Investigating brain aging with Magnetic Resonance Imaging and Convolutional Neural Networks

Sokolova, Ksenia

Awarding institution:
King's College London

Investigating brain aging with Magnetic Resonance Imaging and Convolutional Neural Networks

Ksenia Sokolova
Department of Neuroimaging
King's College London

A thesis submitted for the degree of
Doctor of Philosophy
in Neuroimaging Research

May 29, 2022

To my family

Investigating brain ageing with Magnetic Resonance Imaging and Convolutional Neural Networks

Ksenia Sokolova

Abstract

Throughout human lifetime there are various ageing-related changes occurring. It has been demonstrated that departures from the healthy ageing trajectory can be used as a biomarker for several neurodegenerative conditions. Previous neuroimaging studies have used deep learning algorithms to investigate structural biomarkers including volumetric, microstructural and focal ones, using both cross-sectional and longitudinal cohorts. In this thesis, I investigate deep learning methods for analysing structural Magnetic Resonance Imaging (MRI) data. The methods developed are used to understand brain ageing throughout human adulthood, by identifying *population-wide brain ageing profiles*. In this thesis such profiles are defined by determining periods of the human lifetime which have similar brain age-related features and studying the differences between the features characterising each period. Brain ageing profiles covering human adulthood are investigated as brain ageing occurs during adulthood. Training dataset used in this thesis to train brain age prediction algorithms is the largest up-to-date and, therefore, the algorithms can be considered as generalisable as possible. Therefore, brain ageing profiles are called *"population wide"*.

Population-wide brain ageing profiles are derived and described by developing 5 main contributions in this thesis:

1. A dataset comprising of 10,878 MRI T1-weighted scans acquired in healthy subjects (17 – 96 years of age) was assembled from publicly available data.

2. The dataset was used to develop a convolutional neural network-based ordinal regression model for brain age prediction in order to reflect accumulative nature of brain ageing, using the imaging data with minimal preprocessing (rigid transformation and resampling onto a template).

3. The distribution of predicted ages from both ordinal and metric models were leveraged to propose ageing profile by assuming that the distribution of age predictions represents the distribution of age-related structural features. Ageing profiles were

obtained using 4 different methods – 2 methods were applied for this purpose (Deep Embedded Clustering (DEC) and Preference Ranking Organization METHod for Enrichment of Evaluations (PROMETHEE) II) and 2 methods were developed for this thesis (the method considering the distribution of predictions of a deep learning model for brain age prediction and ordinal DEC). Ordinal DEC was proposed to introduce the concept of ordinality into the standard DEC.

4. The issues of extracting brain ageing-related features describing brain ageing profiles from MRI data were addressed. In order to understand which features of the original inputs which correspond to ageing, five existing saliency mapping methods (vanilla backpropagation, guided backpropagation, the Smoothgrad method, Gradient Class Activation Mapping (Grad-CAM), guided Grad-CAM) were compared, and a methodology for applying such methods to ordinal methods was proposed. As all the existing saliency mapping methods used were originally developed to work on natural images, modifications were needed to adapt them to work on 3D input MRI data.

5. Generalisability of the models presented in this thesis was examined. For this purpose their performance was assessed on an independent clinical dataset - the Institute of Psychiatry, Psychology & Neuroscience (IoPPN) dataset. Further, the performance of brain age prediction models was observed on the Healthy Brain Ageing from Public Sources (HBAPS) dataset with reduced resolution.

Accuracy of the ordinal model implemented for the task of brain age prediction was compared to the deep learning model with metric regression. The ordinal and metric models achieved Mean Absolute Error (MAE) of 3.62 and 3.87 years, respectively, on the test data. On the same test data with repeat scans eliminated, i.e. excluding repeat samples acquired from the subjects contained in the training data, ordinal and metric models achieved MAE of 4.10 and 4.37 years. Both metric and ordinal models achieved better results than previously published experiments using input data with the same level of pre-processing. The performance of both models was shown not to be affected by the fact that the dataset contained data collected using different scanners and protocols. The models' performance was, however, affected for the samples with labels for which the number of samples in the training data was small.

The profiles obtained using the four methods, and both metric and ordinal models, all agreed within models' accuracy defined by MAE achieved. This suggests that they are all being driven by the same underlying brain features, and that these features may be biologically meaningful.

In order to aid interpretation of the age-related features defining the ageing profiles both subject-specific and averaged importance maps were obtained. Further, the importance maps were compared between the ones obtained using ordinal and metric models for brain age prediction. All the results were also interpreted for biological significance. The results agree with existing knowledge on the biology of brain ageing. Among existing saliency mapping methods the methods of guided backpropagation, Grad-CAM and guided Grad-CAM are shown to be usable for studying brain ageing features from MRI data. Saliency maps produced using proposed methodology for saliency mapping for ordinal models also highlighted biologically interpretable features.

The overall aim of this work is not just to develop the methodologies and obtain corresponding results, but also to consider the feasibility of such methods for future use in a clinical setting, in light of the challenges outlined above.

Contents

List of Figures

List of Tables

Acronyms

ADNI Alzheimer's Disease Neuroimaging Initiative. 29, 137, 160, 221

BLSA Baltimore Longitudinal Study of Aging. 134

CAE Convolutional Autoencoder. 60

CAM Class Activation Mapping. 10, 63, 65, 66, 162, 163, 166

CLRP Contrastive Layer-wise Relevance Propagation. 10, 63, 72, 73, 168

CNN Convolutional Neural Network. 6, 7, 9, 10, 30, 34, 39, 41, 45–49, 51–53, 56, 57, 60, 61, 63, 65–68, 71, 72, 101, 104, 105, 129–131, 134, 157, 161, 163, 165, 168, 192

CSF Cerebrospinal Fluid. 36, 82, 103, 134–136, 157, 164, 192

CT Computed Tomography. 29

DEC Deep Embedded Clustering. 4, 7, 13–19, 22, 33, 61, 134, 142–145, 147–149, 152, 154–156, 160, 175–178, 188, 191, 193, 210, 212–218, 222, 225, 226

DGP Deep Gaussian Processes. 30, 105, 190

DL Deep Learning. 21, 29, 30, 39–41, 62, 102, 103, 105, 196

DNN Deep Neural Network. 61, 72

DOFs Degrees of Freedom. 105

DWI Diffusion-weighted Imaging. 36, 99

FDA U.S. Food and Drug Administration. 198

FLAIR Fluid-Attenuated Inversion Recovery. 36, 99

fMRI Functional MRI. 83

GAN Generative Adversarial Network. 160, 226

GAP Global Average Pooling. 10, 65–67

GM Grey Matter. 36, 82, 83, 102–105, 134, 135, 157, 158, 164, 190, 192

GP Gaussian Processes. 30, 103, 104, 164

GPU Graphics Processing Unit. 29, 39, 46, 108

Grad-CAM Gradient Class Activation Mapping. 4, 5, 10, 14–19, 33, 63, 66–69, 161–168, 170, 172, 175–184, 186–193, 197, 198, 212–218, 225

HBAPS Healthy Brain Ageing from Public Sources. 4, 7, 8, 11, 12, 15, 17–23, 87, 88, 95, 97–99, 108–131, 133, 135, 137, 140, 142, 153, 158, 160, 161, 171, 173–175, 192, 195–205, 210–217, 219–222, 224–226

HCP Human Connectome Project. 29

HIPAA Health Insurance Portability and Accountability Act. 20, 88, 95

ILSVRC ImageNet Large-Scale Visual Recognition Challenge. 86

IoPPN Institute of Psychiatry, Psychology & Neuroscience. 4, 7, 8, 11, 18, 19, 23, 99, 100, 195, 196, 198, 206–218, 220–222, 225

KL Kullback-Leibler. 60–62

LRP Layer-wise Relevance Propagation. 10, 14, 20, 33, 63, 69–73, 162, 166, 167, 170, 172, 187, 188

MAE Mean Absolute Error. 4, 5, 14, 18, 19, 21, 23, 30, 102, 104, 108–110, 129–131, 140, 143, 148, 156, 171, 196–199, 201, 202, 206, 209, 210, 213, 216, 219–221, 224

MCDA Multiple-Criteria Decision Analysis. 73, 74

ML Machine learning. 39, 41

MLFFNN Multi-Layer Feed Forward Neural Network. 41

MLP Multi-Layer Perceptron. 9, 39, 41–46, 48, 50, 60, 62, 164, 191

MMSE Mini-mental State Examination. 20, 93–95, 97

MNIST Mixed National Institute of Standards and Technology. 6, 10, 84, 85

MoCA Montreal Cognitive Assessment. 20, 95

MRI Magnetic Resonance Imaging. 3–5, 7, 9, 13, 21, 29–32, 35–37, 46, 81, 82, 84, 87, 99–105, 129, 131, 133–140, 142–144, 148, 157, 158, 161, 164, 165, 167, 192, 193, 195–198, 219, 224–226

MSE Mean Squared Error. 59

NCI National Institutes of Health of the United States Department of Health and Human Services. 198

NFTs Neurofibrillary Tangles. 81

NN Neural Network. 41, 43, 44, 46

PET Positron Emission Tomography. 82, 83

PROMETHEE Preference Ranking Organization METHod for Enrichment of Evaluations. 4, 6, 7, 13, 22, 33, 34, 73–75, 133, 140–142, 149, 154, 155, 225

ReLU Rectified Linear Unit. 9, 10, 42, 47, 54, 64, 66, 67, 147, 166, 167

ResNet Residual Network. 9, 10, 51, 52, 55

ROI Regions of Interest. 103, 105

RVR Relevance Vector Regression. 102, 103, 164

SAE Stacked Autoencoder. 147, 148

SCID Structured Clinical Interview used for DSM Disorders. 20, 94, 95

SCR Subcallosal Region. 82, 135

SE-ResNet Squeeze-and-Excitation Residual Network. 53

SENet Squeeze-and-Excitation Network. 11–19, 21–23, 53, 104, 107–132, 138, 140, 142, 144, 148–152, 154–156, 158–160, 166–168, 171, 173–176, 179–187, 189, 191–193, 197–222

SGD Stochastic Gradient Descent. 44, 45, 108, 148

SGL Sparse Group Lasso. 103, 164

SPECT Single Photon Computed Emission Tomography. 82

SPs Senile Plaques. 81

SVM Support Vector Machine. 56, 102, 103

SVR Support Vector Regression. 163

SWI Susceptibility-weighted Imaging. 36, 99

t-SNE t-distributed Stochastic Neighbour Embedding. 30

WM White Matter. 77, 82, 83, 103–105, 134, 135, 157, 164, 190

Publications

The work presented in this thesis has in part been published in the following conference paper and poster:

- Ksenia Sokolova, Gareth J. Barker, and Giovanni Montana. Convolutional neural-network-based ordinal regression for brain age prediction from MRI scans. *Proceedings of the Medical Imaging 2020: Image Processing Conference*, 113132B, Houston, March 2020. URL https://doi.org/10.1117/12.2549636.

- Ksenia Sokolova, Gareth J. Barker, and Giovanni Montana. Convolutional neural network-based ordinal regression for brain age. Poster presented at the conference "CompAge2020: Computational approaches for ageing and age-related diseases", Paris, France, September 2020.

Chapter 1

Introduction

Developed countries currently face ageing of populations. According to the 2018 Ageing Report by European Commission the old-age dependency ratio, i.e. the ratio of the number of people aged 65 and older to the number of people aged 15 to 64 years old, will increase by 21.6% in 2070 relative to 2016. In UK the old-age dependency ratio will increase by 18.1% in this period [15]. This is expected to lead to an increase in the number of patients with age-related neurodegenerative diseases, e.g. dementia, Parkinson's disease etc. Dementia is a overarching term for a number of progressive diseases affecting patients' cognitive abilities and behaviour and thus leading to decreased patients' ability to perform day-to-day activities. Even though the most important risk factor for developing dementia is age, dementia does not only affect older people. In up to 9% of dementia cases the onset of symptoms occurs before the age of 65 years [16]. In 2015, there were approximately 47 million patients living with dementia worldwide. Due to population ageing, it is predicted that the number of patients in the world with dementia will rise to 75 million by 2030 and to 132 million by 2050 [17]. All these trends in population ageing and prevalence of neurodegenerative diseases highlight the importance of brain ageing research. The early onset of some neurodegenerative diseases also highlights the importance of research of whole brain ageing course over human lifetime.

Modern brain ageing research has three main directions: studying ageing at a cellular level and using genetic data in animals, studying human brain ageing by cognitive assessments, and analysis of neuroimaging data. Animal studies of the cell and genetic samples are performed to understand „hallmarks of ageing“. The term *„hallmarks of ageing“* describes the process of ageing in terms of its main determinants. The hallmarks of

ageing include accumulation of genetic changes during the lifetime, dysfunction of intercellular organoids and their communication, impairment of homeostasis and proteostasis, and decline in the regenerative abilities of tissues [18]. Cognitive neuroscience of ageing encompasses the behavioural research and studies of age-related neural changes. The techniques of cognitive neuroscience study the relation of ageing related cognitive changes to structural and functional brain changes [19]. Cognitive assessment measures include tests of such cognitive abilities as memory, attention, manipulation of information and processing speed [20]. Genetic determinants of ageing can be explored by analysing genetically modified animals in controlled experimental settings. Experiments in controlled environments also help with understanding of genetic or epigenetic nature of ageing determinants [21]. Human gene studies of brain ageing have also been performed, to study risk factors [22], heritability and relationship between ageing and neurodegenerative diseases [23]. Statistical analysis of structural and functional data is another approach used to understand human brain ageing [24, 25]. Structural Magnetic Resonance Imaging (MRI) data is frequently used in academic research of human brain as it typically has higher contrast and spatial resolution compared to Computed Tomography (CT) scans or other neuroimaging techniques, and does not involve ionising radiation, making it particularly appropriate for longitudinal studies.

In recent years machine learning methods has arisen among statistical methods for the analysis of MRI data. Machine learning methods used for the MRI data analysis vary from regression models to Deep Learning (DL) approaches. This has three main reasons. Firstly, the Graphics Processing Unit (GPU), which was initially invented in 1970s to be used in arcade games, was first introduced for use in machine learning for image recognition in 2009 [26]. This application was further developed by Cireşan et al [27]. Secondly, a number of software libraries such as Torch [28], Pytorch [29] or Tensorflow [30], has been introduced and developed to ease the implementation of machine learning methods. Thirdly, emergence of publicly available data-sharing initiatives, such as the Human Connectome Project (HCP) [31] or Alzheimer's Disease Neuroimaging Initiative (ADNI) [32], allowed the researchers all over the world to develop statistical models without actual MRI data acquisition. Data-sharing initiatives are also supported by the development of data-sharing platforms [33, 34].

As statistical methods were developed into machine learning methods, neural network-based classifiers can now be successfully used for brain age prediction from structural MRI

scans [35, 36, 37, 38, 39]. Most of the existing approaches handle pre-processed MRI scans. For example, Popescu et al [38] and Huang et al [40] trained brain age prediction classifier on MRI data segmented into white and grey matter. Huang et al [40] used a Convolutional Neural Network (CNN) and reported a Mean Absolute Error (MAE) of 4.0 years. Feng et al [39] applied an algorithm to extract the brain from a MRI scan before training. Both CNN models and the models derived from CNNs are used to analyse preprocessed MRI data. For instance, CNN can be also used to construct Gaussian Processes (GP) model. Popescu et al [38] used Deep Gaussian Processes (DGP) for voxel-wise interpretation for brain age. They reported their model's MAE as 3.85 years. To date, the lowest MAE reported on MRI data with minimal preprocessing is 4.16 years by Cole et al [35].

However, the DL methods are so-called "black box" methods as their decision logic cannot be traced by a human. For the DL methods to be applicable clinically the DL model results should be as interpretable and explainable as human doctors' decisions are. In order to improve the applicability of these models, various visualisation methods have been proposed. One group of methods, based on t-distributed Stochastic Neighbour Embedding (t-SNE) [41], visualises the distribution of output decisions in order to understand the outliers. Other methods aim to visualise the features which were important in determining the network's final output. Such methods are called *"saliency mapping methods"* [42, 43, 7, 5, 44, 45, 46, 47]. Saliency mapping methods take different approaches for understanding the importance of features in an input. Zeiler and Fergus [47] trace the path of features backwards from a corresponding output by inverting convolutional layers. Another group of authors [42, 43, 5, 45] consider a process of backpropagation, i.e. the process of adjusting network's parameters during its training, and the data contained in these calculations. Montavon et al [7] propagate relevance from an output layer-by-layer. Features extracted using saliency mapping methods may be also useful for the insights they can provide into the nature of tasks themselves. For example, age-related features extracted from a neural network trained for the task of brain age prediction from structural MRI scans may provide understanding of the process of ageing itself. However, the debate regarding the clinical applicability has been deepened by showing that deep neural networks can be easily fooled into producing high confidence decisions from input which look like random noise to a human observer [48, 49, 50].

In order to be applied clinically, the DL methods, which are usually trained either on publicly available data collected for research purposes or limited clinical datasets, should be

generalisable to clinical data and must capture the population-wide effects. Generalisability is a complex issue, due to lack of availability of accurate labelling for most of the diagnostic tasks. The necessary labelling of large datasets by experts may be both expensive and time consuming. The need for labelling the data may be reduced by using unsupervised or semi-supervised learning methods. Furthermore, however, for the models to be used clinically the uncertainty over their decisions should be quantified. There are two types of uncertainty present in neural network predictions – *epistemic* and *aleatoric*. Epistemic uncertainty arises from model's structure and parameters. It is caused by the lack of training data or differences in the distributions between the training and testing datasets. Epistemic uncertainty may lead to complications when a model trained on a research dataset is applied clinically, for example. Aleatoric uncertainty on the other hand arises from noise in the training data or missing data for some labels in the dataset. Currently available MRI datasets, which can be used for studying brain ageing, suffer from large imbalances in the amount of data available per label and this affects generalisability too. This needs to be accounted for to avoid model biases, for example by incorporating measures to alleviate imbalance into loss functions [4].

1.1 Thesis overview

This thesis describes developing deep learning methods for analysing structural MRI, with the overall goal of understanding the process of brain ageing in the form of *population-wide brain ageing profiles*. In this context, I have defined a brain ageing profile by determining periods of the human adulthood which have similar brain age-related features and studying the differences between the features characterising each period. The overall aim of this work is not just to develop the methodologies and obtain corresponding results, but also to consider the credibility of such methods for future use in a clinical settings in light of the challenges outlined above. The main contributions of this thesis are:

1. compiling a dataset of structural MRI scans from publicly available data,
2. implementing and optimising a classifier for predicting brain age from MRI scans,
3. extracting a population-wide brain ageing profile using both existing methods and with two new methods developed specifically for this purpose,

4. obtaining age-related features based on the ageing profile intervals determined by these methods and developing a saliency mapping method suitable for ordinal regression classifier,

5. testing the approach on data representative of that typically collected in a clinical setting.

In order to achieve these goals, a MRI dataset comprising of 10,878 3D T1-weighted MRI scans acquired in healthy subjects was constructed from 34 publicly available datasets. All above mentioned contributions were developed on the dataset with minimal preprocessing applied, i.e. registration using rigid transformation and resampling onto MNI152 template (detailed description in Section 4.2.3). There are three main reasons for using data with minimal preprocessing in this thesis. Firstly, choosing minimal preprocessing allows to make interpretable brain age predictions as this allows relating predicted brain age to important MRI regions. Secondly, a linear registration is chosen as a non-linear registration method would scale same regions in different MRI scans to the same size. Therefore, information on age-related atrophy of various brain regions would be lost in preprocessing. For example, previous published research highlights ventriclular volume increase with age [51] and non-linear registration is unsuitable for further extracting this feature from MRI data. Thirdly, scull stripping was not used in the preprocessing pipeline as that would not allow assessing the size of subarachnoid space and importance of this feature was previously reported by Herent et al [52].

In this thesis *population-wide brain ageing profiles* are investigated as the dataset collated from public datasets is the largest up-to-date used for training. Therefore, the resulting profiles can be considered to be as generalisable as possible and in this thesis the term *"population-wide"* is used as a shorthand to describe the profiles investigated.

The MRI dataset constructed to study human brain ageing is described in Chapter 4 along with the synthetic datasets generated specifically to demonstrate the methods used in this thesis. Chapter 5 describes the development of a classifier for predicting brain age from MRI scans using metric and ordinal regression. Chapter 6 describes extracting population-wide brain ageing profiles from the results obtained in Chapter 5. In order to obtain age-related features based on these ageing profiles, a range of existing saliency mapping methods were implemented and compared; the features obtained are described and compared to existing literature in Chapter 7. Finally, Chapter 8 describes testing of

the approaches presented above on data more representative of the clinical setting, using an independent, locally collected, dataset, in which two test data subsets are used consisting of healthy controls and patients.

1.2 Approvals

1.2.1 Data approvals

Data usage approvals were obtained for all 32 public and 3 private datasets used.

1.2.2 Code approvals

This PhD uses a mixture of locally developed and publicly available software. All code is used in accordance with the corresponding publisher's license. The source of the code used in each section is detailed.

In Chapter 5 feature extraction of the brain age prediction classifiers with both metric and ordinal regression was implemented using squeeze-and-excitation network (SENet) implementation of R. Hataya [53]

In Chapter 6, for extracting population-wide brain ageing profiles following code sources were used:

- the implementation of the Preference Ranking Organization METHod for Enrichment of Evaluations (PROMETHEE) method by A. Dejonckheerey [54],
- the implementation of the Deep Embedded Clustering (DEC) by V. Lukiyanov [55].

In Chapter 7 following implementations of saliency mapping methods were used:

- the implementation of vanilla backpropagation [42], guided backpropagation [46], Gradient Class Activation Mapping (Grad-CAM) [5], guided Grad-CAM and the SmoothGrad methods [45] was based on implementation by U. Ozbulak [56].
- the implementaiton of method of Layer-wise Relevance Propagation (LRP) by [7] was based on work by W. Nam [57].

1.2.3 Figure approvals

For all figures borrowed from existing literature corresponding approvals were received.

Chapter 2

Background

In this chapter the theory and related work useful to understand the content of this PhD thesis are presented. Firstly, the context is set by describing essential imaging terminology. Secondly, the concept of deep learning is introduced, along with the Convolutional Neural Network (CNN) and autoencoder. Two specific aspects of deep learning methodologies, on which the work in this thesis is based, – unsupervised clustering and saliency mapping - are also discussed . Thirdly, the Preference Ranking Organization METHod for Enrichment of Evaluations (PROMETHEE) is described, as it is also used in later chapters of the thesis. Finally, a review is provided of existing knowledge on brain ageing.

2.1 Imaging terminology

In this section essential imaging terminology, which is used throughout the thesis, is described. The terminology is first introduced in the context of natural images and then terms specific to the description of medical images are considered.

Imaging data can be represented in space with more than two dimensions. One-, two-, or three-dimensional spaces are denoted as 1D, 2D and 3D spaces respectively. In medical imaging 4D spaces are used to reflect changes of 3D data over time. The choice of space is dependent on how many parameters are minimally required to describe each point in the data. For instance, natural images are described in a 2D space, as two parameters (x and y position) are required to state position of each element. Each element of a natural image is a *pixel*. A pixel is the smallest unit of an image [58]. *Image size* can be defined in terms of number of pixels - in 2D, the image's width and height. On the other hand, *pixel size* is determined by image size and *resolution* as resolution defines how many pixels are

needed to represent an image of given size at a particular level of detail. Natural images can be either *grayscale* or *colour*. In grayscale images each element is represented solely by its intensity, which is described by a single value. In colour images each pixel is described by a *colour model* of choice. A colour model is a mathematical model defining colours as sets of numbers [59]. For example, frequently used model is *RGB colour model*. In this model each pixel is typically described by three values as every colour can be represented by mixing three primary colours – red, green and blue [60]. These three values describing each pixel are known as *channels*.

Medical images are typically acquired in the form of two- or three-dimensional data. 3D medical images are often displayed and manipulated by considering 2D slices from 3D data to ease visualisation and analysis, or reduce computational costs and time. Conversely, 3D medical data are often created by *stacking* 2D slices, to allow reformatting in orientations other than that of the original acquisition). 4D medical data is also used to capture changes of 3D data over time. In this thesis we consider Magnetic Resonance Imaging (MRI) data which are stored in 3D form, regardless of whether the original data acquisition was 2D or 3D. Each element of 3D data is known as a *voxel*, the 3D equivalent of a 2D pixel. A voxel is the smallest unit of 3D data used to describe it. 3D data size is usually written as a product referred to as the *matrix size* (e.g. $256 \times 256 \times 256$). 3D data size can be defined in terms of number of voxels used to represent a physical 3D volume. *Voxel size* and the size of physical volume are related through *field of view* (FOV). Medical images are also characterised by their FOV. FOV is the overall size, in physical space, of the region which is captured in an image [61]; this is measured in cm or mm, and is typically quoted as 1, 2, or 3 numbers (e.g. $24cm$; $24 \times 22cm$; $24 \times 22 \times 16cm$, with the first of these implicitly implying either a $24 \times 24cm$ 2D image, or a $24 \times 24 \times 24cm$ 3D image, and so on).

Structural MRI data typically has only a single channel, and can be displayed as a grayscale image. MRI data can differ by the sequence type used in acquisition. MRI signal is acquired by applying radio frequency pulses to a tissue to be imaged. The resulting signal decays (usually exponentially) at a rate depending of a number of inherent parameters determined by the tissue structure, with the overall signal intensity depending both on these and on (operator selectable) sequence timing parameters. As a result, sequence parameter settings can be used to change the contrast (i.e. the relative signal form two different tissues) of the resulting scans. For example, there are two frequently used parameter settings, and thus the contrast of resulting scans – T1-weighted and T2-weighted MRI as

shown in Figure 2.1. In T1-weighted scans of brain Grey Matter (GM) appears darker compared to white, while in T2-weighted the relationship is the opposite. In T1-weighted data Cerebrospinal Fluid (CSF) also appears darker compared to the T2-weighted data [62]. There are also many other sequences used. Here are some examples:

- *Fluid-Attenuated Inversion Recovery (FLAIR)* MRI sequence is used to suppress the impact of CSF on resulting images and highlight the lesions [63],
- *Susceptibility-weighted Imaging (SWI)* sequence is used to image venous blood, iron deposition and haemorrhage [64],
- *Diffusion-weighted Imaging (DWI)* sequence is used to image diffusion of water molecules to study tissue structure [65].

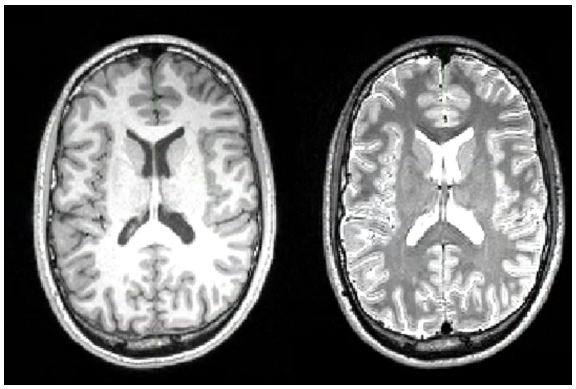

Figure 2.1: Examples of T1-weighted (left) and T2-weighted (right) brain scans.

2.2 MRI data preprocessing

In this thesis, data preprocessing is described in Section 4.2.3 and comprises a registration process which involves *rigid -body transformation* and resampling onto the MNI152 template. Here the concepts of each these steps is described.

2.2.1 Registration

Registration of two MRI scans is a process aligning them such that common features overlap, and the result also reveals differences between the two. The process of image registration is a process of calculating the geometric transformation which minimises the difference between the two images, as assessed by a *cost function* [66]. Figure 2.2.1 shows an example of registration.

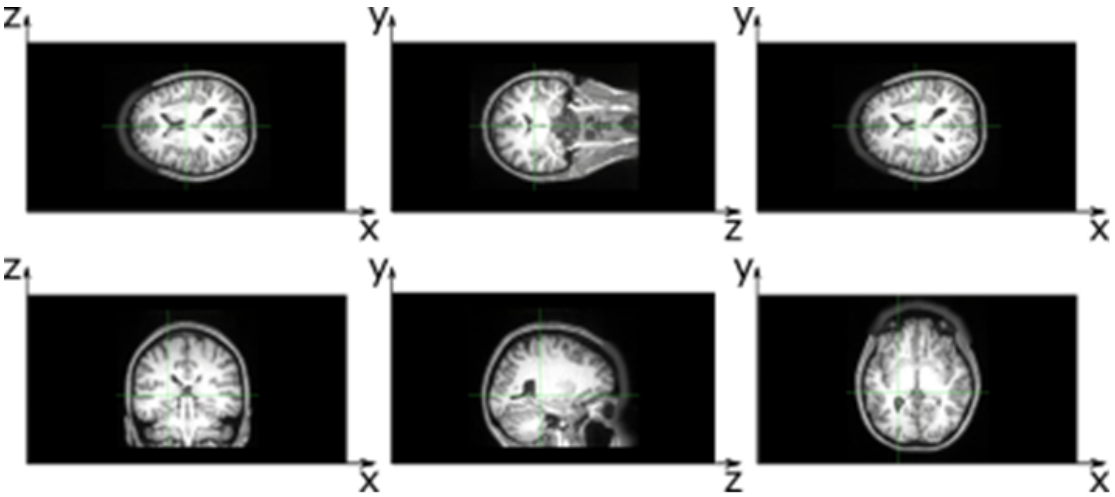

Figure 2.2: Example of registration results. Top row: original MRI scan. Bottom row: the same coordinates in the original scan registered onto MNI152 template.

Rigid body transformation

The 3D rigid body transformation for a point (x_1, x_2, x_3) into new coordinates (y_1, y_2, y_3) is defined by

$$y_1 = m_{11}x_1 + m_{12}x_2 + m_{13}x_3 + m_{14}$$
$$y_2 = m_{21}x_1 + m_{22}x_2 + m_{23}x_3 + m_{24}$$
$$y_3 = m_{31}x_1 + m_{32}x_2 + m_{33}x_3 + m_{34},$$

where m_{ij} are the elements of transformation matrix $\boldsymbol{M}$. Equation 2.1 can be represented by matrix multiplication, $\boldsymbol{y} = \boldsymbol{M}\boldsymbol{x}$:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ 1 \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{bmatrix} \tag{2.1}$$

The form of transformation given in Equation 2.2 is useful, as $\boldsymbol{M}$ can be represented by a combination of matrices corresponding to each part of a transformation – translation and rotation. These transformations can be combined via multiplication of corresponding

matrices. Pure rigid-body transformation involves only translation and rotation. Other transformations may involve zoom and shear operations and such transformations are affine.

Resampling

Resampling is a calculation of voxel intensities in resulting transformed image from corresponding intensities in the original image. This involves interpolation between the voxels. Classical interpolation methods convolve an image locally with a pre-defined interpolant but can be generalised for more effective resampling. Generalised interpolation calculates the intensities in resulting image by considering a combination of pre-defined basis functions. Frequently used basis functions for generalised interpolation are B-splines. In resampling basis function coefficients are first calculated for an input image. Resampling for each voxel value is then a calculation of a linear combination of such basis functions. B-splines form a family of basis functions [67]. Using B-splines of degree 0 or 1 is equivalent to nearest neighbour or linear interpolation. B-spline of degree n is defined as

$$\beta^n(x) = \sum_{j=0}^{n} \frac{(-1)^j (n+1)}{(n+1-j)!j!} max\left(\frac{n+1}{2} + x - j, 0\right)^n. \tag{2.2}$$

2.2.2 MNI152 template

The MNI152 template [68] is a digital brain template which is an anatomical representation of a brain. Such templates are used to allow transformation of the data collected from different subjects into a common (template) space, to allow comparison between equivalent brain regions in each subject. The MNI152 template was developed by the Montreal Neurological Institute (MNI) in 2001, using 3D brain MRI scans of 152 healthy subjects. It was created using Automated Image Registration (AIR) algorithms. A target image was chosen out of 152 data samples and all others were linearly registered onto it using a 9-parameter affine transformation. Following this a non-linear registration was performed to take into account anatomical variability among subjects in the shape and size of brain features, and the differences in scan orientations. The final template thus created is provided with a head mask, i.e. the result of a segmentation performed such that the template's region corresponding to a head has maximum intensity and background has 0 intensity. The MNI152 template was adopted by the International Consortium for Brain Mapping

(ICBM) as a standard; it is used in many neuroimaging analysis tools and protocols, and is included in the FMRIB Software Library (FSL) [69].

2.3 Deep learning

Machine learning (ML) is a family of algorithms having the ability to extract information from input data that can be used for discovering relationships within the data. ML algorithms can be considered to be predictive models capable of processing large datasets of complex nature. The latter is a particular advantage of ML algorithms, as statistical analysis of such datasets can be impractical. Another advantage of ML predictive models is that they are built without needing predefined coding rules.

ML models consist of two main steps – feature extraction and prediction. Performance of the prediction step is heavily dependent on the feature extraction. Feature extraction is a process of extracting relevant information from an input dataset such that predictive ability of a model can be maximised. Deep Learning (DL) has arisen as once common approach to handling the problem of feature extraction in the ML models [70].

DL is a specialisation of ML represented by a family of algorithms consisting of multiple layers of "artificial neurons" so called because of their ability to take one or more data input and combine them to create set of outputs. Figure 2.3 shows a schematic of an artificial neuron. Neurons can be combined into layers to form a Multi-Layer Perceptron (MLP). A MLP is non-linear mapping algorithm from an input to an output [71]. These layers of neurons are able to extract features from input data by passing information deeper through an algorithm. As a DL algorithm "learns" the weights, i.e. parameters, of each layer, the algorithm "learns" to generalise information presented in the input dataset. The more complex is the task, the deeper algorithms are required and the larger should be the training dataset to "learn" the weights and generalise to the data. Once the algorithm is trained to perform a task, i.e. once learning is completed, the algorithm can make predictions on new data present to it [72]. Recently there has been a rapid growth of the DL field due to the development of optimization algorithms, improving hardware for computations (invention of a Graphics Processing Unit (GPU)) and increasing numbers of various public datasets available [73]. In 2012 the deep CNN algorithm emerged as Krizhevsky et al [74] published the AlexNet CNN for natural image classification [75].

