some of these cases is lower than that for the baseline network which utilises the volume of the whole brain, proving that not all parts of the brain are needed for an accurate prediction. Therefore, less computational effort is required for comparable results. On the other hand, patches that include part of the background do not perform as accurately. We observe that although the brain shrinks while age increases, the networks do not manage to capture the changes in the cerebral cortex as accurately. An explanation for this could be the way the images are registered, which reduces size differences between brains, or the relatively small range of existing ages in the dataset, especially of ages beyond 65 years when atrophy tends to accelerate [39]. Furthermore, the use of patches provides a visualisation of the parts of the brain that are valuable for age prediction, leading to a more understandable prediction. Although the patches are big and do not provide detailed localisation, this work is the first step towards an interpretable and localised brain age estimation. More localised detection of a deviation from normal trajectories may also be useful for the automatic detection of more focal pathologies such as ventricular enlargement in hydrocephalus or hippocampal atrophy in temporal lobe epilepsies [89].

We also improve the age prediction by combining the results of a few weak networks. Firstly, we do so by simply averaging the results of all the independent networks. However, this does not lead to any improvement on the predictions (MAE = 2.66 years). Thus, we then leverage the knowledge acquired from the previous experiment and average only the predictions from patches that seemed to include the most relevant information about brain age. There is a major boost in the regression performance (MAE = 2.26 years), showing that by combining results of weaker networks we can achieve a better prediction. Moreover, we aggregate the single-patch models by using a linear regression, as we believe that by giving appropriate weights to the results of every patch, better results could be achieved. Interestingly, in this case, using all the individual models performs similarly (MAE = 2.13 years) to when we use the selected ones (MAE = 2.16 years). The reason is, probably, that by learning the weights, the linear regression model can suppress the effect of the least accurately performing models, while promoting the best performing ones. This results in an age estimator which is more accurate than the baseline model as well as the existing literature.

After bias correction, we notice a considerable increase of the performance of the network

(MAE = 1.96 years). Although using the chronological age as a covariate for the bias correction is not meaningful for a ML model, it could prove important clinically for giving insights for the subject's brain age, as demographics, including chronological age, are usually known.

As our next step, we extracted smaller patches, sized 32x32x32, from the patches that perform better in order to provide even more localised predictions. In this case as well, there seem to be more localised insights as the patches that performed the best were again the ones that included part of the ventricles. However, the method described is not scalable and cannot be applied in a clinical environment.

For this reason, the extraction of random patches that were fed into a single network was explored. Although the network trained or the network patches perform worse than the network of our previous experiment, this method is scalabe.

Comparing the saliency maps created for the baseline model (Figure 4.4), and the model trained on 64x64x64 patches (Figure 4.10), we notice that similar areas seem to be the most important ones. However, for the case of the random patches, the biggest part of the brain has some kind of role in the decision, which is logical since the network learns the local features includes in the various patches and predicts based on these.

From the heatmaps created, we notice that the patches that perform better in the age estimation task are roughly the ones expected from the literature [161]. Although the large 64x64x64 random patches (Figure 4.11) give more accurate results than the 32x32x32 ones (Figure 4.12), they do not provide that localised predictions.

4.5 Summary

In this chapter, we propose the use of patches of the brain, instead of the whole 3D brain scan, as input to separate 3D ResNets for brain age estimation. We explore the use of both patches of a grid and random patches. We combine the results of the networks with an ensemble method, such as linear regression. The method provides accurate results outperforming the state-of-the-art. In addition, we provide localised predictions by specifying the patches that perform better on the task. Precise age prediction in healthy participants, e.g. from UKBB as in this study, can now be used as a baseline to compare trajectories in different pathologies such as Alzheimer's

[39].

A limitation of this work is that the patches used are quite big, thus the extracted maps are coarse. When we use smaller patches, in order to end up to more detailed maps, we increase highly the computational and time costs, since the number of patches increases significantly. In addition, since the pipeline is not trained end-to-end, it is not easily scalable. For this reason, in the next chapter (Chapter 5) we will look more into interpretability techniques for CNNs with the purpose of extracting more detailed attention maps for the task of brain age estimation.

Chapter 5

Voxel-level Importance Maps for Interpretable Brain Age Estimation

In the previous chapter, we focused on the extraction of attention maps that highlight the brain regions that are the most relevant to brain age. However, since the patches used were quite big, the extracted attention maps were coarse. On the other hand, when the use of smaller patches was explored, we noticed that the method is not easily scalable due to computational costs.

In this chapter, we still maintain our attention in the interpretability aspect of Convolutional Neural Networks (CNNs). We work on saliency maps, which have been used to a great extent to provide an indication on the parts of the image that a neural network's prediction is based on.

We focus on the brain age regression task, but most interpretability methods are working along with classification tasks and cannot be directly transferred to regression tasks. For this reason, in this chapter, we work on the task of brain age regression from 3D brain Magnetic Resonance (MR) images using a CNN, and we interpret its predictions by extracting importance maps, using a noise model. By averaging the importance maps of the subjects we end up with a population-based importance map, which displays the regions of the brain that are influential

for the task.

The findings of this chapter were presented and published in [22]:

Bintsi, K., Baltatzis, V., Hammers, A., and Rueckert, D.: Voxel-level importance maps for interpretable brain age estimation. In: Reyes, M., Abreu, P.H., Cardoso, J.S., et al. (eds.) Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data - 4th International Workshop, iMIMIC 2021, and 1st International Workshop, TDA4MedicalData 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings. Lecture Notes in Computer Science, vol. 12929, pp. 65–74. Springer (2021). doi: 10.1007/978-3-030-87444-5_7, https://doi.org/10.1007/978-3-030-87444-5_7

5.1 Introduction

Alzheimer's disease (AD) is the most common cause of dementia [5]. AD leads to the atrophy of the brain more quickly than healthy ageing and is a progressive neurodegenerative disease, meaning that more and more parts of the brain are damaged over time. The atrophy primarily appears in brain regions such as hippocampus, and it afterwards progresses to the cerebral neocortex. At the same time, the ventricles of the brain as well as cisterns, which are outside of the brain, enlarge [161].

Healthy ageing also results in changing of the brain, following specific patterns [2]. Therefore, a possible biomarker used in AD is the estimation of the brain (biological) age of a subject which can then be compared with the subject's real (chronological) age [51]. People at increased risk can be identified by the deviation between these two ages and early computer-aided detection of possible patients with neurodegenerative disorders can be accomplished. For this reason, a large body of research has focused on estimating brain age, especially using Magnetic Resonance (MR) images, which have long been used successfully in the measurement of brain changes related to age [59].

Recently, deep learning models have proved to be successful on the task of brain age estimation, providing relatively high accuracy. They are designed to find correlation and patterns in the input data and, in the supervised learning setting, associate that with a label, which in our

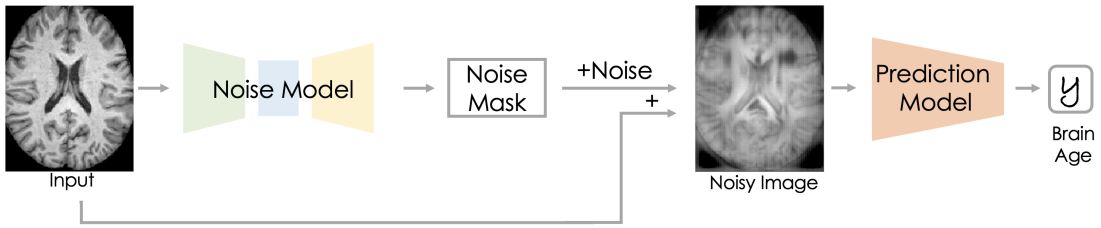


Figure 5.1: Overview of the proposed idea. A noise model is trained with the purpose of adding as much noise as possible to the input. The output is a noise mask, in which noise sampled from the standard normal distribution is added. The result is then added to the input image and is used as an input to a pretrained prediction model with frozen parameters. The aim is to create a noisy image that will maximise the noise level while also not harming the performance of the prediction model.

case is the age of the subject. The models are trained on a dataset of MR images of healthy brains, estimating the expected chronological age of the subject. During training, the difference of the chronological age and the predicted age needs to be as small as possible. During test time, an estimated brain age larger than the subject's chronological age indicates accelerated ageing, thus pointing to a possible AD patient.

Convolutional Neural Networks (CNNs) are used with the purpose of an accurate brain age estimation while using the minimum amount of domain information since they can process raw image data with little to no preprocessing required. Many studies provide very accurate results, with Mean Absolute Error (MAE) as low as around 2.2 years [147, 141, 118]. However, most of these approaches purely focus on minimisation of the prediction error while considering the network as a black box. Recent studies have started to try to identify which regions are most informative for age prediction. For example, in Chapter 4 and in [23] we provided an age prediction for every patch of the brain instead of whole brain. Although the predictions and results presented in [23] were promising, and the localisation was meaningful, the use of large patches meant that the localisation was not very precise. Similar approaches has been explored, such as [11] in which slices of the brain were used instead of patches. In [149], the authors provided an age estimation per voxel but the accuracy of the voxel-wise estimations dropped significantly. In [108], an ensemble of networks and gradient-based techniques [172] were used in order to provide population-based explanation maps. Finally, 3D explanation maps were provided in [15], but the authors used image translation techniques and generative models, such

as VAE-GAN networks.

In computer vision, a common approach to investigate black-box models such as CNNs is to use saliency maps, which show which regions the network focuses on, in order to make the prediction. An overview of the existing techniques for explainability of deep learning models in medical image analysis may be found in [171]. Gradient-based techniques, such as guided backpropagation [175], and Grad-CAM [165], make their conclusions based on the gradients of the networks with respect to a given input. Grad-CAM is one of the most extensively used approaches and usually results in meaningful but very coarse saliency maps. Gradient-based techniques are focused on classification and to our knowledge they do not work as expected for regression task because they detect the features that contributed mostly to the prediction of a specific class. Occlusion-based techniques [195] have also been widely explored in the literature and they can be used both for classification and regression tasks. The idea behind occlusion techniques is very simple: The performance of the network is explored after hiding different patches of the images, with the purpose of finding the ones that affect the performance the most. It is a promising and straightforward approach but bigger patches provide coarse results. On the other hand, the smaller the patches, the greater the computational and time constraints are, which can be a burden in their application.

A recent approach which leverages the advantages of occlusion techniques while also keeping computational and time costs relatively low is U-noise [93] which uses the addition of noise in the images pixel-wise, while keeping the performance of the network unchanged with the purpose of understanding where the deep learning models focus in order to make their predictions. The authors created interpretability maps for the task of pancreas segmentation using as input 2D images using noise image occlusion. In this chapter, inspired by the work described above [93] and the idea that when a voxel is not important for the task, then the addition of noise on this specific voxel will not affect the performance of our network, we make the following contributions: 1) We adapt the architecture of U-noise (Figure 5.1), which was originally used for pancreas segmentation, for the task of brain age regression and visualise the parts of the brain that played the most important role for the prediction by training a noise model that dynamically adds noise to the image voxel-wise, providing localised and fine-grained importance maps. 2) We extend the U-noise architecture to 3D instead of 2D to accommodate training with three-

dimensional volumetric MR images; 3) We propose the use of an autoencoder-based pretraining step on a reconstruction task to facilitate faster convergence of the noise model; 4) We provide a population-based importance map which is generated by aggregating subject-specific importance maps and highlights the regions of the brain that are important for healthy brain ageing.

5.2 Materials & Methods

5.2.1 Dataset & Preprocessing

We use the UK Biobank (UKBB) dataset [3] for estimating brain age and extracting the importance maps. UKBB contains a broad collection of brain MR images, such as T1-weighted, and T2-weighted. The dataset we use in this work includes T1-weighted brain MR images from around 15,000 subjects. The images are provided by UKBB skull-striped and non-linearly registered to the MNI152 space. After removing the subjects lacking the information of age, we end up with 13,750 subjects with ages ranging from 44 to 73 years old, 52,3% of whom are females and 47.7% are males. The brain MR images have a size of 182x218x182 but we resize the 3D images to 140x176x140 to remove a large part of the background and at the same time address the memory limitations that arise from the use of 3D data, and normalise them to zero mean and unit variance.

5.2.2 Brain Age Estimation

We firstly train a CNN, the prediction model, f_θ, with parameters θ, for the task of brain age regression. We use the 3D brain MR images as input to 3D ResNet-18, similar to the one used in [23], which uses 3D convolutional layers instead of 2D [67]. The network is trained with a Mean Squared Error (MSE) loss and its output is a scalar representing the predicted age of the subject in years.

5.2.3 Localisation

An overview of the proposed pipeline is shown in Figure 5.1.

U-Net Pretraining

The prediction model and the noise model have identical architectures (2D U-Net [156]) in [93], as they are both used for image-to-image tasks, which are segmentation and noise mask generation, respectively. Therefore, in that case, the noise model is initialised with the weights of the pretrained prediction model. However, in our case, the main prediction task is not an image-to-image task but rather a regression task and thus, the prediction model's (a 3D ResNet as described above) weights cannot be used to initialise the noise model. Instead, we propose to initially use the noise model f_ψ with parameters ψ, which has the architecture of a 3D U-Net as a reconstruction model, for the task of brain image reconstruction. By pretraining the noise model with a reconstruction task before using it for the importance map extraction task, we facilitate and accelerate the training of our U-noise model, since the network has already learnt features relating to the structure of the brain.

The reconstruction model consists of an encoder and a decoder part. It uses as input the 3D MR volumes and its task is to reconstruct the volumes as well as possible. It is trained with a voxel-wise MSE loss.

Brain Age Importance Map Extraction

For the extraction of the importance maps, we extend the U-noise architecture [93] to 3D. The idea behind U-noise is that if one voxel is important for the prediction task, in our case brain age regression, the addition of noise on it will harm the performance of the prediction model. On the other hand, if a voxel is not relevant for the task, the addition of noise will not affect the performance.

More specifically, after the pretraining phase with the reconstruction loss, the 3D U-Net, i.e. noise model, is used for the task of extraction of a 3D noise mask that provides a noise level for every voxel of the input 3D image. A sigmoid function is applied so the values of the mask are between 0 and 1. The values are scaled to $[v_{min}, v_{max}]$, where v_{min}, v_{max} are hyperparameters. The rescaled mask is then multiplied by $\epsilon \sim \mathcal{N}(0, 1)$, which is sampled from the standard normal distribution and the output is added to the input image element-wise in order to extract the

noisy image. Given an image X, its noisy version can be given by Equation (5.1):

$$X_{noisy} = X + f_{\psi}(X)(v_{max} - v_{min})\epsilon + v_{min}, \tag{5.1}$$

We then use the noisy image as input for the already trained prediction model with frozen weights, and see how it affects its performance. The purpose is to maximise the noise level in our mask, while simultaneously keeping the performance of our prediction model as high as possible. In order to achieve that, we use a loss function with two terms, the noise term, which is given by $-log(f_{\psi}(X))$ and motivates the addition of noise for every voxel, and the prediction term, which is an MSE loss whose purpose is to keep the prediction model unchanged. The logarithm is applied to the noise term to penalise the lack of noise more heavily and ensure a smoother and more stable convergence. The two loss terms are combined with a weighted sum, which is regulated by the ratio hyperparameter r. It is important to note that at this stage the parameters, θ, of the prediction model, f_{θ}, are frozen and it is not being trained. Instead, the loss function $\mathcal{L}$, is driving the training of the noise model f_{ψ}. Given an input image X and label y, the loss function $\mathcal{L}$ takes the form shown in Equation (5.2)

$$\mathcal{L} = \underbrace{(f_{\theta}(X) - y)^2}_{\text{prediction term}} - r \underbrace{log(f_{\psi}(X))}_{\text{noise term}} \tag{5.2}$$

The values v_{min}, v_{max}, as well as r are hyperparameters and are decided based on the performance of our noise model on the validation set.

5.3 Results

From the 13,750 3D brain images, 75% are used for the training set, 10% for the validation set and 15% for the test set. All the networks are trained with backpropagation [159] and adaptive moment estimation (Adam) optimiser [91] with initial learning rate lr=0.0001, reduced by a factor of 10 every 10 epochs. The experiments are implemented on an NVIDIA Titan RTX using the Pytorch deep-learning library [143].

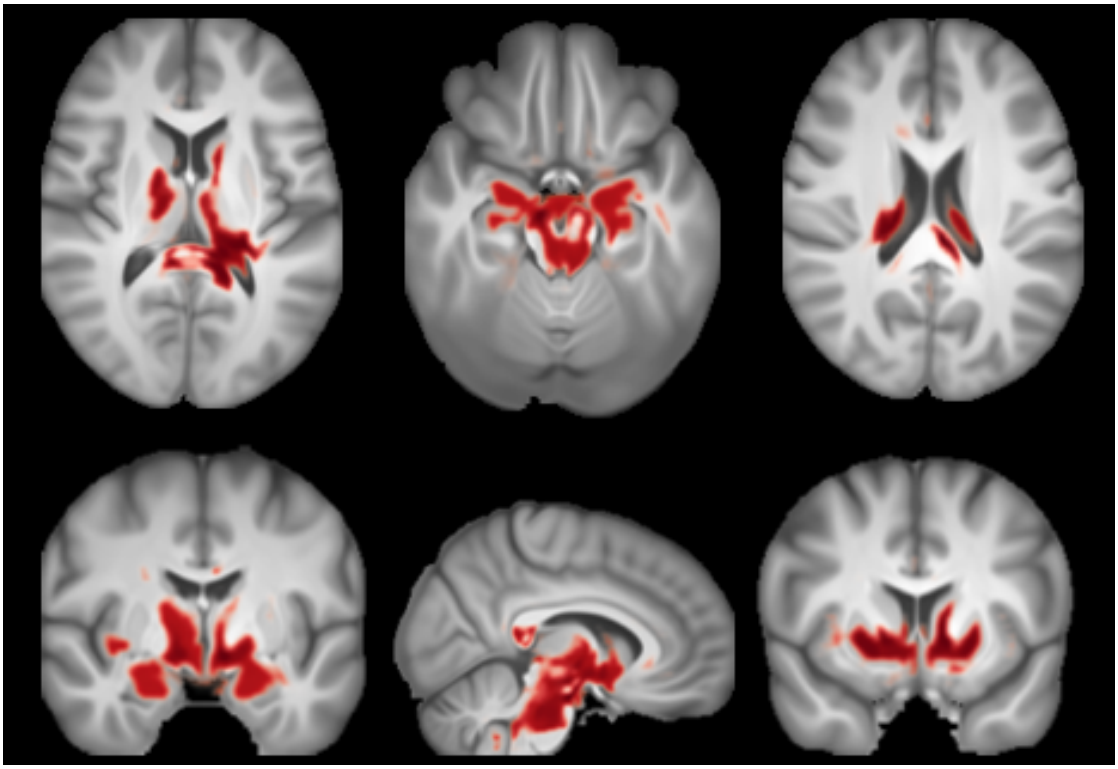


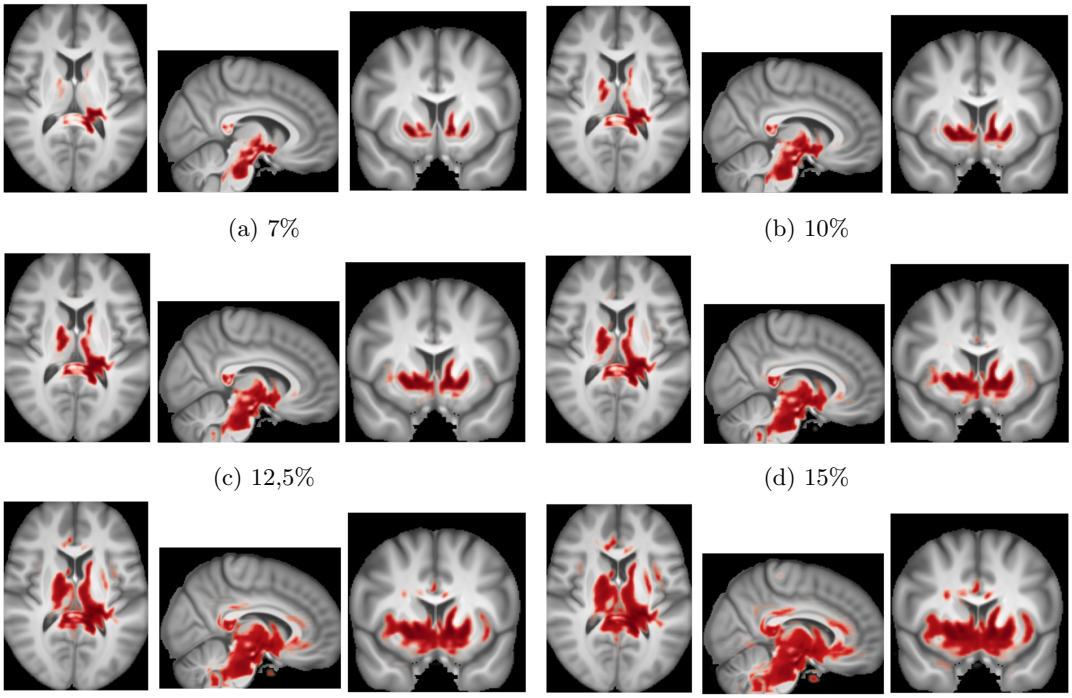
Figure 5.2: Six different slices of the population-based importance map on top of the normalised average of the brain MR Images of the test set. The most important parts of the brain for the model's prediction are highlighted in red. The most significant regions for the task of brain ageing are mesial temporal structures including the hippocampus, brainstem, periventricular and central areas. The results are in agreement with the relevant literature and previous studies.