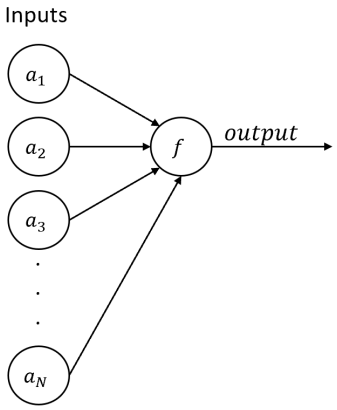

Figure 2.3: Schematic of an artificial neuron. A neuron receives a set of inputs, $\{a_1, a_2, a_3, ..., a_N\}$, and combines them by applying a function, f, to produce an output.

The algorithms of DL can be divided into two main classes – *supervised* and *unsupervised*. In supervised learning some form of ground-truth, such as data labels, is available for training and training aims to find a relationship for inferring these labels from the input data. On the other hand, in unsupervised learning labels are not available, and training aims to extract important features from data independently of data labelling [70].

Recently there has been a growth in data on patients with a range of diseases (plus healthy controls) which has been, collected, systematically stored and made available for additional research purposes. DL algorithms require large amount of data for training, and the increase in available medical data has given rise to an increased DL usage in medical research. This rapid rise in DL applications is also supported by the development of a range of software libraries which provide DL algorithms "off-the-shelf", such as Theano [76], TensorFlow [30], Keras [77] and Pytorch [29].

The applications of DL in the medical field can be divided into 2 main categories – acquiring and improving the data themselves and making decision based on the data. Examples of DL applied at the stage of acquiring data include algorithms for optimizing radiation and contrast dose and scanning time [78]. For the purpose of data preprocessing, DL algorithms have been published for removing image artefacts, normalising the data and improving its quality [79]. Examples of using DL algorithms for making decisions once the data has been acquired include tools for detection [80, 81] and segmentation [82, 83] of anatomical structures and abnormalities, along with providing decision on their correspondence to a particular diagnosis [73].

In order to train for different tasks, the DL algorithms take various forms. The three main categories of the DL architectures of interest for this thesis are feed-forward Neural Network (NN) (or MLP), the CNN and the autoencoder.

2.3.1 Feed-forward neural network

A NN is a ML model consisting of interconnected neurons or perceptrons [84]. The schematic structure of a MLP NN or a Multi-Layer Feed Forward Neural Network (MLFFNN) is presented in Figure 2.5. The MLP model has three components: an input layer, one or more hidden layers and an output layer. In the MLP NN each neuron receives an input either from input nodes or from other neurons, and passes it through hidden and output layers. A schematic of interconnected neurons in a configuration shown in Figure 2.5 is called a fully-connected NN. In such a network the outputs of each neuron are connected to all neurons in a succeeding layer. An output of a neuron is a weighted sum of its inputs:

$$X = \sum_{i=1}^{N} w_{ij} a_i + b_j, \tag{2.3}$$

where

- N is the number of input nodes,
- w_{ij} is a neuron's weight connecting an i^{th} input, a_i, to an j^{th} neuron,
- b_j is a neuron's bias.

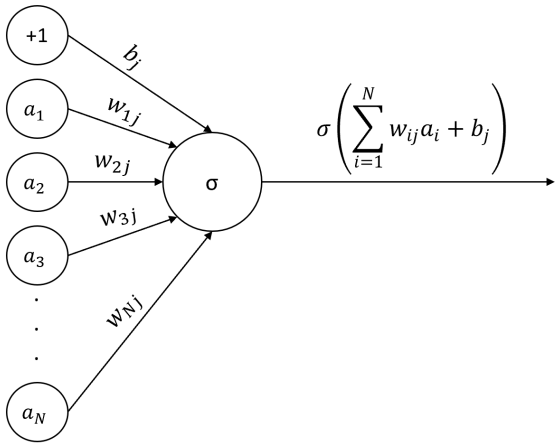

Figure 2.4: Schematic of a neuron. a_n is an n^{th} input to a neuron, w_{ij} is a weight connecting n^{th} input and j^{th} neuron, b_j is a j^{th} neuron's bias and σ is an activation function. Adapted from [1].

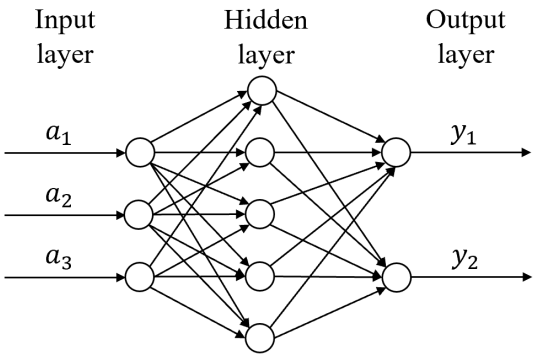

Figure 2.5: Schematic of a MLP with a single hidden layer and a data flow within it. a_n is an n^{th} input into a MLP and y_n is an n^{th} output of a MLP. Adapted from [1].

The output of a neuron propagated deeper into MLP is a result obtained by applying Equation 2.3 put though a non-linear function, $Y = \sigma(X)$. Such non-linear function is called an activation function. Following activation functions (shown in Figure 2.6) will be used in this thesis:

- logistic sigmoid function given by

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \tag{2.4}$$

  which maps real numbers into the range $[0, 1]$;

- hyperbolic tangent function given by

$$\sigma(x) = \tanh(x), \tag{2.5}$$

  which maps real numbers into the range $[-1, 1]$;

- Rectified Linear Unit (ReLU) function given by

$$\sigma(x) = \max(x, 0). \tag{2.6}$$

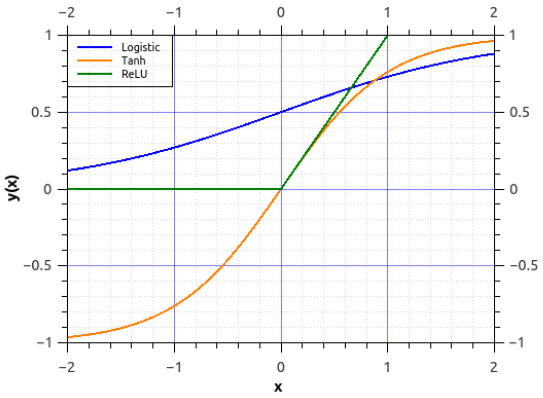

Figure 2.6: Plots of activation functions.

Training a MLP

Training a MLP involves inputting each item of a training dataset, in batches, passing these values forward through the MLP and evaluating the output relative to the desired value through a loss function. During optimisation, the error from a loss function is backpropagated through the NN, i.e. the gradient of the error function with respect to the weights is calculated for each layer, working backwards from the last to the first. The parameters of a MLP are adjusted accordingly to reduce error and improve accuracy [85], and the process continues iteratively until a stopping condition is met. A common stopping condition is absence in validation set error changes. The network's inputs are put through a network in batches for faster training as this approach allows for more weight updates in each training epoch. Each step of the training process will now be described in detail.

Initialisation During training the MLP parameters are iteratively adjusted starting from initial values. Initialisation of parameters is critical to network's performance. If all the MLP's weights are initialised to the same value, then each update iteration would result in the same values for all parameters. Therefore, several parameter initialisation methods have been proposed. Krizhevsky et al [74] proposed to initialise the parameters with small values drawn from a Gaussian distribution. Glorot et al [86] proposed *Xavier initialisation*. In Xavier initialisation variance of each MLP neuron is normalised to 1 as the parameters are drawn:

$$W_{ij} \sim U\left[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right], \tag{2.7}$$

where

- W_{ij} is an element of 2D weight matrix,
- $U\,[-a, a]$ is the uniform distribution defined on the interval $(-a, a)$,
- n is the size of preceding layer or the number of columns of a weight matrix of previous layer.

Further, He et al [87] modified the Xavier initialisation approach to improve the performance of networks, including the MLPs, by considering variance of layers' output. Bernal et al [88] uses the notation of He et al in Equation 2.7 and replace ones in numerator by 2.

Optimisation During the MLP's training, or optimisation, its parameters are iteratively adjusted by computing a loss function and backpropagating the resulting error values. Backpropagation is implemented via gradient descent methods. The details of such methods are dependent on whether the data is presented to a network sample-by-sample, in "mini-batches" or as a complete dataset at once. The gradient descent method for single sample is called *Stochastic Gradient Descent (SGD)*. The optimisation algorithm for mini-batch presentation and whole dataset are, respectively, called mini-batch and batch SGD. Updating values using the whole training dataset can be computationally expensive or even impossible, depending on the size of training data. Therefore, mini-batch gradient descent is typically used [88].

In the mini-batch SGD algorithm, at each iteration a mini-batch, B, of size M training samples is put through a MLP in a forward pass and a backpropagation algorithm is applied to the mini-batch in order to calculate the gradient of the loss function, $\mathcal{L}$, with respect to network's parameters, $\boldsymbol{\theta} - \nabla_{\boldsymbol{\theta}}\mathcal{L}$. In practice, this computation is implemented using the chain rule for partial derivatives, applying it backwards from the loss function to the parameter of interest. The updated value of parameter vector $\boldsymbol{\theta}$ at iteration $t+1$ is given by:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t} \tag{2.8}$$

where η is algorithm's *learning rate* (see below).

In NN training, mini-batch SGD is used by shuffling the training dataset before the training and then taking mini-batches consecutively from the shuffled dataset. Once whole dataset has been used in a number of iterations, a training epoch is completed, and the training data is shuffled again before the next epoch [1].

In order to overcome the problem of SGD becoming trapped in local minima, the learning rate is adjusted as the training progresses. The learning rate may be adjusted in three ways – step, exponential and $1/t$ decays:

- in step decay, the learning rate is reduced by a constant factor every T epochs;
- in exponential decay, the learning rate is adjusted every N epochs according to

$$\eta = \eta_0 e^{-kt}, \tag{2.9}$$

  where
  - η_0 is previous learning rate,
  - η is new learning rate,
  - k is a hyperparameter of adjustment,
  - t is an epoch number [89];
- in $1/t$ decay the learning rate is adjusted every N epochs according to

$$\eta = \frac{\eta_0}{1 + kt}. \tag{2.10}$$

2.3.2 CNN

Training a MLP requires a neuron for each element of the input data (in the context of imaging, each pixel in a 2D image, or voxel in a 3D image). As the result, for 3D inputs the number of parameters to be trained can be prohibitively large. Another issue with MLPs is that they cannot take into account spatial relationships within an input which is essential in order to fully capture the information content of an image. In order to overcome these problems, a CNN architecture extends a MLP as explained below [90]. In order to reduce the number of parameters a CNN uses convolutional layers which connect a neuron to a part of an input as opposed to all of it. This part of an input is known as the neuron's local receptive field. The layers of a CNN are discussed in detail further in this section.

The CNN architecture was first proposed by Fukushima [91]. However, the proposed approach was not widely used as the computational resources required for such method to be trained were not available at this time. Later, in 1998, Lecun et al [92] applied a gradient-based learning strategy to a CNN. The CNN training methodology was further

developed by Hinton et al [93, 94] and together with the development of GPUs this caused a sharp growth in both the application of the CNN to wide range of tasks and further development of its architecture.

CNNs are used for a wide variety of tasks including image recognition, image description for caption generation, speech recognition, handwriting and text generation, prediction of demographics and election results, computer robotics and self-driving vehicles [95]. Medical applications of CNNs include lesion detection and evaluation, estimation of patient survival and segmentation of medical images [85].

CNNs have a wide variety of applications in neuroimaging. At the image acquisition stage CNN can be used for MRI reconstruction [96] and improving image quality [97]. CNN-based techniques can also be used for segmentation of anatomical regions of interest [98, 99, 100]. In the diagnostic arena, CNNs have been used to detect Alzheimer's disease [101], haemorrhage [102] and other neurological diseases [103].

There are several advantages of a CNN compared to a MLP. Firstly, a CNN has more similarity to human's visual processing system and, therefore, is better suited for processing images. Secondly, CNN consists of sparse connections with shared weights and this results in smaller number of parameters to be trained compared to a MLP.

Figure 2.7 shows schematic of the CNN architecture. All CNN architectures can be divided into two main parts – feature extraction and regression or classification. An input is passed through the feature extraction part of a network such that an input to each layer is an output of a previous layer. The output of an individual layer is called a feature map, and features extracted at higher-level are propagated to lower level layers. As the features propagate through a network, from higher to lower level layers, the dimensions of the features reduce. Conversely, as data propagates from higher to lower levels, the number of feature maps increases. The feature extraction part of the network consists of convolutional, max-pooling, batch normalisation and activation layers. All these layers are discussed further below.

The output of the feature extraction part of a CNN forms the input into a regression or classification part. This part of the network is represented by a set of fully-connected layers, which is a MLP. For a classification NN the scores for each class for a given input are calculated using a softmax function in the last layer. The building blocks of a CNN will now be discussed individually, in detail.

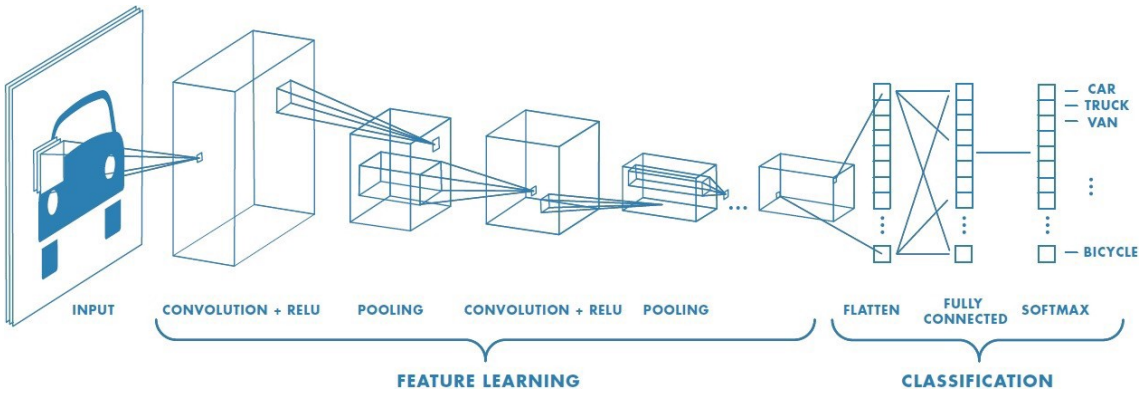

Figure 2.7: Schematic of an example CNN trained for a task of image recognition. In a forward pass, a CNN takes a 2D image as an input and passes it through a feature learning step first. Then the features are flattened and passed through a classification part of the CNN. Feature learning part of the CNN consistsof convolutional, non-linear activation (ReLU) and pooling layers. Classification part first flattens a feature map output of the feature learning step and uses fully-connected layers to produce an output. A soft-max layer produces the output meaning a particular input class [2]. Permission to use illustration obtained from the authors.

CNN layers

Convolutional layer Consider a CNN analysing a 2D data, e.g. a single-channel (grey-scale) image, and denote an input as $\boldsymbol{X} \in \mathbb{R}^{h \times w}$, where h and w are respectively height and width of the input. Now let a convolution kernel matrix to be $\boldsymbol{K} \in \mathbb{R}^{n \times m}$ with n and m being the kernel dimensions and bias to be denoted as $\boldsymbol{b} \in \mathbb{R}^{m}$. The convolution of an input $\boldsymbol{X}$ and a kernel $\boldsymbol{K}$ is given by

$$X'_{ab} = \sigma \left(\sum_{i=1}^{n} \sum_{j=1}^{m} K_{ij} X_{a+i-1,b+j-1} + \boldsymbol{b} \right), \tag{2.11}$$

where σ is one of the activation functions given in Equations 2.4 – 2.6 and X'_{ab} is ab^{th} element of matrix resulting from a convolution operation. This process is then repeated, with the kernel being shifted along the input direction by an amount known as the stride. The principle of the convolution given by Equation 2.11 is illustrated in Figure 2.8 (for a kernel of size 2×2 and a stride of 1).

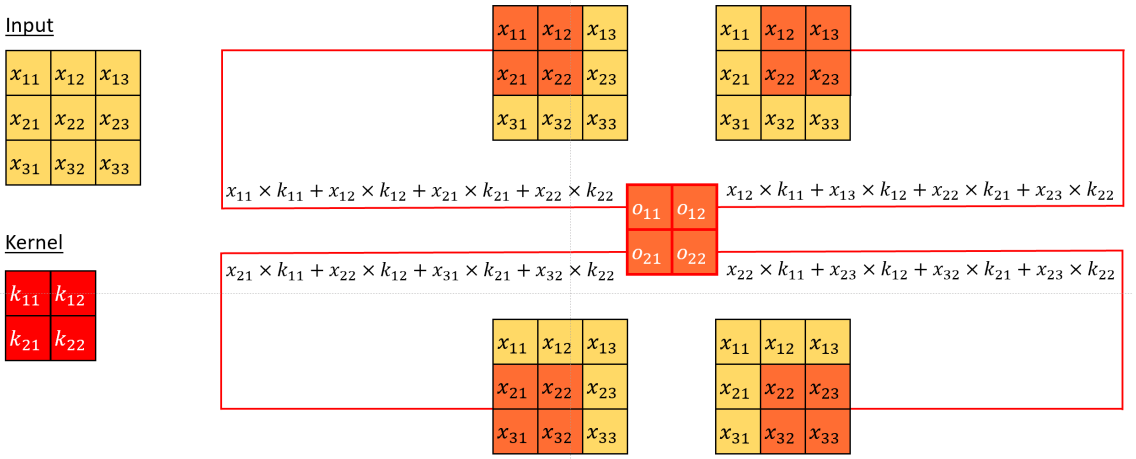

Figure 2.8: 2D convolution operation with stride of 1. Input matrix elements are denoted by x_{ij} and k_{ij} are kernel matrix elements, where i and j are row and column indices respectively. o_{ij} denotes output matrix elements.

Equation 2.11 can be extended to describe the operation of a convolutional layer for a 3D input, again with a single channel, $\boldsymbol{X} \in \mathbb{R}^{h \times w \times d}$, using a kernel, $\boldsymbol{K} \in \mathbb{R}^{n \times m \times l}$, and a bias vector, $\boldsymbol{b} \in \mathbb{R}^{l}$:

$$X'_{abc} = \sigma \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{l} K_{ijk} X_{a+i-1,b+j-1,c+k-1} + \boldsymbol{b} \right). \tag{2.12}$$

Using the operation shown in Equation 2.12 for extracting features exploits the fact that neighbouring regions of an input are more likely to be related to each other than more separated input regions. This makes feature extraction more efficient and prevents loss of information. The convolution operation from Equation 2.12 is also *translation invariant*. In a translation invariant operation, a kernel can be consistently applied to all input regions and, therefore, the operation extracts information about important features independent of their spatial position in an input [1].

Non-linear activation layers In a CNN non-linear activation layers are the same as for the MLP discussed in Section 2.3.1.

Pooling layers A *pooling layer* performs down-sampling of input maps without changing the number of input feature maps. While the convolution operation outputs higher values for the more important features found, dimensionality reduction should preserve the maximum values in a feature map in order to ensure that the most important features are

passed further down the network. Such an operation is called max-pooling and it is the most frequently used type of pooling layer in a CNN. As it is demonstrated in Figure 2.9, the max-pooling layer divides image into $n \times m$ patches and returns the maximum value for each of them. The degree of reduction in the dimensions of the input feature maps depends on pooling kernel size and pooling stride (both defined in a manner analogous to the convolutional layer). Reducing the feature map dimensions frees memory to be used to allow more channels in the feature maps and thus obtain more higher-level features. Therefore, a CNN consists of series of convolution layers separated by max-pooling layers. The layers are repeated until a sufficient number of important features is extracted and the feature map size is small enough to be processed by the fully-connected layers in the regression part of the network [104, 1].

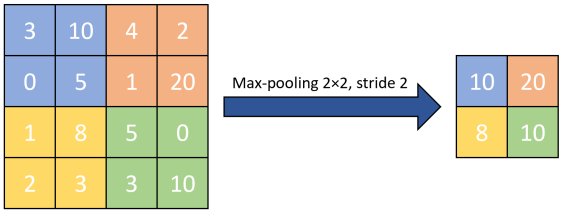

Figure 2.9: Example of applying max-pooling operation on a single channel 2D input with stride 2 in both dimensions and kernel size of 2×2.

Batch normalisation layer During CNN training the distribution of input values into each layer changes on each iteration together with the parameters of a previous layer. This complicates and slows down the training, requiring lower learning rates and making a CNN sensitive to initial parameter choices. This combined effect is called *internal covariate shift*. Using a *batch-normalisation* layers reduces the effect of covariate shift by normalising the data inside a network during training [105]. Batch normalisation is defined by

$$X_i' = \gamma \hat{X}_i + \beta \equiv BN_{\gamma,\beta}\left(X_i\right), \tag{2.13}$$

where γ and β are the parameters of batch-normalisation and $\hat{X}_i$ is the i^{th} normalised input sample from a mini-batch of size M defined as

$$\hat{X}_i = \frac{X_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}. \tag{2.14}$$

X is an input to a batch-normalisation layer, μ_B and σ_B are mean and variance defined as

$$\mu_B = \frac{1}{M}\sum_{i=1}^{M} X_i \tag{2.15}$$

and

$$\sigma_B^2 = \frac{1}{M}\sum_{i=1}^{M} (X_i - \mu_B)^2 . \tag{2.16}$$

Classification or regression part of a CNN The output feature maps of the feature extraction step are first *"flattened"* to reduce them to 1D vectors, as a MLP used in the classification or regression part of a network needs a 1D vector as an input. The flattening operation is illustrated in Figure 2.10.

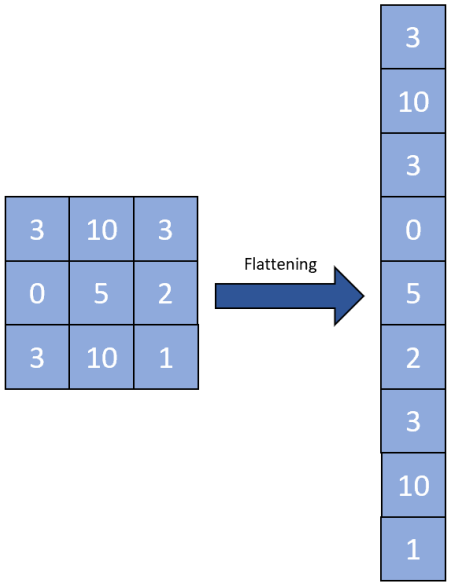

Figure 2.10: Flattening operation.

The flattened input is then fed into a MLP as defined in Section 2.3.1. In a classification network a layer known as a *softmax* layer is applied to the output vector to derive a particular class for an input sample. The softmax function transforms an output vector into normalised probability distribution. Each probability in the resulting vector is proportional to the exponential of an output vector value:

$$\sigma\left(y_i\right) = \frac{e^{y_i}}{\sum_{k=1}^{K} e^{y_i}} \; for \; i = 1, \ldots, \; K, \tag{2.17}$$

where an output vector is denoted by $\boldsymbol{y} = (y_1, y_2, \ldots, y_K) \in \mathbb{R}^K$ [106]. Output of a layer before a softmax layer can include negative numbers or numbers greater than 1. In such case the output does not sum to 1 and, therefore, cannot represent a probability distribution over output classes. After applying the softmax operation as defined in Equation 2.17 each output element is positive and all output elements sum to 1. Once the distribution of probabilities is obtained larger output values correspond to larger probabilities of an output class [107].

Residual network

He et al [108] have shown empirically that as a CNN's depth increases, degradation of network's performance becomes evident. Accuracy reaches an asymptote as number of layers is increased. The experiments by He and Sun show that this degradation is not due to overfitting as they have shown that adding more layers results in higher training error [109]. In order to illustrate network performance degradation He et al [108] propose an example. Consider two networks constructed such that the second one is a deeper counterpart of the first one. Depth of the second network is increased by adding identity mapping layers into the first network. In theory existing of such counterpart for the first network means that the second network should achieve training error no higher than the first network. In practice He et al have not found optimization solution to achieve training performance of the first network [108] in the second network. In order to solve this problem He et al [108] introduced a *Residual Network (ResNet)*. In a ResNet *"skip connections"* are introduced in a form of identity mapping. Skip connections are the connections of layers skipping at least one layer. Identity mapping is used as it does not require additional parameters.

A residual block is shown in Figure 2.11 and is defined by Equation 2.18:

$$\boldsymbol{y}_{res} = F_{res}(\boldsymbol{x}_{res}, W_i) + \boldsymbol{x}_{res}, \tag{2.18}$$

where

- $\boldsymbol{x}_{res}$ and $\boldsymbol{y}_{res}$ are the input and output of a residual block,
- $F_{res}(\boldsymbol{x}_{res}, W_i)$ is a residual mapping to be learned.

The operation $F_{res}(\boldsymbol{x}_{res}) + \boldsymbol{x}_{res}$ is point-wise addition. For the layers which result in F_{res} and $\boldsymbol{x}_{res}$ of different dimensions, e.g. convolutional layer, a linear projection, W_S, is

introduced as

$$\boldsymbol{y}_{res} = F_{res}(\boldsymbol{x}_{res}, W_i) + W_s\boldsymbol{x}_{res}. \tag{2.19}$$

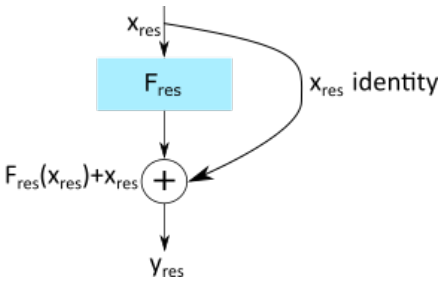

Figure 2.11: Schematic of a residual connection via identity mapping. An input, $\boldsymbol{x_{res}}$, is put through residual module's layers, F_{res}. The output, $F_{res}(\boldsymbol{x_{res}})$, is then added to the corresponding input.

ResNet architecture The *ResNet architecture* can be formed from any CNN architecture by introducing residual modules which "jump" over convolutional blocks, as shown in Figure 2.11. Figure 2.12. shows convolutional part of the ResNet-50 architecture which consists of 50 convolutional and max-pooling layers. In the convolutional part of the network, convolutional and max-pooling layers are followed by 4 convolutional blocks which are repeated 3, 4, 6 and 3 times respectively. Each block uses a convolutional layer with kernel of size 1×1 to ensure that in the residual connection each input can be combined with its output. Each convolutional layer is combined with a batch normalisation layer before activation.

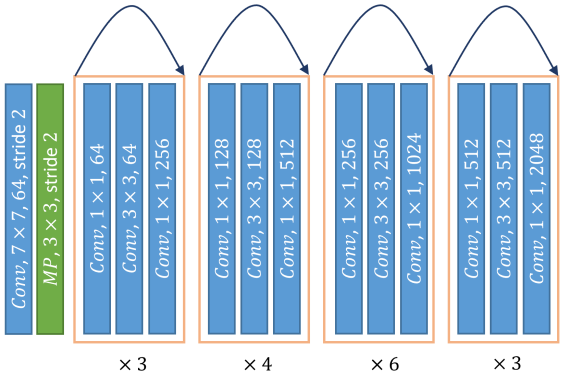

Figure 2.12: ResNet-50 architecture layout. The architecture consists of 50 layers. Each convolutional block (shown by a orange rectangle) consists of three convolutional layers and is repeated number of times shown below each block. Convolutional layers ($Conv$) are shown by blue rectangles and a max-pooling layer (MP) – by green rectangle. For convolutional layers kernel dimensions are followed by the number of channels in an output. Residual connection is shown by a dark blue arrow.

Squeeze-and-Excitation Network architecture

The *Squeeze-and-Excitation Residual Network (SE-ResNet)* is a CNN architecture in which residual blocks are replaced by SE-Res blocks. Squeeze and excitation operation introduced in a SE-block can explore dependencies between channels and "excite" useful features, while suppressing others. This improves network's performance as receptive field of filters in a network decreases deeper into a network and they progressively loose context. In order to describe the Squeeze-and-Excitation Network (SENet) architecture, consider a network taking 2D inputs.

In order to describe the SENet architecture, consider a network taking 2D inputs. First consider a SE-block for a 2D input with a one or more channels, as presented in Figure 2.13. A SE-block performs feature recalibration in the following way. Consider a transformation $\boldsymbol{F_{tr}}$ (which can be a single transformation or a set of transformations) such that $\boldsymbol{F}_{tr} : \boldsymbol{X} \to \boldsymbol{U}, \boldsymbol{X} \in \boldsymbol{R}^{H' \times W' \times C'}, \boldsymbol{U} \in \boldsymbol{R}^{H \times W \times C}$, where $\boldsymbol{X}$ is an input and H, W, C are height, width and a number of channels respectively. In order to demonstrate squeeze operation let $\boldsymbol{F_{tr}}$ be a convolutional transformation, which learns a set of filters $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_C]$ and a vector $\boldsymbol{v}_C$ represents the parameters of c^{th} filter. Then an output of the transformation $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_C]$ is

$$\boldsymbol{u}_c = \boldsymbol{v}_c * \boldsymbol{X} = \sum_{s=1}^{C^i} \boldsymbol{v}_c^s * \boldsymbol{x}^s, \tag{2.20}$$

where $*$ denotes a convolution, $\boldsymbol{v}_c = \left[\boldsymbol{v}_c^1, \boldsymbol{v}_c^2, \ldots, \boldsymbol{v}_c^{C'}\right]$ and $\boldsymbol{X} = \left[\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^{C'}\right]$. As the result of the operation in Equation 2.20. $\boldsymbol{u}_c$ is of size $\boldsymbol{u}_c \in \boldsymbol{R}^{H \times W}$, i.e. $\boldsymbol{u}_c$ has 1 channel. $\boldsymbol{v}_c^s$ is a 2D kernel describing the relationship of channel c in $\boldsymbol{v}_c$ with a channel s in $\boldsymbol{X}$. Output feature maps, $\boldsymbol{u}_c$, are produced by summation of results of convolution channel-wise. Therefore, channel dependencies are encoded in $\boldsymbol{v}_c$. (For simplicity in this example we do not consider the bias term of the convolution operation, but this can be added in the usual way).

A convolutional layer has a local receptive field and its output, $\boldsymbol{U}$, does not taking into account the context of whole input $\boldsymbol{X}$. The squeeze operation in a SE-block aims to resolve this disadvantage by "squeezing" such contextual information into the description of each channel in $\boldsymbol{U}$. For this purpose, a *"squeeze"* operation is applied to $\boldsymbol{U}$. The squeeze operation produces a channel descriptor by extracting channel-wise statistics $\boldsymbol{z} \in \boldsymbol{R}^C, \boldsymbol{z} =$

$[z_1, ..., z_c, ..., z_C]$, as

$$z_c = \boldsymbol{F}_{sq}(\boldsymbol{u}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j). \tag{2.21}$$

In Equation 2.21 global average pooling over each channel of transformation output, $\boldsymbol{U}$, is used to obtain the statistics z_c. Such operation can also be interpreted as "squeezing" the context of input's spatial dimensions $H \times W$ through each channel c.

The "squeeze" operation is followed by the *"excitation"* step. The excitation operation, $\boldsymbol{F}_{ex}$, uses information from squeeze step by applying following "gating" relationship:

$$\boldsymbol{s} = \boldsymbol{F}_{ex}(\boldsymbol{z}, \boldsymbol{W}) = \sigma(g(\boldsymbol{z}, \boldsymbol{W})) = \sigma(\boldsymbol{W}_2 \delta(\boldsymbol{W}_1 \boldsymbol{z})), \tag{2.22}$$

where

- σ denotes sigmoid function,
- δ - ReLU function,
- $\boldsymbol{W}_1 \in \boldsymbol{R}^{\frac{C}{r} \times C}, \boldsymbol{W}_2 \in \boldsymbol{R}^{C \times \frac{C}{r}}$ are weights.

The "gating" mechanism given in Equation 2.22 serves two purposes within the "excitation" operation. Firstly, it is capable of learning non-linear relationships among channels in $\boldsymbol{U}$. Secondly, it allows further enhancement of the importance of multiple channels instead of an emphasis on a single channel. Weights $\boldsymbol{W}_1$ an $\boldsymbol{W}_2$, along with the ReLU operation, form fully-connected layers to perform dimensionality reduction. The final SE-block output is defined as

$$\widetilde{\boldsymbol{x}}_c = \boldsymbol{F}_{scale}(\boldsymbol{u}_c, s_c) = s_c \cdot \boldsymbol{u}_c, \tag{2.23}$$

where $\widetilde{\boldsymbol{X}} = [\widetilde{\boldsymbol{x}}_1, \widetilde{\boldsymbol{x}}_2, \ldots, \widetilde{\boldsymbol{x}}_C]$ and scalar s_c is combined with a feature map $\boldsymbol{u}_c \in \boldsymbol{R}^{H \times W}$ via channel-wise multiplication. In that way $\boldsymbol{U}$ is rescaled with the activations s_c.