5.3.1 Age Estimation

We train the prediction model for 40 epochs using a batch size of 8 using backpropagation. We use MSE between the chronological and biological age of the subject as a loss function. The model achieves a MAE of about 2.4 years on the test set.

5.3.2 Population-based importance maps

The noise model is trained for 50 epochs and with a batch size of 2, on four GPUs in parallel. The number of blocks used, meaning the number of downsample and upsample layers, is 5, while the number of output channels after the first layer is 16. Different values were tested for hyperparameters v_{min}, v_{max} and r. The chosen values, for which the network performed the



(a) 7% (b) 10%

(c) 12,5% (d) 15%

(e) 20% (f) 25%

Figure 5.3: Importance maps after keeping different percentage of important voxels.

best in the validation set, are $v_{min} = 1$, $v_{max} = 5$ and $r = 0.1$. We average the importance maps for all the subjects of the test set, ending up with a population-based importance map. We use a threshold in order to keep only the top 12.5% of the voxels with the lowest tolerated noise levels, meaning the most important ones for brain ageing. We then apply a gaussian smoothing filter with a kernel value of 1. Different slices of the 3D population-based importance map are shown in Figure 5.2. As it can be seen, the areas that are the most relevant for brain ageing according to the model's predictions are mesial temporal structures including the hippocampus, brainstem, periventricular and central areas.

5.3.3 Ablation tests

Percentage of important voxels

We perform an ablation test in order to examine how the choice of the percentage of pixels kept affects the importance maps. Slices of the 3D importance maps with different percentages of

important voxels kept, can be seen in Figure 5.3.

As it can be seen, when keeping only 7% of the voxels, the areas highlighted are correct. However, there is some information lost. On the other hand, when keeping 25%, the importance maps are quite coarse. The balance is important, with 12.5% to 15% of voxels kept resulting in localised importance maps that highlight all the important regions.

U-Net Depth & Pretraining

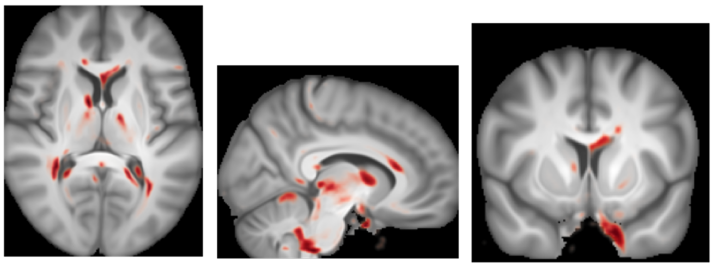
For the case of age regression, we perform an analysis of how the depth of the U-Net used for the Noise Model affects the quality of the extracted maps. For this reason, we train U-Net models with different depths as Noise Models, with depth meaning the number of upsample and downsample layers. We try three different networks, small Noise Model, with depth 3, medium Noise Model, with depth 4, and large Noise Model, which was the one presented in Section 5.3.2, with depth 5. The maps extracted from these cases are shown in Figure 5.4. For all cases, 25% of the most important pixels are kept and visualised.

Regarding the pretraining of the U-Net, the importance maps for both cases are shown in Figure 5.4. It is important to mention here, than a different percentage of important voxels is kept for the pretrained model. More specifically, 25% of the voxels is kept for the small, medium, and large Noise Models, while 12.5% is kept for the pretrained large one. This is due to the fact that models without pretraining did not retain any meaningful information at lower percentages.

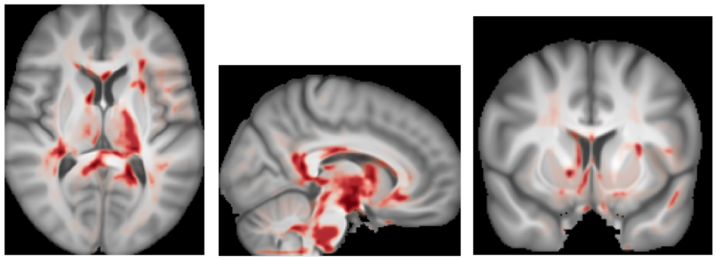
As it can be noticed from Figure 5.4, all of the models work well enough. However, the small model seems to lose some regions that are considered relevant to brain age, while the medium one provides a more generalised mask of relevant regions. The large models, both the not pre-trained and the pre-trained one, provide more localised information without missing important parts, with the pre-trained one being slightly more detailed, as it also gives some attention to the cerebral cortex of the brain.

5.3.4 Comparison with other interpretability methods

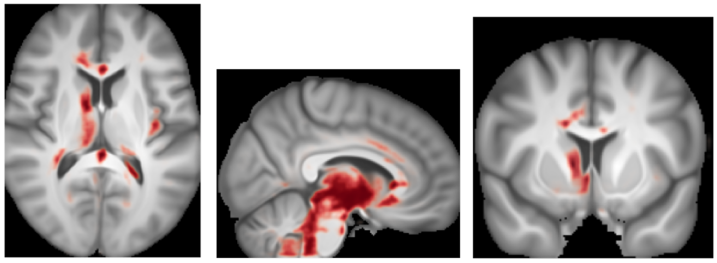
To compare our method with other interpretability techniques, we explored the population-based saliency maps extracted from our utility model using the interpretability techniques that are most



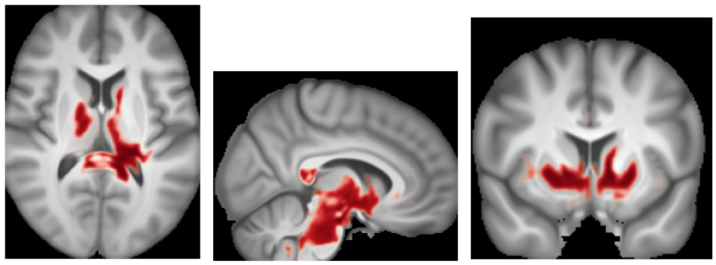
(a) U-Net Small.



(b) U-Net Medium.



(c) U-Net Large.



(d) U-Net Large - Pretrained.

Figure 5.4: Importance maps for the cases of different depths of the Noise model, Small (depth=2), Medium (depth=3), Large (depth=4). For the Large model, the case of the Pretrained Large model is also presented.

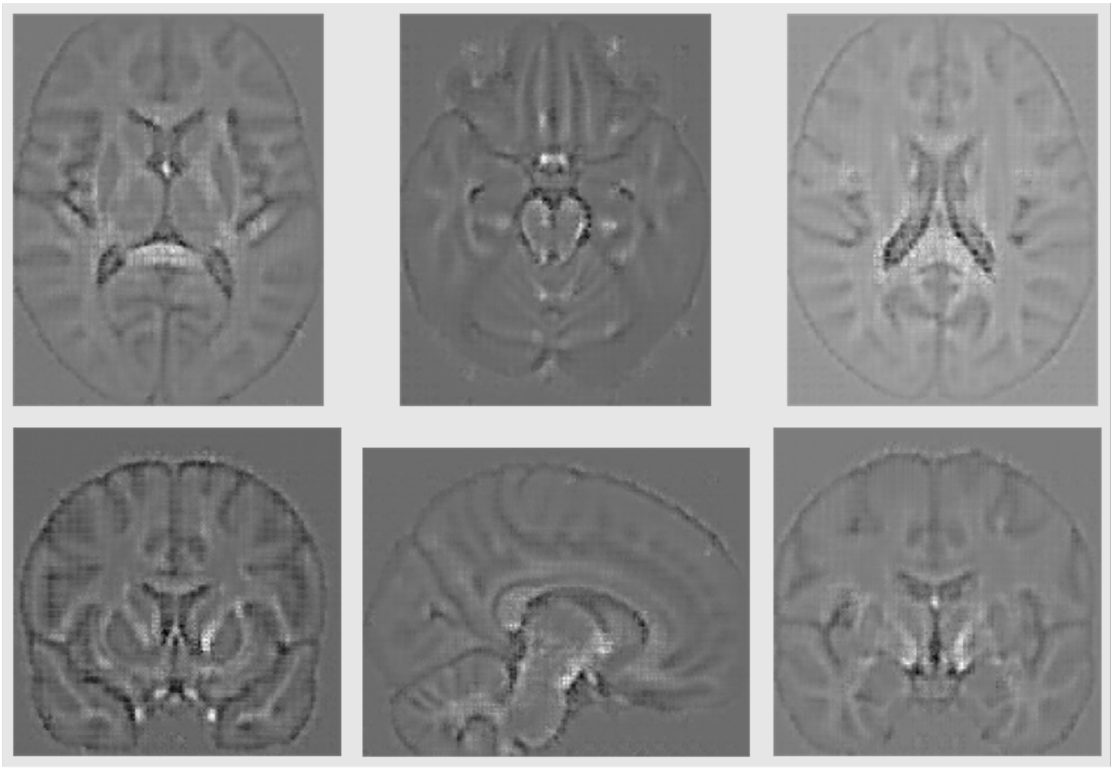


Figure 5.5: Six different slices of the population-based saliency maps using gradient backpropagation. We note that here, only the saliency maps are presented, and there is no overlay of the average brain MR Images of the test set.

commonly used in the literature, namely Grad-CAM [165], and gradient backpropagation [175]. We used the M3d-CAM library [61] for the extraction of the maps. Here, the average of all the subjects of the test set is calculated so the comparison with our population-based importance maps is possible.

In Figures 5.5 and 5.6, the saliency maps extracted using gradient backpropagation and Grad-CAM respectively, may be found. As expected, the saliency maps extracted using the Grad-CAM are not very meaningful, probably because Grad-CAM is designed for classification and cannot be directly applied to regression. On the other hand, gradient backpropagation produces reasonable, but not very informative maps. It is clear that our method (Figure 5.2), produced more meaningful and informative importance maps.

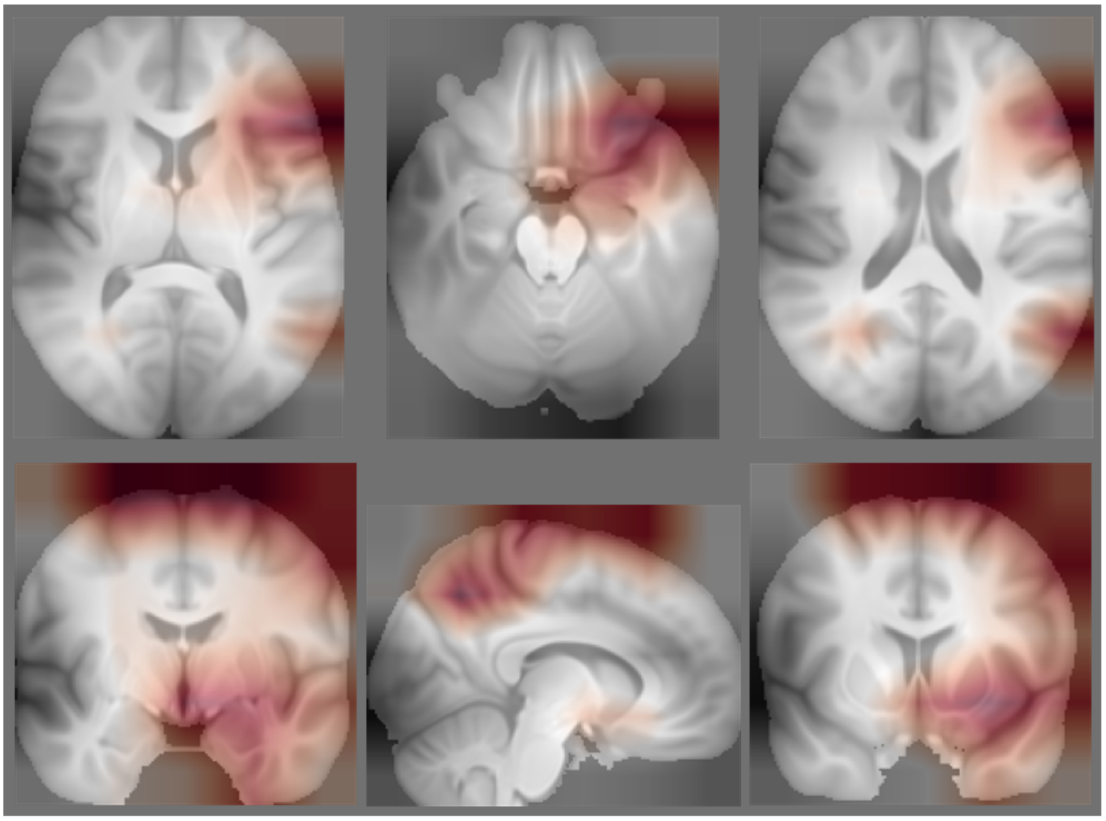


Figure 5.6: Six different slices of the population-based saliency maps using Grad-CAM on top of the normalised average of the brain MR Images of the test set.

5.4 Discussion

Understanding the logic behind a model's decision is very important in assessing trust and therefore in the adoption of the model by the users [57]. For instance, the users should be assured that correct predictions are an outcome of proper problem representation and not of the mistaken use of artifacts. For this reason, some sort of interpretation may be essential in some cases. In the medical domain [72], the ability of a model to explain the reasoning behind its predictions is an even more important property of a model, as crucial decisions about patient management may need to be made based on its predictions.

In this work, we explored which parts of the brain are important for ageing. In order to do so, we made the assumption that unimportant voxels/parts of the brain are not useful for brain age estimation and are not utilised by the prediction model. We trained a prediction model,

which accurately estimated brain ageing, and a noise model, whose purpose is to increase the noise in the input images voxel-wise, while also keeping the performance of the prediction model unaffected. As can be seen from Figure 5.2, our importance maps are in agreement with the existing neuropathology literature [161]. More specifically, it is shown that the hippocampus and parts of the ventricles are where the prediction model focuses to make its decisions.

On the other hand, the differences in the cerebral cortex appear to not be getting captured by the network. In our understanding, there are two reasons behind this. Firstly, the age range of the subjects (44-73 years old) is not large enough for the network to make conclusions. At the same time, the changes in the cerebral cortex are more noticeable after the age of 65 years. In our case, we probably do not have enough subjects in that age range in order to facilitate the network into capturing these differences.

The images that have been used in this study are non-linearly registered, since UKBB provides them ready for use and in the literature more works use the provided preprocessed dataset and therefore, comparison is easier to be done. However, it has been noticed that using non-linearly registered images may lead to the network's missing of subtle changes away from the ventricles, such as cortical changes [45].

5.5 Summary

In this chapter, we extend the use of U-noise [93] for 3D inputs and brain age regression. We use 3D brain MR images to train a prediction model for brain age and we investigate the parts of the brain that play the most important role for this prediction. In order to do so, we implement a noise model, which aims to add as much noise as possible in the input image, without affecting the performance of the prediction model. We then localise the most important regions for the task, by finding the voxels that are the least tolerant to the addition of noise, which for the task of brain age estimation are mesial temporal structures including the hippocampus and periventricular areas.

A limitation of this work, and in general of CNNs, is that they only focus on the image of the subject, without taking into account the associations and the relationships among the subjects. To alleviate this limitation, in Chapter 6 and Chapter 7, we will explore the use of

Graph Neural Networks (GNNs) for the task of age regression, as GNNs can provide an inherent way of combining multi-modal information.

Chapter 6

A comparative study for population-based brain age estimation using Graph Neural Networks

In the previous chapters, we focused on the use of Convolutional Neural Networks (CNNs) for the task of brain age regression, while also making the pipelines as intepretable as possible. However, CNNs use only the imaging information of a subject, without taking into account the non-imaging information as well as the associations among the subjects. Graph Neural Networks (GNNs) provide an inherent way of combining multi-modal information through the use of population graphs. Hence, for the rest of the thesis, we explore how population graphs and GNNs can be applied for age estimation.

In this chapter, we highlight the importance of meaningful graph construction and experiment with different static population-graph construction methods and their effect on GNN performance on brain age estimation. We use the homophily metric and graph visualisations to gain valuable quantitative and qualitative insights on the extracted graph structures.

This chapter is an extension of the work presented in [26]:

Bintsi, K., Mueller, T.T., Starck, S., Baltatzis, V., Hammers, A., and Rueckert, D.: A comparative study of population-graph construction methods and graph neural networks for brain age regression. In: Ahmadi, S. and Pereira, S. (eds.) Graphs in Biomedical Image Analysis, and Overlapped Cell on Tissue Dataset for Histopathology - 5th MICCAI Workshop, GRAIL 2023 and 1st MICCAI Challenge, OCELOT 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, September 23, and October 4, 2023, Proceedings. Lecture Notes in Computer Science, vol. 14373, pp. 64–73. Springer (2023). doi: 10.1007/978-3-031-55088-1_6, https://doi.org/10.1007/978-3-031-55088-1_6

6.1 Introduction

Alzheimer's disease [40], Parkinson's disease [154], and schizophrenia [97], among other neurodegenerative diseases, cause an atypically accelerated ageing process in the brain. Consequently, the difference between an individual's biological brain age and their chronological age can act as an indicator of deviation from the normal ageing trajectory [2], and potentially serve as a significant biomarker for neurodegenerative diseases [34, 51].

GNNs have been recently explored for medical tasks, since graphs can provide an inherent way of combining multi-modal information, and have demonstrated improved performance in comparison to graph-agnostic deep learning models [142, 1]. In most cases, the whole population is represented in a graph, namely population-graph, and the structure of the graph is pre-decided and remains static throughout the training. However, since the structure is not given but has to be inferred from the data, there are multiple ways that a graph can be constructed, which will possibly lead to different levels of performance. Every subject of the cohort is a node of the graph, with the imaging information usually allocated as node features. The interactions and relationships among the subjects of the cohort are represented by the edges. Two nodes are chosen to be connected based on a distance measure, or similarity score, usually based on the non-imaging phenotypes. The population graph is used as input to a GNN for the task of node prediction, most commonly node classification.

Brain MRI data have been widely used in population graphs in combination with GNNs.

Parisot et al. were the first to propose a way to construct a static population graph using a similarity score that takes into account both imaging and non-imaging data [142]. Many works were based on this work and extended the method to other applications [87, 198]. When it comes to brain age regression, which is an inherently more complicated task than classification, there is little ongoing work using population graphs. To our knowledge, only Stankevičiūtė et al. [176] worked on the construction of the static population graph using the non-imaging information, but the predictive performance was relatively low.

In all of the cases mentioned above, the graph is chosen based on some criterion before training and remains static, thus it cannot be changed. However, since there are multiple ways that it can be constructed, there is no way to ensure that the final structure will be optimised for the task. This problem has been identified in the literature and multiple metrics have been described in order to evaluate the final graph structure [200, 122], with homophily being the one most commonly used. A graph is considered homophilic, when nodes are connected to nodes with the same class label, and hence similar node features. Else, it is characterised as heterophilic. It has been found that some GNNs, such as Graph Convolutional Network (GCN) [92], and Graph Attention Network (GAT) [185], are very sensitive to the graph structure, and a meaningless graph along with these models can perform worse than a graph-agnostic model. On the other hand, other models, such as Graph Sample and Aggregated (GraphSage) [66] and Chebyshev [43], are more resilient to the graph structure, and their performance is not affected as much by a heterophilic graph [202].

In this chapter, we implement and evaluate the performance of different static graph construction methods for the task of brain age regression. We test their performance on the UK Biobank (UKBB), which offers a variety of both imaging and non-imaging phenotypes. The extracted population graphs are used along with the most popular GNNs, and more specifically GCN [92] and GAT [185], which are architectures highly sensitive to the graph structure, as well as GraphSage[66] and Chebyshev [43], which have been found to be resilient to the population-graph structure. The quantitative results and the visualisation of the graphs allow us to draw conclusions and make suggestions regarding the use of static graphs for brain age regression.

6.2 Methods

We have a dataset consisting of a set of N subjects, each described by M features. This dataset can be represented as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{N \times M}$, where $\mathbf{x}_i$ represents the feature vector for the i-th subject. Additionally, each subject has a label denoted by $\mathbf{y} \in \mathbb{R}^N$. Every subject i is also characterised by a set of K non-imaging phenotypes $\mathbf{q}_i \in \mathbb{R}^K$.

To establish the relationship between subjects, we introduce a population graph, denoted as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$. The graph consists of two components: $\mathcal{V}$ represents the set of nodes, where each subject corresponds to a unique node, and $\mathcal{E}$ represents the set of edges that define the connectivity between nodes.

In this paper, we explore four different graph construction approaches, which are described in detail below. In all of the graph construction methods, the node features consist of the imaging features described in the Dataset section 6.3.1.

6.2.1 No edges

As the baseline, we consider a graph with no edges among the nodes. This is the equivalent of traditional machine learning, where the node features act as the features used as input in the model. The machine learning model in this case is a Multi-Layer Perceptron (MLP). We refer to this approach as *"No edges"*.

6.2.2 Random Graph

With the next approach, we want to explore whether the way of constructing the graph plays an important role on the performance of the GNN. Thus, we build a random Erdos-Renyi graph [48], where a random number of nodes is chosen as neighbours for every node.

6.2.3 Clinical similarity score (Stankevičiūtė et al.)

An approach of creating a population graph specifically for the task of brain age regression was proposed by [176], and it does not include the imaging features at all in the extraction of the edges. Instead, the edges are decided only from the non-imaging information. More specifically,

the similarity function for two subjects i and j is given by:

$$sim(i,j) = \frac{1}{K}\sum_{k=1}^{K}\mathbf{1}[q_{ik} = q_{jk}], \tag{6.1}$$

where q_{ik} is the value of the k-th non-imaging phenotype for the i-th subject, and $\mathbf{1}$ is the Kronecker delta function. Intuitively, the Kronecker delta function will only return 1 if the values of a particular non-imaging phenotype of two subjects match. Two nodes i and j are connected if $sim(i,j) \geq \mu$, with μ being a similarity threshold decided empirically.

6.2.4 Similarity score (Parisot et al.)

In most of the related works in the literature, the way of creating the adjacency matrix W was originally suggested by [142] and it is calculated by:

$$W(i,j) = Sim(\mathbf{x}_i, \mathbf{x}_j)\sum_{k=1}^{K}\gamma(q_{ik}, q_{jk}), \tag{6.2}$$

where $Sim(\mathbf{x}_i, \mathbf{x}_j)$ is a similarity measure between the node features $(\mathbf{x}_i, \mathbf{x}_j)$ of the subjects i (which are commonly the imaging phenotypes) and j, in our case cosine similarity, $\gamma(\cdot,\cdot)$ is the distance of the non-imaging phenotypes between the nodes, and q_{ik} is the value of the k-th non-imaging phenotype for the i-th subject.

The two terms in Equation (6.2) indicate that both imaging, and non-imaging information are taken into account for the extraction of the edges. The second term is similar to the term $sim(i,j)$ of Equation (6.1).

The computation of $\gamma(\cdot,\cdot)$ is different for continuous and categorical features. For categorical data, $\gamma(\cdot,\cdot)$ is defined as the Kronecker delta function $\mathbf{1}$, as before. For continuous data, $\gamma(\cdot,\cdot)$ is defined as a unit-step function with respect to a specific threshold θ. Intuitively, this means that the output of the γ function will be 1 in case the continuous phenotypes of the two nodes are similar enough.

6.2.5 kNN graph

The last two graph construction approaches are based on the Nearest Neighbours (NN) algorithm. We connect the edges based on a distance function, in this case, cosine similarity. Each node is connected to its 5 closest neighbours.

In the first approach, we use the neuroimaging information for the node features, and we also use the node features in order to estimate the distances of the nodes. We refer to this approach as *"kNN (imaging)"*.

Similarly to before, in the second approach, we use the cosine similarity of a set of features in order to extract the graph structure. The node features incorporate the imaging information as before. The difference here is that we estimate the cosine similarity of the non-imaging phenotypes of the subjects with the purpose of finding the 5 closest neighbours. We refer to this approach as *"kNN (non-imaging)"*.

Finally, we use all the available phenotypes, i.e. both the imaging, and non-imaging information, to connect the nodes, again using cosine similarity. We refer to this approach as *"kNN (all phenotypes)"*.

6.3 Experiments

6.3.1 Dataset

For the experiments of the comparative study we use the UK Biobank (UKBB) [3], which not only offers an extensive range of vital organ images, including brain scans, but also contains a diverse collection of non-imaging information such as demographics, biomedical data, lifestyle factors, and cognitive performance measurements. Consequently, it is exceptionally well-suited for brain age estimation tasks that require integrating both imaging and non-imaging information.

To identify the most important factors influencing brain age in the UKBB, we leverage the work of [33] and select 68 neuroimaging phenotypes and 20 non-imaging phenotypes that were found to be the most relevant to brain age in UKBB. The neuroimaging features are obtained directly from the UKBB and include measurements derived from both structural MRI and diffusion-weighted MRI. A full list of the neuroimaging information and the non-imaging information can

be found in Table 6.5 and Table 6.6, respectively. All phenotypes are standardised to a normalised range between 0 and 1.

The study focuses on individuals aged 47 to 81 years. We include only those subjects who have the necessary phenotypes available, resulting in a group of approximately 6500 subjects. The dataset is split into three parts: 75% for training, 5% for validation, and 20% for testing.

6.3.2 Implementation Details

The static graphs extracted are chosen to be sparse for complexity reasons. The hyperparameters of the different construction methods were selected in such a way that all the extracted graphs have about 40000 to 50000 edges. Regarding the model hyperparameters, every GNN contains a graph convolutional layer consisting of 512 units, followed by a fully connected layer with 128 units, prior to the regression layer. ReLU activation is chosen. The number of layers and their dimensions are determined by conducting a hyperparameter search based on validation performance. During training, the networks are optimised using the AdamW optimiser [119] with a learning rate of 0.001 for 150 epochs, with the best model being saved. For the similarity threshold we use $\mu = 18$ and for the unit-step function threshold, we use $\theta = 0.1$. Both hyperparameters are selected based on validation performance and sparsity requirements. The implementation utilises PyTorch Geometric [49] and a Titan RTX GPU.

6.3.3 Results

The graph construction methods described in the previous section are leveraged by four GNN architectures and evaluated on the task of brain age regression. For every experiment, both spectral and spatial methods, and more specifically, GCN [92], GAT [185], GraphSage [66], and Chebyshev [43], are used. The results for all combinations of graph construction methods and GNN architectures are shown in Table 6.1 and Table 6.2.

As a baseline, we use an MLP (*No edges*) that achieves a MAE of 3.73 years and a R^2 score of 0.56. For the case of GCN and GAT, we notice that even though all of the graph costruction methods manage to extract a graph that outperforms the random graph, none of them manages to outperform the simple MLP. The methods that used only the imaging information, i.e. the

kNN graph (imaging), or all of the phenotypes, such as *Parisot et. al [142]*, extract graphs that when used as input to the GCN or the GAT models, they perform better (MAE of around 4 years) compared to the ones that only use the non-imaging features for the extraction of the graph, such as *kNN graph (non-imaging)* (MAE of 4.5-5 years). The GraphSAGE and Chebyshev models perform on similar levels regardless of the graph construction methods, with a MAE similar to the MAE given by the MLP. The graph construction method that performs the best though, is the *kNN graph (non-imaging)* along with the GraphSAGE, with a MAE of 3.68 years and a R^2 score of 0.59.

Finally, we estimate the homophily ratio of the population graphs extracted from the different construction methods based on [132] (Table 6.3). The most homophilic graphs tend to be the ones that use only imaging or a combination of imaging and non-imaging information for the extraction of the edges.

Table 6.1: MAE (in years) of the different graph construction methods along with various GNNs on the test set. The baseline MLP (*No edges*) has a performance of MAE=3.73 years.

Graph Construction Method	**GCN**	**GraphSAGE**	**GAT**	**Chebyshev**
Random graph	5.19	3.72	5.38	3.83
Parisot [142]	4.21	3.74	4.35	3.77
Stankevičiūtė [176]	4.61	3.73	4.90	3.77
kNN (imaging)	3.89	3.77	4.09	3.75
kNN (non-imaging)	4.76	3.68	4.98	3.72
kNN (all phenotypes)	3.93	3.76	4.07	3.73

Table 6.2: R^2 score of the different graph construction methods along with various GNNs on the test set. The baseline MLP (*No edges*) has a performance of R^2 score of 0.56.

Graph Construction Method	**GCN**	**GraphSAGE**	**GAT**	**Chebyshev**
Random graph	0.26	0.59	0.2	0.54
Parisot [142]	0.49	0.59	0.47	0.55
Stankevičiūtė [176]	0.4	0.59	0.3	0.58
kNN (imaging)	0.56	0.58	0.51	0.59
kNN (non-imaging)	0.38	0.59	0.3	0.55
kNN (all phenotypes)	0.56	0.58	0.53	0.6

6.3.4 Visualisations

Visualising the extracted static graphs, can provide more insights about the performance of the different GNN architectures. Therefore, we visualise the different population graphs coloured

Table 6.3: Homophily ratios for graphs extracted by various graph construction methods.

Graph Construction Method	**Homophily Ratio**
Random graph	0.7495
Parisot [142]	0.7899
Stankevičiūtė [176]	0.7743
kNN (imaging)	0.8259
kNN (non-imaging)	0.7796
kNN (all phenotypes)	0.8191

based on the age of the subjects in Figure 6.1. The various graph construction methods result in very different graph structures. More specifically, the graph construction methods that take into account the imaging phenotypes, either alone or in a combination with the non-imaging ones, provide graphs that are more meaningful, as subjects of similar ages are closer to each other. On the contrary, methods such as Stankevičiūtė et al. [176], create graphs that are similar to the random graph, with the neighbourhoods not being very informative regarding the age of the subjects.

6.3.5 Phenotypes used for the population graph

We perform an ablation test in which we experiment with the performance of the models based on the phenotypes used for the population graph. We test different combinations of phenotypes, both for the node features and the extraction of the edges. More specifically, we test the performance when imaging phenotypes, non-imaging phenotypes, or both of them (named as *all*) along with different combinations are used. For these experiments, we evaluate the performance on a sensitive to the graph structure model (GCN) and on a more resilient to the graph structure one (GraphSage). The rest of the hyperparameters remain the same as in the experiments described above.

For reference, the baseline MLP (*No edges*) has a performance of Mean Absolute Error (MAE)=3.73 years when the imaging phenotypes are used as the node features. When all phenotypes (meaning imaging and non-imaging ones) are used, we get a MAE of 3.47 years, while when only the non-imaging ones are chosen, the performance is MAE=4.91 years.

In Table 6.4, the performance of the two GNN models may be found. The first three rows, are the ones described in the previous section as *kNN (imaging)*, *kNN (non-imaging)*, and *kNN*

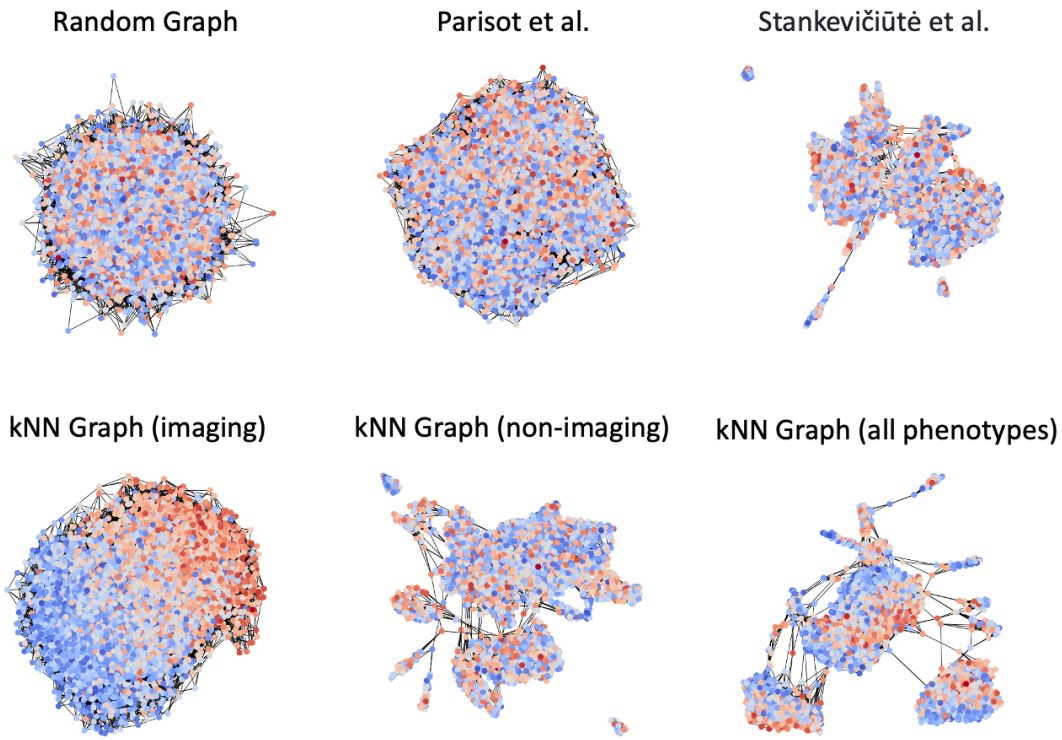


Figure 6.1: Visualisations of the population graphs extracted from the different graph construction methods. The nodes are coloured based on the age of the subject. Colder colours (i.e. blue) indicate subjects with an older age, while warmer colours (i.e. red) correspond to younger subjects.

(all), respectively. We also test the performance in the case that the non-imaging phenotypes are used for the node features, connected based either on imaging or on non-imaging phenotypes. Finally, we evaluate the performance of the models in the cases when we use all of the available phenotypes as node features.

In the case of GCN, we notice than using all of the phenotypes does not improve the performance of the model compared to when we only use the imaging ones. Moreover, using only the non-imaging information, either as node features or for extracting the edges, highly deteriorates the performance of the GCN, which indicates that the non-imaging phenotypes alone do not contain enough information for the task.

In the case of GraphSage, we notice that using all phenotypes improves the overall performance compared to using only the imaging phenotypes as node features. However, it still does

Table 6.4: MAE (in years) of GCN and GraphSage models for a kNN graph that uses different combinations of imaging and non-imaging phenotypes for the nodes and the edges.

Node Features	**Edges**	**GCN**	**GraphSAGE**
Imaging	Imaging	3.89	3.77
Imaging	All	3.93	3.76
Imaging	Non-imaging	4.76	3.68
Non-imaging	Non-imaging	5.08	4.86
Non-imaging	Imaging	4.41	4.18
All	All	3.86	3.51
All	Imaging	3.91	3.49
All	Non-imaging	4.63	3.5

not outperform the baseline MLP (MAE=3.47 years).

6.4 Discussion and Summary

In this chapter, we implement and evaluate static population graphs that are commonly used in the literature for other medical tasks, for brain age regression. We use the extracted graphs along with a number of popular GNN models, namely GCN, GAT, GraphSAGE, Chebyshev in order to get insights about the behavior of both the extraction methods, as well as the performance of the different models. By visualizing the graphs and estimating their homophily, we can provide further intuition in why the different static graphs do not work as expected and we highlight the problem that extracting static graphs from the data is not straightforward and possibly not suitable for brain age regression.

The reported results in Tables 6.1 and 6.2 indicate that the GCN and GAT architectures are highly sensitive to the graph structure, which is in agreement with the relevant literature [202, 120]. It is clear that the graph construction methods that provide graphs of higher homophily (Table 6.3), lead to better performance for GCN and GAT compared to the ones that provide a random-like population graph.

On the contrary, GraphSAGE and Chebyshev are not negatively affected by a graph structure with low homophily. This is because GraphSAGE encodes separately the node's encodings and the neighbours' encodings, which in our case are dissimilar, and hence it is affected less by the graph structure. When it comes to Chebyshev, the model is able to aggregate information from k hops in one layer, while the other GNNs achieve this through multiple layers. Being able to get

information from higher order neighbourhoods allows the model to find more relevant features, which would not be possible in the 1-hop neighbourhood as this is highly heterophilic. But even these GNNs, that are more resilient to the graph structure, cannot distinctly outperform the MLP. This behavior is expected as, according to the relevant literature, these models perform similarly to a MLP, and they can outperform it only after a specific homophily threshold [202]. What is also observed is that the *kNN (non-imaging)* works slightly better than the MLP both for the GraphSAGE and the Chebyshev, probably because the models were able to capture the information added in the edges.