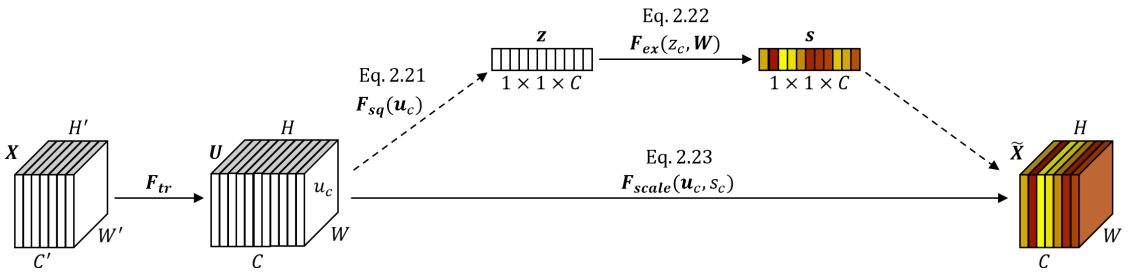

Figure 2.13: A schematic of a SE-block. Arrows show operations applied to whole input, dashed arrows – channel-wise operations. Adopted from Hu et al [3].

For the network used in the experiments described in this thesis, we used SE-Res-blocks to form a SE-ResNet. In a SE-Res-block, $\boldsymbol{F_{tr}}$ is the non-identity branch of a ResNet-block, while squeeze and excitation operations are applied before the summation with identity branch as shown in Figure 2.13. (Hu et al. 2017).

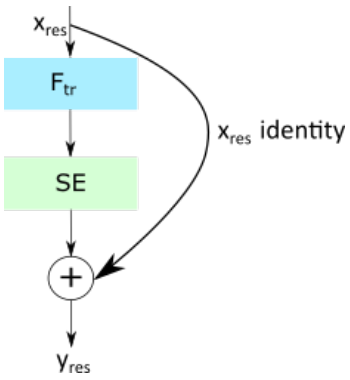

Figure 2.14: Schematic of a SE-Res-block used to form a SE-ResNet. An input, $\boldsymbol{x_{res}}$, is put through residual module's layers, $\boldsymbol{F}_{tr}$. The output, $\boldsymbol{F}_{tr}(\boldsymbol{x_{res}})$, is then forwarded though the squeeze and excitation operations and added to the corresponding input to produce block output $\boldsymbol{y_{res}}$.

Pseudocode for the SE-block is shown in Algorithm 1. Input to the block is as described above and output in step 8 of the pseudocode is defined by Equation 2.23. Global average pooling in step 3 is defined by Equation 2.21. The $view\_as(\cdot)$ and $expand\_as(\cdot)$ are used to ensure that the inputs into the element-wise multiplication in step 7 are of the same size. The $view\_as(\cdot)$ operation returns a tensor with the same number of elements as the tensor it is applied to, but different dimensions supplied as arguments. The $expand\_as(\cdot)$ operation repeats dimensions of the tensor it is applied to return a tensor with the dimensions of the tensor passed as an argument. Fully-connected layer in step 5 is described by Equation 2.22.

Algorithm 1 SE-block

1: Input: $\boldsymbol{X}$ of size $B \times H \times W \times C$
2: Output: $\boldsymbol{Y}$ of size $B \times H \times W \times C$
3: $y \leftarrow global\_average\_pooling(\boldsymbol{X})$
4: $y \leftarrow y.view\_as(B, C)$
5: $y \leftarrow FC(y)$
6: $y \leftarrow y.view\_as(B, C, 1, 1)$
7: $y \leftarrow \boldsymbol{X} * y.expand\_as(\boldsymbol{X})$
8: **return** y

2.3.3 Regression

While classification models output a set of probabilities corresponding to a number of classes, regression models output a quantity of interest. In a *regression CNN*, a set of fully-connected layers in the regression part of a network models an output quantity, as described by Equation 2.3. There are two types of regression – *metric* and *ordinal*. Equation 2.3 represents a model of metric regression, while an ordinal regression model takes into account the relative order of output values, as described further below. Therefore, Equation 2.3 cannot directly represent ordinal regression.

Ordinal regression

Assume an i^{th} input belongs to the input dataset as $x_i \in X$ with corresponding outcome space $y_i \in Y = \{r_1, r_2, \ldots, r_K\}$, where $r_K > r_{K-1} > \ldots > r_1$ are ordered ranks. Solving an ordinal regression problem is equivalent to finding a mapping from the training dataset $D = \{x_i, y_i\}_{i=1}^{N}$ to the ranks $h() : X \to Y$.

There are several ways to implement ordinal regression in the regression part of a CNN. Herbrich et al [110, 111] proposed a modification of the Support Vector Machine (SVM) algorithm to take into account data ordinality. Crammer and Singer [112] analysed the problem of ranking instances and proposed a generalisation of the online perceptron algorithm for ordinal regression. The most relevant approaches for applying to a CNN were proposed by Frank and Hall, Li and Lin, and Niu et al [113, 114, 4]. In all three approaches ordinal regression is represented by a set of binary classifiers.

Niu et al [4] transformed the problem of ordinal regression into a set of binary regression problems: K ranks are represented by $K - 1$ binary classification sub-problems. For each rank $r_k \in \{r_1, r_2, \ldots, r_{K-1}\}$ a binary classifier predicts whether for an i^{th} sample $y_i > r_k$. The architecture used by Niu et al [4] is shown in Figure 2.15.

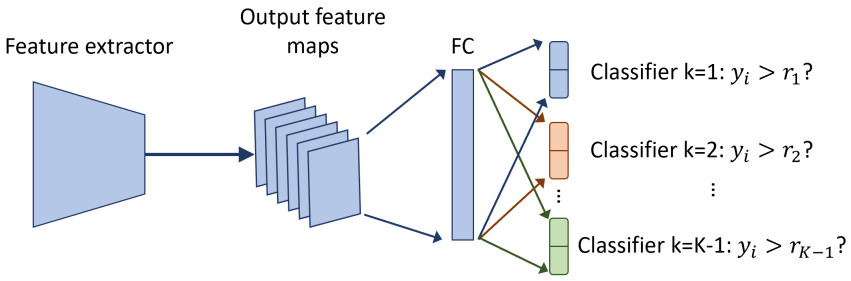

Figure 2.15: Schematic representation of the architecture used by Niu et al [4]. The features of input images are first extracted using convolutional part of the CNN. Resulting feature maps are then fed into a fully-connected layer, results from which is fed into $K-1$ binary classifiers. Each of the $K-1$ classifiers determines whether a label corresponding to an i^{th} sample, y_i, is greater than a value, r_k. Adapted from [4].

Then rank of an unseen sample is predicted based on the classification results of $K-1$ classifiers. Three main steps of the approach are:

- From original dataset $D = \{x_i, y_i\}_{i=1}^{N}$ construct a sub-dataset for each k^{th} binary classification problem as $D^k = \{x_i, y_i^k, w_i^k\}_{i=1}^{N}$, where
  - $y_i^k \in \{0, 1\}$ is a binary classification label indicating whether the rank of a sample y_i is larger than r_k,
  - $w_i^k = |C_{y,k} - C_{y,k+1}|$ is the weight of an i^{th} sample taken from a cost matrix.

  Cost matrix, C, is employed for measuring cost between predicted ranks and ground-truth ranks. C is a $K \times K$ matrix where $C_{y,r}$ is the cost of predicting a sample (x, y) as rank r. Also $C_{y,y} = 0$ and $C_{y,r} > 0$ for $y \neq r$. Niu et al [4] used an absolute cost matrix, with $C_{y,r} = |y - r|$.
- $K-1$ binary classifiers are trained.
- The rank for a sample x' is predicted as

$$h\left(x'\right) = r_q$$

$$q = 1 + \sum_{k=1}^{K-1} f_k(x'), \tag{2.24}$$

  where
  - $h\left(x'\right)$ is a predicted rank for a sample x',

– $f_k(x') \in \{0, 1\}$ is the classification result of the k^{th} binary classifier for a sample x'.

Niu et al [4] used following loss function to combine the results of $K-1$ classifiers and backpropagate:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K-1}\lambda^k 1\left\{o_i^k = y_i^k\right\} w_i^k \log\left(p\left(o_i^k \middle| x_i, W^k\right)\right), \tag{2.25}$$

where

- N – number of samples in the training set,
- o_i^k – a network's output for the i^{th} input and k^{th} task,
- w_i^k – a weight for the i^{th} input in the k^{th} task,
- W^k – the k^{th} task parameters,
- λ^k – an importance coefficient for the k^{th} task:

$$\lambda^k = \frac{\sqrt{N_k}}{\sum_{k=1}^{K}\sqrt{N_k}} \tag{2.26}$$

with N_k being the number of samples of rank k in the training set.

2.3.4 Autoencoder

An *autoencoder* is an unsupervised learning algorithm aiming to encode the input to a network and then fully reconstruct the input from this encoded representation. For simplicity, consider an autoencoder consisting of fully-connected layers taking a 1D input. As shown in Figure 2.16, an autoencoder consists of three main steps:

1. An input $\boldsymbol{x}$ is passed through a series of fully-connected layers described by Equation 2.3. This is the encoder stage.
2. A non-linear activation layer, which scales the encoded representation into range of $[0, 1]$ by applying one of the functions given by Equations 2.4 and 2.5. The activation stage outputs a hidden representation $\boldsymbol{h}$.
3. A decoder, which applies a linear layer given by Equation 2.1 wrapped into an activation function, outputting a decoded representation $\boldsymbol{z}$.

These three steps are applied iteratively during autoencoder's training. At each step of the training an error is measured via the *Mean Squared Error (MSE)* loss function:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{z}-\boldsymbol{x})^2\,, \tag{2.27}$$

where N is the number of input samples in a batch.

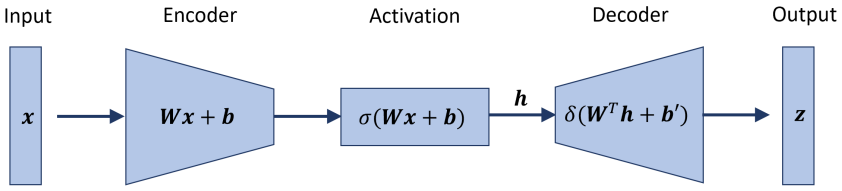

Figure 2.16: A schematic of a simple autoencoder. The encoder and decoder both consist of one linear layer. Activation step is applied between the encoder and decoder. σ and δ denote non-linear activation functions. The output, $\boldsymbol{z}$, represents reconstructed data. $\boldsymbol{W}$, b and b' are the autoencoder's weight and biases respectively.

There are various modifications to the simple autoencoder shown in Figure 2.16. Additional constrains can be applied to the autoencoder training in order to exploit the internal structure of the input data. One of such constraints is the *sparsity constraint*. In fully-connected layers of an autoencoder, a neuron is considered "activated" if its output value is close to 1. Conversely, a neuron with an output close to 0 is considered as inactive. In a sparsely constrained autoencoder, a large proportion of neurons are forced to be inactive, i.e. set to 0, most of the time.

In a sparse autoencoder given a batch input, $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]$, the average activation value of the j^{th} hidden unit, $\hat{\rho}_j$, can be defined as

$$\hat{\rho}_j = \frac{1}{N}\sum_{n=1}^{N}\left[\boldsymbol{h}_j\left(\boldsymbol{x}_n\right)\right]. \tag{2.28}$$

In a sparse autoencoder $\hat{\rho}_j$ is set close to 0 and this forces most of the activations in an autoencoder to be near 0, i.e. inactive. The constraint is implemented by adding a regularization value to Equation 2.27:

$$\mathcal{L}_{sparse} = \mathcal{L}_{MSE} + \beta\sum_{j=1}^{d_h} KL(\rho||\hat{\rho}_j) = \mathcal{L}_{MSE} + \beta\sum_{j=1}^{d_h}\left\{\rho\log\frac{\rho}{\hat{\rho}_j} + (1-\rho)\,log\frac{1-\rho}{1-\hat{\rho}_j}\right\}, \tag{2.29}$$

where

- β is a hyperparameter of the sparse loss function,
- d_h is the dimensionality of the j^{th} hidden layer,
- $KL(\cdot)$ is a Kullback-Leibler (KL) divergence function.

Good data encoding needs to be able to robustly extract stable structures from unknown distributions and to ensure that these structures are stable towards slight disturbances. This can be achieved by training a *denoising autoencoder* such as that illustrated in Figure 2.17. An input is first partially disturbed by means of a stochastic mapping. Such disturbance typically come from two distributions:

1. Binary noise: in each input a fixed number of values are set to 0.
2. Gaussian noise: for each input a fixed number of values are drawn from a Gaussian distribution and added to the input.

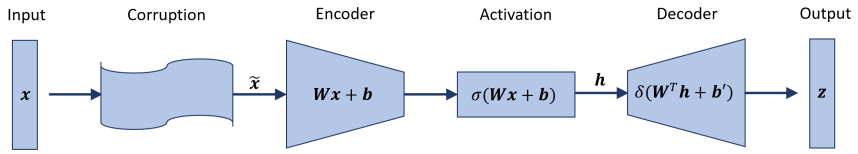

Figure 2.17: A schematic of a denoising autoencoder. An input $\boldsymbol{x}$ is first corrupted before being input to an encoder.

A *Convolutional Autoencoder (CAE)* extends the functionality of an autoencoder in the same way that a CNN allows the spatial distribution of features in an input to be taken into account better than a simple MLP. In a CAE, the features are extracted by convolutional layers (defined by Equation 2.12) instead of fully-connected layers [115].

2.3.5 Unsupervised clustering

An important field within *unsupervised learning* is *unsupervised clustering*, which is the process of grouping samples based on the similarity of features extracted. In unsupervised learning, ground-truth labels are (by definition) not available and clustering aims to label the data based on the clusters obtained instead. The clusters themselves, the metric of similarity used and cluster validation methods may vary significantly between implementation, but relatively little work has been undertaken into the impact of these variations.

The *K-means clustering* algorithm [116] uses Euclidean distance as a similarity metric. It is very inefficient to compare large 2D or 3D inputs by Euclidean distance, however, so the feature extraction step and metric choice are both crucial.

Van der Maaten and Hinton [117] proposed minimizing the KL divergence between the input data distribution and the latent representation to be used for dimensionality reduction. Van der Maaten and Hinton [117] also used a Deep Neural Network (DNN) for obtaining a latent representation of the input data. Further some authors [118, 119, 120, 121, 122] used DNNs along with KL error to develop architectures capable of simultaneous feature extraction and clustering. Atul Shah and Koltun [118] used pairwise error measurements in the calculations of a KL-based loss function. Caron et al [119] first extracted important features using a CNN, and then performed clustering and input the result of this clustering into a sample classification function with known per class labels; the CNN was then trained by backpropagation from the sample classification function. Xie et al [122] used a KL divergence loss based on cluster centres and referred to this approach as Deep Embedded Clustering (DEC).

DEC

In order to describe the DEC approach consider the problem of clustering a set of N points, $\{\boldsymbol{x}_i \in \boldsymbol{X}, y_i \in Y\}_{i=1}^{N}$, into K clusters, each represented by a centroid $\boldsymbol{\mu}_j, j = \{1, ..., K\}$. y_i is a feature of interest (for example, age). $\boldsymbol{x}_i$ is first mapped into an embedding space, $f_\theta \colon \boldsymbol{x}_i \rightarrow \boldsymbol{z}_i$, where

- θ are learnable parameters,
- $\boldsymbol{z}$ denotes embedding of $\boldsymbol{x}_i$.

The DEC approach consists of two main steps. First an autoencoder is trained to simultaneously encode and decode $\boldsymbol{x}_i$ and predict y_i. In this stage, the parameters θ of function f are learned. In the second stage, the clustering capability of a network is trained. Clustering is applied to the encoded representation of $\boldsymbol{x}_i$ - $\boldsymbol{z}_i$.

Once an autoencoder has been pretrained (i.e. given a pretrained mapping, f_θ, and initial cluster centroids, $\{\boldsymbol{\mu}_j \in Z\}^K, j = \{1, ..., K\}$), clusters are initialized by k-means clustering and the cluster centres, $\{\boldsymbol{\mu}_j \in Z\}^K, j = \{1, ..., K\}$, are then learned by iteratively computing an auxiliary target distribution and minimizing the KL divergence. Training is performed by iterating between two main steps. In the first step, a soft assignment between

the embedded points and the cluster centroids is computed. In the second step, the deep mapping, f_θ, is updated and the cluster centroids are refined by learning from current high confidence assignments using an auxiliary target distribution. This process is repeated until a convergence criterion is met (for example, a change in the error calculated on validation dataset is below pre-set threshold for a pre-set of consecutive training epochs). Xie et al [122] used the students t-distribution as a kernel to measure the similarity between the embedded point, $\boldsymbol{z}_i$, and centroid, $\boldsymbol{\mu}_j$:

$$q_{ij} = \frac{(1 + \|\boldsymbol{z}_i - \boldsymbol{\mu}_j\|^2)^{-1}}{\sum_j (1 + \|\boldsymbol{z}_i - \boldsymbol{\mu}_j\|^2)^{-1}} \tag{2.30}$$

where

- $\boldsymbol{z}_i = f_\theta(\boldsymbol{x}_i) \in Z$ corresponds to $\boldsymbol{x}_i \in \boldsymbol{X}$ after embedding,
- q_{ij} can be interpreted as the probability of assigning a sample i to a cluster j (i.e. a soft assignment).

The model is trained by matching the soft assignment to the target distribution using a KL divergence loss between q_{ij} and the auxiliary distribution p_{ij} as follows:

$$\mathcal{L}_C = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}, \tag{2.31}$$

where P is the target distribution, defined as

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_j q_{ij}^2 / \sum_i q_{ij}}. \tag{2.32}$$

2.3.6 Saliency mapping

DL algorithms are well-suited for large amounts of healthcare data and perform better than traditional statistical models at the expense of sacrificing interpretability for predictive power. For example, Choi et al developed a predictive DL-model for patient's status from electronic health records. The model outperformed traditional statistical methods such as logistic regression and MLP [123]. Popularity of the DL-based methods in medicine is constantly increasing, but their "black-box" nature represents an adoption barrier as *interpretability* and *explainability* should be first ensured [44]. Interpretability of DL-methods was previously defined by Doshi-Velez and Kim as "the ability to explain or to

present in understandable terms to a human" [124]. For this purpose, various visualisation methods are used.

One of the approaches for ensuring interpretability and explainability of the algorithms is understanding which regions of an input are of importance for making a decision [125, 126, 7, 43, 44, 42, 45, 46, 127, 47, 5, 128]. Such methods either use input occlusion techniques or apply calculations to the parameter gradients inside a network to calculate an *"importance"* score to the input regions. The "importance" score of a region corresponds to its influence on a network's output [45]. Methods applying the calculations to network's gradients are also called *saliency mapping methods*. Further, in this thesis the terms "saliency map" and "importance map" are used interchangeably. A subset of saliency maps are *"occlusion maps"* resulting from applying occlusion techniques. Here, a number of methods relevant to this thesis are presented – gradient visualization, SmoothGrad, guided backpropagation, Class Activation Mapping (CAM), Gradient Class Activation Mapping (Grad-CAM), guided Grad-CAM, Layer-wise Relevance Propagation (LRP) and Contrastive Layer-wise Relevance Propagation (CLRP).

Gradient Visualisation

The method of *gradient visualisation* for producing saliency maps was introduced by Simonyan et al [42]. Simonyan et al's [42] method creates class-specific saliency maps using the gradient of input in a CNN. The hypothesis behind using a gradient in this way is that a map of its magnitude indicates those pixels in the input which need to change least to affect an output the most. Therefore, for a classification problem, these regions of the input correspond to the location of an object of interest.

In describing the method, a score function for class c is considered:

$$S_c(I) = w_c^T I + b_c, \tag{2.33}$$

where

- I is a vectorised input,
- w_c and b_c are weights and bias of the model respectively.

For such simple linear model it is clear that the weight, w_c, defines importance of various regions of an input for scoring. Further, a non-linear model, S_c, calculated for an input sample, X, can be approximated by its first-order Taylor expansion: $L(X) \approx wX + b$,

where w is a derivative of loss function with respect to an input image and b is a bias matrix such that

$$w = \frac{\partial L}{\partial L}\bigg|_X. \tag{2.34}$$

Computing a saliency map For simplicity, consider a saliency map for a 2D grayscale input I of size $m \times n$ and of class c. A saliency map corresponding to this class, $M \in \boldsymbol{R}^{m \times n}$, is computed by first finding w from back-propagation and then calculating

$$M_{ij} = |w_{h(i,j)}|, \tag{2.35}$$

where $h\,(i,j)$ is an index of the element of w representing saliency for the i^{th} row and j^{th} column element of I.

SmoothGrad

Saliency maps produced by the gradient visualisation method suffer from noise, because the score function S_c may change sharply for small magnitude changes in an input. This means that a gradient forming a saliency map has meaningless variations, which result in the noise observed. Such meaningless regions result from backpropagation through non-continuously differential functions, e.g. ReLU, and making S_c non-continuously differentiable.

Smilkov et al [45] propose the *SmoothGrad method* to overcome this, based on the fact that the meaningless regions in a saliency map can be reduced by average pooling of such regions. This is achieved by combining a gradient kernel from Equation 2.35 with a Gaussian smoothing kernel. The value of an element of a saliency map produced by applying a smoothing kernel, $\hat{M}_{ij}(x)$, can be defined as

$$\hat{M}_{ij}(x) = \frac{1}{N}\sum_{n=1}^{N} M_{ij} + \mathcal{N}(0, \sigma^2), \tag{2.36}$$

where

- N sample values are taken from a neighbourhood of M_{ij},
- $\mathcal{N}(0, \sigma^2)$ defines a Gaussian distribution with 0 mean and σ^2 standard deviation [45]).

CAM

Producing a saliency map by finding a *CAM* was proposed by Zhou et al [5]. A class activation map for a particular category indicates the discriminative image regions used by the CNN to identify that category. The architecture is presented in Figure 2.18. Global Average Pooling (GAP) outputs the spatial average of the feature map of each unit at the last convolutional layer. A weighted sum of these feature maps is used to generate the final output with weights being corresponding parameters in the last fully-connected layer. This has also been described for a CNN used for classification and using a softmax layer to output class probabilities. The same technique can be applied to regression and other approaches.

Computing a saliency map Consider a saliency map for a given 2D input. $f_k(x, y)$ is an activation of the last convolutional layer unit k at a pixel (x, y) of an input. The global average pooling layer after the last convolution operation thus generates $F^k = \sum_{x,y} f_k(x, y)$. Thus for a given class c, the input to the softmax function is $S_c = \sum_k w_k^c F_k$. Therefore, $S_c = \sum_{x,y} \sum_k w_k F_k$ and we can define an activation map for prediction c as $M_c = \sum_k w_k^c f_k(x, y)$.

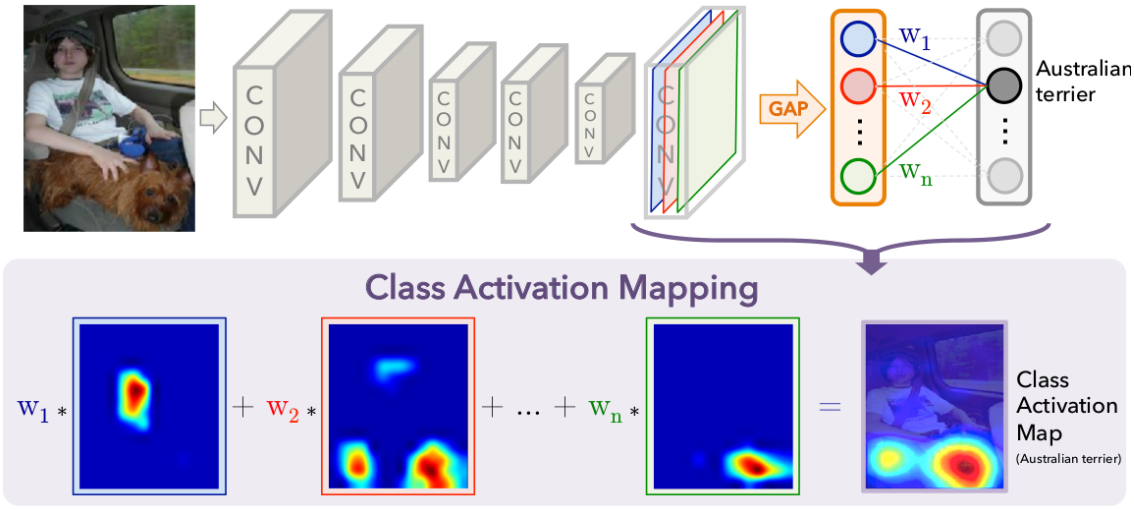

Figure 2.18: Schematic of an example CNN classifying images into classes corresponding to a dog pictured and an illustration of constructing a CAM. Figure from Zhou et al [5].

Grad-CAM

The CAM approach has a significant disadvantage of being applicable only to the CNN models without fully-connected layers in the classification or regression step. An universal approach including other architectures was proposed by Selvaraju et al [43], and is known as Grad-CAM. Here, for simplicity, consider producing a saliency map for a network analysing a 2D input. The Grad-CAM method defines a class-specific saliency map, $L^c_{Grad-CAM} \in \mathbb{R}^{u \times v}$, of dimensions u and v and for class c, as

$$L^c_{Grad-CAM} = ReLU\left(\sum_k \alpha^c_k A^k\right), \tag{2.37}$$

where

- A^k is a ^{kth} feature map,
- α^c_k is a weight representing an "importance" of k^{th} feature map for a class c:

$$\alpha^c_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}}. \tag{2.38}$$

Here $\frac{\partial y^c}{\partial A^k_{ij}}$ are gradients obtained from backpropagation, and the summation over i and j represents a GAP operation. The Grad-CAM method is presented in Figure 2.19.

The ReLU operation in Equation 2.37 removes negative values which correspond to classes other than c. Therefore, applying ReLU extracts the values which have positive influence of choosing class c and whose values are expected to increase as the probability of choosing class c increases.

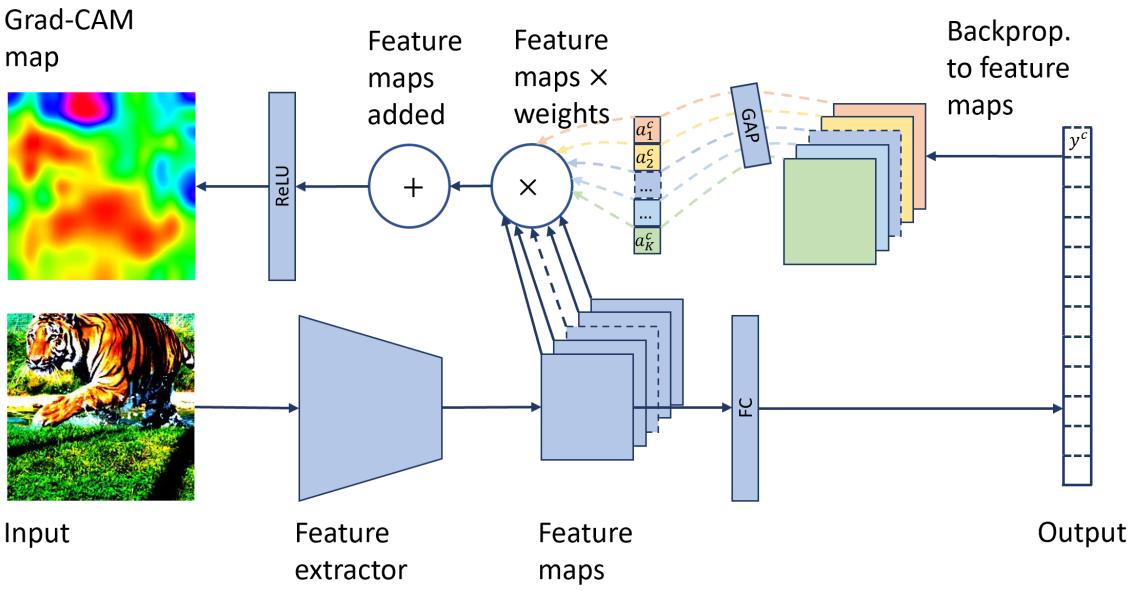

Figure 2.19: Grad-CAM method. In a forward-pass, an input is first put through a convolutional part of a CNN, i.e. a feature extractor, to obtain feature maps, which are input into a fully-connected layer to get an output vector of the probabilities for each class. In a backward-pass, backpropagation is performed from a class of interest, y^c, to the feature maps. Then the gradients are used to calculate the weights by GAP as defined in Equation 2.38. The feature maps are multiplied by the weight, summed and put through a ReLU as shown in Equation 2.37 to produce the final saliency map.

Guided Backpropagation

Guided propagation [46] is another method to obtain a saliency map for a given input. In this approach, the authors modify the course of backpropagation through non-linear layers such as ReLU defined in Equation 2.6. In backpropagation, the gradient at ReLU layer l is defined as

$$R_i^l = \left(f_i^l > 0\right) \times R_i^{l+1}, \tag{2.39}$$

where

- $R_i^l = \frac{\partial f^{out}}{\partial R_i^l}$ denotes an i^{th} element of a backpropagation result for an output activation out,
- f_i^l is an i^{th} element of an input f to the l^{th} layer.

In Equation 2.39 the first term indicates that only bottom signal, i.e. an input to a layer, is taken into account during the backpropagation. In the deconvolution approach [47] the backpropagation, on the other hand, considers only top gradient signal, i.e. gradient of a

deeper layer (deeper meaning the layers closer to an output layer):

$$R_i^l = \left(R_i^{l+1} > 0\right) \times R_i^{l+1}. \tag{2.40}$$

The guided backpropagation approach combines Equations 2.39 and 2.40 and masks out values for which bottom signal and/or top data values are 0:

$$R_i^l = \left(f_i^l > 0\right) \left(R_i^{l+1} > 0\right) \times R_i^{l+1}. \tag{2.41}$$

The guided backpropagation approach includes additional guidance from top layers compared to simple backpropagation alone, and only require storing the same information from a forward pass as for the deconvolution approach [46].

Guided Grad-CAM

The Grad-CAM [43]) method produces saliency maps of size of the feature map resulting from the last convolution, with maps then being scaled to the size of the input. In order to produce saliency maps of better resolution, the Grad-CAM method can be combined with the method of guided backpropagation, with the overall method being known as *guided Grad-CAM*. This is represented in Figure 2.20.

An input is first put through a CNN feature extractor to obtain a set of feature maps, which are input into a task specific part of the network to produce an output. In the classification output, the probability of the class of interest is then set to 1 while all other probabilities are set to 0 and backpropagation is performed to the feature maps of interest to obtain a Grad-CAM saliency map. Finally, this saliency map is combined via pointwise multiplication with the result of guided backpropagation to obtain guided Grad-CAM result.

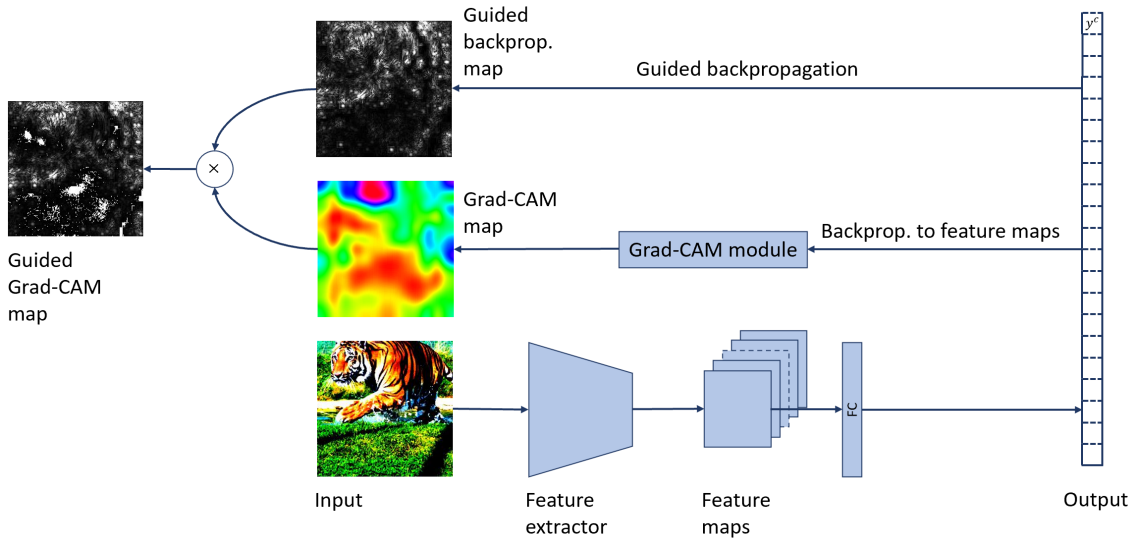

Figure 2.20: Schematic representation of the guided Grad-CAM method.

LRP

LRP exploits neural network's graph structure to produce salient explanations. The LRP rules for producing saliency maps are derived analogously to the Kirchoff's law for electrical circuits, as in a backward pass through the network a signal received by a neuron from an upper layer, i.e. deeper layer meaning the one closer to an output layer, is redistributed to the neurons down the network. Consider the neurons denoted by j and k in adjacent layers. R_j propagates relevance from the neuron k, R_k, to the relevance on neuron j according to the rule

$$R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k, \tag{2.42}$$

where z_{jk} is a function modelling how much the contribution from k neuron influences j neuron; it can be defined by a number of LRP rules, as described below. The denominator in Equation 2.42 ensures the principle of signal conservation is preserved. In order to produce saliency map, such a rule is applied consecutively layer-by-layer from the final layer back to the input.

Basic LRP rule (LRP-0). Since neural network layers can be represented by the form given in Equation 2.3, Equation 2.42, can be defined as

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k, \tag{2.43}$$

where a_j is the activation of j^{th} neuron and w_{jk} is the weight connecting the a_j. The sum in the denominator includes 0 in order to take into account bias term in the layer of neuron j. This rule is equivalent to computing a saliency map by masking an input by a gradient with respect to it. In order to show this, rewrite the denominator as $\sum_{0,j} a_j w_{jk} = a_k$ and consider a relevance score being propagated through two layers consecutively:

$$R_i = \sum_k \frac{a_i w_{ij}}{a_j} \frac{a_j w_{jk}}{a_k} R_k. \tag{2.44}$$

Now assume k is the last layer and set $a_k = R_k$ such that the relevance of neurons in the final layer is equal to its own activations. This leads to the following relationship between R_i and R_k:

$$R_i = \sum_k a_i w_{ij} w_{jk}. \tag{2.45}$$

The relevance score in Equation 2.45 is equivalent to multiplying an activation a_i by the gradient of a_k with respect to a_i [44]. Unfortunately, the experiments by [7] show that such saliency maps are affected by noise.

Epsilon LRP Rule (LRP-ϵ). As it can be seen in Equation 2.42 if $(a_j = 0)$ or $(w_j = 0)$ then the relevance R_j will be 0. As the result of this the concepts of zero weight, deactivation, and absence of connection lead to the same relevance score. In order to resolve cases such as these for which the activations contributing to the k activation are weak or confusing, Equation 2.42 is modified as

$$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \sum_{0,j} a_j w_{jk}} R_k. \tag{2.46}$$

For small values of ϵ, the formulation in Equation 2.46 will absorb weak contributions to the k activation and, therefore, weak contributions do not continue to affect the final result. As the ϵ value is increased, only more salient features will not be absorbed and saliency map will become sparser and less noisy.

Gamma LRP Rule (LRP-γ). Salient features may be further enhanced by preferring positive contributions over negative ones by reformulating R_j as

$$R_j = \sum_k \frac{a_j (w_{jk} + \gamma w_{jk}^+)}{\sum_{0,j} a_j (w_{jk} + \gamma w_{jk}^+)} R_k. \tag{2.47}$$

Rule	**Formula**	**Layers to be used at**
LRP-0	$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$	Upper
LRP-ϵ	$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \sum_{0,j} a_j w_{jk}} R_k$	Middle
LRP-γ	$R_j = \sum_k \frac{a_j (w_{jk} + \gamma w_{jk}^+)}{\sum_{0,j} a_j (w_{jk} + \gamma w_{jk}^+)} R_k$	Lower
LRP-$\alpha\beta$	$R_j = \sum_k \left(\alpha \frac{(a_j w_{jk})^+}{\sum_{0,j} (a_j w_{jk})^+} + \beta \frac{(a_j w_{jk})^-}{\sum_{0,j} (a_j w_{jk})^-} \right) R_k$	Lower
Flat	$R_j = \sum_k \frac{1}{\sum_j 1} R_k$	Lower
w^2-rule	$R_j = \sum_k \frac{w_{jk}^2}{\sum_{0,j} w_{jk}^2} R_k$	First
z^B-rule	$R_j = \sum_k \frac{x_j w_{jk} \; - \; l_j w_{jk}^+ \; - \; h_j w_{jk}^-}{\sum_j x_j w_{jk} \; - \; l_j w_{jk}^+ \; - \; h_j w_{jk}^-} R_k$	First

Table 2.1: Summary of existing LRP rules. The righthand column shows the layers of neural network the rule is most usable in the composite LRP method. For the LRP-$\alpha\beta$ following notation is used: $(\cdot)^+ = max(0, \cdot)$ and $(\cdot)^- = min(0, \cdot)$. α, β are subject to conservation rule $\alpha + beta = 1$. In the $z^{\boldsymbol{B}}$-rule the parameters l_i and h_i are the domain constraints ($\forall_j : l_j \leq x_j \leq h_j$). Lower layers correspond to the shallower ones, i.e. close to the first layer, while upper layers are the deeper ones, i.e. closer to an output layer [7].

The γ parameter controls the degree to which positive features are preferred over negative ones. This makes resulting saliency maps more stable and less noisy. LRP-γ rule is the basis of a family of LRP-$\alpha\beta$ rules and the z^+-rule, as shown in Table 2.1, which summarises all the LRP rules.

Applying LRP to deep neural networks. Montavon et al [7] propose the method of composite LRP for application to deep neural networks, by considering the properties of LRP rules discussed above. In general, in the upper layers, i.e. deeper ones closer to the output layer, of a CNN the features related to different classes of inputs are entangled, i.e each output feature of a convolutional part of a network is connected to each of the output classes. LRP-0 is used to separate the target class from all the others in the upper layers in relevance backpropagation. In order to illustrate this, consider an example – a network for binary classification. For a particular input the network outputs a vector $[2.0 \; -2.0]$. According to Equation 2.43 denominator will be the same for calculating relevance score for both output classes. Then relevance for the first class is $2D$, where $D \; = \sum_k \frac{w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$, and for the second one - $(-2D)$.

As the features are forwarded into shallower layers, in the middle layers on the network the features of different classes become progressively more separated and the LRP-ϵ can

be applied to compute the relevance score. Finally, the LRP-γ rule is most appropriate for the lower layers of a CNN, as it tends to spread the relevance score uniformly across an importance map. While the same general principles apply to all networks, Montavon et al note that the application of each LRP-rule should be manually adjusted for any particular CNN [7].

Contrastive LRP

In order to describe *CLRP* [126], consider saliency mapping of a DNN trained for a classification task. Such a DNN takes an input, $\boldsymbol{X} = \{x_1, x_2, ..., x_N\}$, and produces an output, $\boldsymbol{Y} = \{y_1, y_2, ..., y_M\}$, where M is the number of classes present in the training dataset. For each output neuron the network predicts an activation value, S_{y_m}, before the softmax layer. Applying the LRP method to the network's output produces a set of relevance scores for all input neurons, $\boldsymbol{R} = \{r_1, r_2, ..., r_N\}$. In LRP $\boldsymbol{R}$ is calculated by passing the relevance backwards through the network from S_{y_m} of a target class to $\boldsymbol{X}$. As a result, the values of $\boldsymbol{R}$ are different for different target classes even if the inputs are the same, as the weight connections for the classes will differ. Conversely, for each target class, the patterns of non-zero entries are similar in $\boldsymbol{R}$ for different inputs, as calculations are starting from the same value. Therefore, the LRP method generates similar importance maps for different output classes. The CLRP method introduces an additional calculation step into the LRP method to avoid this disadvantage of LRP for classification networks.

In a classification network each class is described by a distinctive set of input neurons such that $\boldsymbol{X}_{dis} \subset \boldsymbol{X}$. The input neurons in $\boldsymbol{X}_{dis}$ for the target class receive higher relevance scores than the same neurons in $\boldsymbol{R}$ produced for all other classes. Therefore, the input neurons in $\boldsymbol{X}_{dis}$ can be identified by comparing the explanation corresponding to $\boldsymbol{X}_{dis}$ to all the other explanations. In CLRP, a *virtual class* is constructed to represent all the explanations not corresponding to the target class.

A schematic representation of the CLRP method is shown in Figure 2.21. Consider the weights connecting the m-th class of output neuron to an input in a network consisting of L layers - $\boldsymbol{W} = \{\boldsymbol{W}^1, \boldsymbol{W}^2, \ldots, \boldsymbol{W}^{L-1}, \boldsymbol{W}^L_m\}$, where

- $\boldsymbol{W}^L$ are the weights connecting the $(L-1)^{th}$ layer and the L^{th} layer,
- $\boldsymbol{W}^L_m$ are the weights connecting the $(L-1)^{th}$ layer to the neuron in the L^{th} layer corresponding to the m^{th} class, y_m.

The LRP backward pass for an output neuron, y_m, can be defined as $\boldsymbol{R} = f_{LRP}(\boldsymbol{X}, \boldsymbol{W}, S_{y_m})$. Now define a *virtual concept* O as the features of $\boldsymbol{X}$ which describe the class m and are modelled by the neuron y_m. Then a *dual virtual concept* $\bar{O}$ can be defined as the features corresponding to all the classes except the target class. From this definition it follows that the concept, $\bar{O}$, is opposite to the concept O. For example, in Figure 2.21. the concept O represents an object of class 2 and $\bar{O}$ models all other classes (including classes 1 and 3). The concept $\bar{O}$ is described by LRP backward pass through the following set of weights - $\bar{\boldsymbol{W}} = \{\boldsymbol{W}^1, \boldsymbol{W}^2, ..., \boldsymbol{W}^{L-1}, \boldsymbol{W}^L_{-m}\}$, where $\boldsymbol{W}^L_{-m}$ are the weights connecting the $(L-1)^{th}$ layer to the neurons in the L^{th} layer corresponding to all classes except the m^{th} class neuron, y_m. In Figure 2.21 red lines represent the LRP backward passes for all the classes except 2. LRP backward pass for the visual concept $\bar{O}$ can be defined as $\boldsymbol{R_{dual}} = f_{LRP}(\boldsymbol{X}, \bar{\boldsymbol{W}}, S_{y_m})$. Then a relevance vector for the CLRP method can be defined as

$$\boldsymbol{R}_{CLRP} = max\left(0, (\boldsymbol{R} - \boldsymbol{R}_{dual})\right), \tag{2.48}$$

where the function $max(0, \cdot)$ is used to ensure that all negative values are set to 0. Considering the difference $\boldsymbol{R} - \boldsymbol{R}_{dual}$ in Equation 2.48 cancels the common features and ensures that only most descriptive neurons are included in $\boldsymbol{X}_{dis}$.

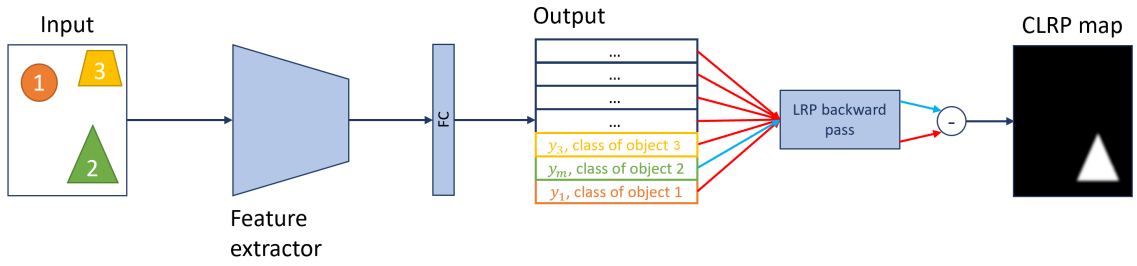

Figure 2.21: Schematic of the CLRP saliency mapping method. For the illustration of how CLRP distinguishes between the features corresponding to different classes in a classification network, a synthetic input containing the objects of three classes are presented. The input is put through a convolutional feature extractor and a fully-connected layer to produce an output. Then, from an activation value, S_{y_m}, corresponding to each output class a LRP backward pass is performed as described in Section 2.3.6. To obtain a CLRP map a subtraction is then performed of the maps not corresponding to a target class (2) from the map corresponding to that class.

2.4 PROMETHEE II method

PROMETHEE II is a widely used Multiple-Criteria Decision Analysis (MCDA) method.

2.4.1 Multicriteria Problems

Consider a general multicriteria problem:

$$max\{g_1(a), g_2(a), \ ..., g_j(a), ..., g_k(a)|a \in A\}, \tag{2.49}$$

where

- A – a finite set of N alternatives, $\{a_1, a_2, ..., a_N\}$,
- G – a finite set of K evaluation criteria, $\{g_1(\cdot), g_2(\cdot), ..., g_K(\cdot)\}$.

The objective of MCDA is to find an alternative optimising all the criteria. In most cases it is an ill-posed mathematical problem as there is no single alternative which optimises all criteria simultaneously. MCDA provides the methods to find an optimal solution in such cases. In order to perform the optimization additional information is needed. Such additional information can be divided into two categories – information between and information within the criteria.

2.4.2 Additional information

Information between the criteria is represented by the criteria importance weights which form a finite set as $W = \{w_1, w_2, ..., w_K\}$. The weights are normalised such that $\sum_{k=1}^{K} w_k = 1$. The higher the weight of a criterion, the higher is its importance.

The PROMETHEE methods are ranking methods, i.e. they utilise pair-wise comparisons among the alternatives. A ranking is constructed by considering the difference between evaluations of the alternatives for a particular criterion., i.e. information within the criteria. To obtain the full ranking this comparison process is performed for all criteria. Depending on the size of a pair-wise difference, a decision can be made – one alternative can be preferred over another or not. The larger is the difference, the greater is the preference of one alternative over the other. For convenience, the preference can be represented by a real number varying from 0 to 1. overall, this can be described by the following function:

$$P_k(a, b) = F_k[d_k(a, b)] \ \forall a, b \in A, \tag{2.50}$$

where

- $d_k(a, b) = g_k(a) - g_k(b)$ is the difference between evaluations of two alternatives on the same criterion,

- F_k is a function to scale d_k into an interval such that $0 \leq P_K \leq 1$.

A number of forms have been proposed for the P_k function including a constant function, a Heaviside function and functions involving exponential decay [129]. In this thesis a V-shaped function for P_K is used as defined in Equation 2.51 and shown in Figure 2.22:

$$P(d) = \begin{cases} 0 & d \leq 0 \\ \frac{d-q}{p-q} & q < d \leq p \\ 1 & d > p \end{cases} \tag{2.51}$$

where p and q are the parameters to be set. One of the choices is to set $q = min(g(a), g(b))$ and $p = max(g(a), g(b))$ as $d = g(a) - g(b)$. A V-shaped function is used as the only constraint for the $P(d)$ function is that larger d should lead to larger value of $P(d)$. q and p are chosen as discussed above for $P(d)$ to vary from 0 to 1 for any distance d.

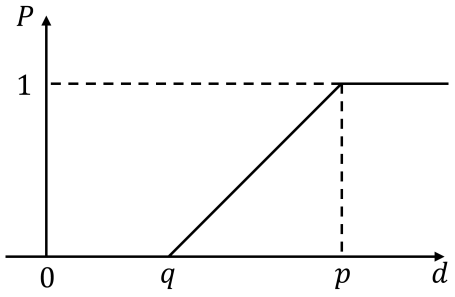

Figure 2.22: V-shaped form of the P_K function. q and p are defined as in Equation 2.51.

2.4.3 PROMETHEE II method

In the PROMETHEE II method, an aggregated preferences index, $\pi(a, b)$, describes the overall preference across all criteria for alternative a over b. $\pi(a, b)$ is defined for all a and b as

$$\begin{cases} \pi(a,b) = \sum_{k=1}^{K} w_k P_k(a,b), \\ \pi(b,a) = \sum_{k=1}^{K} w_k P_k(b,a). \end{cases} \tag{2.52}$$

As can be seen from Equation 2.52 the greater the $\pi(a, b)$, the greater is the preference of a over b. This aggregated preferences index can thus be used to summarise the degree to which one alternative is better than the others. The positive outranking flow, φ^+, defines

how an alternative, a, is better than all the others:

$$\varphi^+ = \frac{1}{N-1} \sum_{x \in A} \pi(a, x). \tag{2.53}$$

On the other hand, the negative outranking flow describes how an alternative, a, is worse than all the others:

$$\varphi^- = \frac{1}{N-1} \sum_{x \in A} \pi(x, a). \tag{2.54}$$

Positive and negative outranking flows can be used to obtain a complete ranking of all alternatives [129], with the final ranking of each alternative being determined by the net outranking flow:

$$\varphi(a) = \varphi^+(a) - \varphi^-(a). \tag{2.55}$$

Chapter 3

Biology of brain ageing

Ageing can be defined as "time-dependent functional decline" [18]. Ageing effects on brain and cognition are spread throughout the brain and have various manifestations. Ageing affects brain on molecular and cellular levels as well as its vasculature, morphology and human cognition. For instance, age-related morphological changes cause brain volume to shrink. Vacsculature changes occurring with age lead to White Matter (WM) lesions, i.e. the regions of demyelination of neural axons. Cognitive changes result in decline in memory function [130]. Brain ageing is researched by studying cognitive abilities, genetics, structural and functional changes in brain throughout human lifetime. Molecular and cellular level ageing-related changes are also studied using animal models. In this chapter mechanisms of brain ageing are discussed.

Studies of cognitive function show that ageing-related effects in older-subjects particularly affect memories of life experiences and situations, i.e. episodic memory. Episodic memory is studied by presenting subjects with tasks considering retrieving non-verbal and verbal material, i.e. list of words [131]. Effect of ageing on cognitive abilities is also measured by evaluating the effect on human intellect. For that purpose intelligence theory based on the concepts of crystallised and fluid intelligence is used. Crystallised intelligence is a concept describing human knowledge accumulated as the result of educational and cultural experiences and exposure to information. The concept of fluid intelligence refers to cognitive abilities engaged in solving any new problem present to a human. Previous research shows that crystallised and fluid intelligence are influenced by ageing-related changes differently. Fluid intelligence starts to decline in early adulthood. Age-related changes in fluid intelligence are measured by using cross-sectional and longitudinal data [132]. Fluid

intelligence describes intellectual skills acquired during lifetime and among such skills speed of information processing, working memory, visuo-spatial ability and abstract reasoning are affected the most [133]. Crystallised intelligence is found to be either stable throughout human lifetime or be affected by a small changes compared to the changes in fluid intelligence on old age, after late sixties [132]. For example, certain aspects of semantic memory representing knowledge about surrounding world and emotional regulation is not subject to age-related effects. However, some aspects of semantic memory such as verbal abilities are known to be even improving with age, but a decline is still observed in the very old age [134]. Nyberg et al [135] used longitudinal cohort and identified average decline of episodic memory to be on average around the age of 60. Rönnlund et al [136] also studied trajectories of cognitive changes based on changes in episodic and semantic memory both on cross-sectional dataset and cohort-matched longitudinal dataset. The results obtained on cross-sectional and longitudinal datasets did not agree. On cross-sectional data episodic memory was found to be gradually declining with age, while on longitudinal data there was not changes found in episodic memory before the age of 60. Semantic memory was studied using longitudinal dataset and minor declines were found to occur before the age of 55 with decline slower in older ages compared to episodic memory. The authors explained the discrepancies between the results obtained on cross-sectional and longitudinal datasets by the differences in educational level of participants in these two datasets. Overall Rönnlund et al highlight the need of retesting in cross-sectional and longitudinal datasets [136].

However, it should be noted that the research by Rahhal et al [137] has shown that the magnitude of age differences observed in memory task studies is dependent on experimental setup. Moreover, in studying brain ageing by observing cognitive trajectories *differentiation and dedifferentiation hypotheses* should be taken into account. *Differentiation hypothesis* was stated by Garrett [138]: "abstract or symbol intelligence changes in its organization as age increases from a fairly unified and general ability to a loosely organized group of abilities or factors". On the other hand, *dedifferentiation hypothesis* states that cognitive abilities become more and more interrelated during adulthood [139]. Recent advances in the field of adult developmental theory, however, study dedifferentiation hypothesis in both cross-sectional and longitudal cohorts. The studies, however, produce mixed support for the dedifferentiation hypothesis [140]. Also recent systematic review of 37 studies of cognitive ageing trajectories by Wu et al [141] showed that it is not possible to define specific cognitive trajectories for which different studies would provide consistent evidence.

Therefore, other experimental data sources are used to understand the mechanisms in brain which underlie age-related cognitive changes [142].

In order to understand mechanisms of human brain ageing cellular and molecular mechanisms underlying it are studied. At a cellular level mitochondrial dysfunction is studied as mitochondria are part of the cells which make up dendrites and axons of neurons in a brain. Dendrites and axons are parts of nerve cells which serve the signal propagation between the cells [143]. Mitochondria in dendrites and axons play important role as they are cell "power plants" [144] and generate energy required for signal transmission, cell maintenance and repair. Mitochondria are involved in several processes inside cells. Mitochondria are capable of mitochondrial biogenesis (growth in size and division) and mitophagy (removal of mitochondria by degradation). Mitochondria also play role in cell homeostasis and apoptosis, which is one of the processes of programmed cell death [145] occuring in normal brain development and in pathological ageing. Mitochondrial changes associated with ageing are studied by extracting mitochondria from animal brain tissue. For instance, the studies on mice brain tissue show that most cell type in the brain accumulate dysfunctional mitochondria with ageing [146, 147]. Since neurons are integrated into neuronal networks throughout the lifetime of the organisms, the ability of cells comprising the neurons to remove damaged and dysfunctional molecules is of importance. Experimental evidence, however, also shows that this ability of cells also decays with age [148, 149, 150] [151]. Other molecular age-related changes include decline in brain aerobic glycolysis. Glucose uptake in brain is found to be decreasing with age. As the result, glucose uptake decrease exceeds oxygen use and decline in brain aerobic glycolysis occurs [152].

Brain ageing is also studied at the level of brain neuronal networks. During brain ageing integrity of neuronal networks' activity is increasingly perturbed. Neuronal networks are affected by hyperexcitability and excitotoxic damage with ageing [151]. Excitotoxic damage of nerve cells is present whenever the levels of neurotransmitters, which are safe and necessary for the work of neuronal network, are pathologically high and this leads to overstimulation of receptors [153]. Neurotransmitters are produced by nerve cells specifically for signal transmission [154]. Neurotransmitters are also important for following brain functions - learning, memory, decision making and regulating mood. Thus dysregulation of neurotransmitters can lead to cognitive decline in ageing individuals [155, 151].

There is also a number of other molecular mechanisms proposed in existing literature to explain the course of human brain ageing. It should be noted that none of this mechanisms

cannot fully explain the process of ageing, but some mechanisms are supported by extensive experimental evidence. Here known mechanisms are briefly presented. One of the theories explains cellular damage in healthy ageing by increasing production of free radicals with age. Free radicals are atoms and molecules with unpaired electrons in outer electron shell [156]. Production of free radicals leads to mitochondrial dysfunction which has been shown to be be related to brain ageing [157]. Existing literature also contains supportive evidence for the calcium hypothesis of brain ageing. Ageing is related to changes in calcium homoeostasis and it is involved in the number of cellular functions in brain including neurotransmitter synthesis and release [29]. Another molecular theory explains ageing by secretion of glucocorticoids. Secretion of glucocorticoids influences the process of brain ageing as it affects hippocampus and hippocampus is related to the performance in learning and memory tasks [158].

Study into molecular mechanisms explaining brain ageing also includes studies of age-related gene expressions. Age-related gene expressions are studied by profiling over human lifetime. This provides understanding of biological processes which are activated or suppressed at various stages. In previous research it has been highlighted that genes regulating neuroinflammation and immune system performance are of particular importance in the ageing process. Other gene expressions are found to be influencing synapses and this leads to cognitive decline in older adults. There is also existing research into protective genetic mechanisms. Studying gene expressions having both "negative" and "positive" effect on the process of brain ageing allows to extract healthy and disease related ageing trajectories [159]. Genetic studies of brain ageing also involve studies focused on telomere shortening. Telomeres consists of repeated nucleotide (molecules, monomeric units of DNA) DNA sequences at the ends of chromosomes. Human body cells are experiencing telomere shortening and cell senescence in the process of ageing. As cells divide, telomeres can shorten. During cell's interphase (the phase of cell's life, in which it spends most of its life [160]) telomeric DNA is conserved by telomere-associated proteins. Excessive telomere shortening disrupts the protection mechanism of telomeric DNA. This can lead to DNA damage and further cell apoptosis or senescence. Since neurons do not perform mitosis (division of a cell into two genetically identical cells [161]), neuron's cell telomeres cannot shorten, but still can be damaged during neuron's life cycle. In mice models it was also previously shown that deficit of telomerase leads to the problems with spatial learning and memory [151].

Post-mortem studies are also used to study biochemical age-related changes such as iron accumulation with age [162]. Another manifestation of biochemical age-related changes is occurrence of Senile Plaques (SPs), which are extracellular protein deposits [163]. Evidence in existing research shows that SPs appear both in the course of healthy ageing and in neurodegenerative conditions such as Alzheimer's disease [164]. SPs are considered to be of two types – dense and pale diffuse SPs. These two types of SPs are originating from different biochemical processes. However, in existing research there is no definitive conclusion whether SPs are related to normal ageing or not as conflicting results are published. There are research results indicating that SPs are infrequent in subjects who died without dementia symptoms [44]. Other results show occurrence of SPs is unavoidable in subjects living long enough [165]. Biochemical age-related changes are also reflected by changes in Neurofibrillary Tangles (NFTs). NFTs mostly consist of tau protein which are known to be associated with the ageing process [156]. NFTs are found to be accumulating with age locally in healthy subjects, while in Alzheimer's disease patients NFTs accumulate in whole brain cortex [166].

Post-mortem studies also investigate the changes in synaptic density with age. The results of such studies show decline in synaptic density, but also indicate that synaptic loss may be compensated [167]. In previous research one of the reasons named for such changes are cerebrovascular changes. Cerebrovascular changes manifest themselves in decreasing density of capillaries [168] and increasing microvasculature deformations. After 50 years of age the amount of microvasculature deformations constantly increases [169].

Further, structural changes in brain have been studied to understand its ageing. Post-mortem studies in humans are used to understand structural changes occurring with age. Post-mortem studies of brain atrophy reveal 5% reduction in whole brain volume per decade after 40 years of age [156]. Post-mortem studies also show that white matter is affected more in healthy ageing compared to grey matter [170]. These findings of post-mortem studies are in agreement with Magnetic Resonance Imaging (MRI) data studies (for example, [51, 171]). Post-mortem study have also identified to be in accordance with the findings of studying cognitive ageing. Henstridge et al [172] found that morphological and molecular changes associated with ageing occurred in brain regions associated with cognition. This post-mortem study using the Lothian Birth Cohort 1936 also found that synaptic structure is more preserved in healthy ageing compared to the subjects affected by Alzheimer's diseases [172].

A number of studies have investigated structural brain changes occurring with ageing using MRI data acquired with T1-weighted sequences. All such studies use volumetric measures to assess changes in brain structures during human lifetime. MRI data is first segmented into separate brain regions for its volume to be assessed. Studies using both longitudinal [51] and cross-sectional data exist [171]. The studies consistently report decline in total brain volume, WM and Grey Matter (GM) with age, while ventricular and Cerebrospinal Fluid (CSF) volumes increase [51, 173, 171, 173, 174, 175, 176]. Raz et al [177] used longitudinal dataset with two data points per subject to also observe shrink in brain region volumes. Review by Fama and Sullivan [178] describe thalamic volume decreases with age and their relation to decline in cognitive performance. Salat et al [179] also report cortical thinning occurring with age in all brain regions. Vinke et al [171] reported volumetric decrease with age in putamen, amygalda, pallidium and caudate. Wei et al [180] used T1-weighted MRI cross-sectional data in order to measure the changes in Subcallosal Region (SCR) volume during adulthood. The authors found the volume of SCR to be decreasing.

A number of studies also used T1-weighted sequences MRI data in order to investigate which regions of the brain are associated with brain ageing, but instead of measuring volumes of the brain regions used statistical methods. Some studies defined the ageing-related regions by averaging important features over whole age range present in cross-sectional dataset used. For instance, Fujimoto et al [181] and Kondo et al [182] have reported that frontal association area, the Wernicke's area, the angularis gyrus and the primary motor cortex are associated with the brain ageing process. Amoroso et al [13] also found frontal gyrus to be important along with Heschl's gyrus, anterior cingulate parahippocampal gyrus, cingulate gyri, precentral gyri and temporal lobe. Statistical methods were also applied to derive age-related features in the form of trajectories, i.e. separately for a number of age ranges over human lifetime. Feng et al [39] presented average maps separately for each life decade from 20s to 80s. In these results the importance of frontal lobes increased with age [39]. Bermudez et al [183] defined three age ranges over adulthood and concluded that GM volume decreases with age in accordance with volumetric studies. Detailed discussion on structural brain changes relating to age can be found in Chapters 6 and 7.

Functional imaging is used also to study brain ageing. For this purpose Single Photon Computed Emission Tomography (SPECT) [184] and Positron Emission Tomography

(PET) [185, 186] imaging are applied. PET imaging was also used to identify decline of cerebral blood flow and metabolism in GM with age. However, PET technique studies also identify no change in blood flow in WM [156]. Functional imaging studies of cognitive performance, memory in particular, show decline in older subjects compared to the young ones [187]. In retrieval tasks, older subjects involve wider networks. This may be explained by inefficiency of networks leading to their widening [156]. Functional MRI (fMRI) is also used to study brain ageing including vascular age-related changes [188]. One of the studies using fMRI has identified reduction in cerebral blood flow responsiveness to visual stimuli in healthy ageing [189]. Li et al have shown that regional brain activation in fMRI analysis change with ageing and can be related to behavioural outcomes in research subjects. Older subjects were found to have hypo-activation in the visual network and hyper-activation in the frontoparietal and default networks (brain regions which show spontaneous activation during passive periods in fMRI [190]) compared to younger ones [191].

In conclusion, it can be said that existence of age-related brain changes is evident from previous research. Age-related brain changes affecting human cognition and behaviour originate from the changes on intercellular and intracellular, molecular, tissue and organ levels. However, the rate of change of these changes and the processes behind should be studied further.

Chapter 4

Datasesets

In this chapter the datasets used in subsequent chapters are described. In Section 4.1 the non-MR image datasets used are described and Magnetic Resonance Imaging (MRI) datasets are presented in Section 4.2.

4.1 Image datasets

4.1.1 Mixed National Institute of Standards and Technology (MNIST)

The MNIST dataset consists of the images of handwritten digits [192]. The MNIST dataset is widely used as a benchmark dataset in image recognition tasks for understanding algorithms' performance. In the dataset each grayscale image of a handwritten digit of size 28×28 pixels is correspondingly labelled from 0 to 9, resulting in 10 classes. The examples of images are given in Figure 4.1.

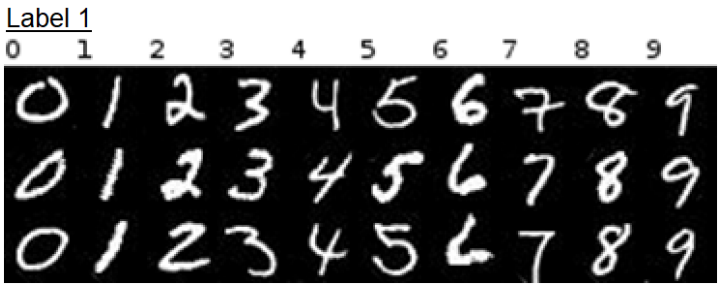

Figure 4.1: Examples of the MNIST dataset samples for each of 10 classes.

4.1.2 Ordinal MNIST

The Ordinal MNIST dataset was created by modifying the MNIST dataset such that the images have two labels. The first label is taken from the MNIST dataset labelling

convention. The second label is 0 for label 1 values from 0 to 2, 1 – from 3 to 5, 2 – from 6 to 9, resulting in 3 classes. All the images with second label 1 were multiplied by an image drawn from normal distribution with mean 0 and standard deviation 0.25 of the same size as MNIST samples. For the images with second label being 1 mean of a normal distribution was 0.5 and for 2 – 1.0. Example of the images in the dataset are shown in Figure 4.2.

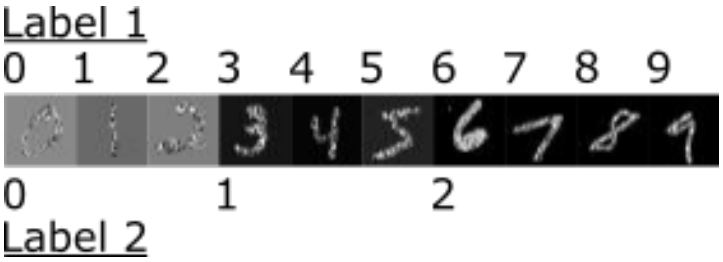

Figure 4.2: Examples of the ordinal MNIST dataset samples.

4.1.3 Two-variable ordinal dataset

Two-variable ordinal dataset was created by generating images of a circle on black background with 2 labels each – one label corresponds to intensity of a circle in a grayscale image and another to a circle radius. In this dataset first label has 14 classes which correspond to constant change in circle radius from class to class. Second label has three classes corresponding to a mean of normal distribution from which the values are drawn to multiply white circle pixels by. In Table 4.1 the variable values corresponding to the two labels are listed. Image size in the dataset is 100×100 pixels and a circle is drawn with its centre at random position coordinates drawn from a 2D uniform whole number distribution of range $100 - r$ for each dimension, where r is a radius corresponding to a class of label 1. For each of the 14 classes 200 images were generated for training and 20 for testing. Example of the images in the dataset are shown in Figure 4.3.

Label 1	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Radius	2	4	6	8	18	20	22	24	34	36	38	40	42	44
Distribution mean	0.08	0.12	0.16	0.2	0.44	0.48	0.52	0.56	0.80	0.84	0.88	0.92	0.96	1.0
Label 2	1				2				3					

Table 4.1: Variable values defining the classes for two labels in continuous ordinal dataset. The classes corresponding to label 1 are defined by white circle radius, while the classes corresponding to label 2 are defined by a mean of intensity of normal distribution from which the values are drawn to multiply white circle pixels by.

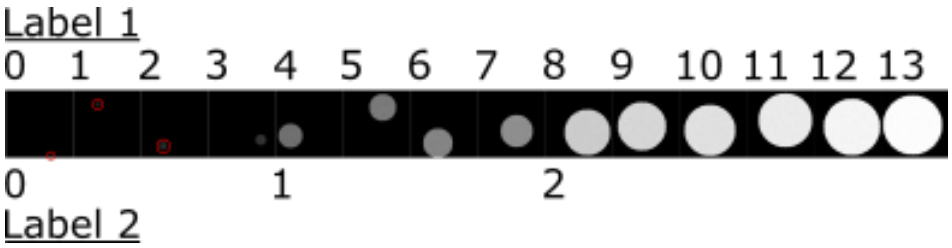

Figure 4.3: Examples of the two-variable ordinal dataset samples. Red circles are drawn for visualisation purposes in the classes 0 – 2 and are not present in the dataset images.