To further explore the behaviour of the graph construction methods and the population-graph they produce, we calculate the homophily ratio and we visualise the graphs. The graphs that are more homophilic perform better along with the models that are more sensitive to the graph structure, such as the GCN and the GAT. We note here that all of the homophily ratios are higher than expected compared to the homophily reported in classification tasks, since we would expect that the random graph would have a homophily of 0.5. This possibly occurs because of the implementation of homophily for regression, as well as due to the imbalanced nature of the dataset. However, the trend is very clear and in agreement with the graph visualisations and the performance of the models.

All in all, the extraction of static population graphs for brain age regression, and in general for medical tasks for which the graph is not given, does not look very promising. In our opinion, there are multiple directions that should be explored. Firstly, one approach could be to learn the edges of the graph along with the training of the GNN, which allows the extraction of an optimised graph for the specific task at-hand. There is some ongoing work on this adaptive graph learning [189, 85, 38, 24], but more focus should be given. In addition, the creation of GNN models for graphs with high heterophily [200, 202], or the exploitation of graph rewiring techniques that could make the existing GNNs work better on heterophilic graphs [21], have proved to be useful for classification tasks. In the case of medical tasks, it might also be beneficial to incorporate the existing medical insights along with the above. It is also important to make the models and the graphs interpretable, as interpretability can be vital in healthcare, and it is something that is not currently widely explored in heterophilic graphs [200]. Last but not least, introducing metrics that evaluate population graphs is of high importance, since this can help us understand better

the structure of the graph. Even though there is some ongoing work when it comes to node classification [120, 200], metrics regarding node regression have only recently started to being explored [132].

For the final contibution chapter of this thesis (Chapter 7), we will investigate the first one of the proposed directions mentioned above. More specifically, we will propose an adaptive graph learning pipeline, designed for brain age estimation, that is able to learn the optimised edges of the population graph through training.

Table 6.5: Neuroimaging phenotypes

Neuroimaging phenotype	UK Biobank data field #
Volume of grey matter (normalised for head size)	25005
Volume of peripheral cortical grey matter (normalised for head size)	25001
Volume of ventricular cerebrospinal fluid (normalised for head size)	25003
Volume of white matter (normalised for head size)	25007
Volume of brain, grey+white matter (normalised for head size)	25009
Volume of brain stem + 4th ventricle	25025
Volume of grey matter in Putamen (left, right)	25882, 25883
Volume of thalamus (left, right)	25011, 25012
Volume of putamen (left, right)	25015, 25016
Volume of hippocampus (left, right)	25019, 25020
Volume of grey matter in Hippocampus (left, right)	25886, 25887
Volume of amygdala (left, right)	25021, 25022
Volume of grey matter in Amygdala (left, right)	25888, 25889
Volume of grey matter in Thalamus (left, right)	25878, 25879
Total volume of white matter hyperintensities (from T1 and T2 FLAIR images)	25781
Weighted-mean ICVF in tract forceps major	25660
Weighted-mean ICVF in tract forceps minor	25661
Weighted-mean FA in tract anterior thalamic radiation (left, right)	25490, 25491
Mean ISOVF in fornix on FA skeleton	25445
Mean FA in middle cerebellar peduncle on FA skeleton	25056
Mean FA in cerebral peduncle on FA skeleton (left, right)	25071, 25070
Mean FA in superior cerebellar peduncle on FA skeleton (left, right)	25069, 25068
Mean L1 in middle cerebellar peduncle on FA skeleton	25200
Mean L3 in posterior thalamic radiation on FA skeleton (left, right)	25325, 25324
Mean L2 in fornix cres+stria terminalis on FA skeleton (left, right)	25287, 25286
Mean ICVF in body of corpus callosum on FA skeleton	25347
Mean L1 in anterior limb of internal capsule on FA skeleton (left, right)	25217, 25216
Mean MO in fornix cres+stria terminalis on FA skeleton (left, right)	25191, 25190
Mean OD in superior longitudinal fasciculus on FA skeleton (left, right)	25433, 25432
Weighted-mean L3 in tract uncinate fasciculus (left, right)	25648, 25649
Mean MD in cingulum hippocampus on FA skeleton (left, right)	25141, 25140
Mean L2 in splenium of corpus callosum on FA skeleton	25252
Weighted-mean MO in tract anterior thalamic radiation (left, right)	25544, 25545
90th percentile of BOLD effect (in group-defined mask) for shapes activation	25761
90th percentile of z-statistic (in group-defined mask) for shapes activation	25762
Weighted-mean MO in tract forceps major	25552
Weighted-mean MO in tract forceps minor	25553
Median T2star in putamen (left, right)	25030, 25031
Median BOLD effect (in group-defined mask) for faces-shapes contrast	25048
Median z-statistic (in group-defined mask) for faces-shapes contrast	25050
Weighted-mean L1 in tract parahippocampal part of cingulum (left, right)	25575, 25576
Mean L2 in cingulum cingulate gyrus on FA skeleton (left, right)	25283, 25282
Mean L2 in cingulum hippocampus on FA skeleton (left, right)	25285, 25284
Weighted-mean L1 in tract anterior thalamic radiation (left, right)	25571, 25572

Table 6.6: Non-imaging phenotypes

Phenotype	UK Biobank data field #
Sex	50
Weight	21002
Height	50
Body mass index (BMI)	21001
Systolic blood pressure	4080
Diastolic blood pressure	4079
College education	845
Smoking status	20116
Alcohol intake frequency	1558
Stroke	4056
Diabetes	2443
Duration of walks	874
Duration of moderate activity	894
Duration of vigorous activity	914
Fluid intelligence	20016
Tower rearranging: number of puzzles correct	6382
Trail making task: duration to complete numeric path trail 1	6348
Trail making task: duration to complete alphanumeric path trail 2	6350
Matrix pattern completion: number of puzzles correctly solved	6373
Matrix pattern completion: duration spent answering each puzzle	6333

Chapter 7

Multimodal brain age estimation using interpretable adaptive population-graph learning

In Chapter 6 we explored how static population graphs work along with Graph Neural Networks (GNNs). We highlighted that a pre-defined population graph based on a chosen criterion, usually a similarity metric of the non-imaging information, is in most cased not optimised for age estimation and it might end up performing worse than even a simple graph-agnostic model.

In this chapter, we propose a framework that learns a population graph structure optimised for the downstream task. An attention mechanism assigns weights to a set of imaging and non-imaging features (phenotypes), which are then used for edge extraction. The resulting graph is used to train the GNN. Here, we use Graph Convolutional Network (GCN) as our model, since in Chapter 6 we concluded that GCN is the GNN that is the most sensitive to the graph structure. The entire pipeline can be trained end-to-end. Additionally, by visualising the attention weights that were the most important for the graph construction, we increase the interpretability of the graph.

Part of the work of this chapter was presented at MICCAI 2023 and published in [24]:

Bintsi, K., Baltatzis, V., Potamias, R.A., Hammers, A., and Rueckert, D.: Multimodal brain age estimation using interpretable adaptive population-graph learning. In: Greenspan, H., Madabhushi, A., Mousavi, P., et al. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2023 - 26th International Conference, Vancouver, BC, Canada, October 8-12, 2023, Proceedings, Part VIII. Lecture Notes in Computer Science, vol. 14227, pp. 195–204. Springer (2023). doi: 10.1007/978-3-031-43993-3_19, https://doi.org/10.1007/978-3-031-43993-3_19

7.1 Introduction

Healthy brain ageing follows specific patterns [2]. However, various neurological diseases, such as Alzheimer's disease [40], Parkinson's disease [154], and schizophrenia [97], are accompanied by an abnormal accelerated ageing of the human brain. Thus, the difference between the biological brain age of a person and their chronological age can show the deviation from the healthy ageing trajectory, and may prove to be an important biomarker for neurodegenerative diseases [34, 51].

Recently, graph-based methods have been explored for brain age estimation as graphs can inherently combine multimodal information by integrating the subjects' neuroimaging information as node features and, through a similarity metric, the associations among subjects through as edges that connect these nodes [142]. However, in medical applications, the construction of a population-graph is not always simple as there are various ways subjects could be considered similar.

GCNs [92] have been extensively [1] used in the medical domain for node classification tasks, such as Alzheimer's prediction [142, 87] and Autism classification [7]. They take graphs as input and, in most cases, the graph structure is predefined and static. GCN performance is highly dependent on the graph structure. This has been correlated in related literature with the heterophily of graphs in the semi-supervised node classification tasks, which refers to the case when the nodes of a graph are connected to nodes that have dissimilar features and different class labels [200]. It has been shown that if the homophily ratio is very low, a simple Multi-Layer Perceptron (MLP) that completely ignores the structure, can outperform a GCN [202].

A way to address this problem is through adaptive graph learning [189], which learns the

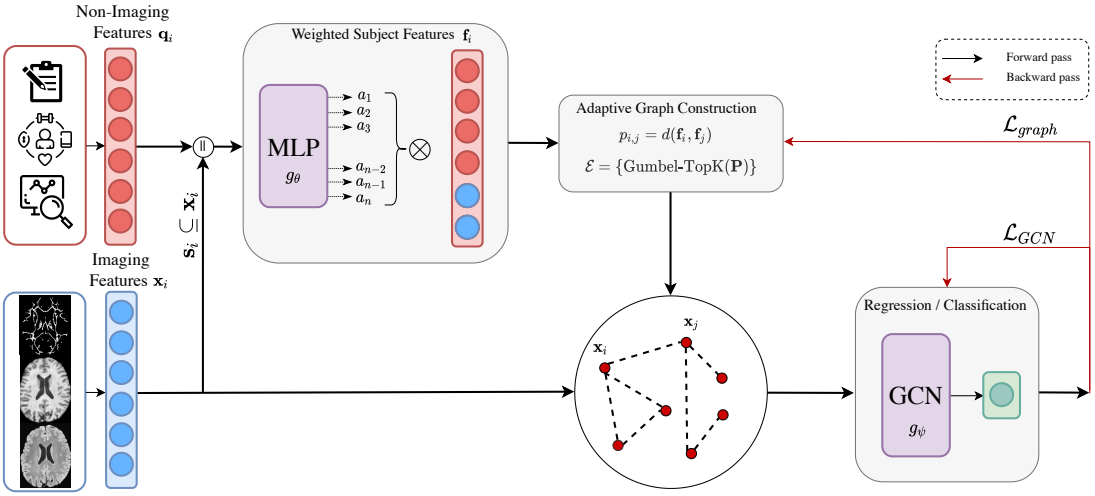


Figure 7.1: Proposed methodology: The non-imaging features $\mathbf{q}_i$ and a subset of the imaging features $\mathbf{s}_i \subseteq \mathbf{x}_i$ per subject i are used as input to a MLP, which produces attention weights for each one of these features. Based on these, the edges of the graph are stochastically sampled using the Gumbel-Top-k trick. The constructed graph, which uses the imaging features $\mathbf{X}$ as node features is used to train a GCN for brain age estimation tasks.

graph structure through training. In the medical domain, there is little ongoing research on the topic [85, 199]. However, the adaptive graphs in [85, 86, 38] are connected based on the imaging features and do not take advantage of the associations of the non-imaging information for the edges. In [76], non-imaging information is used for graph connectivity, but the edges are already pre-pruned similar to [142], and only the weights of the edges can be modified. In [188], even though the graph is dynamic, it is not being learnt. While the extracted graphs in these studies are optimised for the task at hand, there is no explanation for the node connections that were proposed. Given that interpretability is essential in the medical domain [166], since the tools need to be trusted by the clinicians, an adaptive graph learnt during the training, whose connections are also interpretable and clear can prove important. Additionally, most existing works focus on brain age classification in four bins, usually classifying the subjects per decade [85, 86, 38]. Age regression, which is a more challenging task, has not been extensively explored with published results not reaching sufficient levels of performance [176].

This chapter has the following contributions: 1) We combine imaging and non-imaging information in an attention-based framework to learn adaptively an optimised graph structure for brain age estimation. 2) We propose a novel graph loss that enables end-to-end training for

the task of brain age regression. 3) Our framework is inherently interpretable as the attention mechanism allows us to rank all imaging and non-imaging phenotypes according to their significance for the task. 4) We evaluate our method on the UK Biobank (UKBB) and achieve state-of-the-art results on the tasks of brain age regression and classification.

7.2 Methods

Given a set of N subjects with M features $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N] \in \mathbb{R}^{N \times M}$ and labels $\mathbf{y} \in \mathbb{R}^N$, a population graph is defined as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is a set of nodes, one per subject, and $\mathcal{E}$ is a set of paired nodes that specifies the connectivity of the graph, meaning the edges of the graph. To create an optimised set of edges for the task of brain age estimation, we leverage a set of non-imaging phenotypes $\mathbf{q}_i \in \mathbb{R}^Q$ and a set of imaging phenotypes $\mathbf{s}_i \in \mathbb{R}^S$ per subject i through an attention-based framework. The imaging phenotypes are a subset of the imaging features $\mathbf{s}_i \subseteq \mathbf{x}_i$. The phenotypes are selected according to [33]. The resulting graph is used as input in a GCN that performs node-level prediction tasks. Here, we give a detailed description of the proposed architecture, in which the connectivity of the graph $\mathcal{E}$, is learnt through end-to-end training in order to find the optimal graph structure for the task of brain age estimation. An outline of the proposed pipeline may be found in Figure 7.1.

Attention Weights Extraction

Based on the assumption that not all phenotypes are equally important for the construction of the graph, we train a MLP g_θ, with parameters θ, which takes as input both the non-imaging features $\mathbf{q}_i$ and the imaging features $\mathbf{s}_i$ for every subject i and outputs an attention weight vector $\mathbf{a} \in \mathbb{R}^{Q+S}$, where every element of $\mathbf{a}$ corresponds to a specific phenotype. Intuitively, we expect that the features that are relevant to brain age estimation will get attention weights close to 1, and close to 0 otherwise. Since we are interested in the overall relevance of the phenotypes for the task, the output weights need to be global and apply to all of the nodes. To do so, we average the attention weights across subjects and normalise them between 0 and 1.

Edge Extraction

The weighted phenotypes for each subject i are calculated as $\mathbf{f}_i = \mathbf{a} \odot (\mathbf{q}_i \,\|\, \mathbf{s}_i)$, where $\mathbf{f}_i \in \mathbb{R}^{Q+S}$, $\mathbf{a}$ are the attention weights produced by the MLP g_θ, $(\cdot \,\|\, \cdot)$ denotes the concatenation function between two vectors and $\odot$ denotes the Hadamard product. We further define the probability $p_{ij}(\mathbf{f}_i, \mathbf{f}_j; \theta, t)$ of a pair of nodes $(i, j) \in \mathcal{V}$ to be connected in Equation (7.1):

$$p_{ij}(\mathbf{f}_i, \mathbf{f}_j; \theta, t) = e^{-td(\mathbf{f}_i, \mathbf{f}_j)^2} \tag{7.1}$$

where t is a learnable parameter and d is a distance function that calculates the distance between the weighted phenotypes of two nodes. To keep the memory cost low, we create a sparse k-degree graph. We use the Gumbel-Top-k trick [95], which acts as a stochastic relaxation of the kNN rule, in order to sample k edges for every node according to the probability matrix $\mathbf{P} \in \mathbb{R}^{N \times N}$. Since there is stochasticity in the sampling scheme, multiple runs are performed at inference time and the predictions are averaged.

Optimisation

The extracted graph is used as input, along with the imaging features $\mathbf{X}$, which are used as node features, to a GCN g_ψ, with parameters ψ, which comprises of a number of graph convolutional layers, followed by fully connected layers. The pipeline is trained end-to-end with a loss function $\mathcal{L}$ that consists of two components and is defined as in Equation (7.2).

$$\mathcal{L} = \mathcal{L}_{GCN} + \mathcal{L}_{graph}. \tag{7.2}$$

The first component, $\mathcal{L}_{GCN}$, is optimising the GCN, g_ψ. For regression we use the Huber loss [77], while for classification we use the Cross Entropy loss function. The second component, $\mathcal{L}_{graph}$, optimises the MLP g_θ, whose output are the phenotypes' attention weights. However, the graph is sparse with discrete edges and hence the network cannot be trained with backpropagation as is. To alleviate this issue, we formulate our graph loss in a way that rewards edges that lead to correct predictions and penalise edges that lead to wrong predictions. Inspired by [85], where a similar approach was used for classification, the proposed graph loss function is designed for

regression instead and is defined in Equation (7.3):

$$\mathcal{L}_{graph} = \sum_{(i,j)\in\mathcal{E}} \rho(y_i, g_\psi(\mathbf{x}_i))\, log(p_{ij}), \tag{7.3}$$

where $\rho(\cdot,\cdot)$ is the reward function which is defined in Equation (7.4):

$$\rho(y_i, g_\psi(\mathbf{x}_i)) = |y_i - g_\psi(\mathbf{x}_i)| - \varepsilon \tag{7.4}$$

Here ε (or ε_0) is initialised as the null model's prediction (i.e. the average brain age of the training set), and it is then updated based on $\varepsilon_t = 0.95\varepsilon_{t-1} + 0.05\text{MAE}_t$, with t being the current epoch. Intuitively, when the prediction error $|y_i - g_\psi(\mathbf{x}_i)|$ is smaller than the null model's prediction then the reward function will be negative. In turn, this will encourage the maximisation of p_{ij} so that $\mathcal{L}_{graph}$ is minimised.

7.3 Experiments

7.3.1 Dataset

The proposed framework is evaluated on the UKBB [3], which provides not only a wide collection of images of vital organs, including brain scans, but also various non-imaging information, such as demographics, and biomedical, lifestyle, and cognitive performance measurements. Hence, it is perfectly suitable for brain age estimation tasks that incorporate the integration of imaging and non-imaging information. Here, we use 68 neuroimaging phenotypes and 20 non-imaging phenotypes proposed by [33] as the ones most relevant to brain age in UKBB. The neuroimaging features are provided by UKBB, and include measurements extracted from structural MRI and diffusion weighted MRI. All phenotypes are normalised from 0 to 1. The age range of the subjects is 47-81 years. We only keep the subjects that have available the necessary phenotypes ending up with about 6500 subjects. We split the dataset into 75% for the training set, 5% for the validation set, and 20% for the test set. Our pipeline has been primarily designed to tackle the challenging regression task and therefore the main experiment is brain age regression. We also evaluate our framework on the 4-class classification task that has been used in other related

papers [85, 38].

7.3.2 Baselines

Given that a GCN trained on a meaningless graph can perform even worse than a simple regressor/classifier, our first baseline is a Linear/Logistic Regression model. For the second baseline, we construct a static graph based on a similarity metric, more specifically cosine similarity, of a set of features (either node features or non-imaging and imaging phenotypes) using the kNN rule with $k = 5$ and train a GCN on that graph. Using euclidean distance as the similarity metric leads to worse performance and is therefore not explored for the baselines. We also compare our method with DGM, which is the state-of-the-art on graph learning for medical applications [85]. Since this work is only applicable for classification tasks, we extend it to regression, implementing the graph loss function we used for our pipeline as well.