4.1.4 ImageNet

ImageNet is a publicly available dataset consiting of over 15 million natural images having over 22,000 labels. The images were collated from the web. Labels in the dataset were chosen by humans through the Amazon's Mechanical Turk tool. The ImageNet-ILSVRC dataset in a subset of the ImageNet dataset formed for the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). The ImageNet-ILSVRC dataset consists of images with 1000 labels and over 1000 images per label [74]. Pytorch library [29] includes a number of models, e.g. AlexNet model [74], pretrained on the ImageNet-ILSVRC dataset.

In this thesis a subset of the ImageNet-ILSVRC dataset was used. This subset included the images with two labels, for which pretrained AlexNet model achieved the best accuracy – "tiger" and "zebra". In original dataset images are of variable size. Pretrained AlexNet resize each image first to the size 256×256 and then applies central crop to size 224×224. Examples of images in this subset dataset are shown in Figure 4.4.

Figure 4.4: Examples of the subset of ImageNet-ILSVRC dataset samples. The subset contains the images of two classes – "zebra" and "tiger". Pretrained AlexNet has the best accuracy on these two classes. Upper rows shows images "easier" for classification and bottom row – "harder".

4.2 MRI datasets

The datasets were obtained after seeking permission from data owners and agreeing to the terms and conditions of data usage. The study protocols of all the datasets used include obtaining informed consent from the participants both for the original use of the data, and for data re-use in projects such as the current one, either as part of normal study procedures, or as a separate, optional, step.

All MRI data is presented in this thesis using radiological convention.

4.2.1 Healthy Brain Ageing from Public Sources (HBAPS) dataset

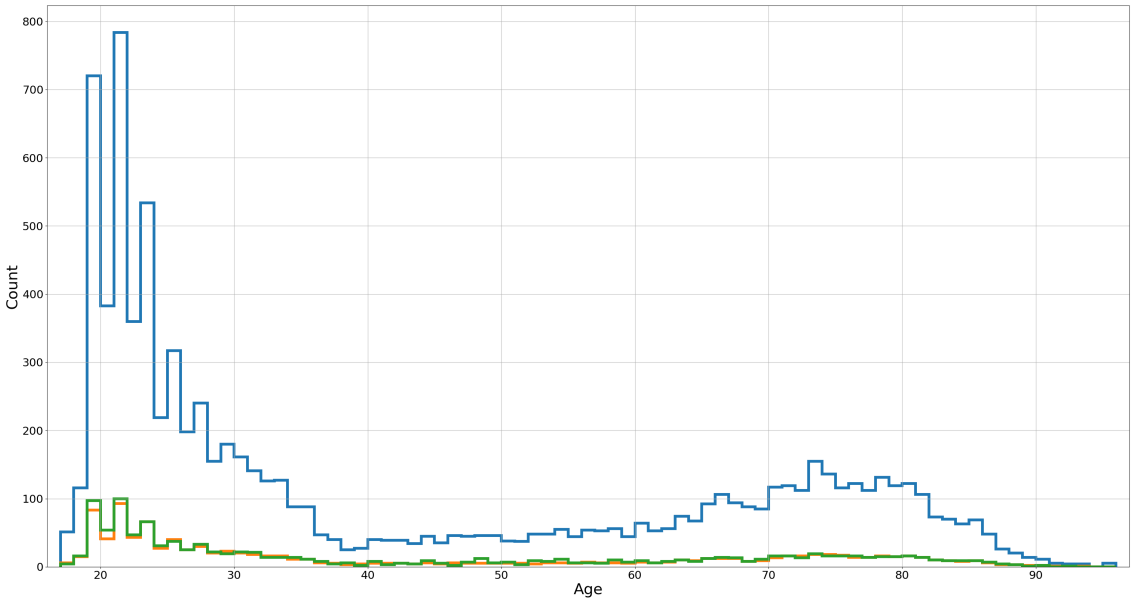

Figure 4.5: The age distribution of samples ($N = 10,876$) over the age range of 17 – 96 years in the HBAPS dataset. Blue corresponds to training data, green - validation, orange - testing.

As shown in Table 4.2 the HBAPS dataset was compiled from 36 public and 2 non-public datasets available through various data-sharing projects. The cohort consists of 8631 subjects (male/female = 3732/4249, age range 17 – 96 years). In the HBAPS dataset there are 2.8% of subjects with repeat scans. In order to increase the overall size of the dataset, multiple scans from the same subject were considered as unrelated samples. There are several reasons for this. Firstly, scans from the same subject acquired at different times are as representative of the population as scans from different subjects. Secondly, the models used in Chapter 5 do not contain information about interactions between scans acquired from the same subjects at different time. Thirdly, consider an "absurd" dataset, in which

all samples are collected from the same subject. In this case a brain age prediction model would only be capable of predicting brain age of that one subject. Therefore, if the brain age predicting model is trained on the dataset, in which only some number of subjects have repeat samples and number of such subjects is much smaller compared to the total number of subjects, it can be considered that the brain age prediction model captures brain ageing features as close to population-wide effects as big is the training dataset.

As 2.8% of the HBAPS dataset subjects have repeat scans, this results in 10878 scans (male/female = 4876/5352, mean age = 42.1 ± 23.2 years, age range 17 – 96 years). 650 subjects belonging to the for2017 dataset do not have information on subjects' sex and mean subject age for each sex is not given. The distribution of samples over the age range is shown in Figure 4.5. All data were collected using either 1.5T or 3T scanners using T1-weighted sequences (details are shown in Table 4.2).

Only healthy subjects were included in the dataset. Disease status of the subjects was determined from information provided by each study. This information is summarised in Table 4.3. For the 1000 FCP, ABIDE I, ABIDE II and NKI-RS datasets no protected health information is provided in compliance with the Health Insurance Portability and Accountability Act (HIPAA) statute. In United States the HIPAA statute regulates how personal information should be handled in healthcare. For the DLBS, for2017, HNU, IBATRT, IXI, SWU, VPT and XHCUMS datasets only information on healthy or diseased status of subjects is available.

Table 4.2: Data sources for the HBAPS dataset.

Cohort	**№** Image	**№** Subject	**Age** $\bar{x}(\sigma^2)$	**Age**	**Sex** (M/F)	**Repo**	**Scanner** (Field strengtht)	**Scan**	**Voxel** Size(mm)
1000 FCP 1000 Functional Connectomes Project	836	834	30.5 (14.8)	17-85	377/459	INDI	* (1.5T, 3T)	*	*
ABIDE I Autism Brain Imaging Data Exchange	177	177	25.2 (7.3)	17-56	154/23	INDI	Various (all 3T)	MPRAGE	Various
ABIDE II	144	144	27.1 (10.3)	17-64	36/108	INDI	Various (all 3T)	MPRAGE	Various
ADHD Attention Deficit Hyperactivity Disorder	53	53	18.0 (1.1)	17-21	22/31	INDI	Siemens (3T)	MPRAGE	Various
ADNI Alzheimer's Disease Neuroimaging Initiative	1608	267	76.7 (6.5)	56-95	808/800	LONI IDA	Various (1.5T, 3T)	MPRAGE	Various
AIBL Australian Imaging Biomarkers and Lifestyle	396	348	73.7 (7.0)	55-96	378/418	LONI IDA	Various (1.5T, 3T)	MPRAGE	Various
BNU 1 Beijing Normal University	107	57	23.0 (2.3)	19-30	56/51	INDI	Siemens Tim Trio(3T)	MPRAGE	$1.3 \times 1.0 \times 1.3$
BNU 2	122	61	20.9 (0.9)	19-23	66/56	INDI	Siemens Tim Trio(3T)	MPRAGE	Various
BNU Enhanced Sample	180	180	21.2 (1.9)	17-28	73/107	INDI	Siemens (3T)	MPRAGE	$1.33 \times 1.0 \times 1.0$
BNU Eyes Open Eyes Closed (EOEC) Sample	180	48	22.5 (2.2)	18-30	24/24	INDI	Siemens Tim Trio(3T)	MPRAGE	$1.3 \times 1.0 \times 1.3$
BNU Short TR Sample	28	28	24.5 (7.5)	18-46	14/14	INDI	Siemens Tim Trio(3T)	MPRAGE	$1.3 \times 1.0 \times 1.3$
CamCAN Cambridge Centre for Ageing and Neuroscience	652	652	54.3 (18.6)	18-88	322/330	CamCAN**	Siemens Tim Trio(3T)	MPRAGE	$1.0 \times 1.0 \times 1.0$

continued ...

. . . continued

Cohort	№ Image	№ Subject	Age $\bar{x}(\sigma^2)$	Age	Sex (M/F)	Repo	Scanner (Field strengtht)	Scan	Voxel Size[mm]
COBRE Centre for Biomedical Research Excellence	74	74	35.8 (11.6)	17-65	11/20	INDI	Siemens Tim Trio(3T)	MPRAGE	$1.0 \times 1.0 \times 1.0$
DLBS Dallas Lifespan Brain Study	315	315	54.6 (20.1)	20-89	117/198	INDI	Philips (3T)	MPRAGE	$1.0 \times 1.0 \times 1.0$
For2017***	650	650	32.6(12.4)	18-65	***	Private**	***	***	***
GSP Brain Genomics Superstruct Project	1634	1634	21.5 (2.9)	19-35	698/936	LONI IDA	Siemens Tim Trio(3T)	MPRAGE	$1.2 \times 1.2 \times 1.2$
HCP Human Connectome Project	1113	1113	28.8 (3.7)	22-37	507/606	LONI IDA	Siemens Magnetom(3T)	MPRAGE	$1.0 \times 1.0 \times 1.0$
HNU Hangzhou Normal University	300	30	24.4 (2.4)	20-30	150/150	INDI	MR750 (3T)	SPGR	$1.0 \times 1.0 \times 1.0$
IBATRT Intrinsic Brain Activity (Test-Retest)	50	36	28.0 (7.6)	19-48	23/27	INDI	Siemens Tim Trio(3T)	MPRAGE	$1.0 \times 1.0 \times 1.0$
IPCAS Institute of Psychology Chinese Academy of Sciences	39	35	21.2 (1.9)	17-35	14/25	INDI	Siemens Tim Trio(3T)	MPRAGE	$1.0 \times 1.0 \times 1.0$
IPCAS 4	20	20	23.2 (1.6)	21-28	10/10	INDI	GE Discovery(3T)	SPGR	$1.0 \times 1.0 \times 1.0$
IPCAS 5	22	22	18.3 (0.5)	18-19	22/0	INDI	Siemens Tim Trio(3T)	MPRAGE	$1.0 \times 1.0 \times 1.0$
IPCAS 6	2	2	23.0 (2.8)	21-25	1/1	INDI	Siemens Tim Trio(3T)	MPRAGE	$1.0 \times 1.0 \times 1.0$
IPCAS 8	26	13	57.6 (3.6)	50-61	12/14	INDI	Siemens Tim Trio(3T)	MPRAGE	$1.3 \times 1.0 \times 1.3$

continued . . .

. . . continued

Cohort	№ Image	№ Subject	Age $\bar{x}(\sigma^2)$	Age	Sex (M/F)	Repo	Scanner (Field strengtht)	Scan	Voxel Size[mm]
IXI**** Information eXtraction from Images	563	583	49.5 (16.8)	19-86	328/255	****	Philips Intera(1.5T, 3T)	MPRAGE, T1-FFE	$0.94 \times 0.94 \times 1.2$
NKI-RS Nathan Kline Institute Rockland Sample	341	341	45.88 (18.8)	17-85	117/224	INDI	Siemens Tim Trio(3T)	MPRAGE	$1.0 \times 1.0 \times 1.0$
PPMI***** Parkinson's Progression Markers Initiative	245	175	60.3 (11.6)	31-83	154/91	LONI IDA	*****	MPRAGE, SPGR	$1.0 \times 1.0 \times 1.2$
SWU 1 Southwest University	59	20	21.5 (1.7)	19-24	21/38	INDI	Siemens Tim Trio(3T)	MPRAGE	$1.0 \times 1.0 \times 1.0$
SWU 2	27	27	21.0 (1.7)	18-24	7/20	INDI	Siemens Tim Trio(3T)	MPAGE	$1.0 \times 1.0 \times 1.0$
SWU 3	46	23	20.4 (1.6)	18-25	16/30	INDI	Siemens Tim Trio(3T)	MPRAGE	$1.0 \times 1.0 \times 1.0$
SWU 4	464	234	20.0 (1.3)	17-27	236/228	INDI	Siemens Tim Trio(3T)	MPRAGE	$1.0 \times 1.0 \times 1.0$
VPT******	28	28	28.9 (1.6)	26-32	14/14	Private******	******	******	******
XHCUMS Xuanwu Hospital, Capital University of Medical Sciences	78	23	53.6 (5.5)	36-62	41/37	INDI	Siemens Tim Trio(3T)	MPRAGE	$1.0 \times 1.0 \times 1.0$

*1000 FCP dataset: information on scanner model, scan type and voxel dimension is not available.

ABIDE consortium comprises data from various sites with different scanners/parameters.

**CamCAN study uses it own data repository which is accessible on request (https://www.cam-can.org/index.php?content=dataset)

***For2017 is a non-publicly available dataset obtained from the machine learning group at the Münster University. Information on scanner model, scan type, voxel dimensions and subjects' sex is not provided.

INDI = International Neuroimaging Data-sharing Initiative (http://fcon_1000.projects.nitrc.org)

****IXI study data is accessible on project's page (https://brain-development.org/ixi-dataset/) LONI IDA = Laboratory of NeuroImaging Image & Data Archive (https://ida.loni.usc.edu)

*****PPMI: information on scanner model is not provided.

******VPT is a non-publicly available dataset obtained from the NIHR Maudsley Biomedical Research Centre at King's College London. Information on scanner model, scan type, and voxel dimensions is not provided.

Dataset	Healthy subjects information
1000 FCP	No protected health information is included [193].
ABIDE I	No protected health information is included [194].
ABIDE II	No protected health information is included [194].
ADNI	See Table 4.4 .
AIBL	Inclusion criteria for healthy subjects: age >= 60 years, Mini-mental State Examination (MMSE) score > 24, no history of stroke/transient ischemic attack or serious head injury [195].
BNU 1	Healthy subjects have no history of neurological and psychiatric disorders [196].
BNU 2	Healthy subjects have no history of neurological and psychiatric disorders [197].
BNU Enhanced	Healthy subjects have no history of neurological and psychiatric disorders [198] .
BNU Eyes Open Eyes Closed	Healthy subjects have no history of neurological and psychiatric disorders [199].
BNU Short TR	Healthy subjects have no history of neurological and psychiatric disorders [198] .
CamCAN	Inclusion criteria for healthy subjects: MMSE score > 24, no MRI contraindications, no neurological or serious psychiatric conditions, subjects meet hearing, vision, and English language ability criteria necessary for completing experimental tasks [200].

COBRE	Inclusion criteria for healthy subjects: no history of neurological disorders, mental retardation, severe head trauma with more than 5 minutes loss of consciousness, substance abuse or dependence within the last 12 months. Diagnostic information was collected using the Structured Clinical Interview used for DSM Disorders (SCID) [201, 202].
DLBS	- [203]
For2017	
GSP	Exclusion criteria for healthy subjects: self-reported health information indicates current/past history of Axis I pathology or neurological disorder, current psychotropic medication usage and/or acute physical illness or displayed atypical brain anatomy [204].
HCP	Exclusion criteria for healthy subjects: age < 22 or > 35 years, history of psychiatric disorder, substance abuse, neurological or cardiovascular diseases, two or more seizures after age of 5 or a diagnosis of epilepsy, any genetic disorder, multiple sclerosis, cerebral palsy, brain tumour, stroke, head injuries, premature birth, radiation or chemotherapy, claustrophobia. Other exclusion criteria include pregnancy, MMSE score $<= 25$, unsafe metal in body. Following treatments also lead to exclusion: thyroid hormone treatment, diabetes treatment or using prescribed medication in the past month [205].
HNU	- [198]
IBATRT	- [198]
IPCAS	Inclusion criteria for healthy subjects: no history of severe physical or mental injuries [206].
IXI	- [207]
NKI-RS	No protected health information is included [208].

PPMI	Healthy controls inclusion criteria: age >= 30 years. Healthy controls exclusion criteria: current or active clinically significant neurological disorder, first degree relative with Parkinson's disease, Montreal Cognitive Assessment (MoCA) score >= 26 [209].
SWU	- [198]
VPT	-
XHCUMS	- [210]

Table 4.3: Information on inclusion and exclusion criteria for healthy subjects contained in the 34 publicly available datasets used in the HBAPS dataset. For the 1000 FCP, ABIDE I, ABIDE II and NKI-RS datasets no protected health information is included in compliance with t he HIPAA. The HIPAA is a United federal statute which describes handling personal information in healthcare [8]. For the DLBS, for2017, HNU, IBATRT, IXI, SWU, VPT and XHCUMS datasets information only healthy or diseased status of subjects is available. In the COBRE study healthy or diseased status of subjects was determined using the SCID [9] (DSM is Diagnostic and Statistical Manual of Mental Disorders - a classification of mentral disorders [10]). SCID is a diagnostic examination for Axis I diseases [9]. According to DSM Axis I disorders include most common psychiatric diseases [10]. In the AIBL and CamCAN studies cognitevely healthy subjects are determined using the MMSE test, while in the PPMI study MoCA test is used. MMSE and MoCA tests are used to detect cognitive impairment [11, 12]. For the datasets for which inclusuion and exclusion criteria are not stated only information on healthy or diseased status of subjects is available. The datasets for2017 and VPT are non-public datasets and information on subjects' health is not accessible.

General inclusion criteria

- Subjects' age >= 55 years and <= 90
- Geriatric Depression Scale score 6
- History of at least six grades of education or extensive work history to exclude signs of mental retardation
- Hachinski Ischemic Score = 4
- Self-reported good general health, no history of disease known to impede the study results

General exclusion criteria

- Baseline MRI scan revealing history of infection, infarction or any focal lesions
- History of lacunae in memory structures
- Contra-indications for MRI
- History of diagnosed major depression, bipolar disorder within the last year, alcohol or substance abuse within the past 2 years and schizophrenia according to DSM-IV
- History of psychotic features, agitation or behavioural problems within the last 3 months leading to subject's inability to participate in the study
- Any medical condition leading to subject's inability to participate in the study
- Any biochemical blood parameters which may impede the study results
- History of use of specific psychoactive medications which may impede the study results
- History of cancer in the last five years except of skin melanoma

Inclusion criteria for CN
- No history of memory problems beyond the ones normal for given age (verified by a study partner) - MMSE score = 24 - Clinical Dementia Rating = 0 - Memory Box score = 0 - Cognitively normal, no history of significant cognitive impairment leading to interruption of daily living activities - Stable intake of the medication considered not interfering with the study
Exclusion criteria for CN
- History of significant neurological diseases or head trauma leading to structural brain abnormalities

Table 4.4: Information on inclusion and exclusion criteria for normal controls (CN), i.e. healthy subjects, contained in the ADNI dataset. Geriatric Depression Scale is a scale for self-reporting of the levels of depression in adults. Hachinski Ischemic Score is a scale for differentiating different types of dementia or its absence. DSM-IV is a tool for classification of mental disorders.

Preparing the HBAPS dataset for machine learning purposes

Machine learning models are defined by a set of parameters which control the models' complexity. The models with higher complexity have higher discriminating power, but are also more susceptible to overfitting. Overfitting occurs when the model does not generalise well, i.e. the model performs well on training samples, but its performance on unknown samples is much worse compared to training. During the training in order to find the set of model parameters, which is balanced with respect to achieving the best possible performance and not overfitting, the dataset is split into training, validation and testing data. Training data is used to learn model parameters and validation data is used to assess its performance in order to avoid overfitting during training. Validation set is used in the process of training to measure model's performance. Model's accuracy can be assessed as

validation samples originate from the same dataset as training data due to initial split, but their labels are not known to a model. Therefore, predictions on the validation dataset can be used to assess model's performance. As the result of training process, optimal set of model's parameters is chosen corresponding to the lowest validation error based on the errors obtained on the validation data. Latter selection of the optimal model is called model selection. Testing data is used to measure optimal model's performance after training. Testing data can originate from original dataset and be separated at the splitting stage or be completely independent from the original dataset [211].

HBAPS dataset is split into training, validation and test data in the $80\% - 10\% - 10\%$ proportion. The same data split is also used in a number of similar published experiments, e.g. Cole et al [35], Feng et al [39] and Ren et al [212]. This proportion is chosen due to complexity and depth of the models used further in this thesis as described in Chapter 5. HBAPS dataset is split using stratified strategy. In case of stratified split the $80\% - 10\% - 10\%$ proportion for training, validation and testing data is maintained for each label. For example, HBAPS dataset contains 147 samples labelled to be 18 years old and with this label there are 116, 15, 16 samples in the training, validation and testing data respectively. The distributions of samples per label for the split used are shown in Figure 4.5. The same data split of the HBAPS dataset is used in all experiments in this thesis involving this dataset.

The HBAPS dataset - testing data

Further in this thesis two representations of the testing data of the HBAPS dataset are used - whole test data and test data only consisting of independent subjects, i.e. containing only the samples originating from the subjects not presented in the training data. This is motivated by the fact that although repeat scans acquired at different times from the same subject are not exactly the same, they are still more similar to each other than the scans acquired from different subjects. Therefore, excluding repeat scans from the HBAPS test data is considered.

Whole HBAPS test dataset consists of 1071 subjects (male/female = 480/526, mean age = 41.0 ± 22.4 years, age range 17 – 92 years). In order to increase the overall size of the dataset, multiple scans from the same subject were considered as unrelated samples resulting in 1140 scans (male/female =360/411, mean age = 42.0 ± 23.0 years). HBAPS test data only consisting of independent subjects contains 836 scans (mean age = $37.9.0 \pm 20.3$

years). The subjects belonging to the for2017 dataset do not have information on subjects' sex. The distribution of samples over the age range for both whole HBAPS test data and HBAPS test data including only independent samples are shown in Figure 4.6.

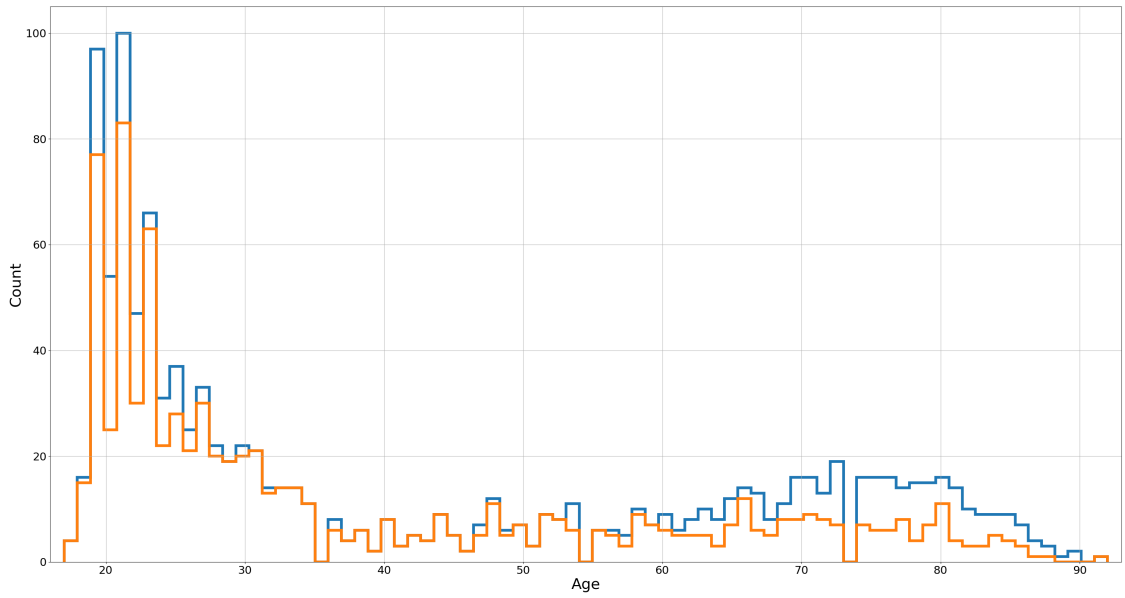

Figure 4.6: The age distribution of samples in whole HBAPS test data ($N = 1140$, blue histogram) and HBAPS test data including only independent samples ($N = 836$, orange histogram) over the age range of 17 – 92 years.

4.2.2 Institute of Psychiatry, Psychology & Neuroscience (IoPPN) data

The IoPPN dataset contains both healthy and diseased subjects. The samples in the IoPPN dataset are also acquired using a number different imaging sequences including T1- and T2-weighted imaging, Diffusion-weighted Imaging (DWI), Susceptibility-weighted Imaging (SWI) and Fluid-Attenuated Inversion Recovery (FLAIR) MRI. The sequences are described in Section 2.1. These sequences result in different appearance of the same type of tissues in MRI scans. Since in this thesis the models for brain age prediction described in Chapter 5 are trained using the HBAPS dataset consisting of T1-weighted samples collected from healthy subjects, the two subsets of the IoPPN data are used further in this thesis - only samples acquired with T1-weighted sequences from healthy and diseased subjects respectively.

All scanners used for data collection were produced by General Electric – 2 were SIGNA HDx 1.5T scanners and 2 were MR750 3.0T scanners. Information on the consent of participants and data usage agreement is as stated for the HBAPS dataset.

The subset of the IoPPN dataset containing only healthy subjects imaged using T1-weighted sequences consists of 1347 subjects (male/female = 751/596, mean age = 28.2 ± 10.3, age range 17 – 80 years). For augmenting the data multiple scans from the same subject were considered as unrelated samples resulting in 1799 scans (male/female = 995/804, mean age = 27.0 ± 9.7, age range 17 – 80 years). The subset of the IoPPN dataset containing only diseased subjects consists of 1887 subjects (male/female = 920/967, mean age = 48.7 ± 19.6, age range 17 – 96 years). The same approach for sample augmentation results in 1979 scans (male/female = 966/1013, mean age = 48.6 ± 19.6, age range 17 – 96 years) for the subset of diseased subjects who were also imaged using T1-weighted sequences. The distribution of samples over the age range for both subsets of the IoPPN dataset are shown in Figure 4.7.

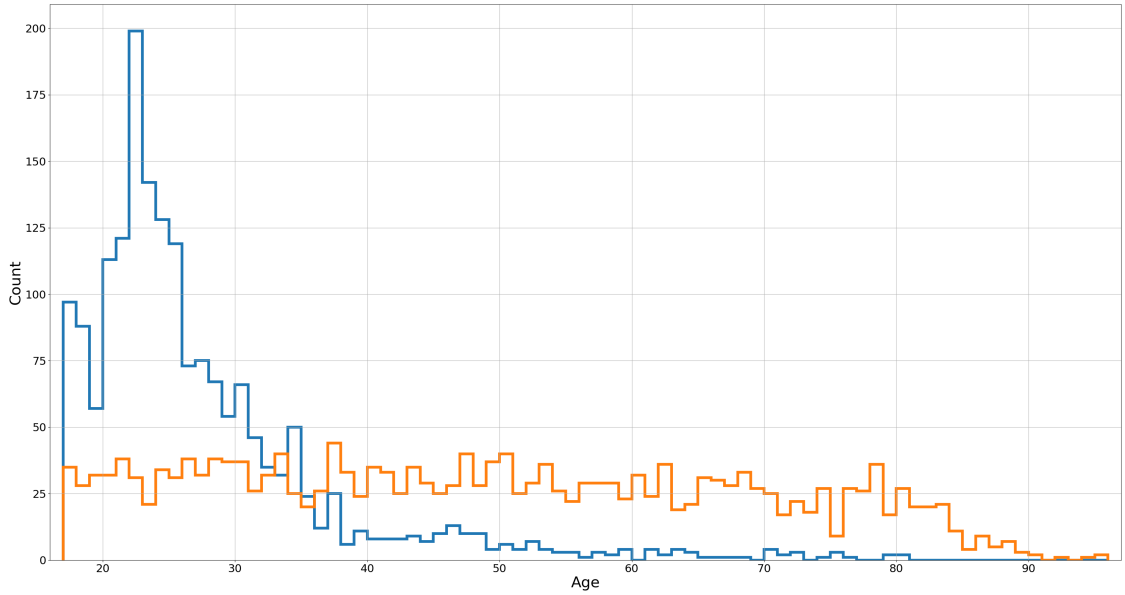

Figure 4.7: The age distribution of samples in the subset of IoPPN data consisting of healthy subjects ($N = 1799$, blue histogram) and the subset of IoPPN data consisting of diseased subjects ($N = 1979$, orange histogram) over the age range of 17 – 96 years. Blue corresponds to the subset of only healthy subjects and green to the subset of samples originating from only diseased subjects.

4.2.3 Preprocessing of MRI data

Raw T1-weighted MRI scans underwent rigid transformation and resampling onto the MNI152 template using FSL FLIRT software [213] as described in Section 2.2. After preprocessing all samples in the dataset have the same orientation in three dimensions and are of the same size - $196 \times 232 \times 188$.

Chapter 5

Deep ordinal regression on structural MRI data for brain age prediction

5.1 Main contributions

In this chapter it is shown that a Convolutional Neural Network (CNN) with ordinal regression as described in Section 2.3.3 with end-to-end learning of imaging features through convolutional filters can be used effectively for the task of brain age prediction on Magnetic Resonance Imaging (MRI) scans with minimal preprocessing. In order to analyse ordinal regression model on the brain age prediction task we compare it to the model based on metric regression. Ordinal CNN model performance is studied in light of following aspects:

- the model accuracy comparable to the results published in literature can be achieved with minimal preprocessing of input data;
- the model accuracy is not affected by the fact that the data comes from a number of independent sources as described in Section 4.2.1;
- the model accuracy is not affected by non-homogeneous distribution of training data over age range.

5.2 Introduction

It was previously shown that human brain age can be predicted from structural MRI data [212, 35, 214, 215, 216, 24, 39, 38, 13, 23, 217, 218, 14, 219, 220, 221, 183, 36, 40, 222, 223]. This information can be used to calculate the *brain age gap* – the difference

between subject's chronological age and predicted brain age. There are two main reasons for calculating human brain age from MRI data and measuring brain age gap. Firstly, studying brain age gap helps with understanding healthy brain ageing trajectory over human lifetime. Secondly, brain age gap is a promising biomarker for a number of disesses including Alzheimer's disease [224], epilepsy [225] and HIV [226].

All existing approaches for brain age prediction can be divided into two groups – the ones that use Deep Learning (DL) approaches and the others that not. Here only previous work using T1-weighted MRI data is presented. Previous works not using the DL methods are summarised in Table 5.1. While some authors [24, 219, 220, 221] applied non-DL-based machine learning methods such as Support Vector Machine (SVM) or Relevance Vector Regression (RVR), Beheshti et al [222] used MRI data segmented into patches with cortical labels as an input to a grading procedure to determine subject's brain age. Beheshti et al derived brain age from cortical age metric measured by the grading procedure. Cortical age metric was calculated by measuring similarity between the cortical labels of a test sample and known cortical labels of a sample belonging to the training set. Latter was calculated using a machine learning-based method. As it can be seen from Table 5.1 while most of the published works using non-DL-based methods used the datasets with age range covering adulthood, the datasets cannot be considered to represent population-wide ageing effects due to size.

In Table 5.2 DL-based published works are presented. A number of works presented in Table 5.2 [13, 14, 36, 40, 223] has achieved a Mean Absolute Error (MAE) comparable to other works in the field, but due to the size of the dataset used the results cannot be considered to study population-wide ageing effects. Due to vast anatomical variability and variations in ageing trajectory over human lifetime present in the population, the size of dataset used for training is a crucial aspect for both developing a clinical biomarker for neurodegenerative conditions and understanding brain ageing. Both these research aims also require the results to be interpretable, i.e. the model's output should be explainable in terms of input data features. For the model output to be interpretable data should be input into model with the fewest possible preprocessing. This would allow to connect the output directly to anatomical features of a subject. The model's architecture should also allow for the information inside it to be interpretable and visualisable. Ren et al [212] used a model allowing for interpretability, but used GM maps as an input. In this case the interpretation would only allow to assess GM degradation as the result of ageing.

Author, year	Citation	MAE	Dataset			Preprocessing	Prediction model
			N		Age range		
			Training	Testing			
Franke et al., 2010	[24]	5.0	410	137	19-86	Bias-field correction, spatial normalisation, segmentation into Grey Matter (GM), White Matter (WM), and Cerebrospinal Fluid (CSF) (correction for partial volume effects); GM maps were used.	RVR
Fujimoto et al., 2016	[181]	3.59	600	499	20-80	Normalised, modulated, segmentation into GM, WM and CSF, parcellated into local regions of each type.	RVR
Aycheh et al., 2018	[220]	4.05	1 895	810	45-91	Volumetric segmentation, cortex parcellation of 148 Regions of Interest (ROI), constructing the outer and inner cortical surface meshes.	Sparse Group Lasso (SGL), Gaussian Processes (GP)
Lancaster et al., 2018	[221]	5.08	2 003	648	16-90	GM segmentation, normalisation, non-linear registration on MNI152, resampling, modulation and spatial smoothing.	SVM
Beheshti et al., 2019	[222]	1.66	100	78	19-61	Segmentation to produce cortical labels, linear registration of the labels onto MNI305 pseudo-Talairach space.	Patch-based grading procedure

Table 5.1: Previous works published on brain age prediction of MRI data not using DL-based approaches. The Montreal Neurosciences Institute (MNI) atlases are average atlases based on 152 (MNI152) or 305 (MNI305) subjects.