7.3.3 Implementation Details

The GCN architecture uses ReLU activations and consists of one graph convolutional layer with 512 units and one fully connected layer with 128 units before the regression/classification layer. The number and the dimensions of the layers are determined through hyperparameter search based on validation performance. The networks are trained with the AdamW optimiser [119], with a learning rate of 0.005 for 300 epochs and utilise an early stopping scheme. The average brain age of the training set, which is used in Equation (7.3), is $\varepsilon = 6$. We use PyTorch [143] and a Titan RTX GPU. The reported results are computed by averaging 10 runs with different initialisations.

7.4 Results

7.4.1 Regression

For the evaluation of the pipeline in the regression task, we use Mean Absolute Error (MAE), and the Pearson Correlation Coefficient (r score). A summary of the performance of the proposed and competing methods is available at Table 7.1. Linear regression (MAE=3.82) outperforms

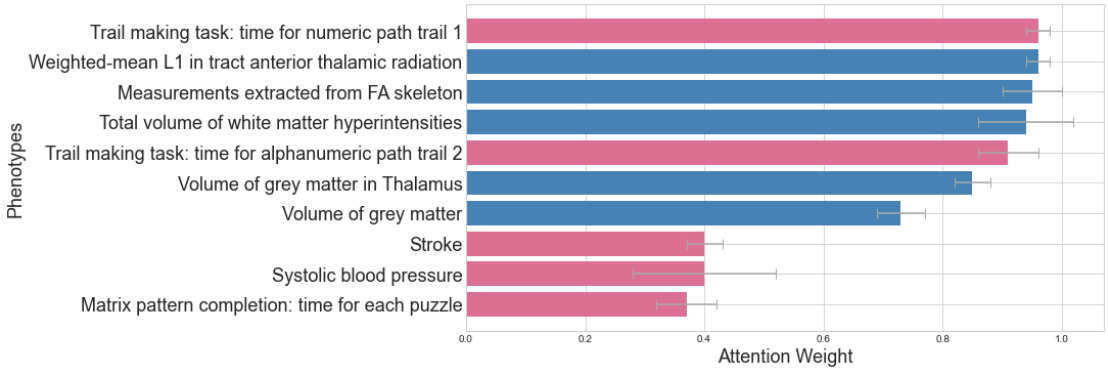


Figure 7.2: Attention weights for the top 5 imaging (blue) and top 5 non-imaging phenotypes (pink) averaged across 10 runs of the pipeline.

Table 7.1: Performance of the proposed approach in comparison with a traditional machine learning model, static graph baselines, and the state-of-the-art for age regression.

Method	MAE (years)	r score
Linear Regression	3.82	0.75
Static (Node features)	3.87	0.75
Static (Phenotypes)	3.98	0.74
DGM	3.72	0.75
Proposed	**3.61**	**0.79**

GCNs trained on static graphs whether these are based on the node features (MAE=3.87) or leverage the phenotypes (MAE=3.98). The DGM outperforms the baselines (MAE=3.72). The proposed method outperforms all others (MAE=3.61).

7.4.2 Classification

For the classification task, we divide the subjects into four balanced classes. The four bins that the subjects are split into are the following: 47-56, 57-63, 64-68, and 69-80 years. The metrics used for the classification task to evaluate the performance of the model are accuracy, the area under the ROC curve (AUC), which is defined as the average of the AUC curve of every class, and the Macro F1-score (Table 7.2). A similar trend to the regression task appears here, with the proposed method reaching top performance with 58% accuracy.

Our hypothesis that the construction of a pre-defined graph structure is suboptimal, and might even hurt performance, is confirmed. Both for the classification and regression tasks,

Table 7.2: Performance of the proposed approach in comparison with a traditional machine learning model, static graph baselines, and the state-of-the-art for and 4-class age classification (right).

Method	Accuracy	AUC	F1
Logistic Regression	0.54	0.79	0.54
Static (Node features)	0.53	0.75	0.52
Static (Phenotypes)	0.52	0.75	0.51
DGM	0.55	**0.80**	0.55
Proposed	**0.58**	**0.80**	**0.56**

Table 7.3: Ablation study: Effect of the choice of phenotypes used for the connection of the edges.

Number of phenotypes	MAE (years)	r score
20 (non-imaging only)	4.63	0.68
35 (non-imaging & imaging)	**3.61**	**0.79**
50 (non-imaging & imaging)	3.65	0.78
68 (imaging only)	3.73	0.77

GCNs trained on static graphs do not outperform a simple linear model that ignores completely the structure of the graph. The proposed approach proves that not all phenotypes are equally important for the construction of the graph, and that giving attention weights accordingly increases the performance for both tasks.

7.4.3 Ablation studies

Number of phenotypes

We perform an ablation test to investigate the effect of the number of features used for the extraction of the edges. We used only non-imaging phenotypes, only imaging phenotypes, or a combination of both (Table 7.3). The combination of imaging and non-imaging phenotypes (MAE=3.61) performs better than using either one of them, while adding more imaging phenotypes does not necessarily improve performance.

Distance Metrics

Moreover, we explore how different distance metrics affect performance. Here, euclidean, cosine, and hyperbolic [99] distances are explored. We also include a random edge selection to examine

Table 7.4: Ablation study: Performance of the proposed using different distances on the phenotypes to connect the edges.

Similarity Metric	MAE (years)	r score
Random	5.59	0.61
Euclidean	**3.61**	**0.79**
Cosine	3.7	0.78
Hyperbolic	4.08	0.72

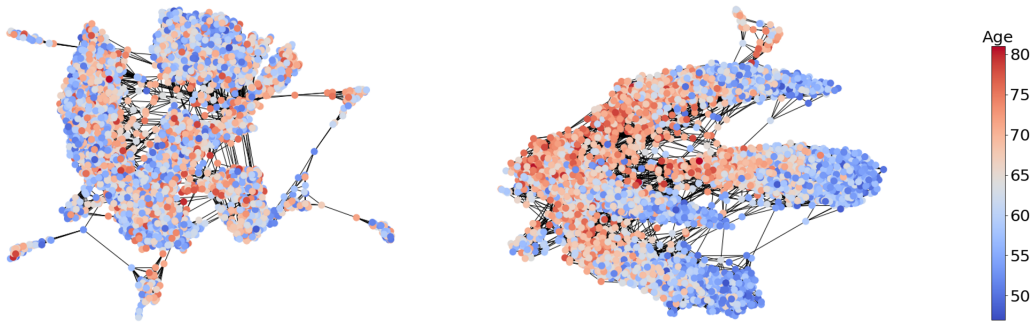


Figure 7.3: Population-graph visualisations. Left: graph constructed based on the cosine similarity of the phenotypes. Right: graph constructed based on the cosine similarity of the weighted phenotypes, where the attention weights are the ones extracted from the proposed pipeline. The colour of the nodes corresponds to the subject's age. It is evident that using the attention weights increases the homophily of the graph.

whether using the phenotypic information improves the results. The results (Table 7.4) indicate that euclidean distance performs the best, closely followed by cosine similarity. Hyperbolic performed a bit worse (MAE=4 years), while random edges gave a MAE of 5.59. Regardless of the distance metric, phenotypes do incorporate valuable information as performance is considerably better than using random edges.

Table 7.5: Homophily ratios for the population graphs visualised in Figure 7.3. *Static (Phenotypes)* refers to the graph on the left side of the figure, while *Proposed* refers to the graph on the right side of figure.

Method	Static (Phenotypes) (Left)	Proposed (Right)
Homophily ratio	0.81	0.84

Table 7.6: List of phenotypes used for the connectivity of the graph along with the attention weights (mean & std after 10 runs with different initialisations) given from the pipeline.

Index	Phenotypes	Attention Weights
1	Sex	0.21 ± 0.2
2	Height	0.22 ± 0.05
3	Body mass index (BMI)	0.03 ± 0.02
4	Systolic blood pressure	0.4 ± 0.12
5	Diastolic blood pressure	0.14 ± 0.17
6	College education	0.1 ± 0.18
7	Smoking status	0.1 ± 0.16
8	Alcohol intake frequency	0.02 ± 0.02
9	Stroke	0.40 ± 0.03
10	Diabetes	0.18 ± 0.18
11	Moderate per week	0.02 ± 0.03
12	Vigorous per week	0.03 ± 0.03
13	Fluid intelligence	0.03 ± 0.03
14	Tower rearranging: number of puzzles correct	0.03 ± 0.03
15	Trail making task: duration to complete numeric path trail 1	0.96 ± 0.02
16	Trail making task: duration to complete alphanumeric path trail 2	0.91 ± 0.05
17	Matrix pattern completion: number of puzzles correctly solved	0.02 ± 0.02
18	Matrix pattern completion: duration spent answering each puzzle	0.37 ± 0.05
19	Volume of grey matter (normalised for head size)	0.73 ± 0.04
20	Volume of brain stem + 4th ventricle	0.47 ± 0.03
21	Volume of grey matter in Putamen (right)	0.1 ± 0.09
22	Volume of thalamus (right)	0.34 ± 0.03
23	Volume of putamen (left)	0.28 ± 0.03
24	Volume of grey matter in Thalamus (right)	0.85 ± 0.03
25	Total volume of white matter hyperintensities	0.94 ± 0.08
26	Weighted-mean ICVF in tract forceps minor	0.95 ± 0.04
27	Weighted-mean L1 in tract anterior thalamic radiation (right)	0.98 ± 0.02
28	Mean FA in cerebral peduncle on FA skeleton (left)	0.95 ± 0.02
29	Mean FA in superior cerebellar peduncle on FA skeleton (left)	0.90 ± 0.21
30	Mean L1 in middle cerebellar peduncle on FA skeleton	0.96 ± 0.02
31	Mean ICVF in body of corpus callosum on FA skeleton	0.98 ± 0.01
32	Mean L2 in fornix cres+stria terminalis on FA skeleton (left)	0.98 ± 0.01
33	Weighted-mean L3 in tract uncinate fasciculus (left)	0.72 ± 0.06
34	Mean ISOVF in fornix on FA skeleton	0.33 ± 0.02
35	Weighted-mean L1 in tract parahippocampal part of cingulum (left)	0.43 ± 0.1

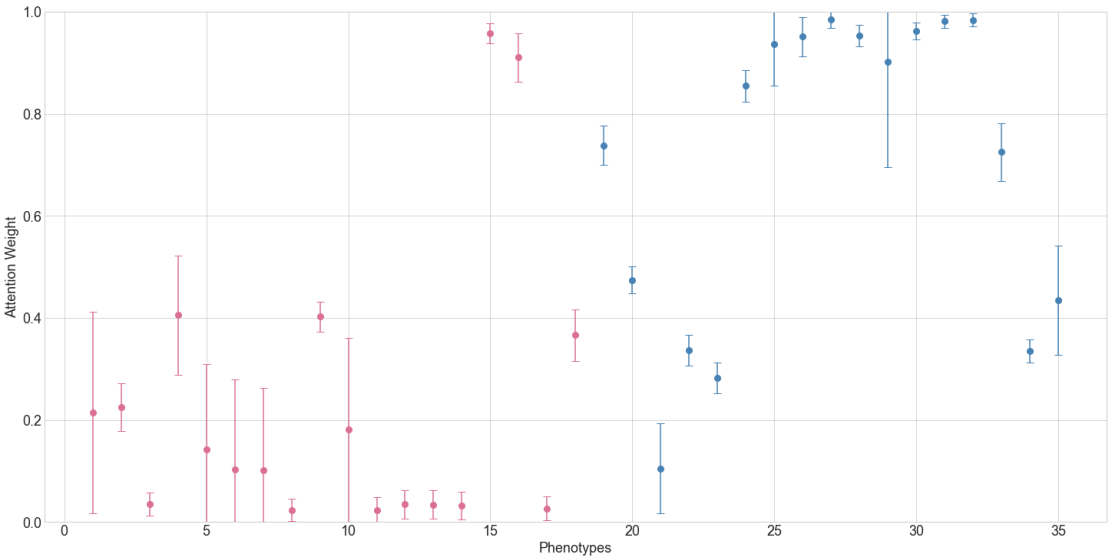


Figure 7.4: Attention weights extracted from the pipeline per phenotype. The mean and std of the attention weights from 10 runs with different initialisations are shown. The phenotypes are indexed 1-35. Full list of the phenotypes and the attention scores can be found in Table 7.6. The non-imaging phenotypes are coloured in pink, and the neuroimaging phenotypes are coloured in blue. The highest weights are given to two cognitive tasks, namely trail making tasks numeric (15) and alphanumeric (16), and measurements extracted from FA skeleton (28-32).

7.5 Interpretability

A very important advantage of the pipeline is that the graph extracted through the training is interpretable, in terms of why two nodes are connected or not. Visualising the attention weights given to the phenotypes, we get an understanding of the features that are the most relevant to brain ageing. The regression problem is clinically more important, thus we will focus on this in this section. A similar trend was presented for the classification task as well.

The imaging and non-imaging phenotypes that were given the highest attention scores in the construction of the graph can be seen in Figure 7.2. We colour the non-imaging phenotypes in pink, and the imaging ones in blue. A detailed list of the names of the phenotypes along with the attention weights given by the pipeline can be found in Table 7.6 and Figure 7.4. The non-imaging phenotypes that were given the highest attention weights, were two cognitive tasks, the numeric and the alphanumeric trail making tasks. Systolic blood pressure, and stroke, were the next most important non-imaging phenotypes, even though they were not as important as

some of the imaging features. Various neuroimaging phenotypes were considered important, such as information regarding the tract anterior thalamic radiaton, the volume of white matter hyperintensities, grey matter volumes, as well as measurements extracted from the FA (fractional anisotropy, a measure of the connectivity of the brain) skeleton. Our findings are in agreement with the relevant literature [33].

Apart from the attention weights, we also visualise the population graph that was used as the static graph of the baseline (Figure 7.3(left)). This population graph is constructed based on the cosine similarity of the phenotypes, with all the phenotypes playing an equally important role for the connectivity. In addition, we visualise the population graph using the cosine similarity of the weighted phenotypes (Figure 7.3(right)), with the attention weights provided by our trained pipeline. In both of the graph visualisations, the colour of each node corresponds to the subject's age. It is clear that learning the graph through our pipeline results in a population graph where subjects with similar ages end up in more compact clusters, whereas the static graph does not demonstrate any form of organisation. This can be further confirmed by estimation the homophily ratios for the two graphs. The homophily ratios may be found in Table 7.5. Since GCNs are affected by a graph construction with low homophily, it is reasonable that the static graphs perform worse than a simple machine learning method, and why the proposed approach manages to produce state-of-the-art results. Using a different GNN architecture that is not as dependent on the constructed graph's homophily could prove beneficial [66, 200].

7.6 Summary

In this final contribution chapter, we propose an end-to-end pipeline for adaptive population-graph learning that is optimised for brain age estimation. We integrate multimodal information and use attention scores for the construction of the graph, while also increasing interpretability. The graph is sparse, which minimises the computational costs. We implement the approach both for node regression and node classification and show that we outperform both the static graph-based and state-of-the-art approaches. We finally provide an insight into the most important phenotypes for graph construction, which are in agreement with related neurobiological literature.

Chapter 8

Conclusions and Future Directions

This chapter offers a summary of the research points introduced in the dissertation, highlighting the current work's limitations and suggesting potential directions for future research. Ultimately, conclusive remarks are presented.

8.1 Summary

This thesis aimed to investigate the application of deep learning in estimating brain age. Various deep learning techniques were examined, with a specific focus on Convolutional Neural Networks (CNNs) in Chapters 4 and 5, as well as Graph Neural Networks (GNNs) in Chapters 6 and 7. In addition, a notable emphasis throughout the thesis was placed on ensuring the interpretability of the proposed methods.

In Section 2, we started by exploring how the use of patches, instead of the whole brain Magnetic Resonance (MR) images, may benefit the performance of CNNs. Our goal extended beyond achieving accurate brain age estimation to delivering more localised results. By combining the results of the different CNNs using an ensemble model, we managed to outperform the state-of-the-art, generating attention maps highlighting the brain regions most meaningful to brain

age. This marked our **first contribution**. We explored both 2D and 3D data, extracting patches of varying sizes either in a grid pattern or randomly.

Despite the promising outcomes in Chapter 4, the use of relatively large patches led to somewhat coarse attention maps. Consequently, as the **second contribution** of this thesis, we shifted our focus towards enhancing the interpretability of CNNs to produce more detailed attention maps. Existing interpretability methods usually focus on classification tasks, and they do not work as expected for regression tasks. Based on the assumption that not all voxels are equally important for brain age estimation, we implemented a pipeline involving a noise model and a brain age regression model—specifically, Residual Network (ResNet). The objective was to introduce maximum noise to the input image without impacting the model's performance. Regions resistant to noise addition were identified as the most crucial, and by averaging saliency maps across different subjects in the test set, we generated population-based maps pinpointing brain regions most relevant to age estimation. Our results aligned with existing neuropathology literature, underscoring the significance of the hippocampus and ventricles in the ageing process.

CNNs solely consider imaging data and overlook non-imaging information and inter-subject relationships within the population. To overcome this constraint, Chapters 6 and 7 delve into the utilisation of Graph Neural Networks (GNNs) in conjunction with population graphs. This approach harnesses the intrinsic ability of GNNs to integrate multi-modal information, aiming to enhance the accuracy of brain age estimation.

In Chapter 6, our objective was to derive a meaningful population graph for the downstream task. The **third contribution** of this thesis involved the exploration of various methods of constructing static graphs and assessed their impact on the performance of various GNN models, including Graph Convolutional Network (GCN) and Graph Sample and Aggregated (GraphSage). The findings underscored the critical role of graph construction, leading us to the conclusion that static population graphs are often suboptimal for brain age regression, since they performed worse than even a simple graph-agnostic model.

These findings led us to the **fourth contribution** of the thesis, presented in Chapter 7, an end-to-end framework designed to learn an optimised population graph structure for brain age estimation using adaptive graph learning. Based on the assumption that not all phenotypes are equally important for brain age, we implemented a pipeline incorporating an attention mecha-

nism to assign weights to both imaging and non-imaging phenotypes. Intuitively, we expected significant weights (close to 1) for crucial phenotypes and minimal weights (close to 0) for less influential ones, which was confirmed by our findings. These attention phenotypes were then used to derive the edges of the population graph. The resulting graph, coupled with GCNs, the GNN that is the most sensitive to the graph structure, outperformed both baseline static population graphs and the state-of-the-art method. Furthermore, by visualising the attention weights, we provided insights into the most critical phenotypes for brain age, aligning with neurobiological literature and enhancing the interpretability of the graph.

8.2 Limitations and Future Directions

In this section, we delve into the limitations inherent in the current thesis and propose potential avenues for future research to overcome these constraints.

In Chapter 4, we employed brain patches, rather than the entire brain scan, as input to CNNs. Our exploration included both 2D and 3D data, with patches extracted in a grid or randomly. Through a linear regression combination of patch outputs, we outperformed state-of-the-art methods in brain age regression. However, a limitation lies in the relatively large size (64x64x64) of the used patches, resulting in accurate yet coarse attention maps. Reducing patch size could enhance granularity, but the current non-end-to-end training approach hinders scalability. A promising future direction involves efforts to identify crucial patches and enable end-to-end training for improved scalability and precision.