Popescu et al [38] used a deep GP model allowing for a voxel-wise brain age prediction and, therefore, interpretation, but the input data was segmented into GM and WM and this limited the interpretation. Bermudez et al [183] achieved MAE of 5.0 years on the dataset with minimal preprocessing as described in Section 4.2.3. This may be due to the fact that the dataset included subjects as young as 4 years old. As human brain of such young subjects is still developing as opposed to ageing in subjects of age 94, which were also included in the dataset, this may have affected CNN's training process. Peng et al [227] achieved MAE of 2.71 years, but the dataset did not include subjects younger than 44 years of age. This may have potential implications as previous research has shown that first signs of brain ageing occur after 30 years of age.

All models shown in Tables 5.1 and 5.2 are metric regression models. Human brain ageing, however, is an ordinal process, i.e. brain can only get older with time. Brain ageing is also accumulative process and features related to every age does not appear on certain point during human lifetime but develop continuously through it. Therefore, using a model with ordinal regression may both improve model's interpretability and generalisability.

5.3 CNN for brain age prediction

The model for brain age prediction consists of convolutional and regression components. The convolutional part takes as an input structural MRI scans of size $196 \times 188 \times 232$ voxels. In practice, the Squeeze-and-Excitation Network (SENet) was adopted, as described in Section 2.3.2. The SENet architecture was chosen as the analysis of 3D MRI data presents two main challenges. Firstly, the statistical contrast of age-related features in the MRI data is likely to be low for the subjects of similar ages. Secondly, the analysis of 3D data imposes computational and memory demands compared to the analysis of 2D data. Ren et al [3] show on a number of computer vision tasks that the SENet uses statistics extracted from the data inside a network for improving performance in a computationally cheap manner.

In the SENet implementation used SE blocks are incorporated into the ResNet-50 network [14]. The basic principles behind the ResNet were presented in Section 2.3.2. The convolutional component of the models consisted of 50 layers in which every convolutional layer was followed by a batch normalization before a non-linear activation. In the networks used in the experiments SE-Res block defined in Figure 5.1 was used to form SE-ResNet defined in Table 5.4. The SE-Res block takes an input, $\boldsymbol{X}$, of size $N \times C \times H \times W \times Z$,

Article	Citation	MAE	Dataset			Preprocessing	Prediction model
			N		Age range		
			Training	Testing			
Ren et al, 2019	[212]	3.33	8 840	1 250	45-79	Cropping, bias-field correction, registration onto MNI152, brain extraction, tissue-type segmentation; GM maps were used.	CNN
Popescu et al, 2018	[38]	3.85	2001	653	18-90	Segmentation into GM and WM, non-linear registration using DARTEL algorithm onto MNI152.	Deep Gaussian Processes (DGP)
Amoroso et al., 2019	[13]	4.7	484	-*	7-80	Intensity normalisation, skull-stripping, non-linear registration.	Multiplex network
Richard et al., 2020	[14]	5.5	628	68	18-87	Brain extraction, intensity normalization, automated tissue segmentation, generation of white and pial surfaces, extraction of mean cortical thickness, area and volumes from 180 ROIs per hemisphere.	Xgboost
Sturmfels et al., 2018	[223]	1.43	524	100	8-21	Bias field correction, brain extraction, and spatially normalizing to MNI152 template space.	CCNN
Cole et al., 2017	[228]	4.65	1601	200	18-90	Rigid transformation (6 Degrees of Freedom (DOFs)) onto MNI152 template and resampling.	CNN
Huang et al., 2017	[40]	4.00	600	499	20-80	Aligned and cropped with ICBM152 template.	CNN
Qi et al., 2018	[36]	4.28	491	90	19-86	Realignement and normalisation. Training data is augumented to 6000 by applying shift, flipping and rotation.	CNN
Bermudez et al., 2019	[183]	5.00	4038	1010	4-94	Affine transformation onto MNI152 template and resampling.	CNN
Peng et al., 2019	[227]	2.71	12 949	1036	44-80	Affine transformation onto MNI152 template and resampling.	CNN

Table 5.2: Previous works published on brain age prediction of MRI data not using DL-based approaches. The ICBM152 atlas is average atlas based on 152 subjects. *Training-testing split was not stated by [13]. Richard et al [14] used parallel tree boosting (Xgboost) for predicting brain age.

where N is a batch size, C is a number of channels, H, W and Z are height, width and depth of a sample respectively. $\boldsymbol{X}$ is put through a set of layers as described in Table 5.3 Output of this branch is then multiplied with $\boldsymbol{X}$ to form block output $\widetilde{\boldsymbol{X}}$.

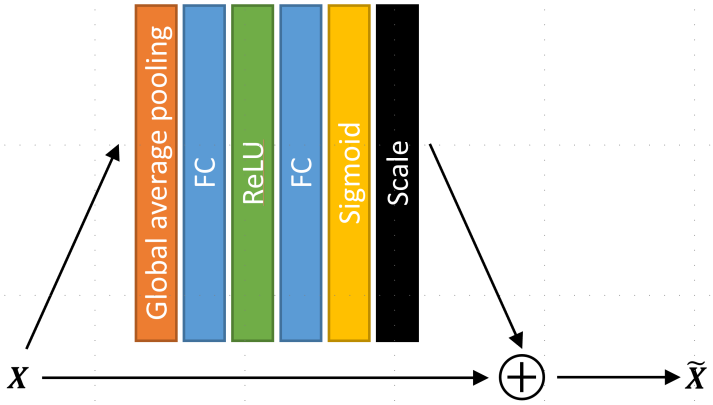

Figure 5.1: Schematic representation of a SE-Res block. Input, $\boldsymbol{X}$, is put through a set of layers and is then scaled to be added onto initial $\boldsymbol{X}$ to produce block's output $\widetilde{\boldsymbol{X}}$. Fully-connected layer is denoted by FC, a rectified linear unit – ReLU. (Based on Figure 3 in [3]).

Layer	Output size
Global average pooling	$N \times C \times 1 \times 1 \times 1$
FC	$N \times C_1 \times H \times 1 \times 1$
ReLU	$N \times C_1 \times H \times 1 \times 1$
FC	$N \times C \times H \times 1 \times 1$
ReLU	$N \times C \times 1 \times 1 \times 1$
Sigmoid	$N \times C \times 1 \times 1 \times 1$
Scale	$N \times C \times H \times W \times Z$

Table 5.3: The layers sequence applied to an input $\boldsymbol{X}$ in SE-Res block. After global average pooling is applied a tensor is shaped as $N \times C$. After sigmoid is applied a tensor is first reshaped as $N \times C \times 1 \times 1 \times 1$ and then expanded as $N \times C \times H \times W \times Z$. The number of channels C_1 is calculated as C divided by a reduction factor set to 16 and rounding down to closest whole number.

Output size	Layer
$\frac{H}{2} \times \frac{W}{2} \times \frac{Z}{2}$	*Conv* , 7×7, 64, stride 2
$\frac{H}{4} \times \frac{W}{4} \times \frac{Z}{4}$	*Max-pooling* , 3×3, stride 2
	$\begin{bmatrix} Conv, 1 \times 1, 64 \\ Conv, 3 \times 3, 64 \\ Conv, 1 \times 1, 256 \\ SE - Res, [16, 256] \end{bmatrix} \times 3$
$\frac{H}{8} \times \frac{W}{8} \times \frac{Z}{8}$	$\begin{bmatrix} Conv, 1 \times 1, 128 \\ Conv, 3 \times 3, 128 \\ Conv, 1 \times 1, 512 \\ SE - Res, [32, 512] \end{bmatrix} \times 3$
$\frac{H}{16} \times \frac{W}{16} \times \frac{Z}{16}$	$\begin{bmatrix} Conv, 1 \times 1, 256 \\ Conv, 3 \times 3, 256 \\ Conv, 1 \times 1, 1024 \\ SE - Res, [64, 1024] \end{bmatrix} \times 3$
$\frac{H}{32} \times \frac{W}{32} \times \frac{Z}{32}$	$\begin{bmatrix} Conv, 1 \times 1, 512 \\ Conv, 3 \times 3, 512 \\ Conv, 1 \times 1, 2048 \\ SE - Res, [128, 2048] \end{bmatrix} \times 3$

Model	X	*Global average pooling*
SENet-MR	1×1	*FC*, $[X, 1]$
SENet-OR	$2 \times (K-1)$	*FC*, $[X, 2 \times (K-1)]$

Table 5.4: SE-ResNet architecture. Output size for layer blocks are shown in terms of initial input dimensions (H – sample height, W – width, Z - depth). *Conv* denotes 3D covolutional layer and parameters are given in order – kernel size, output channel and stride. *Max-pooling* parameters are kernel size and stride. Repeated blocks are denoted by multiplication. SE-Res denotes a block defined in Figure 5.1. Last fully-connected layers (*FC*) are different for the SENet-MR and SENet-OR architectures. The SENet-MR architecture output is age (1×1). The SENet-OR architecture outputs probabilities for $K-1$ binary classifiers by applying a logistic regression function (defined in Equation 2.4) to the output of these binary classifiers.

5.3.1 Metric regression

The model with metric regression (SENet-MR) is implemented by attaching a fully-connected layer after a convolutional part of the model, to output brain age prediction. In this implementation the architecture is combined with the MAE loss function:

$$\alpha_{MAE} = \frac{1}{N} \sum_{n=1}^{N} \left\lfloor y_n - y'_n \right\rfloor , \tag{5.1}$$

where y_n and y'_n are the target and predicted ages respectively, and N is the number of samples in a batch.

5.3.2 Ordinal regression

The model with ordinal regression (SENet-OR) was implemented as described in Section 2.3.2. We trained a model on the Healthy Brain Ageing from Public Sources (HBAPS) dataset described in Section 4.2.1, which has samples aged from 17 to 96 years old. Therefore, each sample belonging to the data $x_i \in X$ has a corresponding outcome space $y_i \in Y = \{r_1, r_2, \ldots, r_K\}$, where $r_1 = 17$ and $r_K = 96$. Therefore, 80 age classes are represented by 79 binary classification sub-problems. In testing stage Equation 2.24 is modified to reflect this range of samples as

$$h\left(x'\right) = r_q$$

$$q = 17 + \sum_{k=1}^{79} f_k(x'), \tag{5.2}$$

where

- $h(x')$ is a predicted rank of a sample x',
- $f_k\left(x'\right) \in \{0, 1\}$ is the classification result of the k^{th} binary classifier for a sample x'.

5.3.3 Classifier implementation

Both classifiers were trained with the same parameters. The network's weights were initialised using the approach described by He et al [87]. Optimisation was performed using Stochastic Gradient Descent (SGD) with momentum 0.9 and a mini-batch size of 4 due to Graphics Processing Unit (GPU) memory constraints. The initial learning rate was set to

0.1 and decreased by a factor of 10 every 66 epochs. The initial learning rate and its rate of decrease was set empirically by running "test" epochs making sure the model is training. Models were trained for 200 epochs from scratch. The epoch with the lowest validation MAE was used for testing. In the testing stage the input images were normalised as in the training.

5.4 Results

After training the models, they were tested on whole HBAPS testing data and the distribution of all predicted ages is shown in Figure 5.2 for both the SENet-MR and SENet-OR models. Performance of the models is further examined by considering theHBAPS test dataset with the subjects used in training data excluded. Details on HBAPS test data are presented in Section 4.2.1. The distribution of predictions for both models on independent test data is shown in Figure 5.3. MAE and the Pearson's correlation coefficient, R, of chronological and predicted ages for both models are presented in Table 5.5. In Table 5.5 models' performance is measured on whole HBAPS test dataset, HBAPS test data excluding repeat scans, i.e. the scans of subjects also included in the training data, and only repeated scans contained in the HBAPS test data.

Model	Test data	MAE [years]	R	RMSE [years$^{0.5}$]	Error variance [years]	σ [years$^{0.5}$]
SENet-MR	Whole	3.87	0.96	6.88	32.36	5.69
	Independent subjects	4.37	0.94	6.18	29.63	5.44
	Repeat scans	2.48	0.97	7.36	37.35	6.11
SENet-OR	Whole	3.62	0.97	5.91	21.79	4.67
	Independent subjects	4.10	0.95	2.02	23.65	4.86
	Repeat scans	2.31	0.99	1.52	14.35	3.79

Table 5.5: MAE and the Pearson's correlation coefficient, R, of chronological and predicted ages, Root Mean Square Error (RMSE), error variance and standard deviation on MAE, σ, for the SENet-MR and SENet-OR model predictions on whole HBAPS test dataset, on HBAPS test dataset containing only independent subjects, i.e. not used in HBAPS training data, and only repeat scans contained in whole HBAPS dataset.

The linear regression fit for each model's predictions is shown by the dashed dark blue line in Figures 5.2 and 5.3, along with the 95% intervals (dashed light blue-green lines) and the line of identity (in orange). Parameters of the linear fit models for both SENet-MR

and SENet-OR models are represented in Table 5.6 for both whole HBAPS test dataset and the HBAPS test dataset with repeated subjects excluded.

Model	Test data	a	95% confidence interval bounds on a		p_a	c	95% confidence interval bounds on c		p_c
			Lower	Upper			Lower	Upper	
SENet-MR	Whole	0.928	0.911	0.944	0.0005	3.493	2.687	4.298	0.0005
	Independent subjects	0.898	0.876	0.920	0.0005	4.177	3.218	5.137	0.0005
SENet-OR	Whole	0.951	0.936	0.965	0.001	1.174	0.481	1.867	0.0005
	Independent subjects	0.928	0.907	0.948	0.0005	1.747	0.867	2.628	0.0005

Table 5.6: Parameters of the linear regression fit into the distributions of predictions by the SENet-MR and SENet-OR models for whole HBAPS test dataset and the HBAPS test data with repeat scans excluded. Parameters of the linear regression fit models are represented by a and c such that $y_p = a \times y_t + c$, where y_p and y_t are predicted and true sample ages respectively. For both a and c values $p-$values are given corresponding to p_a and p_c along with 95% confidence interval lower and upper bounds.

For both SENet-MR and SENet-OR models MAE on only repeat scans from HBAPS test data is lower than MAE on whole HBAPS test dataset for which MAE is lower compared to MAE on HBAPS test data containing only scans independent from training data. This is expected result as other scans from the subjects contained in the repeat scans from HBAPS test data are also part of the HBAPS training data. Whole HBAPS test dataset also contains repeat scans of subjects included in HBAPS training data. As the result, in the testing stage the model is shown scans of the same subjects which are contained in the HBAPS training data and MAE is lower compared to the HBAPStest data containing only independent subjects. Even though for both whole HBAPS test dataset and the HBAPS dataset with excluded repeat scans the Pearson's correlation coefficients for the both models are very close, the SENet-OR model's linear regression fit and its 95% confidence interval are closer to the line of identity than the SENet-MR model's fit. Therefore, it can be concluded that while both models had similar overall performance, the SENet-OR model generalised to the data better.

As it can be seen in Table 5.5 for both whole HBAPS test dataset and HBAPS test data excluding repeated subjects, while the means of two error distributions from the SENet-MR and SENet-OR models' predictions are close, their variances, and therefore, standard deviations vary significantly. The SENet-OR model predictions are more concentrated around the mean. From Figures 5.2 and 5.3 it can be observed that SENet-OR performed

better on the youngest subjects (younger than 22 years old) and specifically for the samples with age of 17 years (only 4 of whom were present in the testing set). However, for the subjects older than 87 years, the SENet-MR model had better performance. This is further explored in Figures 5.4 and 5.5.

Figures 5.6 and 5.7 show the plots of prediction residuals by the SENet-MR and SENet-OR models on the whole HBAPS test data and the HBAPS test data including only independent subjects. Prediction residual is defined as a deviation of predicted age from a linear fit into the plot of the corresponding model predictions vs true age. Coefficients for describing the fits represented by blue lines are presented in Table 5.7.

Model	**Test data**	**Linear fit**		**Quadratic fit**		
		b_l	c_l	a_q	b_q	c_q
SENet-MR	Whole	-0.049	2.797	0.000	-0.049	2.797
	Independent subjects	0.000	0.000	0.001	-0.078	1.428
SENet-OR	Whole	0.000	0.000	0.001	-0.111	2.052
	Independent subjects	0.000	0.000	0.002	-0.169	3.118

Table 5.7: Parameters of the linear and quadratic regression fits into the distributions of prediction residuals by the SENet-MR and SENet-OR models for whole HBAPS test dataset and the HBAPS test data with repeat scans excluded. Parameters of the linear regression fit models are represented by b_l and c_l such that $y_r = b_l \times y_t + c_l$, where y_r and y_t are prediction residual and true sample age respectively. Parameters of the quadratic regression fit models are represented by a_q, b_q and c_q such that $y_r = a_q \times y_t^2 + b_q \times y_t + c_q$.

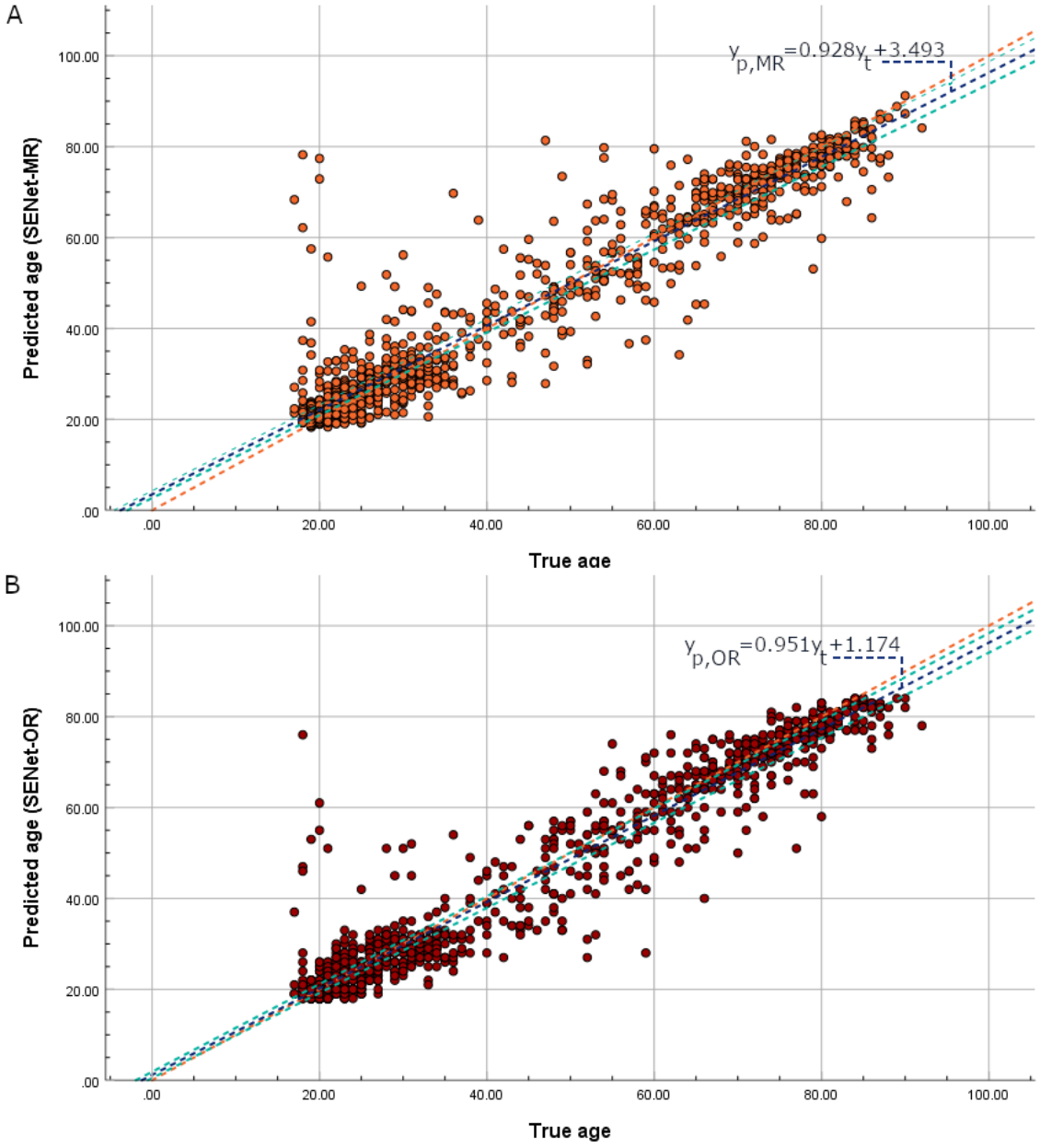

Figure 5.2: Distribution of predictions of the SENet-MR (A) and SENet-OR (B) models on whole HBAPS test dataset. The brain age predicted by the models on this test data (y-axis) is plotted against chronological age label (x-axis). Orange line denotes the $y = x$ function, blue – linear regression fit function, green – the lines representing the lower and upper bounds of the 95% confidence interval over linear regression fit parameters.

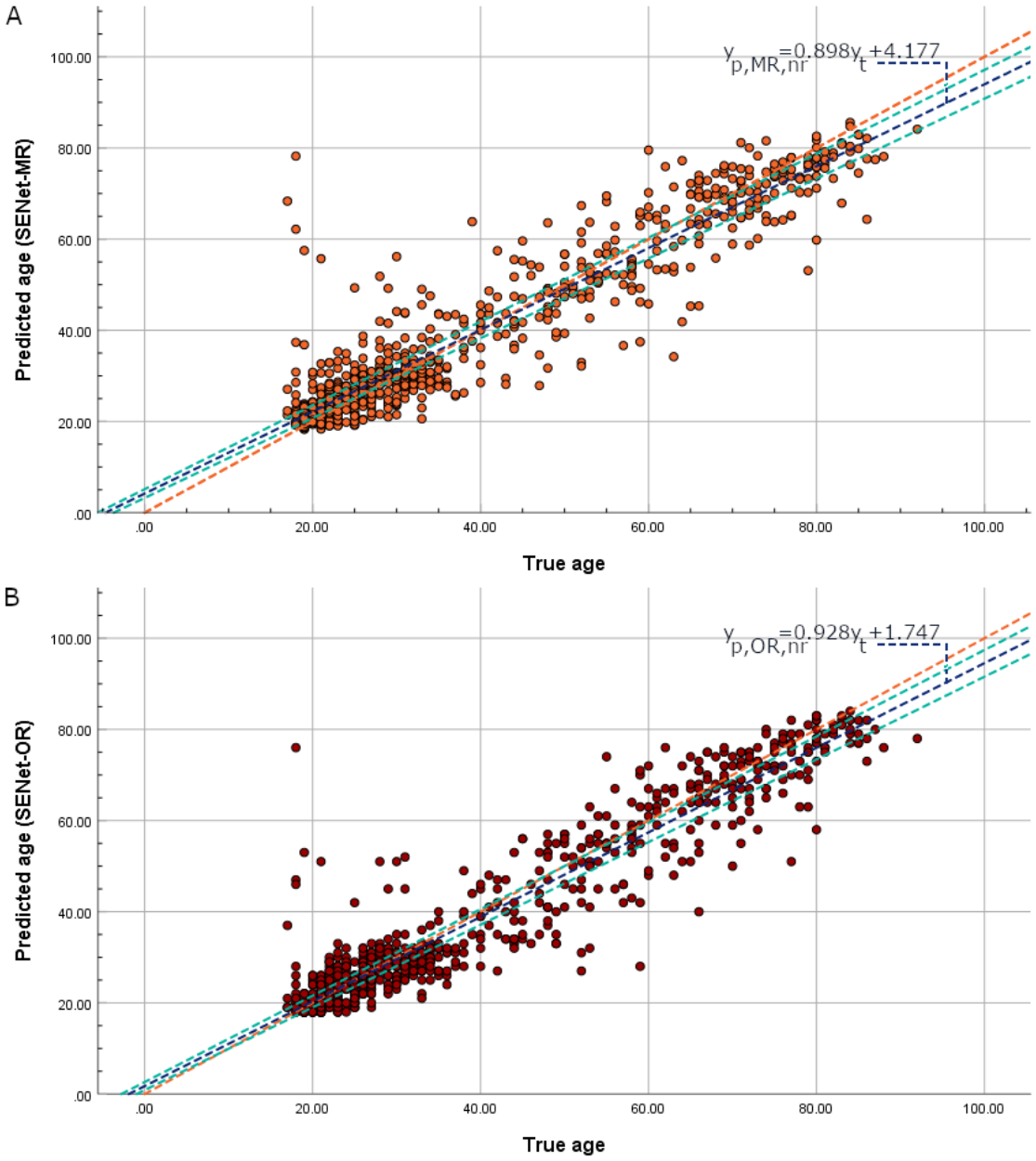

Figure 5.3: Distribution of predictions of the SENet-MR (A) and SENet-OR (B) models on HBAPS test dataset with excluded repeat scans. The brain age predicted by the models on this test data (y-axis) is plotted against chronological age label (x-axis). Orange line denotes the $y = x$ function, blue – linear regression fit function, green – the lines representing the lower and upper bounds of the 95% confidence interval over linear regression fit parameters.

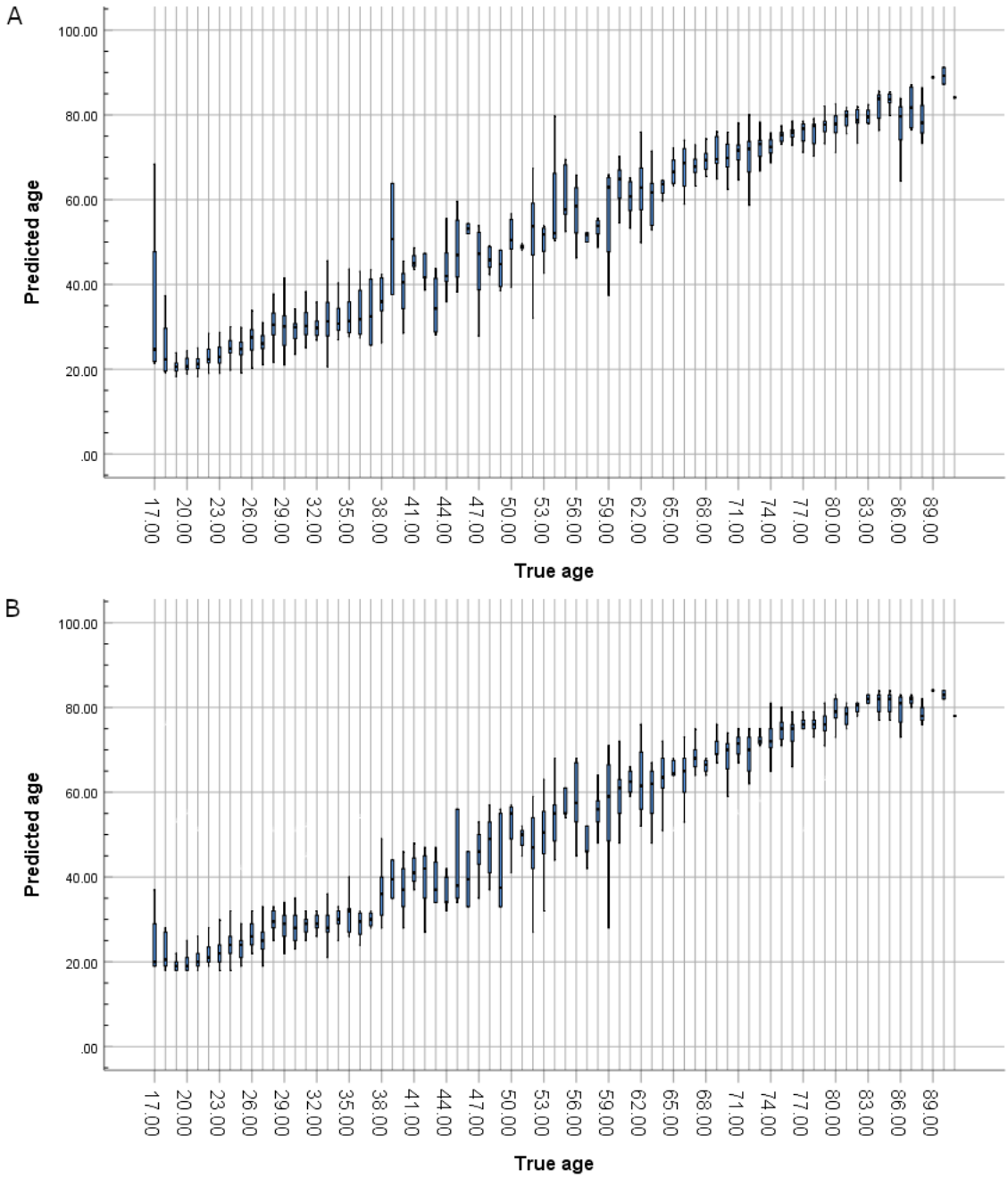

Figure 5.4: Box plot of the predictions by the SENet-MR (A) and SENet-OR (B) models with respect to the true age labels for the whole HBAPS test data.

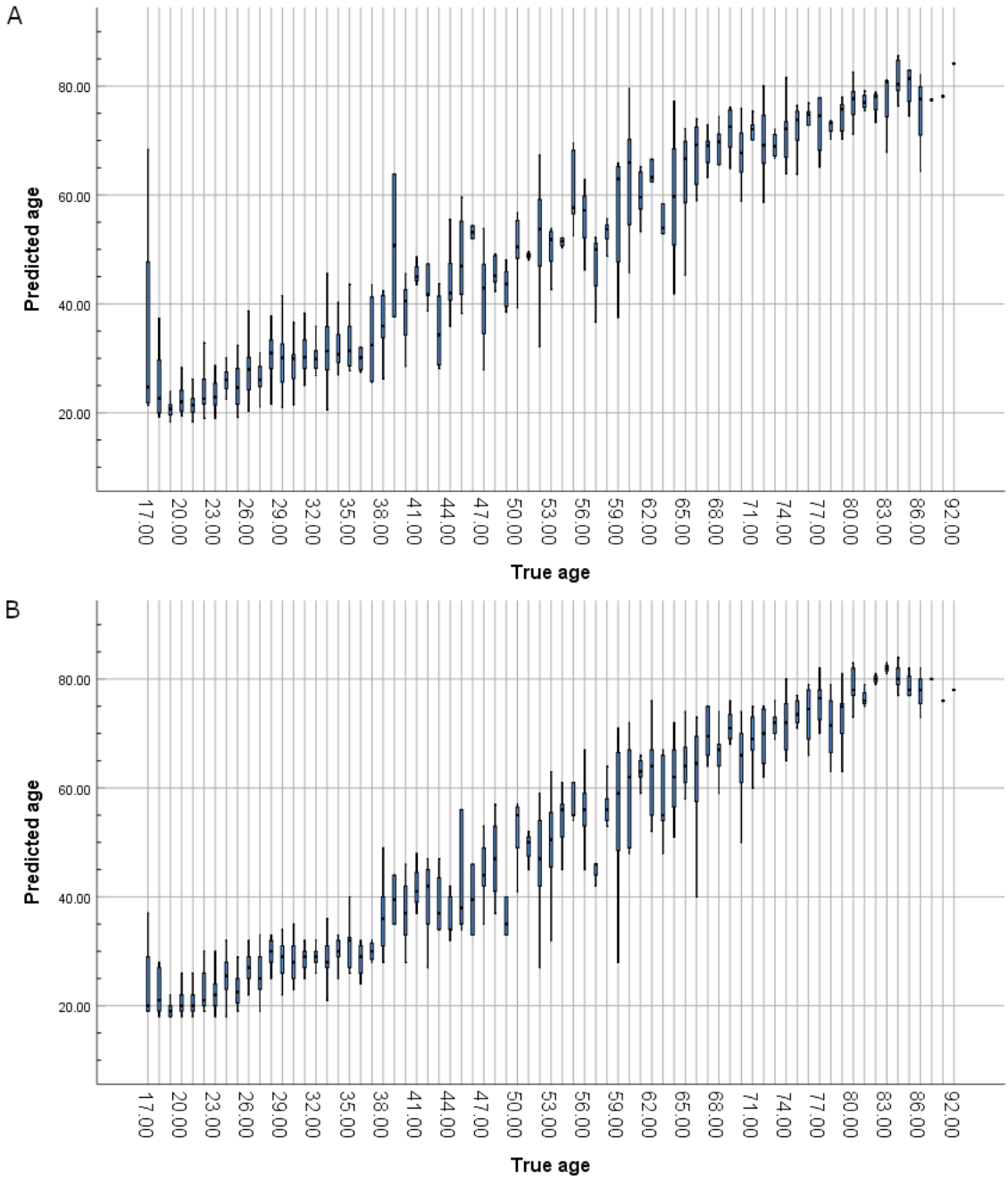

Figure 5.5: Box plot of the predictions by the SENet-MR (A) and SENet-OR (B) models with respect to the true age labels for the HBAPS test data including only independent subjects.