The ultimate goal of integrating machine learning into medical imaging is its adoption by clinicians. Building trust among clinicians is pivotal, and interpretability plays a crucial role in achieving this. In Chapter 5, we shifted our focus from the performance of age regression models to enhancing interpretability, aiming to provide more detailed attention maps. Our proposed framework employed a noise model to generate saliency maps. Through the aggregation of these maps across the test set, we crafted population-based importance maps, pinpointing brain regions most relevant to age estimation. Our finding were consistent with existing literature. An interesting future direction includes the extension of this framework to other regression tasks and diverse brain datasets, such as Alzheimer's Disease Neuroimaging Initiative (ADNI)

(https://adni.loni.usc.edu/). Additionally, it would be beneficial to expand the framework to accommodate classification tasks, including Alzheimer's Disease (AD) classification. In this context, comparing the generated maps with existing saliency maps from the literature, which are traditionally tailored for classification tasks, will be made possible.

One constraint of CNNs is their exclusive focus on imaging information, overlooking inter-subject associations within the population and non-imaging data. Leveraging the inherent capacity of population graphs and GNNs to incorporate multi-modal information, we sought to address this limitation in the context of brain age regression—a particularly challenging task—aiming to explore their potential effectiveness in this domain.

In Chapter 6, we initially delved into static population graphs, exploring various construction methods proposed in the literature and testing their performance with multiple GNN models. Our evaluation, using metrics and graph visualisations, provided both quantitative and qualitative insights into model performance. However, we observed that the extracted static population graphs exhibited high heterophily and did not show promising results for brain age regression.

Several potential future directions could be pursued. Firstly, incorporating adaptive graph learning by jointly training graph edges with the GNN could optimise the graph structure for downstream tasks. Ongoing work on adaptive graph learning [85, 38, 76] exists, warranting further exploration. Additionally, developing GNN models tailored for graphs with high heterophily is an area of interest, with ongoing efforts [200, 202] requiring more attention. Considering medical tasks, incorporating domain-specific insights into proposed pipelines could enhance performance. Given the importance of interpretability in healthcare, constructing interpretable graphs should be a focal point [200]. Finally, introducing metrics tailored for evaluating population graphs is a crucial future direction. While existing works focus on classification tasks [120, 200], exploration in node regression tasks has recently begun [132], necessitating further attention.

In the final Chapter 7, we attempted to address some of the raised points. Our focus centered on adaptive graph learning, presenting a framework that optimally learns a graph structure for brain age regression. Leveraging an attention mechanism for imaging and non-imaging phenotypes, we extracted edges to construct the graph, which, in conjunction with GCNs, facilitated a visual representation of the most relevant brain age-related phenotypes, enhancing interpretability. Notably, a limitation lies in the current inductive nature of the pipeline, restricting the

addition of new nodes post-graph construction. An intriguing future direction involves extending the framework to a transductive environment, allowing the addition of new nodes directly during test time. Furthermore, the framework's encouraging results justify further investigation into its performance across a variety of medical imaging datasets and tasks.

8.3 Conclusion

This thesis embarked on a comprehensive exploration of deep learning applications in brain age estimation, emphasising interpretability throughout the investigative journey. Key contributions emerged from distinct chapters, each addressing specific challenges and advancing the field. Initially, the focus centered on CNNs, investigating the impact of patch-based approaches on brain age estimation. The proposed approach surpassed existing benchmarks, generating attention maps highlighting critical brain regions. Despite promising outcomes, the need for more detailed attention maps led to the introduction of a pipeline with a noise model and a prediction model for the extraction of attention maps. This approach revealed the significance of hippocampal and ventricular regions in ageing, aligning with neurobiological literature. Acknowledging the limitations of CNNs in considering non-imaging information and inter-subject relationships, we shift the focus on the integration GNNs and population graphs. We highlighted the suboptimal nature of static population graphs for brain age regression, prompting a deeper investigation into adaptive graph learning. We, then, introduced an end-to-end framework with an attention mechanism, assigning weights to phenotypes. This dynamic approach outperformed static graphs and state-of-the-art methods, providing both accuracy and interpretability. Visualisation of attention weights offered insights into critical phenotypes, further reinforcing the alignment with neurobiological literature. In essence, this thesis not only advanced brain age estimation methodologies but also emphasised the significance of interpretability, paving the way for more accurate and interpretable models in neuroimaging research.

Bibliography

[1] Ahmedt-Aristizabal, D., Armin, M.A., Denman, S., Fookes, C., and Petersson, L.: Graph-based deep learning for medical diagnosis and analysis: Past, present and future. Sensors **21**(14), 4758 (2021). doi: 10.3390/S21144758, https://doi.org/10.3390/s21144758

[2] Alam, S.B., Nakano, R., Kamiura, N., and Kobashi, S.: Morphological changes of aging brain structure in MRI analysis. In: 2014 Joint 7th International Conference on Soft Computing and Intelligent Systems (SCIS) and 15th International Symposium on Advanced Intelligent Systems (ISIS), Kita-Kyushu, Japan, December 3-6, 2014. pp. 683–687. IEEE (2014). doi: 10.1109/SCIS-ISIS.2014.7044901, https://doi.org/10.1109/SCIS-ISIS.2014.7044901

[3] Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., et al.: Image processing and quality control for the first 10, 000 brain imaging datasets from UK biobank. NeuroImage **166**, 400–424 (2018). doi: 10.1016/J.NEUROIMAGE.2017.10.034, https://doi.org/10.1016/j.neuroimage.2017.10.034

[4] Aljabar, P., Heckemann, R.A., Hammers, A., Hajnal, J.V., and Rueckert, D.: Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. NeuroImage **46**(3), 726–738 (2009). doi: 10.1016/J.NEUROIMAGE.2009.02.018, https://doi.org/10.1016/j.neuroimage.2009.02.018

[5] Alzheimer's Association: 2019 Alzheimer's disease facts and figures includes a special report on Alzheimer's detection in the primary care setting: connecting patients and physicians.

Tech. rep. (2019), https://www.alz.org/media/Documents/alzheimers-facts-and-figures-2019-r.pdf

[6] Anderson, P., He, X., Buehler, C., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 6077–6086. Computer Vision Foundation / IEEE Computer Society (2018). doi: 10.1109/CVPR.2018.00636, http://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Bottom-Up_and_Top-Down_CVPR_2018_paper.html

[7] Anirudh, R. and Thiagarajan, J.J.: Bootstrapping graph convolutional neural networks for autism spectrum disorder classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019. pp. 3197–3201. IEEE (2019). doi: 10.1109/ICASSP.2019.8683547, https://doi.org/10.1109/ICASSP.2019.8683547

[8] Aramadaka, S., Mannam, R., Sankara Narayanan, R., et al.: Neuroimaging in Alzheimer's Disease for Early Diagnosis: A Comprehensive Review. Cureus **15**(5), e38544 (May 2023), https://dx.doi.org/10.7759/cureus.38544

[9] Ashburner, J. and Friston, K.J.: Voxel-based morphometry–the methods. Neuroimage **11**(6 Pt 1), 805–821 (Jun 2000), https://dx.doi.org/10.1006/nimg.2000.0582

[10] Avberšek, L.K. and Repovš, G.: Deep learning in neuroimaging data analysis: Applications, challenges, and solutions. Frontiers in neuroimaging **1**, 981642 (2022)

[11] Ballester, P.L., da Silva, L.T., Marcon, M., et al.: Predicting Brain Age at Slice Level: Convolutional Neural Networks and Consequences for Interpretability. Frontiers in Psychiatry **12** (2 2021). doi: 10.3389/fpsyt.2021.598518, https://www.frontiersin.org/articles/10.3389/fpsyt.2021.598518/full

[12] Baltatzis, V., Bintsi, K., Folgoc, L.L., et al.: The pitfalls of sample selection: A case study on lung nodule classification. In: Rekik, I., Adeli, E., Park, S.H., and Schnabel, J.A. (eds.) Predictive Intelligence in Medicine - 4th International Workshop, PRIME 2021,

Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings. Lecture Notes in Computer Science, vol. 12928, pp. 201–211. Springer (2021). doi: 10.1007/978-3-030-87602-9_19, https://doi.org/10.1007/978-3-030-87602-9_19

[13] Baltatzis, V. and Costabello, L.: Kgex: Explaining knowledge graph embeddings via subgraph sampling and knowledge distillation. CoRR **abs/2310.01065** (2023). doi: 10.48550/ARXIV.2310.01065, https://doi.org/10.48550/arXiv.2310.01065

[14] Baltatzis, V., Folgoc, L.L., Ellis, S., et al.: The effect of the loss on generalization: Empirical study on synthetic lung nodule data. In: Reyes, M., Abreu, P.H., Cardoso, J.S., et al. (eds.) Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data - 4th International Workshop, iMIMIC 2021, and 1st International Workshop, TDA4MedicalData 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings. Lecture Notes in Computer Science, vol. 12929, pp. 56–64. Springer (2021). doi: 10.1007/978-3-030-87444-5_6, https://doi.org/10.1007/978-3-030-87444-5_6

[15] Bass, C., da Silva, M., Sudre, C.H., et al.: Icam-reg: Interpretable classification and regression with feature attribution for mapping neurological phenotypes in individual scans. IEEE Trans. Medical Imaging **42**(4), 959–970 (2023). doi: 10.1109/TMI.2022.3221890, https://doi.org/10.1109/TMI.2022.3221890

[16] Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 3319–3327. IEEE Computer Society (2017). doi: 10.1109/CVPR.2017.354, https://doi.org/10.1109/CVPR.2017.354

[17] Becker, B.G., Klein, T., Wachinger, C., Initiative, A.D.N., et al.: Gaussian process uncertainty in age estimation as a measure of brain abnormality. Neuroimage **175**, 246–258 (2018)

[18] Beheshti, I., Nugent, S., Potvin, O., and Duchesne, S.: Bias-adjustment in neuroimaging-based brain age frameworks: A robust scheme. Neuroimage: Clinical **24**, 102063 (2019)

[19] Bello, G.A., Dawes, T.J.W., Duan, J., et al.: Deep-learning cardiac motion analysis for human survival prediction. Nat. Mach. Intell. **1**(2), 95–104 (2019). doi: 10.1038/S42256-019-0019-2, https://doi.org/10.1038/s42256-019-0019-2

[20] Bessadok, A., Mahjoub, M.A., and Rekik, I.: Graph neural networks in network neuroscience. IEEE Trans. Pattern Anal. Mach. Intell. **45**(5), 5833–5848 (2023). doi: 10.1109/TPAMI.2022.3209686, https://doi.org/10.1109/TPAMI.2022.3209686

[21] Bi, W., Du, L., Fu, Q., Wang, Y., Han, S., and Zhang, D.: Make heterophily graphs better fit GNN: A graph rewiring approach. CoRR **abs/2209.08264** (2022). doi: 10.48550/ARXIV.2209.08264, https://doi.org/10.48550/arXiv.2209.08264

[22] Bintsi, K., Baltatzis, V., Hammers, A., and Rueckert, D.: Voxel-level importance maps for interpretable brain age estimation. In: Reyes, M., Abreu, P.H., Cardoso, J.S., et al. (eds.) Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data - 4th International Workshop, iMIMIC 2021, and 1st International Workshop, TDA4MedicalData 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings. Lecture Notes in Computer Science, vol. 12929, pp. 65–74. Springer (2021). doi: 10.1007/978-3-030-87444-5_7, https://doi.org/10.1007/978-3-030-87444-5_7

[23] Bintsi, K., Baltatzis, V., Kolbeinsson, A., Hammers, A., and Rueckert, D.: Patch-based brain age estimation from MR images. In: Kia, S.M., Mohy-ud-Din, H., Abdulkadir, A., et al. (eds.) Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology - Third International Workshop, MLCN 2020, and Second International Workshop, RNO-AI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4-8, 2020, Proceedings. Lecture Notes in Computer Science, vol. 12449, pp. 98–107. Springer (2020). doi: 10.1007/978-3-030-66843-3_10, https://doi.org/10.1007/978-3-030-66843-3_10

[24] Bintsi, K., Baltatzis, V., Potamias, R.A., Hammers, A., and Rueckert, D.: Multi-modal brain age estimation using interpretable adaptive population-graph learning. In: Greenspan, H., Madabhushi, A., Mousavi, P., et al. (eds.) Medical Image Computing and

Computer Assisted Intervention - MICCAI 2023 - 26th International Conference, Vancouver, BC, Canada, October 8-12, 2023, Proceedings, Part VIII. Lecture Notes in Computer Science, vol. 14227, pp. 195–204. Springer (2023). doi: 10.1007/978-3-031-43993-3_19, https://doi.org/10.1007/978-3-031-43993-3_19

[25] Bintsi, K.M., Jarolim, R., Tremblay, B., et al.: Sunerf: Validation of a 3d global reconstruction of the solar corona using simulated euv images (2022)

[26] Bintsi, K., Mueller, T.T., Starck, S., Baltatzis, V., Hammers, A., and Rueckert, D.: A comparative study of population-graph construction methods and graph neural networks for brain age regression. In: Ahmadi, S. and Pereira, S. (eds.) Graphs in Biomedical Image Analysis, and Overlapped Cell on Tissue Dataset for Histopathology - 5th MICCAI Workshop, GRAIL 2023 and 1st MICCAI Challenge, OCELOT 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, September 23, and October 4, 2023, Proceedings. Lecture Notes in Computer Science, vol. 14373, pp. 64–73. Springer (2023). doi: 10.1007/978-3-031-55088-1_6, https://doi.org/10.1007/978-3-031-55088-1_6

[27] Bologna, G. and Hayashi, Y.: Characterization of symbolic rules embedded in deep DIMLP networks: A challenge to transparency of deep learning. J. Artif. Intell. Soft Comput. Res. **7**(4), 265 (2017). doi: 10.1515/JAISCR-2017-0019, https://doi.org/10.1515/jaiscr-2017-0019

[28] Bonner, S., Barrett, I.P., Ye, C., et al.: A review of biomedical datasets relating to drug discovery: a knowledge graph perspective. Briefings Bioinform. **23**(6) (2022). doi: 10.1093/BIB/BBAC404, https://doi.org/10.1093/bib/bbac404

[29] Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P.: Geometric deep learning: Going beyond euclidean data. IEEE Signal Process. Mag. **34**(4), 18–42 (2017). doi: 10.1109/MSP.2017.2693418, https://doi.org/10.1109/MSP.2017.2693418

[30] Bunge, S.A. and Whitaker, K.J.: Brain imaging: Your brain scan doesn't lie about your age. Current Biology **22**(18), R800–R801 (sep 2012). doi: 10.1016/j.cub.2012.07.032, https://www.sciencedirect.com/science/article/pii/S0960982212008573

[31] Cherubini, A., Caligiuri, M.E., Péran, P., Sabatini, U., Cosentino, C., and Amato, F.: Importance of multimodal MRI in characterizing brain tissue and its potential application for individual age prediction. IEEE J. Biomed. Health Informatics **20**(5), 1232–1239 (2016). doi: 10.1109/JBHI.2016.2559938, https://doi.org/10.1109/JBHI.2016.2559938

[32] Chu, L.: Alzheimer's disease: early diagnosis and treatment. Hong Kong Medical Journal **18**(3), 228 (2012)

[33] Cole, J.H.: Multimodality neuroimaging brain-age in uk biobank: relationship to biomedical, lifestyle, and cognitive factors. Neurobiology of aging **92**, 34–42 (2020)

[34] Cole, J.H. and Franke, K.: Predicting age using neuroimaging: innovative brain ageing biomarkers. Trends in neurosciences **40**(12), 681–690 (2017)

[35] Cole, J.H., Poudel, R.P.K., Tsagkrasoulis, D., et al.: Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. NeuroImage **163**, 115–124 (2017). doi: 10.1016/J.NEUROIMAGE.2017.07.059, https://doi.org/10.1016/j.neuroimage.2017.07.059

[36] Cole, J.H., Raffel, J., Friede, T., et al.: Longitudinal assessment of multiple sclerosis with the brain-age paradigm. Annals of neurology **88**(1), 93–105 (2020)

[37] Cole, J.H., Ritchie, S.J., Bastin, M.E., et al.: Brain age predicts mortality. Molecular psychiatry **23**(5), 1385–1392 (2018)

[38] Cosmo, L., Kazi, A., Ahmadi, S., Navab, N., and Bronstein, M.M.: Latent-graph learning for disease prediction. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., et al. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part II. Lecture Notes in Computer Science, vol. 12262, pp. 643–653. Springer (2020). doi: 10.1007/978-3-030-59713-9_62, https://doi.org/10.1007/978-3-030-59713-9_62

[39] Coupé, P., Catheline, G., Lanuza, E., Manjón, J.V., and Initiative, A.D.N.: Towards a unified analysis of brain maturation and aging across the entire lifespan: A MRI analysis. Human brain mapping **38**(11), 5501–5518 (2017)

[40] Davatzikos, C., Bhatt, P., Shaw, L.M., Batmanghelich, K.N., and Trojanowski, J.Q.: Prediction of mci to ad conversion, via MRI, csf biomarkers, and pattern classification. Neurobiology of aging **32**(12), 2322–e19 (2011)

[41] Davatzikos, C., Xu, F., An, Y., Fan, Y., and Resnick, S.M.: Longitudinal progression of Alzheimers-like patterns of atrophy in normal older adults: The SPARE-AD index. Brain **132**(8), 2026–2035 (aug 2009). doi: 10.1093/brain/awp091, http://www.ncbi.nlm.nih.gov/pubmed/19416949

[42] De Lange, A.M.G., Kaufmann, T., Van Der Meer, D., et al.: Population-based neuroimaging reveals traces of childbirth in the maternal brain. Proceedings of the National Academy of Sciences of the United States of America **116**(44), 22341–22346 (2019). doi: 10.1073/pnas.1910666116

[43] Defferrard, M., Bresson, X., and Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.) Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain. pp. 3837–3845 (2016), https://proceedings.neurips.cc/paper/2016/hash/04df4d434d481c5bb723be1b6df1ee65-Abstract.html

[44] Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA. pp. 248–255. IEEE Computer Society (2009). doi: 10.1109/CVPR.2009.5206848, https://doi.org/10.1109/CVPR.2009.5206848

[45] Dinsdale, N.K., Bluemke, E., Smith, S.M., et al.: Learning patterns of the ageing brain in MRI using deep convolutional networks. NeuroImage **224**, 117401 (2021). doi: 10.1016/J.NEUROIMAGE.2020.117401, https://doi.org/10.1016/j.neuroimage.2020.117401

[46] Duchi, J.C., Hazan, E., and Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**, 2121–2159 (2011). doi: 10.5555/1953048.2021068, https://dl.acm.org/doi/10.5555/1953048.2021068

[47] Eickhoff, C.R., Hoffstaedter, F., Caspers, J., et al.: Advanced brain ageing in parkinson's disease is related to disease duration and individual impairment. Brain communications **3**(3), fcab191 (2021)

[48] Erdős, P., Rényi, A., et al.: On the evolution of random graphs. Publ. math. inst. hung. acad. sci **5**(1), 17–60 (1960)

[49] Fey, M. and Lenssen, J.E.: Fast graph representation learning with pytorch geometric. CoRR **abs/1903.02428** (2019), http://arxiv.org/abs/1903.02428

[50] Francis, P.T.: The interplay of neurotransmitters in alzheimer's disease. CNS spectrums **10**(S18), 6–9 (2005)