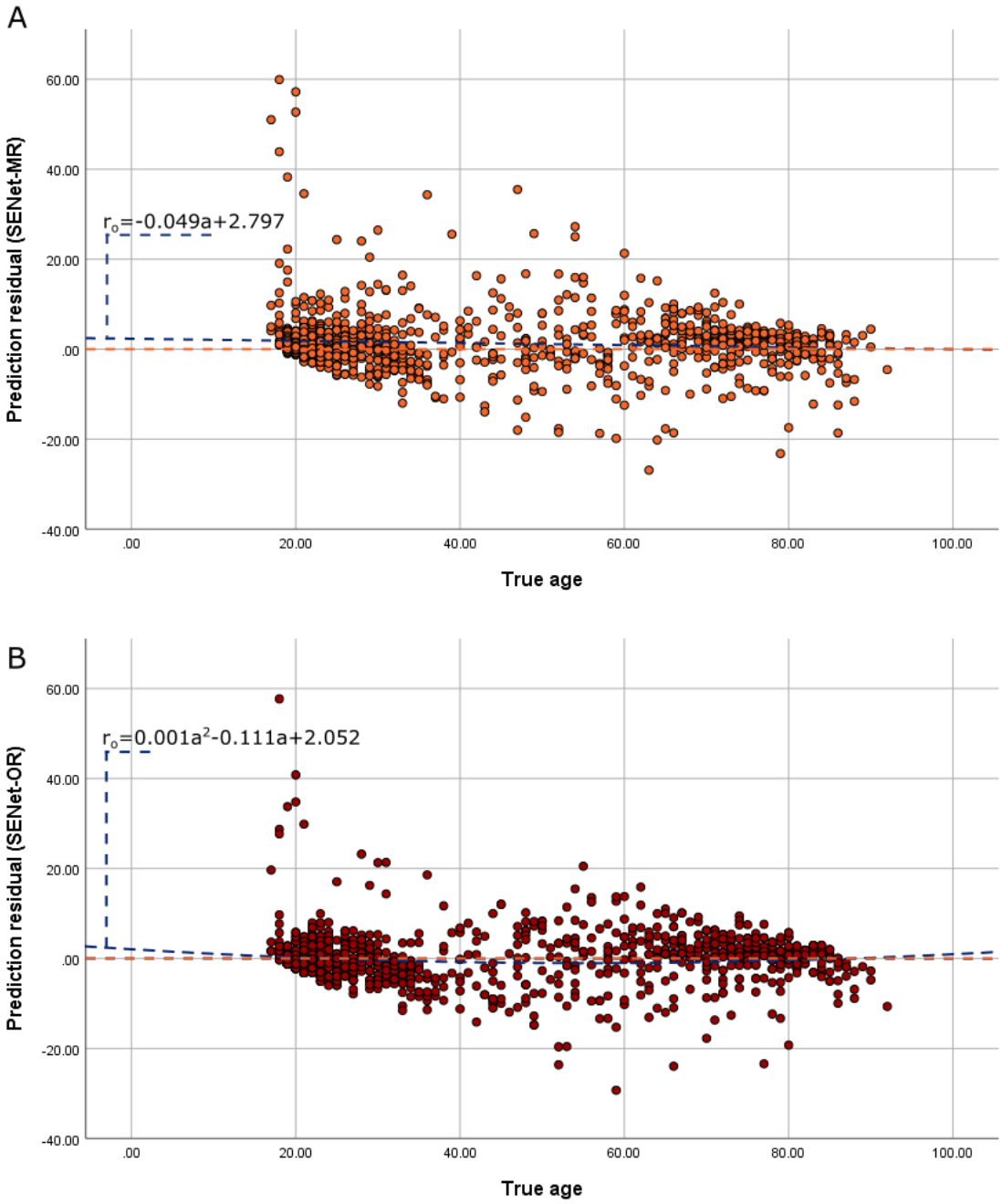

Figure 5.6: Distribution of the prediction residuals with respect to true age labels for the SENet-MR (A) and SENet-OR (B) predictions on the whole HBAPS test data. Prediction residual is defined as a deviation of predicted age from a linear fit into the plot of the corresponding model predictions with respect to true age. Orange line denotes the $y = 0$ function for reference and blue line corresponds to linear (A) and quadratic (B) fits into the plot of prediction residuals with respect to true age labels. Coefficients for describing the fits represented by blue lines are presented in Table 5.7. Linear fit is not shown for the SENet-OR prediction residuals (B) as it was close to the $y = 0$ function.

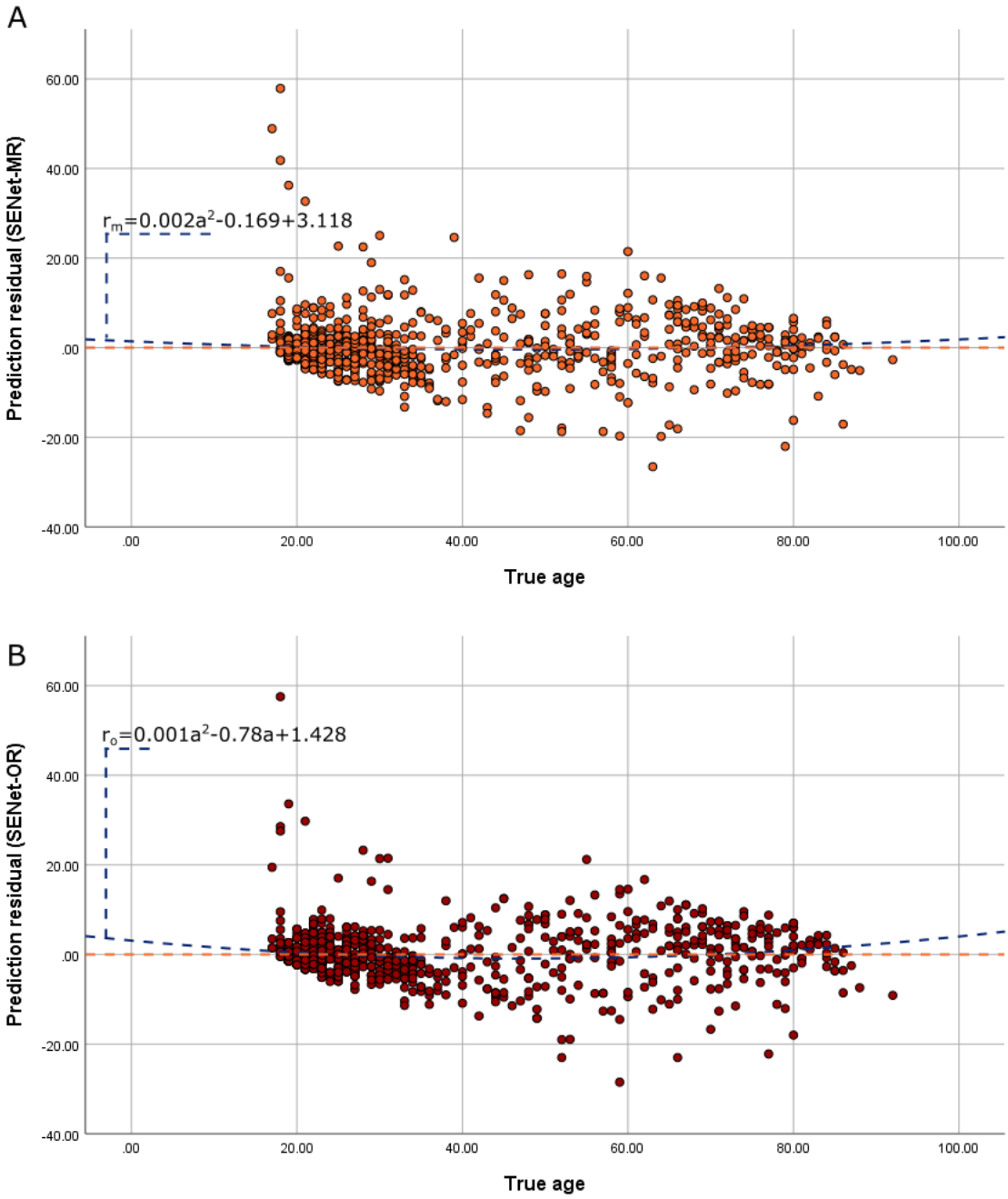

Figure 5.7: Distribution of the prediction residuals with respect to true age labels for the SENet-MR (A) and SENet-OR (B) predictions on the HBAPS test data including only independent subjects. Prediction residual is defined as a deviation of predicted age from a linear fit into the plot of the corresponding model predictions with respect to true age. Orange line denotes the $y = 0$ function for reference and blue line corresponds to quadratic fits into the plot of prediction residuals with respect to true age labels. Coefficients for describing the fits represented by blue lines are presented in Table 5.7. Linear fit is not shown on the graphs as it was close to the $y = 0$ function.

As training dataset was highly unbalanced (see Section 4.2.1) it was important to make sure that the models were able to make as accurate predictions as possible for all ages present in the dataset. In Figure 5.8 the average prediction error for each age in the dataset is plotted for each model, along with the corresponding number of samples in the training dataset. In Table 5.8 the correlation coefficient between the number of samples and average of mean errors per age label along with variances over the average of mean errors per age label are presented for both the SENet-MR and SENet-OR models and two test datasets - whole HBAPS test dataset and HBAPS test data excluding repeat scans. Table 5.8 also shows that the variances of the average of mean errors per age label are lower for the SENet-OR model predictions on both modifications of the HBAPS test data.

Model	**Test data**	**Correlation**	**Variance** [years]
SENet-MR	Whole	-0.36	9.15
	Independent subjects	-0.39	9.41
SENet-OR	Whole	-0.44	6.74
	Independent subjects	-0.41	6.65

Table 5.8: Correlation coefficient between the number of samples and average of mean errors per age label and variance over the average of mean errors per age label are presented for both the SENet-MR and SENet-OR models and two test datasets - whole HBAPS test dataset and HBAPS test data excluding repeat scans

Both the SENet-MR and SENet-OR models underperformed at the ends of the age range, i.e. on the subjects younger than 19 years of age and older than 85 years of age on both whole HBAPS test data and HBAPS test data with only independent subjects included. It can also be seen from Figure 5.8 that in these test datasets there is very little data for these parts of the age range – 20 and 18 samples in total are in the age ranges younger than 19 years and older than 85 years respectively. As the result the largest prediction outliers occur for the samples with the true age labels in range younger than 21 years of age. This can also be observed in Figures 5.2 and 5.3. Number of samples and average of mean errors per age label for the prediction distributions by SENet-MR and SENet-OR models for the two test datasets are summarised in Table 5.9. In whole HBAPS test data in the age range from 20 to 39 years of age the total number of samples is 2.4

and 2.6 times larger respectively compared to the age ranges of 40 to 69 years and 70 to 84 years of age. When repeat samples are excluded from the HBAPS test data the ratios of number of samples for these true age label ranges are 2.48 and 4.7 respectively. The ratios of average mean errors per age label of the SENet-MR and SENet-OR models for the two test datasets in the age ranges presented in Table 5.9 are different from the ratios of number of samples present in these age ranges. For instance, the ratio of the average mean errors per age label of the SENet-MR model on the HBAPS dataset not containing repeat samples on the age ranges of 20 to 39 and 40 to 69 years of age is 0.72. The ratio of the number of samples for the same age ranges is 2.48. From Table 5.9 it can be observed that the performance of SENet-MR and SENet-OR models is comparable on both whole HBAPS test data and HBAPS test data with repeat scans excluded. For both models the difference in average of mean errors per age label increased with increased difference in the number of samples for a given age range between whole HBAPS test data and HBAPS test data containing only independent samples.

Number of samples in age range	**Test data**		**Average of mean errors per age label [years]**			
			SENet-MR		**SENet-OR**	
	Whole	**Independent subjects**	**Whole**	**Independent subjects**	**Whole**	**Independent subjects**
20 to 39	550	452	4.61	4.62	3.96	3.98
40 to 69	232	182	5.99	6.42	5.91	6.44
70 to 84	214	96	3.04	4.82	3.13	4.49

Table 5.9: Number of samples and average of mean errors per age label for three age ranges - 20 to 39, 40 to 69 and 70 to 84 years of age. Number of samples is presented for two test datasets - whole HBAPS test data and HBAPS test data containing only independent subjects. Average of mean errors per age label is calculated using the distributions of predictions by the SENet-MR and SENet-OR models.

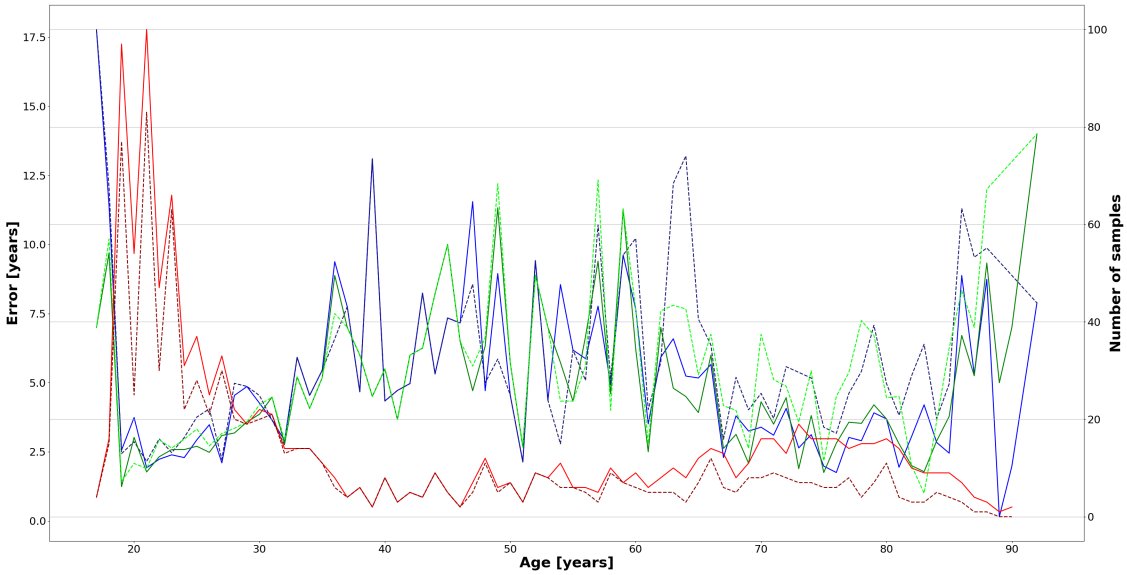

Figure 5.8: The number of samples per age in the whole HBAPS test data (continuous red line) and HBAPS test data including only independent subjects (dashed red line) (right y-axis) is plotted against chronological age label (x-axis). The average prediction error (left y-axis) for the SENet-MR (blue line) and SENet-OR (green line) measured on whole HBAPS test dataset (continuous line) and HBAPS test data with repeated scans excluded (dashed line) are plotted against chronological age label (x-axis).

In Tables 5.10 and 5.11 mean error is presented for all data sources in the whole HBAPS test dataset and HBAPS test data with repeat scans excluded for the predictions using the SENet-MR and SENet-OR models (details on the data sources are presented in Section 4.2.1) together with summary statistics. For both whole HBAPS test dataset and HBAPS test dataset including only independent subjects mean error did not correlate with the number of samples provided by each data source. This indicates that the models did not differ in their ability to work on a dataset including data from multiple sources. It can also be observed that averages over all mean errors for both models were similar. However, for whole HBAPS test dataset the SENet-OR model predictions resulted in lower variance over mean errors – 13.07 and 7.00 years respectively. For the HBAPS test dataset with repeat samples excluded variance over mean errors for the SENet-OR model was higher compared to the one for the SENet-MR model. The difference in variances was, however, smaller compared to whole HBAPS test dataset.

Dataset	Num. of samples	Model	
		SENet-MR	SENet-OR
1000 FCP	91	5.15	4.38
ABIDE I	24	14.88	11.88
ABIDE II	12	2.91	4.75
ADNI	152	1.66	2.08
AIBL	98	4.15	3.66
BNU1	10	1.94	2.20
BNU2	15	1.15	1.13
BNU Enhanced	20	2.21	1.55
BNU Eyes Open Eyes Closed	6	2.27	2.27
BNU Short TR	4	2.25	2.25
CamCAN	66	4.46	4.76
COBRE	6	5.29	8.67
DLBS	33	8.67	5.73
for2017	69	4.30	4.13
GSP	152	1.95	1.95
HCP	114	2.88	3.20
HNU	42	0.62	1.24
IBATRT	8	2.23	2.75
IPCAS	9	5.50	5.78
IXI	63	6.06	5.78
NKI-RS	41	6.74	6.46
PPMI	15	6.47	7.60
SWU	76	2.93	2.16
VPT	2	2.67	1.50
XHCUMS	12	14.33	5.92
Statistic			
R		-0.20	-0.21
Mean error [years]		4.55	4.15
Variance [years]		13.07	7.00

Table 5.10: Mean error was measured for all data sources in the whole HBAPS test dataset for the predictions using the SENet-MR and SENet-OR models. This is presented along with the number of samples coming from each data source. Summary statistics is presented at the bottom. R denotes Pearson's correlation coefficient between the number of samples from a data source and mean error. Details on the data sources are presented in Section 4.2.1.

Dataset	Num. of samples	Model	
		SENet-MR	SENet-OR
1000 FCP	91	5.15	4.38
ABIDE I	24	14.88	11.88
ABIDE II	12	2.91	4.75
AIBL	86	4.31	3.66
BNU1	5	1.94	3.20
BNU2	4	1.02	1.00
BNU Enhanced	20	2.21	1.55
BNU Eyes Open Eyes Closed	6	2.27	2.27
BNU Short TR	4	2.25	2.25
CamCAN	66	4.46	4.76
COBRE	6	5.29	2.75
DLBS	33	8.67	5.73
for2017	69	4.30	4.13
GSP	152	1.95	1.95
HCP	114	2.88	3.20
IBATRT	4	2.23	2.57
IPCAS	6	5.50	6.17
IXI	63	6.06	5.78
NKI-RS	41	6.74	6.46
PPMI	7	8.72	10.14
SWU	21	1.49	1.57
VPT	2	2.67	1.50
Statistic			
R		-0.04	-0.08
Mean error [years]		4.51	4.60
Variance [years]		9.96	10.70

Table 5.11: Mean error was measured for all data sources in the HBAPS test dataset with repeat scans excluded for the predictions using the SENet-MR and SENet-OR models. This is presented along with the number of samples coming from each data source. Summary statistics is presented at the bottom. R denotes Pearson's correlation coefficient between the number of samples from a data source and mean error. Details on the data sources are presented in Section 4.2.1.

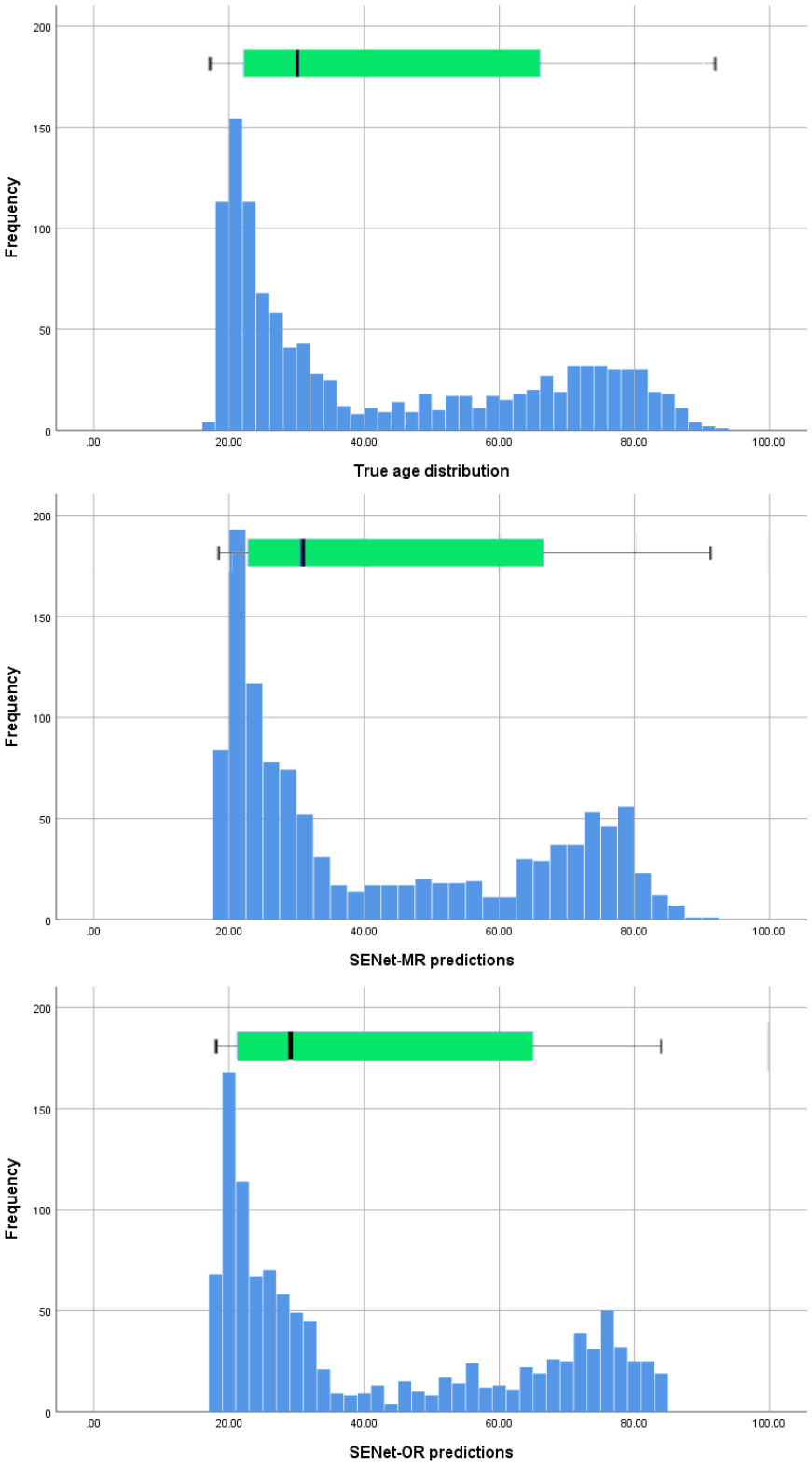

Figure 5.9: Box plot comparing predictions using the SENet-MR and SENet-OR models with the distribution of true age labels in whole HBAPS test dataset.

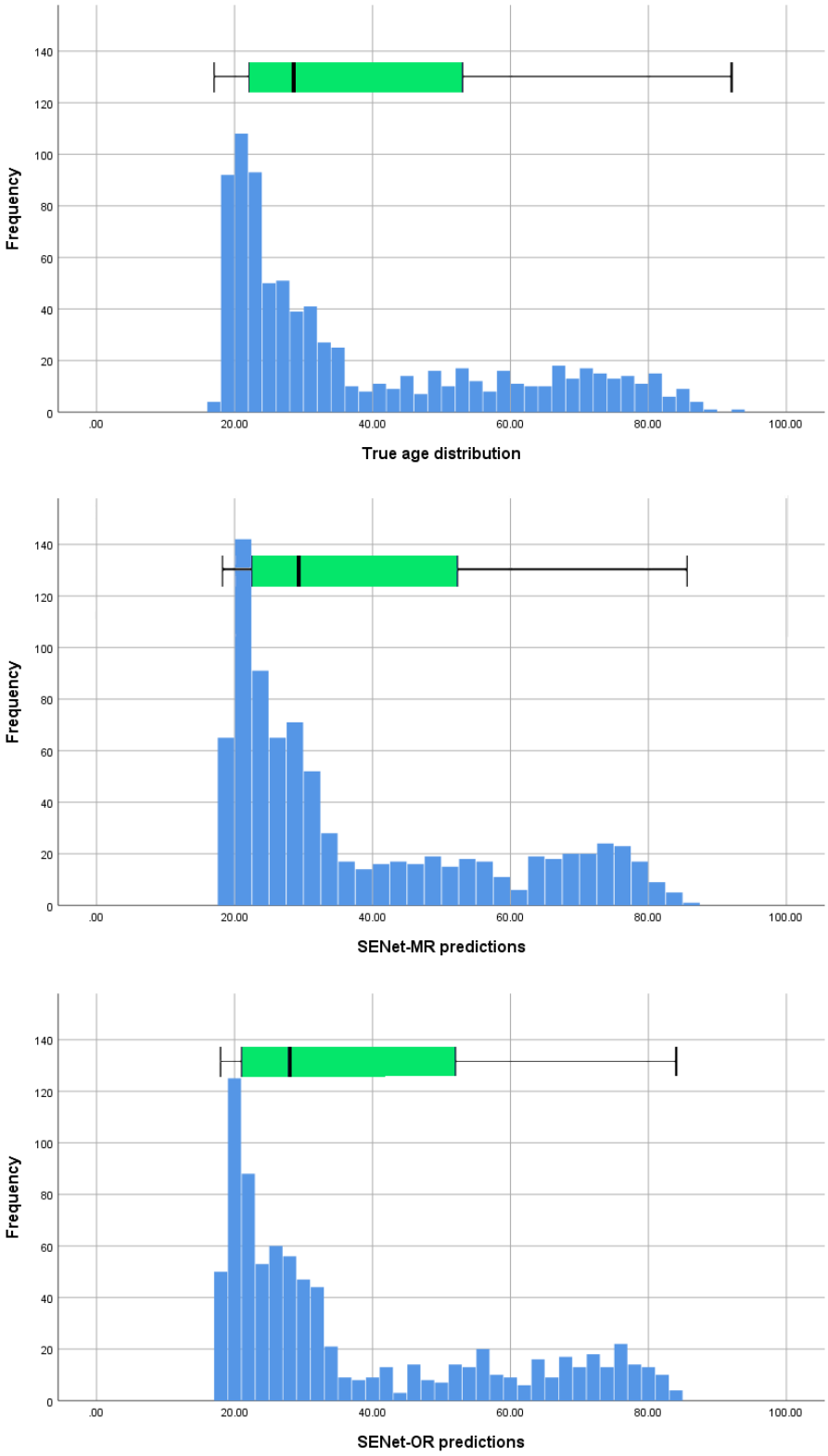

Figure 5.10: Box plot comparing predictions using the SENet-MR and SENet-OR models with the distribution of true age labels in the HBAPS test dataset with repeat scans excluded.

Box plots and histograms in Figures 5.9 and 5.10 further show that both models well captured true data distribution on both whole HBAPS test dataset and HBAPS test dataset containing only independent subjects, but the SENet-OR model underperformed in the older subjects compared to the SENet-MR model. For the subjects younger than 40 years of age both models predicted similar samples distribution comparable in shape with true age distribution. The SENet-MR model predictions distribution for the age range of 40 to 60 years of age was close to uniform, while the SENet-OR model predictions resembled true distribution more. The subjects from 78 to 80 years of age were predicted to be older by the SENet-MR model. For the subjects older then 80 years of age the SENet-OR model predicted all the subjects to be younger or of right age, but the resulting distribution from the SENet-MR model followed the true distribution closer. Descriptive statistics for the three distributions and two modifications of the test dataset are presented in Table 5.12. On both modifications of the test dataset the predictions distributions for both models captured well the distribution in testing dataset.

Statistic	Test data					
	Whole			Independent subjects		
	Distribution					
	Prediction		Actual	Prediction		Actual
	SENet-MR	SENet-OR		SENet-MR	SENet-OR	
Mean [years]	42.3	41.1	42.0	38.2	36.9	37.90
Median [years]	30.8	29.0	30.0	29.3	28.0	28.5
Minimum [years]	18.3	18.0	17.0	18.31	18.0	17.0
Maximum [years]	91.2	84	92.0	85.6	84	92.0
Range [years]	72.9	66.0	75.0	67.3	66.0	75.0
Interquartile range [years]	43.8	44.0	44.0	29.9	31.0	31.0
Variance [years$^{0.5}$]	501.0	512.3	530.8	375.9	391.2	411.1
Standard deviation [years]	22.4	22.6	23.0	19.4	19.8	20.28

Table 5.12: Descriptive statistics for the predicted distributions by the SENet-MR and SENet-OR models compared to the true distribution on whole HBAPS test data and HBAPS test data with repeat scans excluded.

Figure 5.11 shows the SENet-MR model predictions plotted against the SENet-OR predictions, for both whole HBAPS test dataset and HBAPS test dataset containing only independent subjects. The correlation coefficient for the two models' predictions was 0.973 and 0.962 for whole HBAPS test dataset and HBAPS test dataset excluding subjects from the training data respectively. Parameters of the linear fit models for both SENet-MR

Test data	a_o	95% confidence interval bounds on a_o		p_{a_o}	b_o	95% confidence interval bounds on b_o		p_{b_o}
		Lower	Upper			Lower	Upper	
Whole	0.985	0.971	0.998	0.0005	-0.689	-1.136	-0.042	0.037
Independent subjects	0.981	0.962	1.000	0.0005	-0.589	-1.401	0.222	0.155

Table 5.13: Parameters of the linear regression fit into the plot of the SENet-MR predictions (metric) vs the SENet-OR predictions (ordinal) for both whole HBAPS test dataset and HBAPS test dataset containing only independent subjects. Linear regression fit into the plot is represented by a function $y_o = a_o \times y_m - b_o$, where y_m and y_o are predictions be the SENet-MR and SENet-OR models respectively and a_o and b_o correspond to the slope and intercept of the function. For both a_o and b_o values $p-$values are given corresponding to p_{a_o} and p_{b_o} along with 95% confidence interval lower and upper bounds.

and SENet-OR models are represented in Table 5.6 for both versions of the HBAPS test data. It should be noted that the upper bound linear fit 95% confidence interval closely follows line of identity in Figure 5.11 for the whole HBAPS test dataset and its subset consisting of only independent subjects. Therefore, it can be concluded that predictions by the SENet-MR and SENet-OR models are in agreement with each other.

However, from Figures 5.8 and 5.9 it is also clear that there is a difference in performance by the SENet-MR and SENet-OR models for various age intervals. This can be further explored using the Bland-Altman plots in Figure 5.12. Mean difference over whole age range was 1.4 years for both versions of the HBAPS test data. Most of the samples with prediction difference outside the range of $\mu_d \pm 1.96\sigma_d$ (μ_d – mean difference, σ_d – standard deviation of difference) are the samples with mean predictions between 30 and 70 years, while ages below and above this are in general predicted relatively well. It should be noted that in the Bland-Altman plot for the whole HBAPS test dataset there are more samples, for which prediction difference is outside the range of $\mu_d \pm 1.96\sigma_d$. Therefore, it can be concluded that most of these samples are repeated scans.

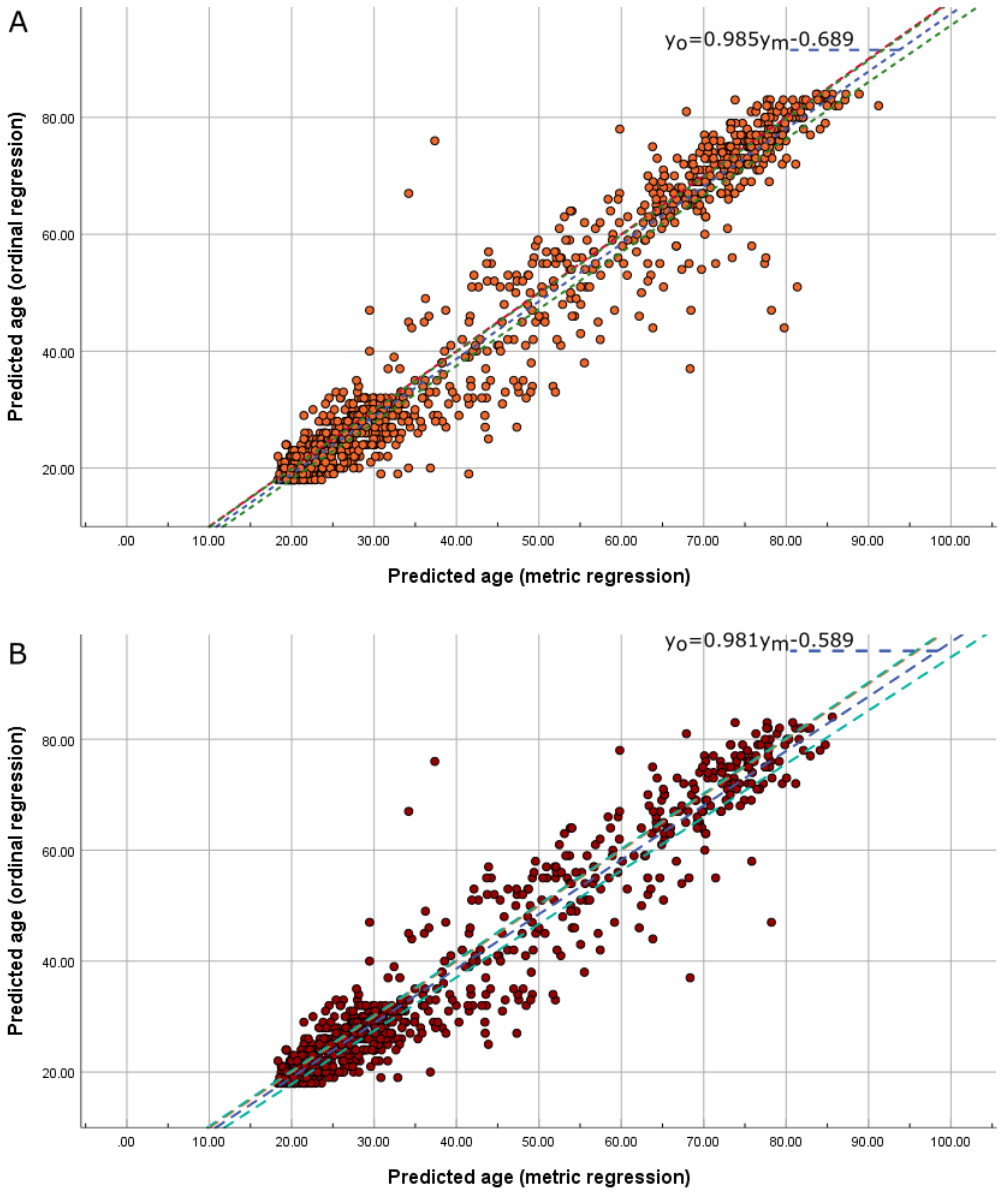

Figure 5.11: Plot of the SENet-MR predictions (metric) vs the SENet-OR predictions (ordinal) for both whole HBAPS test dataset (A) and HBAPS test dataset containing only independent subjects (B). Red line denotes the $y = x$ function, blue line represents linear regression fit into the plot represented by a function $y_o = a_o y_m - b_o$, where y_m and y_o are predictions be the SENet-MR and SENet-OR models respectively and a_o and b_o correspond to the slope and intercept of the function. Green lines represent lower and upper bound linear fit curves for the 95% confidence interval. For the whole HBAPS test dataset (A) linear fit takes the form $y_o = 0.984y_m - 0.689$ and for HBAPS test dataset containing only independent subjects (B) - $y_o = 0.981y_m - 0.589$. Lower and upper bound linear fit curves for the 95% confidence interval on the whole HBAPS test dataset are represented by $y_o = 0.971y_m - 1.336$ and $y_o = 0.998y_m - 0.042$ respectively. For the HBAPS test dataset containing only independent subjects the lower and upper bound linear fit curves are $y_o = 0.962y_m - 1.401$ and $y_o = y_m + 0.222$.