[51] Franke, K. and Gaser, C.: Ten years of brainage as a neuroimaging biomarker of brain aging: What insights have we gained? Frontiers in neurology **10**, 789 (2019)

[52] Franke, K., Ziegler, G., Klöppel, S., and Gaser, C.: Estimating the age of healthy subjects from t_1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. NeuroImage **50**(3), 883–892 (2010). doi: 10.1016/J.NEUROIMAGE.2010.01.005, https://doi.org/10.1016/j.neuroimage.2010.01.005

[53] Fukushima, K.: Cognitron: A self-organizing multilayered neural network. Biological cybernetics **20**(3), 121–136 (1975)

[54] Gaugler, J., James, B., Johnson, T., et al.: 2022 alzheimer's disease facts and figures. Alzheimers & Dementia **18**(4), 700–789 (2022)

[55] Gaugler, J., James, B., Johnson, T., Scholz, K., and Weuve, J.: 2016 Alzheimer's disease facts and figures. Alzheimer's and Dementia **12**(4), 459–509 (mar 2016). doi: 10.1016/j.jalz.2016.03.001, https://www.sciencedirect.com/science/article/pii/S1552526018300414

[56] Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., and Dahl, G.E.: Neural message passing for quantum chemistry. In: Precup, D. and Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia,

6-11 August 2017. Proceedings of Machine Learning Research, vol. 70, pp. 1263–1272. PMLR (2017), http://proceedings.mlr.press/v70/gilmer17a.html

[57] Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M.A., and Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: Bonchi, F., Provost, F.J., Eliassi-Rad, T., Wang, W., Cattuto, C., and Ghani, R. (eds.) 5th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2018, Turin, Italy, October 1-3, 2018. pp. 80–89. IEEE (2018). doi: 10.1109/DSAA.2018.00018, https://doi.org/10.1109/DSAA.2018.00018

[58] Glasser, M.F., Smith, S.M., Marcus, D.S., et al.: The human connectome project's neuroimaging approach. Nature neuroscience **19**(9), 1175–1187 (2016)

[59] Good, C.D., Johnsrude, I.S., Ashburner, J., Henson, R.N.A., Friston, K.J., and Frackowiak, R.S.J.: A voxel-based morphometric study of ageing in 465 normal adult human brains. NeuroImage **14**(1), 21–36 (2001). doi: 10.1006/NIMG.2001.0786, https://doi.org/10.1006/nimg.2001.0786

[60] Gori, M., Monfardini, G., and Scarselli, F.: A new model for learning in graph domains. In: Proc. 2005 IEEE Int. joint conf. on neural networks. vol. 2(2005), pp. 729–734 (2005)

[61] Gotkowski, K., González, C., Bucher, A., and Mukhopadhyay, A.: M3d-cam - A pytorch library to generate 3d attention maps for medical deep learning. In: Palm, C., Deserno, T.M., Handels, H., Maier, A., Maier-Hein, K.H., and Tolxdorff, T. (eds.) Bildverarbeitung für die Medizin 2021 - Proceedings, German Workshop on Medical Image Computing, Regensburg, March 7-9, 2021. pp. 217–222. Informatik Aktuell, Springer (2021). doi: 10.1007/978-3-658-33198-6_52, https://doi.org/10.1007/978-3-658-33198-6_52

[62] Groeschel, S., Vollmer, B., King, M.D., and Connelly, A.: Developmental changes in cerebral grey and white matter volume from infancy to adulthood. International Journal of Developmental Neuroscience **28**(6), 481–489 (oct 2010). doi: 10.1016/j.ijdevneu.2010.06.004, https://www.sciencedirect.com/science/article/pii/S0736574810000985?via{%}3Dihub

[63] Gunes, S., Aizawa, Y., Sugashi, T., Sugimoto, M., and Rodrigues, P.P.: Biomarkers for alzheimer's disease in the current state: a narrative review. International journal of molecular sciences **23**(9), 4962 (2022)

[64] Gutiérrez-Becker, B., Klein, T., and Wachinger, C.: Gaussian process uncertainty in age estimation as a measure of brain abnormality. NeuroImage **175**, 246–258 (2018). doi: 10.1016/J.NEUROIMAGE.2018.03.075, https://doi.org/10.1016/j.neuroimage.2018.03.075

[65] Habes, M., Janowitz, D., Erus, G., et al.: Advanced brain aging: Relationship with epidemiologic and genetic risk factors, and overlap with Alzheimer disease atrophy patterns. Translational Psychiatry **6**(4), e775–e775 (apr 2016). doi: 10.1038/tp.2016.39, http://www.ncbi.nlm.nih.gov/pubmed/27045845

[66] Hamilton, W.L., Ying, Z., and Leskovec, J.: Inductive representation learning on large graphs. In: Guyon, I., von Luxburg, U., Bengio, S., et al. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 1024–1034 (2017), https://proceedings.neurips.cc/paper/2017/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html

[67] He, K., Zhang, X., Ren, S., and Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 770–778. IEEE Computer Society (2016). doi: 10.1109/CVPR.2016.90, https://doi.org/10.1109/CVPR.2016.90

[68] Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., and Hammers, A.: Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage **33**(1), 115–126 (2006). doi: 10.1016/J.NEUROIMAGE.2006.05.061, https://doi.org/10.1016/j.neuroimage.2006.05.061

[69] Heckemann, R.A., Hammers, A., Rueckert, D., Aviv, R.I., Harvey, C.J., and Hajnal, J.V.: Automatic volumetry on MR brain images can support diagnostic decision making. BMC

Medical Imaging **8**, 9 (2008). doi: 10.1186/1471-2342-8-9, https://doi.org/10.1186/1471-2342-8-9

[70] Heckemann, R.A., Keihaninejad, S., Aljabar, P., Rueckert, D., Hajnal, J.V., and Hammers, A.: Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation. NeuroImage **51**(1), 221–227 (2010). doi: 10.1016/J.NEUROIMAGE.2010.01.072, https://doi.org/10.1016/j.neuroimage.2010.01.072

[71] Høgestøl, E.A., Kaufmann, T., Nygaard, G.O., et al.: Cross-sectional and longitudinal mri brain scans reveal accelerated brain aging in multiple sclerosis. Frontiers in neurology **10**, 444539 (2019)

[72] Holzinger, A., Biemann, C., Pattichis, C.S., and Kell, D.B.: What do we need to build explainable AI systems for the medical domain? CoRR **abs/1712.09923** (2017), http://arxiv.org/abs/1712.09923

[73] Howard, A.G., Zhu, M., Chen, B., et al.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR **abs/1704.04861** (2017), http://arxiv.org/abs/1704.04861

[74] Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 2261–2269. IEEE Computer Society (2017). doi: 10.1109/CVPR.2017.243, https://doi.org/10.1109/CVPR.2017.243

[75] Huang, T.W., Chen, H.T., Fujimoto, R., et al.: Age estimation from brain MRI images using deep learning. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). pp. 849–852. IEEE (2017)

[76] Huang, Y. and Chung, A.C.S.: Edge-variational graph convolutional networks for uncertainty-aware disease prediction. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., et al. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part

VII. Lecture Notes in Computer Science, vol. 12267, pp. 562–572. Springer (2020). doi: 10.1007/978-3-030-59728-3_55, https://doi.org/10.1007/978-3-030-59728-3_55

[77] Huber, P.J.: Robust estimation of a location parameter. Breakthroughs in statistics: Methodology and distribution pp. 492–518 (1992)

[78] Hurtado-Puerto, A.M., Russo, C., and Fregni, F.: Alzheimer's disease. In: Neuromethods, vol. 138, pp. 297–338. Elsevier (mar 2018). doi: 10.1007/978-1-4939-7880-9_9, https://www.sciencedirect.com/science/article/pii/S0140673610613499?via{%}3Dihub

[79] Ioffe, S.: Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. In: Guyon, I., von Luxburg, U., Bengio, S., et al. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 1945–1953 (2017), https://proceedings.neurips.cc/paper/2017/hash/c54e7837e0cd0ced286cb5995327d1ab-Abstract.html

[80] Jarolim, R., Tremblay, B., Muñoz-Jaramillo, A., et al.: Sunerf: 3d reconstruction of the solar euv corona using neural radiance fields. The Astrophysical Journal Letters **961**(2), L31 (2024)

[81] Jernigan, T.L., Archibald, S.L., Fennema-Notestine, C., et al.: Effects of age on tissues and regions of the cerebrum and cerebellum. Neurobiology of aging **22**(4), 581–594 (2001)

[82] Jetley, S., Lord, N.A., Lee, N., and Torr, P.H.S.: Learn to pay attention. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018), https://openreview.net/forum?id=HyzbhfWRW

[83] Jonsson, B.A., Bjornsdottir, G., Thorgeirsson, T., et al.: Brain age prediction using deep learning uncovers associated sequence variants. Nature communications **10**(1), 1–10 (2019)

[84] Juntu, J., Sijbers, J., Dyck, D.V., and Gielen, J.: Bias field correction for MRI images. In: Kurzynski, M., Puchala, E., Wozniak, M., and Zolnierek, A. (eds.) Computer Recognition Systems, Proceedings of the 4th International Conference on Computer Recogni-

tion Systems, CORES'05, May 22-25, 2005, Rydzyna Castle, Poland. Advances in Soft Computing, vol. 30, pp. 543–551. Springer (2005). doi: 10.1007/3-540-32390-2_64, https://doi.org/10.1007/3-540-32390-2_64

[85] Kazi, A., Cosmo, L., Ahmadi, S., Navab, N., and Bronstein, M.M.: Differentiable graph module (DGM) for graph convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **45**(2), 1606–1617 (2023). doi: 10.1109/TPAMI.2022.3170249, https://doi.org/10.1109/TPAMI.2022.3170249

[86] Kazi, A., Farghadani, S., Aganj, I., and Navab, N.: IA-GCN: interpretable attention based graph convolutional network for disease prediction. In: Cao, X., Xu, X., Rekik, I., Cui, Z., and Ouyang, X. (eds.) Machine Learning in Medical Imaging - 14th International Workshop, MLMI 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 8, 2023, Proceedings, Part I. Lecture Notes in Computer Science, vol. 14348, pp. 382–392. Springer (2023). doi: 10.1007/978-3-031-45673-2_38, https://doi.org/10.1007/978-3-031-45673-2_38

[87] Kazi, A., Shekarforoush, S., Krishna, S.A., et al.: Inceptiongcn: Receptive field aware graph convolutional network for disease prediction. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., and Bao, S. (eds.) Information Processing in Medical Imaging - 26th International Conference, IPMI 2019, Hong Kong, China, June 2-7, 2019, Proceedings. Lecture Notes in Computer Science, vol. 11492, pp. 73–85. Springer (2019). doi: 10.1007/978-3-030-20351-1_6, https://doi.org/10.1007/978-3-030-20351-1_6

[88] Keicher, M., Burwinkel, H., Bani-Harouni, D., et al.: U-GAT: multimodal graph attention network for COVID-19 outcome prediction. CoRR **abs/2108.00860** (2021), https://arxiv.org/abs/2108.00860

[89] Keihaninejad, S., Heckemann, R.A., Gousias, I.S., et al.: Classification and lateralization of temporal lobe epilepsies with and without hippocampal atrophy based on whole-brain automatic MRI segmentation. PloS one **7**(4), e33096 (2012)

[90] Kim, D. and Oh, A.: How to find your friendly neighborhood: Graph attention design with self-supervision. In: 9th International Conference on Learning Representations, ICLR

2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=Wi5KUNlqWty

[91] Kingma, D.P. and Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y. and LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980

[92] Kipf, T.N. and Welling, M.: Semi-supervised classification with graph convolutional networks. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net (2017), https://openreview.net/forum?id=SJU4ayYgl

[93] Koker, T., Mireshghallah, F., Titcombe, T., and Kaissis, G.: U-noise: Learnable noise masks for interpretable image segmentation. In: 2021 IEEE International Conference on Image Processing, ICIP 2021, Anchorage, AK, USA, September 19-22, 2021. pp. 394–398. IEEE (2021). doi: 10.1109/ICIP42928.2021.9506345, https://doi.org/10.1109/ICIP42928.2021.9506345

[94] Kondo, C., Ito, K., Wu, K., et al.: An age estimation method using brain local features for t1-weighted images. In: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2015, Milan, Italy, August 25-29, 2015. pp. 666–669. IEEE (2015). doi: 10.1109/EMBC.2015.7318450, https://doi.org/10.1109/EMBC.2015.7318450

[95] Kool, W., van Hoof, H., and Welling, M.: Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In: Chaudhuri, K. and Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research, vol. 97, pp. 3499–3508. PMLR (2019), http://proceedings.mlr.press/v97/kool19a.html

[96] Kossaifi, J., Lipton, Z.C., Kolbeinsson, A., Khanna, A., Furlanello, T., and Anandkumar, A.: Tensor regression networks. J. Mach. Learn. Res. **21**, 123:1–123:21 (2020), http://jmlr.org/papers/v21/18-503.html

[97] Koutsouleris, N., Davatzikos, C., Borgwardt, S., et al.: Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders. Schizophrenia bulletin **40**(5), 1140–1153 (2014)

[98] Kramarow, E.A. and Tejada-Vera, B.: Dementia mortality in the united states, 2000-2017. National Vital Statistics Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System **68**(2), 1–29 (2019)

[99] Krioukov, D.V., Papadopoulos, F., Kitsak, M., Vahdat, A., and Boguñá, M.: Hyperbolic geometry of complex networks. CoRR **abs/1006.5169** (2010), http://arxiv.org/abs/1006.5169

[100] Krizhevsky, A., Sutskever, I., and Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017). doi: 10.1145/3065386, https://doi.org/10.1145/3065386

[101] Kuo, C.J., Zhang, M., Li, S., Duan, J., and Chen, Y.: Interpretable convolutional neural networks via feedforward design. J. Vis. Commun. Image Represent. **60**, 346–359 (2019). doi: 10.1016/J.JVCIR.2019.03.010, https://doi.org/10.1016/j.jvcir.2019.03.010

[102] de Lange, A.M.G. and Cole, J.H.: Commentary: Correction procedures in brain-age prediction. Neuroimage: Clinical **26** (2020)

[103] de Lange, A.M.G., Kaufmann, T., van der Meer, D., et al.: Population-based neuroimaging reveals traces of childbirth in the maternal brain. Proceedings of the National Academy of Sciences **116**(44), 22341–22346 (2019)

[104] Le, T.T., Kuplicki, R.T., McKinney, B.A., Yeh, H.W., Thompson, W.K., and Paulus, M.P.: A Nonlinear Simulation Framework Supports Adjusting for Age When Analyzing BrainAGE. Frontiers in Aging Neuroscience **10** (oct 2018). doi: 10.3389/fnagi.2018.00317, https://www.frontiersin.org/article/10.3389/fnagi.2018.00317/full

[105] LeCun, Y., Boser, B., Denker, J.S., et al.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**, 541–551 (1989)

[106] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998). doi: 10.1109/5.726791, https://doi.org/10.1109/5.726791

[107] Lepri, B., Oliver, N., Letouzé, E., Pentland, A., and Vinck, P.: Fair, Transparent, and Accountable Algorithmic Decision-making Processes. Philosophy & Technology **31**(4), 611–627 (2017). doi: 10.1007/s13347-017-0279-x, https://www.whitehouse.gov/blog/2016/05/03/preparing-future-artificial-intelligence

[108] Levakov, G., Rosenthal, G., Shelef, I., Raviv, T.R., and Avidan, G.: From a deep learning model back to the brain—Identifying regional predictors and their relation to aging. Human Brain Mapping **41**(12), 3235–3252 (8 2020). doi: 10.1002/hbm.25011

[109] Li, F.F., Wu, J., and Gao, R.: Cs231n convolutional neural networks for visual recognition (2022), https://cs231n.github.io/neural-networks-1/

[110] Lian, C., Liu, M., Zhang, J., and Shen, D.: Hierarchical fully convolutional network for joint atrophy localization and alzheimer's disease diagnosis using structural MRI. IEEE Trans. Pattern Anal. Mach. Intell. **42**(4), 880–893 (2020). doi: 10.1109/TPAMI.2018.2889096, https://doi.org/10.1109/TPAMI.2018.2889096

[111] Liem, F., Varoquaux, G., Kynast, J., et al.: Predicting brain-age from multimodal imaging data captures cognitive impairment. NeuroImage **148**, 179–188 (2017). doi: 10.1016/J.NEUROIMAGE.2016.11.005, https://doi.org/10.1016/j.neuroimage.2016.11.005

[112] Lim, D., Li, X., Hohne, F., and Lim, S.: New benchmarks for learning on non-homophilous graphs. CoRR **abs/2104.01404** (2021), https://arxiv.org/abs/2104.01404

[113] Litjens, G., Kooi, T., Bejnordi, B.E., et al.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (dec 2017). doi: 10.1016/j.media.2017.07.005, https://www.sciencedirect.com/science/article/pii/S1361841517301135

[114] Liu, M., Zhang, J., Adeli, E., and Shen, D.: Landmark-based deep multi-instance learning for brain disease diagnosis. Medical image analysis **43**, 157–168 (2018)

[115] Liu, M., Zhang, J., Adeli, E., and Shen, D.: Joint classification and regression via deep multi-task multi-channel learning for alzheimer's disease diagnosis. IEEE Trans. Biomed. Eng. **66**(5), 1195–1206 (2019). doi: 10.1109/TBME.2018.2869989, https://doi.org/10.1109/TBME.2018.2869989

[116] Liu, R.S.N., Lemieux, L., Bell, G.S., et al.: A longitudinal study of brain morphometrics using quantitative magnetic resonance imaging and difference image analysis. NeuroImage **20**(1), 22–33 (2003). doi: 10.1016/S1053-8119(03)00219-2, https://doi.org/10.1016/S1053-8119(03)00219-2

[117] Liu, T., Zhang, P., Huang, W., Zha, Y., You, T., and Zhang, Y.: How does layer normalization improve batch normalization in self-supervised sound source localization? Neurocomputing **567**, 127040 (2024). doi: 10.1016/J.NEUCOM.2023.127040, https://doi.org/10.1016/j.neucom.2023.127040

[118] Liu, Z., Cheng, J., Zhu, H., Zhang, J., and Liu, T.: Brain Age Estimation from MRI Using a Two-Stage Cascade Network with Ranking Loss. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **12267 LNCS**, 198–207 (2020). doi: 10.1007/978-3-030-59728-3{_}20

[119] Loshchilov, I. and Hutter, F.: Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019), https://openreview.net/forum?id=Bkg6RiCqY7

[120] Luan, S., Hua, C., Lu, Q., Zhu, J., Chang, X., and Precup, D.: When do we need GNN for node classification? CoRR **abs/2210.16979** (2022). doi: 10.48550/ARXIV.2210.16979, https://doi.org/10.48550/arXiv.2210.16979

[121] Luan, S., Hua, C., Lu, Q., et al.: Is heterophily A real nightmare for graph neural networks to do node classification? CoRR **abs/2109.05641** (2021), https://arxiv.org/abs/2109.05641

[122] Ma, Y., Liu, X., Shah, N., and Tang, J.: Is homophily a necessity for graph neural networks? In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022), https://openreview.net/forum?id=ucASPPD9GKN

[123] Maas, A.L., Hannun, A.Y., Ng, A.Y., et al.: Rectifier nonlinearities improve neural network acoustic models. In: Proc. icml. vol. 30, p. 3. Atlanta, Georgia, USA (2013)

[124] Marcus, D.S., Fotenos, A.F., Csernansky, J.G., Morris, J.C., and Buckner, R.L.: Open access series of imaging studies: longitudinal mri data in nondemented and demented older adults. Journal of cognitive neuroscience **22**(12), 2677–2684 (2010)

[125] Marsh, J. and Alifragis, P.: Synaptic dysfunction in alzheimer's disease: the effects of amyloid beta on synaptic vesicle dynamics as a novel target for therapeutic intervention. Neural regeneration research **13**(4), 616 (2018)

[126] Mason, J.P., Bintsi, K.M., Gilly, C., et al.: Leveraging artificial intelligence to enhance the science return of 4π solar constellations. Bulletin of the American Astronomical Society **55**(3), 269 (2023)

[127] Matsuda, H.: Voxel-based morphometry of brain MRI in normal aging and Alzheimer's disease. Aging and Disease **4**(1), 29–37 (2013)

[128] Millar, P.R., Gordon, B.A., Luckett, P.H., et al.: Multimodal brain age estimates relate to alzheimer disease biomarkers and cognition in early stages: a cross-sectional observational study. Elife **12**, e81869 (2023)

[129] Mittelstadt, B.D., Russell, C., and Wachter, S.: Explaining explanations in AI. CoRR **abs/1811.01439** (2018), http://arxiv.org/abs/1811.01439

[130] Montavon, G., Samek, W., and Müller, K.: Methods for interpreting and understanding deep neural networks. CoRR **abs/1706.07979** (2017), http://arxiv.org/abs/1706.07979

[131] Morris, C., Ritzert, M., Fey, M., et al.: Weisfeiler and leman go neural: Higher-order graph neural networks. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI

2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. pp. 4602–4609. AAAI Press (2019). doi: 10.1609/AAAI.V33I01.33014602, https://doi.org/10.1609/aaai.v33i01.33014602

[132] Mueller, T.T., Starck, S., Feiner, L.F., Bintsi, K., Rueckert, D., and Kaissis, G.: Extended graph assessment metrics for regression and weighted graphs. In: Ahmadi, S. and Pereira, S. (eds.) Graphs in Biomedical Image Analysis, and Overlapped Cell on Tissue Dataset for Histopathology - 5th MICCAI Workshop, GRAIL 2023 and 1st MICCAI Challenge, OCELOT 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, September 23, and October 4, 2023, Proceedings. Lecture Notes in Computer Science, vol. 14373, pp. 14–26. Springer (2023). doi: 10.1007/978-3-031-55088-1_2, https://doi.org/10.1007/978-3-031-55088-1_2

[133] Mueller, T.T., Zhou, S., Starck, S., et al.: Body fat estimation from surface meshes using graph neural networks. arXiv preprint arXiv:2308.02493 (2023)

[134] Müller, T.T., Starck, S., Bintsi, K.M., et al.: Are population graphs really as powerful as believed? Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=TTRDCVnbjI

[135] Müller, T.T., Starck, S., Dima, A., et al.: A survey on graph construction for geometric deep learning in medicine: Methods and recommendations. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=sWlHhfijcS, survey Certification

[136] Murphy, K.P.: Machine learning - a probabilistic perspective. Adaptive computation and machine learning series, MIT Press (2012)

[137] Nesterov, Y.: A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2). In: Doklady an ussr. vol. 269, pp. 543–547 (1983)

[138] Nwankpa, C., Ijomah, W., Gachagan, A., and Marshall, S.: Activation functions: Comparison of trends in practice and research for deep learning. CoRR **abs/1811.03378** (2018), http://arxiv.org/abs/1811.03378

[139] Paetzold, J.C., McGinnis, J., Shit, S., et al.: Whole brain vessel graphs: A dataset and benchmark for graph learning and neuroscience (vesselgraph). CoRR **abs/2108.13233** (2021), https://arxiv.org/abs/2108.13233

[140] Paquerault, S.: Battle against Alzheimer's Disease. Academic Radiology **19**(5), 509–511 (2012). doi: 10.1016/j.acra.2012.02.003, http://www.cdc.gov/

[141] Pardakhti, N. and Sajedi, H.: Brain age estimation based on 3d MRI images using 3d convolutional neural network. Multim. Tools Appl. **79**(33-34), 25051–25065 (2020). doi: 10.1007/S11042-020-09121-Z, https://doi.org/10.1007/s11042-020-09121-z

[142] Parisot, S., Ktena, S.I., Ferrante, E., et al.: Disease prediction using graph convolutional networks: Application to autism spectrum disorder and alzheimer's disease. Medical Image Anal. **48**, 117–130 (2018). doi: 10.1016/J.MEDIA.2018.06.001, https://doi.org/10.1016/j.media.2018.06.001

[143] Paszke, A., Gross, S., Massa, F., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., and Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. pp. 8024–8035 (2019), https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

[144] Pawlowski, N. and Glocker, B.: Is texture predictive for age and sex in brain mri? CoRR **abs/1907.10961** (2019), http://arxiv.org/abs/1907.10961

[145] Pei, H., Wei, B., Chang, K.C., Lei, Y., and Yang, B.: Geom-gcn: Geometric graph convolutional networks. CoRR **abs/2002.05287** (2020), https://arxiv.org/abs/2002.05287

[146] Pellegrini, C., Kazi, A., and Navab, N.: Unsupervised pre-training on patient population graphs for patient-level predictions. CoRR **abs/2203.12616** (2022). doi: 10.48550/ARXIV.2203.12616, https://doi.org/10.48550/arXiv.2203.12616

[147] Peng, H., Gong, W., Beckmann, C.F., Vedaldi, A., and Smith, S.M.: Accurate brain age prediction with lightweight deep neural networks. Medical Image Analysis **68**, 101871 (2021). doi: 10.1016/j.media.2020.101871, https://doi.org/10.1016/j.media.2020.101871

[148] Petersen, R.C., Aisen, P.S., Beckett, L.A., et al.: Alzheimer's disease neuroimaging initiative (adni) clinical characterization. Neurology **74**(3), 201–209 (2010)

[149] Popescu, S.G., Glocker, B., Sharp, D.J., and Cole, J.H.: A U-NET MODEL OF LOCAL BRAIN-AGE. bioRxiv (2021). doi: 10.1101/2021.01.26.428243, https://doi.org/10.1101/2021.01.26.428243

[150] Potamias, R.A., Neofytou, A., Bintsi, K., and Zafeiriou, S.: Graphwalks: Efficient shape agnostic geodesic shortest path estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022. pp. 2967–2976. IEEE (2022). doi: 10.1109/CVPRW56347.2022.00335, https://doi.org/10.1109/CVPRW56347.2022.00335

[151] Qi, Q., Du, B., Zhuang, M., Huang, Y., and Ding, X.: Age estimation from MR images via 3d convolutional neural network and densely connect. In: Cheng, L., Leung, A.C., and Ozawa, S. (eds.) Neural Information Processing - 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13-16, 2018, Proceedings, Part VII. Lecture Notes in Computer Science, vol. 11307, pp. 410–419. Springer (2018). doi: 10.1007/978-3-030-04239-4_37, https://doi.org/10.1007/978-3-030-04239-4_37

[152] Qian, N.: On the momentum term in gradient descent learning algorithms. Neural Networks **12**(1), 145–151 (1999). doi: 10.1016/S0893-6080(98)00116-6, https://doi.org/10.1016/S0893-6080(98)00116-6

[153] Rajmohan, R. and Reddy, P.H.: Amyloid-beta and phosphorylated tau accumulations cause abnormalities at synapses of alzheimer's disease neurons. Journal of Alzheimer's Disease **57**(4), 975–999 (2017)

[154] Reeve, A., Simcox, E., and Turnbull, D.: Ageing and parkinson's disease: why is advancing age the biggest risk factor? Ageing research reviews **14**, 19–30 (2014)

[155] Ribeiro, M.T., Singh, S., and Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the Demonstrations Session, NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016. pp. 97–101. The Association for Computational Linguistics (2016). doi: 10.18653/V1/N16-3020, https://doi.org/10.18653/v1/n16-3020

[156] Ronneberger, O., Fischer, P., and Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., III, W.M.W., and Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III. Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer (2015). doi: 10.1007/978-3-319-24574-4_28, https://doi.org/10.1007/978-3-319-24574-4_28

[157] Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review **65 6**, 386–408 (1958), https://api.semanticscholar.org/CorpusID:12781225

[158] Ruder, S.: An overview of gradient descent optimization algorithms. CoRR **abs/1609.04747** (2016), http://arxiv.org/abs/1609.04747

[159] Rumelhart, D.E., Hinton, G.E., and Williams, R.J.: Learning representations by back-propagating errors. Nature **323**(6088), 533–536 (1986). doi: 10.1038/323533a0, https://www.nature.com/articles/323533a0

[160] Savelieff, M.G., Nam, G., Kang, J., Lee, H.J., Lee, M., and Lim, M.H.: Development of multifunctional molecules as potential therapeutic candidates for alzheimer's disease,

parkinson's disease, and amyotrophic lateral sclerosis in the last decade. Chemical reviews **119**(2), 1221–1322 (2018)

[161] Savva, G.M., Wharton, S.B., Ince, P.G., Forster, G., Matthews, F.E., and Brayne, C.: Age, neuropathology, and dementia. New England Journal of Medicine **360**(22), 2302–2309 (2009)

[162] Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., and Monfardini, G.: The graph neural network model. IEEE Trans. Neural Networks **20**(1), 61–80 (2009). doi: 10.1109/TNN.2008.2005605, https://doi.org/10.1109/TNN.2008.2005605

[163] Scheltens, P.: Imaging in alzheimer's disease. Dialogues in clinical neuroscience **11**(2), 191–199 (2009)

[164] Schlemper, J., Oktay, O., Schaap, M., et al.: Attention gated networks: Learning to leverage salient regions in medical images. CoRR **abs/1808.08114** (2018), http://arxiv.org/abs/1808.08114

[165] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 618–626. IEEE Computer Society (2017). doi: 10.1109/ICCV.2017.74, https://doi.org/10.1109/ICCV.2017.74

[166] Shaban-Nejad, A., Michalowski, M., and Buckeridge, D.L.: Explainability and interpretability: keys to deep medicine. Explainable AI in healthcare and medicine: Building a culture of transparency and accountability pp. 1–10 (2021)

[167] Shrikumar, A., Greenside, P., and Kundaje, A.: Learning important features through propagating activation differences. CoRR **abs/1704.02685** (2017), http://arxiv.org/abs/1704.02685

[168] Silk, T.J. and Wood, A.G.: Lessons about neurodevelopment from anatomical magnetic resonance imaging (feb 2011). doi: 10.1097/DBP.0b013e318206d58f, http://www.ncbi.nlm.nih.gov/pubmed/21200332

[169] Simonyan, K., Vedaldi, A., and Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Bengio, Y. and LeCun, Y. (eds.) 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings (2014), http://arxiv.org/abs/1312.6034

[170] Simonyan, K. and Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y. and LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1409.1556

[171] Singh, A., Sengupta, S., and Lakshminarayanan, V.: Explainable deep learning models in medical image analysis. J. Imaging **6**(6), 52 (2020). doi: 10.3390/JIMAGING6060052, https://doi.org/10.3390/jimaging6060052

[172] Smilkov, D., Thorat, N., Kim, B., Viégas, F.B., and Wattenberg, M.: Smoothgrad: removing noise by adding noise. CoRR **abs/1706.03825** (2017), http://arxiv.org/abs/1706.03825

[173] Smith, S.M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T.E., and Miller, K.L.: Estimation of brain age delta from brain imaging. Neuroimage **200**, 528–539 (2019)

[174] Sone, D. and Beheshti, I.: Neuroimaging-based brain age estimation: A promising personalized biomarker in neuropsychiatry. Journal of Personalized Medicine **12**(11), 1850 (2022)

[175] Springenberg, J.T., Dosovitskiy, A., Brox, T., and Riedmiller, M.A.: Striving for simplicity: The all convolutional net. In: Bengio, Y. and LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings (2015), http://arxiv.org/abs/1412.6806

[176] Stankeviciute, K., Azevedo, T., Campbell, A., Bethlehem, R., and Lio, P.: Population graph gnns for brain age prediction. In: ICML Workshop on Graph Representation Learning and Beyond (GRL+). pp. 17–83 (2020)

[177] Sturmfels, P., Rutherford, S., Angstadt, M., Peterson, M., Sripada, C.S., and Wiens, J.: A domain guided CNN architecture for predicting age from structural brain images. In: Doshi-Velez, F., Fackler, J., Jung, K., et al. (eds.) Proceedings of the Machine Learning for Healthcare Conference, MLHC 2018, 17-18 August 2018, Palo Alto, California. Proceedings of Machine Learning Research, vol. 85, pp. 295–311. PMLR (2018), http://proceedings.mlr.press/v85/sturmfels18a.html

[178] Szegedy, C., Liu, W., Jia, Y., et al.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 1–9. IEEE Computer Society (2015). doi: 10.1109/CVPR.2015.7298594, https://doi.org/10.1109/CVPR.2015.7298594

[179] Tan, M. and Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri, K. and Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114. PMLR (2019), http://proceedings.mlr.press/v97/tan19a.html

[180] Thompson, P., Moussai, J., Zohoori, S., et al.: Cortical variability and asymmetry in normal aging and alzheimer's disease. Cerebral Cortex (New York, NY: 1991) **8**(6), 492–509 (1998)

[181] Tohka, J., Moradi, E., and Huttunen, H.: Comparison of feature selection techniques in machine learning for anatomical brain MRI in dementia. Neuroinformatics **14**(3), 279–296 (2016). doi: 10.1007/S12021-015-9292-3, https://doi.org/10.1007/s12021-015-9292-3

[182] Tremblay, B., Jarolim, R., Jungbluth, A., et al.: Ai to enhance the capabilities of euv-observing satellites and estimate spectral irradiance. Tech. rep., Copernicus Meetings (2024)

[183] Ulyanov, D., Vedaldi, A., and Lempitsky, V.S.: Instance normalization: The missing ingredient for fast stylization. CoRR **abs/1607.08022** (2016), http://arxiv.org/abs/1607.08022

[184] Varatharajah, Y., Baradwaj, S., Kiraly, A., et al.: Predicting brain age using structural neuroimaging and deep learning. BioRxiv p. 497925 (2018)

[185] Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y.: Graph attention networks. CoRR **abs/1710.10903** (2017), http://arxiv.org/abs/1710.10903

[186] Wang, B. and Pham, T.D.: MRI-based age prediction using hidden Markov models. Journal of Neuroscience Methods **199**(1), 140–145 (jul 2011). doi: 10.1016/j.jneumeth.2011.04.022, https://www.sciencedirect.com/science/article/pii/S0165027011002263

[187] Wang, J., Li, W., Miao, W., Dai, D., Hua, J., and He, H.: Age estimation using cortical surface pattern combining thickness with curvatures. Medical and Biological Engineering and Computing **52**(4), 331–341 (apr 2014). doi: 10.1007/s11517-013-1131-9, http://link.springer.com/10.1007/s11517-013-1131-9

[188] Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., and Solomon, J.M.: Dynamic graph CNN for learning on point clouds. CoRR **abs/1801.07829** (2018), http://arxiv.org/abs/1801.07829

[189] Wei, S. and Zhao, Y.: Graph learning: A comprehensive survey and future directions. CoRR **abs/2212.08966** (2022). doi: 10.48550/ARXIV.2212.08966, https://doi.org/10.48550/arXiv.2212.08966

[190] Weiner, M.F., Hynan, L.S., Bret, M.E., and White, C.: Early behavioral symptoms and course of Alzheimer's disease. Acta Psychiatr Scand **111**(5), 367–371 (2005), https://dx.doi.org/10.1111/j.1600-0447.2004.00472.x

[191] Wolz, R., Aljabar, P., Hajnal, J.V., Hammers, A., Rueckert, D., and Weiner, M.W.: LEAP: learning embeddings for atlas propagation. Neuroimage **49**(2), 1316–1325 (2010), https://dx.doi.org/10.1016/j.neuroimage.2009.09.069

[192] Wu, Y. and He, K.: Group normalization. In: Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII. Lecture Notes in Com-

puter Science, vol. 11217, pp. 3–19. Springer (2018). doi: 10.1007/978-3-030-01261-8_1, https://doi.org/10.1007/978-3-030-01261-8_1

[193] Xu, K., Hu, W., Leskovec, J., and Jegelka, S.: How powerful are graph neural networks? In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019), https://openreview.net/forum?id=ryGs6iA5Km

[194] Yin, X., Qiu, Y., Zhao, C., Zhou, Z., Bao, J., and Qian, W.: The Role of Amyloid-Beta and Tau in the Early Pathogenesis of Alzheimer's Disease. Med Sci Monit **27**, e933084 (2021), https://dx.doi.org/10.12659/MSM.933084

[195] Zeiler, M.D. and Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D.J., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.) Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I. Lecture Notes in Computer Science, vol. 8689, pp. 818–833. Springer (2014). doi: 10.1007/978-3-319-10590-1_53, https://doi.org/10.1007/978-3-319-10590-1_53

[196] Zeisel, J., Bennett, K., and Fleming, R.: World alzheimer report 2020: Design, dignity, dementia: Dementia-related design and the built environment (2020), https://www.alzint.org/resource/world-alzheimer-report-2020/

[197] Zhang, Q. and Zhu, S.: Visual interpretability for deep learning: a survey. Frontiers Inf. Technol. Electron. Eng. **19**(1), 27–39 (2018). doi: 10.1631/FITEE.1700808, https://doi.org/10.1631/FITEE.1700808

[198] Zhao, X., Zhou, F., Ou-Yang, L., Wang, T., and Lei, B.: Graph convolutional network analysis for mild cognitive impairment prediction. In: 16th IEEE International Symposium on Biomedical Imaging, ISBI 2019, Venice, Italy, April 8-11, 2019. pp. 1598–1601. IEEE (2019). doi: 10.1109/ISBI.2019.8759256, https://doi.org/10.1109/ISBI.2019.8759256

[199] Zheng, S., Zhu, Z., Liu, Z., et al.: Multi-modal graph learning for disease prediction. IEEE Trans. Medical Imaging **41**(9), 2207–2216 (2022). doi: 10.1109/TMI.2022.3159264, https://doi.org/10.1109/TMI.2022.3159264

[200] Zheng, X., Liu, Y., Pan, S., Zhang, M., Jin, D., and Yu, P.S.: Graph neural networks for graphs with heterophily: A survey. CoRR **abs/2202.07082** (2022), https://arxiv.org/abs/2202.07082

[201] Zhou, Y. and Chellappa, R.: Computation of optical flow using a neural network. In: Proceedings of International Conference on Neural Networks (ICNN'88), San Diego, CA, USA, July 24-27, 1988. pp. 71–78. IEEE (1988). doi: 10.1109/ICNN.1988.23914, https://doi.org/10.1109/ICNN.1988.23914

[202] Zhu, J., Yan, Y., Zhao, L., Heimann, M., Akoglu, L., and Koutra, D.: Beyond homophily in graph neural networks: Current limitations and effective designs. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020), https://proceedings.neurips.cc/paper/2020/hash/58ae23d878a47004366189884c2f8440-Abstract.html