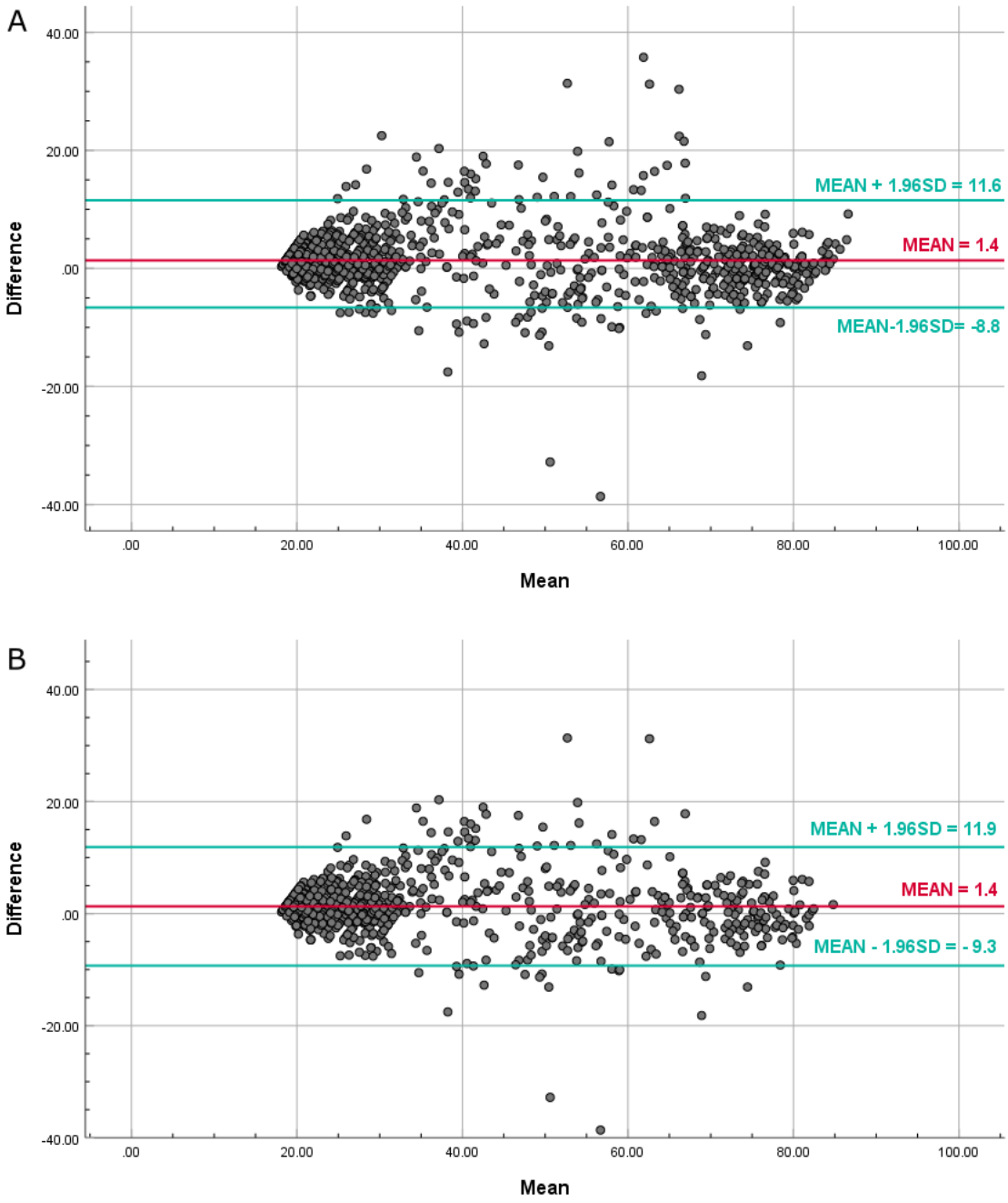

Figure 5.12: Bland-Altman plot for the predictions of the SENet-MR and SENet-OR models for the whole HBAPS test data (A) and HBAPS test data excluding repeated scans (B). Difference on the y-axis is calculated by subtracting the prediction by the SENet-MR model from the prediction by the SENet-OR model. Red line indicates mean difference over all predictions. Green lines indicate mean difference plus 1.96 standard deviations and minus 1.96 standard deviations respectively.

5.5 Discussion and conclusion

In this chapter we assessed a CNN with ordinal regression on the task of brain age prediction from MRI scans with minimal preprocessing. The ordinal regression model's performance was compared to a metric regression model on the same task. Ordinal regression model achieved comparable performance to metric regression model and both models achieved comparable MAE with minimal preprocessing of input data up-to-date in comparison to existing literature. Both models were also not affected by the fact that the dataset used included data from 34 sources as described in Section 4.2.1.

Below, we additionally compare the ordinal model's performance with respect to the literature. Here the literature considered should fulfil certain criteria for fair comparison. First, the model described should be trained on a dataset of comparable size to that used to train our SENet-MR and SENet-OR models, in order to ensure that the model captures the population-wide ageing effects. This is important as in this thesis the dataset used for training is the largest used up-to date and, therefore, the SENet-MR and SENet-OR models can be considered to be capturing the ageing effects as close as possible to the population-wide ageing effects. Secondly, the model for comparison should be trained on MRI data which has been minimally preprocessed, with a pipeline similar to the one described in Section 4.2.3. Thirdly, the model should have been trained on data collected from healthy subjects only. Using such models for comparison also allows us to perform the comparison without the need to re-implementing and training the model. From the works represented in Table 5.1 three examples fulfil all the conditions – the model by Cole et al [35], Bermudez et al [183] and Peng et al [227]. Cole et al [229] used the dataset with fewer samples ($N = 2001$) compared to the HBAPS dataset ($N = 10,873$), but the age range (18-90 years of age) was similar to the HBAPS dataset (17-96 years of age). The SENet-MR and SENet-OR models achieved MAE comparable to the one achieved by the Cole et al's model. The dataset used by Cole et al [229] was collated from the datasets not containing any repeat scans, i.e , multiple scans collated from the same subject. As the result, the test dataset ($N = 200$) used by Cole et al [229] did not contain the samples from the subjects also present in the training data ($N = 1601$). Cole et al achieved MAE of 4.28 years, while the SENet-OR model achieved MAE of 4.10 years on the HBAPS test data excluding repeat samples. On the other hand, Cole et al used not residual and shallower CNN compared to the SENet-MR and SENet-OR models. Residual connections are required to enable training of deeper networks [108]. Cole et al have also reported networks inability

to perform equally well on the data from different acquisitions sources. Here for both the SENet-MR and SENet-OR models standard deviation of average error over data sources was 3.62 and 2.64 years respectively, while correlation between the number of samples in a dataset used and average error was -0.20 and -0.21 respectively. On the HBAPS test dataset excluding repeat samples standard deviation of average error over data sources was 3.16 and 3.27 years for the SENet-MR and SENet-OR models respectively, while corresponding correlations between the number of samples in a dataset used and average error were -0.04 and -0.08. Therefore, deeper residual network allows to perform training on larger datasets collated from different sources using different acquisition protocols. Bermudez et al [183] achieved MAE of 5.00 years using a dataset approximately half a size of the HBAPS dataset. Bermudez et al also included subjects as young as 4 years old. Peng et al [227] achieved MAE of 2.71 years using an ensemble model and 2.80 years using a single CNN model. From this it can be concluded that using ensemble CNN models does not introduce significant performance gain. Peng et al used the UK Biobank dataset for training with the age range of 44 to 88 years old. This may affect model's performance as first ageing signs were reported to occur at age of 30. Finally, it can be said that both the SENet-MR and SENet-OR models achieved performance comparable to the results previously obtained in the field.

The HBAPS dataset used for training (described in Section 4.2.1) is highly unbalanced over samples' age range. The SENet-MR model was trained using this raw data distribution. The SENet-OR model was trained using a sampling technique described in Section 2.3.3. Both models achieved similar MAE on both versions of the HBAPS test data, but in Figure 5.8 it can be seen that average error was higher for the youngest and the oldest subjects for both models. Also in Figure 5.8 it is shown that the ages for which average error was higher, corresponded to those for which the number of samples was lower than elsewhere. Therefore, two conclusions can be drawn. Firstly, sampling technique did not have a significant effect on the training process. Secondly, uneven sample distribution can introduce additional bias to a model.

Also from the point of view of using such models in clinical environment the advantage of training the models to achieve good performance by providing many examples in the age range younger than 30 years of age is not particularly useful compared to middle aged and older groups of subjects. From the Bland-Altman plot in Figure 5.12 it can be observed that most of the samples, for which the predicted age was outside of the range of $\mu_d \pm 1.96\sigma_d$

(μ_d – mean difference, σ_d – standard deviation of difference), are between the 30 and 70 years of age on both whole HBAPS test data and the HBAPS test data excluding repeat samples. The average error per age did not, however, correlate with the number of data sources contributing to this age. It can also be observed from Table 5.9 that average of mean errors per age label was the largest from the age range from 40 to 69 years of age for both metric and ordinal brain age prediction models on both versions of the HBAPS test data, while the largest number of samples was present in the age range from 20 to 39 years of age. However, the smallest number of samples was present in the age range from 70 to 84 years of age and this range also corresponded to the smallest average of mean errors per age label for both brain age prediction models.

Feng et al [39] achieved MAE of 4.06 years using a shallower network without residual connections on the dataset of comparable size with applied brain extraction. The SENet-MR and SENet-OR models achieved MAE of 3.87 and 3.62 years respectively on whole HBAPS test dataset with minimal preprocessing applied (as described in Section 4.2.3). On the HBAPS test data including only independent subjects MAE achieved by the SENet-MR and SENet-OR models was 4.37 and 4.10 years respectively. Feng et al [39] also used data collected from 4 different scanners, while in the HBAPS dataset the data originates from 34 public data sources. In this work comparable MAE was achieved using more heterogeneous dataset of less preprocessed data coming from a number of independent sources compared to the work by Feng et al using, however, deeper and residual networks. Jonsson et al [218] also concluded that residual networks improve performance. In the work by Jonsson et al [218] the performance of the CNNs with and without residual connections were compared on the brain age prediction task from skull-stripped MRI data non-linearly registered onto the MNI152 template. This is, however, concluded on the data with much more involved preprocessing procedure compared to the approach in this thesis.

It should be noted that further work is needed for brain age prediction model. Models' generalisability can be further assessed using independent dataset. Further comparison with Peng et al [227] work could also be useful and fruitful for improvements. For instance, Peng et al used shallower CNN similar to Cole et al [35], but achieved significantly lower results compared to Cole et al. The SENet-MR and SENet-OR models could also be examined using deconvolutional network approach introduced by Zeiler and Fergus [47] to understand which network parts contributed the most to accurate predictions and what have they detected at each step during a forward pass. This would allow making more

educated choice of improved architecture as opposed to simply using shallower networks. Once network could be made shallower and training would be less time-consuming, more extensive network hyperparameter optimization would be possible for further results improvement.

In conclusion, the accuracy of both the SENet-MR and SENet-OR models achieved better results compared to other published experiments on the input data with the same level of pre-processing. The model was, however, affected by the small number of samples for the age ranges of subjects younger than 20 and over 85 years of age even though the SENet-OR model used the sampling technique to remove such bias from the model. The models were not affected by the fact that the dataset was collated using 34 independent sources.

Chapter 6

Extracting population-wide brain ageing profile

6.1 Main contributions

In this chapter a population-wide brain ageing profile is considered. A brain ageing profile can be defined by determining periods of lifetime which have similar brain age-related features. The resulting brain ageing profiles are derived using the Healthy Brain Ageing from Public Sources (HBAPS) dataset described in Section 4.2.1, which is the largest dataset used to date for this type of analysis. Therefore, these brain ageing profiles can be considered to represent population-wide effects. As the subjects present in the HBAPS dataset are at least 17 years old, brain ageing profiles describe the changes related to brain ageing during adulthood. The profiles can be considered to be describe whole-brain effects as age-related features are extracted from Magnetic Resonance Imaging (MRI) samples as a whole using only minimal preprocessing.

In this chapter, four methods to derive a population-wide whole brain ageing profile are presented:

- Firstly, a method considering the distribution of errors from an underlying assessment of brain age is proposed.
- Secondly, application of the Preference Ranking Organization METHod for Enrichment of Evaluations (PROMETHEE) II method for obtaining a profile is discussed. The PROMETHEE II ranking is constructed based on the similarity of features extracted from the MRI samples. The features are extracted using a convolutional part

of a Convolutional Neural Network (CNN) model trained on the task of brain age prediction from MRI data.

- Thirdly, a profile is derived using the method of Deep Embedded Clustering (DEC) described in Section 2.3.5. In order to extract the brain ageing profile using DEC it is hypothesized that all features defining a brain age are contained in an MRI scan, and that profile intervals can be defined by considering the similarity of such features.

- Finally, a modification of DEC is proposed in order to create an ordinal DEC method. The ordinal method is developed as DEC itself cannot take into account ordinal nature of data. This is crucial for brain age prediction problem as brain ageing process can only proceed in one direction.

6.2 Introduction

Structural brain changes during human lifetime are studied both for the purpose of understanding the healthy course of ageing and to examine the departures from it to understand various diseases. Previous studies have investigated the trajectories of different structural biomarkers including volumetric, microstructural and focal ones, using both cross-sectional and longitudinal approaches [230]. Nevertheless, the variability of onset age for age-related brain diseases suggests that the effects of ageing on a human brain show significant differences both between individuals and during human lifetime. In this light, understanding of the population-wide brain ageing trajectory is crucial.

The ageing trajectories created to date were derived by considering volumetric changes of separate brain regions as well as that of a whole brain [51, 173, 231, 178, 180, 179, 171, 232]. Here only the approaches which analysed more global measures instead of regional ones are discussed – White Matter (WM), Grey Matter (GM), Cerebrospinal Fluid (CSF), ventricular and total brain volumes. The work discussed here is also restricted to trajectories derived from MRI data. Such approaches can be divided into two main groups – the ones using longitudinal datasets and the ones using cross-sectional data.

Kuo et al [51] assessed brain ageing trajectories using longitudinal data from the Baltimore Longitudinal Study of Aging (BLSA). The BLSA cohort consisted of 1581 subjects who underwent assessment with a frequency dependent on their age – once every 4 years for under 60 years of age, once every 2 years between 60 and 79 years of age, and yearly after 80 years. Total brain volume, along with the volumes of WM, GM and ventricular

volume, were calculated from the MRI data. The volumes were averaged for each decade from 50 to 90 years of age. Kuo et al [51] found total brain, WM and GM volumes to be linearly decreasing over lifetime, while ventricular volume increased. The rates of decline of GM and increase of ventricular volume also increased with age. Fama and Sullivan [178] also used longitudinal data collected from 122 subjects aged from 20 to 85 years of age. The subjects were scanned 1 to 6 times over a period of between 1 and 8 years. The authors identified that brain ageing is associated with a reduction in thalamic volume, and that this reduction accelerates after 60 years of age.

Good et al [173] inferred the trajectory using cross-sectional data from a cohort of 465 healthy subjects, using measurements of the volumes of WM, GM and CSF from MRI scans. Wei et al [231] described the Southwest University (SWU) cohort consisting of 494 subjects aged from 19 to 80 years.[1]The authors calculated the correlation coefficient between GM volume and age and found them to be negatively correlated. This analysis was, however, limited by the fact that a correlation can only identify a linear relationship. However, it should be noted that other authors have also found an inverse linear relationship between GM volume and age [51, 173]. Wei et al [180] assessed the relationship between the volume of Subcallosal Region (SCR) and age in cross sectional dataset from a cohort including 112 healthy subjects in the age range of 19 to 72 years. The authors identified an inverse linear relationship between SCR volume and age, i.e. SCR volume decreased during human lifetime. Salat et al [179] used a cross-sectional dataset consisting of 106 subjects in the age range from 18 to 93 years to study cortical thinning in relation to age. Overall cortical volume was found to decrease with age.

The biggest cohort among the published work on brain ageing trajectories to date was used by Vinke et al [171]. Vinke et al performed the Rotterdam study consisting of 5286 subjects of which 57 had four MRI scans, 1921 – three scans, 1456 – two, 1852 – one (10755 scans in total). For the subjects for whom repeated scans were performed, the inter-scan interval was 3 to 4 years. Therefore, the derived trajectories were based on longitudinal and cross-sectional data simultaneously. The resulting trajectories described the relationships between 45 and 95 years of age. The GM, WM and total brain volumes all decreased with age. While GM volume had a linear relationship with age, for the WM and total brain volumes the relationship was found to be non-linear, with the slope of the curve increasing

[1]SWU cohort is also contained in the HBAPS dataset described in Section 4.2.1.

with age. The volume of CSF also increased with age and rate of increase also grew with age.

It should be noted that there are a number of additional articles in this field (for example [174, 175, 176]) which are not discussed and compared to here as no meaningful comparison can be made to the work described here due to differences in methodologies and data used for the research. The articles not discussed here either use solely volumetric measures to measure the ageing process, use small cohort datasets or present the results in a way which makes the comparison with the results of this chapter complicated. For some articles there is a combination of factors which resulted in complications for comparison. For example, Pfefferbaum et al [174] applied semiautomated segmentation on MRI data collected from 73 males aged 21 to 70 years and used volumetric measures of separate brain regions to assess the process of ageing. This means that it is not possible to meaningfully compare these results to the results of this chapter as the cohort used by Pfefferbaum et al cannot be considered to represent population-wide effects and trajectories of separate regions cannot be compared to whole brain measures used in this chapter. For the same reason the results are not compared to the results of Lim et al [175] as only 8 subjects were used and ageing process was assessed for separate brain regions. Comparison with the results of Gur et al [176] was also not meaningful as they also used only 69 subjects and focused on understanding the differences in ageing between men and women instead changes over human lifetime. All these papers provide evidence of changes in brain features over the human lifetime, but are not considered further here.

The approaches described above, i.e. assessing volumetric brain changes from MRI data, have both advantages and disadvantages. The approaches using cross-sectional data cannot be considered to produce true ageing trajectories (for which multiple measures on the same individual are necessary), while longitudinal studies suffer from lack of availability of consistent data over the complete human lifetime. Longitudinal studies to date also suffer from attempting to describe the whole trajectory using at most 4 time points per subjects [171]. For example, Raz et al [177] applied statistical methods to multiple brain regions on MRI data, but limited the research to two time points. The studies analysing volume changes in separate brain regions have advantages for understanding brain ageing in terms of its relationship to cognitive abilities of subjects, as separate regions are known to be related to different cognitive functions. Assessing brain ageing via considering anatomical brain regions has, however, a disadvantage of not representing the ageing as

a process affecting whole brain. Longitudinal data, however, are hard to collect and only a few datasets are available, for example the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [32].

Since derivation of an ageing trajectory per se is not possible from cross-sectional data, Shamir and Long [6] proposed to instead derive a brain ageing profile. Shamir and Long use a cohort of 416 subjects aged 18 to 96 years of age, with total 436 T1-weighted MRI scans. The authors compute the profile by measuring similarity between MRI scans. The feature extraction uses the Wndchrm software [233]. Wndchrm software extracts features from images by considering the measures describing image content and image transformations including compound image transformations. Wndchrm extracts 11 sets of features including the ones extracted using Gabor filters and Chebyshev statistics (full list and details on feature extraction can be found in [233]). These features are then used to choose the most descriptive ones and obtain a feature vector for each image which can be used for similarity measurements. The profile is obtained by dividing the age range present in the dataset into several intervals, calculating average predicted age for each interval and plotting these values against the intervals. In such a plot, regions of slower and faster changes may be identified by considering the slope of a graph. The authors showed that the two periods of most rapid age-related changes were around chronological age of 55 and 65 years [6]. Shamir and Long [6], however, did not use age-specific features and only used a k-means clustering algorithm without extracting the ranking of samples.

In this thesis a population-wide brain ageing profile assessing whole-brain ageing effects is considered instead of a brain ageing trajectory per se. A brain ageing profile can be defined by determining periods of lifetime which have similar brain age-related features. The resulting brain ageing profiles, including the ones presented here as (which is derived using the HBAPS dataset described in Section 4.2.1, which is the largest dataset used to date for this type of analysis) can be considered to represent population-wide effects. The profiles presented here assess whole-brain effects, as features are extracted from MRI samples as a whole, after minimal preprocessing.

In this chapter the terminology of an ageing profile consisting of intervals and bins is used interchangeably.

6.3 Methods

6.3.1 Obtaining a brain ageing profile using the error distribution of a brain age prediction model

Brain age prediction models, SENet-MR an SENet-OR based on Squeeze-and-Excitation Network (SENet) architecture described in Section 5.3, output an age for an input MRI scan. Consider a dataset of N MRI samples, $\{\boldsymbol{x}_i \in \boldsymbol{X}, y_i \in Y\}_{i=1}^{N}$, where $\boldsymbol{x}_i$ is a MRI input sample and $y_i \in Y = \{a_1, a_2, \ldots, a_K\}$, is the corresponding age, where $\{a_1, a_2, \ldots, a_K\}$ is the age range represented in the data.

The ageing profile bins are searched by iteratively moving the bin walls one-by-one through the age range. The range is first searched to set the first wall. The first wall is chosen from a set $\{a_2, a_3, \ldots, a_{K-1}\}$. Then each of the resulting bins is sub-divided recursively until the model resolution, i.e. the smallest bin size, is reached. For each bin $B = \{b_1, ..., b_I\}$, where I is the number of bins and each bin contains set of ages such that $b_l = \{a_1, ..., a_L\}$, where L is the bin length. Any bin can be sub-divided into two bins by a wall, with a set of possible "walls" is considered such that $W_l = \{a_2, ..., a_{L-1}\}$. Figure 6.1 shows an example of age range being divided into three intervals. A bin B is divided by a "wall" at a_k by summing the probabilities for each sample to be classified into an incorrect bin (i.e. a bin to which its true age does not belong) given a wall at this position, and then finding the value of k which minimises these incorrect assignments. For example, consider dividing the range of K ages (17 – 92 years) into 2 bins of the ageing profile. In order to find the "wall" age it is iteratively moved from a_2 to a_{K-1} ($a_2 = 18,\ a_2 = 91$). For each sample the probability of predicting a sample age to be outside a bin, where its true age label lies is assessed. Then the probabilities are summed over all samples in the test dataset.

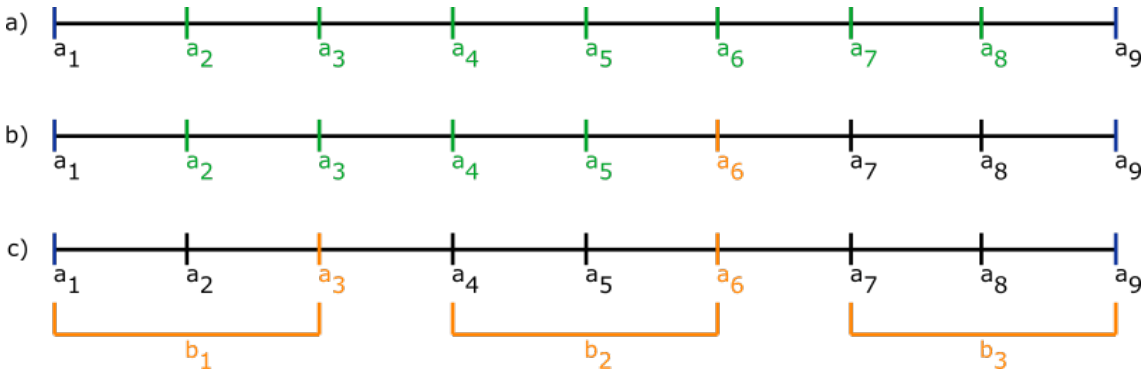

Figure 6.1: Example of dividing age range Y into three intervals by considering the distribution of predictions of a brain age prediction model. Consider a dataset of N MRI samples, $\{\boldsymbol{x}_i \in \boldsymbol{X}, y_i \in Y\}_{i=1}^{N}$, where $\boldsymbol{x}_i$ is a MRI input sample and $y_i \in Y = \{a_1, a_2, \ldots, a_9\}$. a) First wall, w_1, is chosen from a set of ages – $W_1 = \{a_2, a_3, \ldots, a_8\}$. b) Set first wall $w_1 = a_6$ and search for w_2 to divide first bin including $\{a_1, a_3, \ldots, a_6\}$ into two bins recursively. w_2, is chosen from a set of ages – $W_2 = \{a_2, a_3, \ldots, a_5\}$. c) As the results the age range $Y = \{a_1, a_2, \ldots, a_9\}$ is divided into three bins, $I = 3$ and $B = \{b_1, b_2, b_3\}$. Then $b_1 = \{a_1, a_2, a_3\}$, $b_2 = \{a_4, a_5, a_6\}$, $b_3 = \{a_7, a_8, a_9\}$.

Two approaches were considered when calculating the probabilities needed to set the bin "walls". In the first approach, the mean and variance of the predictions were calculated for samples of each true age. Predictions for each true age were then assumed to be normally distributed, and fully described by this mean and variance.

For some true ages the distribution of predictions strongly deviated from a normal one, however, so a second approach was developed in which we considered the distribution of predictions directly, without this normal approximation. An example of how the predictions are distributed is presented in Figure 6.2 for the true age of 76 years. For each true age, the prediction distribution was obtained by dividing the number of predictions for each label in the test set by the number of samples with a given true age:

$$p_{y,\hat{y}} = \frac{N_{\hat{y}}}{N_y}, \tag{6.1}$$

where

- $p_{y,\hat{y}}$ is the probability of predicting the samples with true age y to be of age $\hat{y}$,
- N_y is the number of samples with true age label y,
- $N_{\hat{y}}$ is the number of samples of true age label y predicted to have age of label $\hat{y}$.

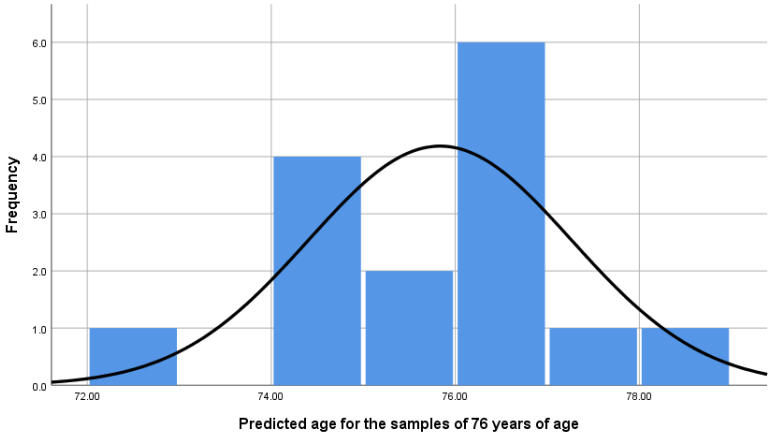

Figure 6.2: Histogram representing the distribution of predictions by the SENet-MR model for the subjects of age 76. Black line indicates normal distribution.

Implementation

In order to obtain a population-wide ageing profile we used two trained models - SENet with metric regression (SENet-MR) and SENet with ordinal regression (SENet-OR). A detailed description of SENet is provided in Section 2.3.2. Both networks were evaluated on the testing HBAPS dataset described in Section 4.2.1. The Mean Absolute Error (MAE) achieved by the SENet with metric and ordinal regression was 3.6 and 3.8 years respectively, giving a model resolution of 4 for both models. The smallest possible bin size in the profile extraction procedure was therefore set to 4.

6.3.2 PROMETHEE II method

The PROMETHEE II method is a ranking method, i.e. it utilises pair-wise comparisons among the inputs. Ranking is constructed by considering the difference in evaluations of the inputs on a particular criterion. Depending on a pair-wise difference measured a decision can be made – one sample can be preferred over another or not with respect to a criterion depending on the size of a difference measured. The larger is the difference measured, the larger is the preference of one alternative over another.

The PROMETHEE II method is described in Section 2.4.3. As in previous section consider a dataset of N MRI samples, $\{\boldsymbol{x}_i \in \boldsymbol{X}, y_i \in Y\}_{i=1}^{N}$, where $\boldsymbol{x}_i$ is a MRI input sample and y_i is the corresponding age. Then consider the features describing the samples in $\boldsymbol{X}$, $Z = \{\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_N\}$, where $\boldsymbol{z}_n$ is a set of features extracted from the n^{th} sample by a feature extractor. Each sample has a label from the set $Y = \{y_1, y_2, ..., y_K\}$. In order

to initialise the PROMETHEE II method k-means clustering is performed first on Z to obtain M clusters, $C = \{\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_M\}$. Based on these M clusters, M criteria can then be stated such that $G = \{g_1, g_2, ..., g_M\}$ and G is a set of criteria. Relationship of a sample $\boldsymbol{x}_i$ and a criterion g_m can be formulated in form of a question – "Does a sample $\boldsymbol{x}_i$ belong to a cluster $\boldsymbol{c}_m$?". Here the differences between the samples with respect to a criterion m are quantified using a following function:

$$P_m(d) = \begin{cases} 0 & d_m \leq 0 \\ \frac{d_m - q}{p - q} & q < d \leq p \\ 1 & d_m > p \end{cases} \tag{6.2}$$

where

- $d_m(\boldsymbol{z}_i, \boldsymbol{z}_j) = g_m(\boldsymbol{z}_i) - g_m(\boldsymbol{z}_j)$ is a difference between evaluating the features of two input samples, $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$, on criterion m with g_m defined as

$$g_m(\boldsymbol{z}_i) = \sum_{\boldsymbol{z}_k \in \boldsymbol{c}_m} \|\boldsymbol{z}_i - \boldsymbol{z}_k\|, \tag{6.3}$$

  $\|\cdot\|$ denotes Euclidean distance between a pair of sample features. Equation 6.3 represents an evaluation of an alternative $\boldsymbol{z}_i$ with respect to a criterion g_m. It is represented by a sum over Euclidean distances calculated between a sample features of interest and all other samples assigned to a corresponding cluster at the initialisation step.

- $q = min(g_m(\boldsymbol{z}_i), g_m(\boldsymbol{z}_j))$
- $p = max(g_m(\boldsymbol{z}_i), g_m(\boldsymbol{z}_j))$

Once the differences between input samples are quantified the ranking is constructed as detailed in Section 2.4.3. Final cluster assignment for a sample is determined by a criterion with respect to which the sample ranked the highest. The intervals of ageing profile are determined by considering a mode cluster assignment for each label in Y. A wall is set at the first age label (starting from the lowest) at which the mode interval label changes. An example of defining intervals over age range Y is shown in Figure 5.3.

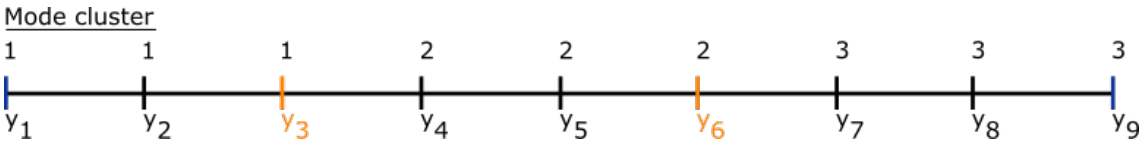

Figure 6.3: Example of dividing age range Y into three intervals using the PROMETHEE II method. As in previous section consider a dataset of N MRI samples, $\{\boldsymbol{x}_i \in \boldsymbol{X}, y_i \in Y\}_{i=1}^{N}$, where $\boldsymbol{x}_i$ is a MRI input sample, y_i is the corresponding age, $Y \in \{y_1, y_2, ..., y_9\}$ and $y_1 < y_2 < ... < y_9$. Once PROMETHEE II ranking is obtained, a cluster assignment is predicted for each sample by considering its highest ranking. Mode cluster assignment is determined for each age in Y. First wall is set at y_3 as it is the first age label at which mode cluster value changes. Mode cluster value also changes at y_6 and second wall is set at this age label.

Implementation

The feature extractor used here is the convolutional part of the SENet-MR and SENet-OR models trained on the task of brain age prediction from MRI data as described in Chapter 5. An ageing profile was obtained by applying the PROMETHEE II method to the HBAPS dataset described in Section 4.2.1. The whole dataset could be used, including the subsets which would usually be reserved for training, validation and testing, as training is not required for the PROMETHEE II method and the parameters of a feature extractor are not adjusted in the process.

6.3.3 DEC

The DEC method is described in Section 2.3.5. Consider a set of N MRI scans, $\{\boldsymbol{x}_i \in \boldsymbol{X}, y_i \in Y\}_{n=1}^{N}$, where $\boldsymbol{x}_n$ is an input MRI scan and y_n is the true age, which we wish to divide into I intervals, each represented by a centroid $\boldsymbol{\mu}_i, i = \{1, ..., I\}$. $\boldsymbol{x}_n$ is first mapped into an embedding space, $f_\theta : \boldsymbol{x}_n \rightarrow \boldsymbol{z}_n$, where θ are learnable parameters of a feature extractor f and $\boldsymbol{z}_n$ denotes an embedding of $\boldsymbol{x}_n$.

The DEC approach consists of two main steps. First a feature extractor, f, is pretrained to extracted features from age-related features from input MRI scans. This is achieved by using a feature extractor from a model trained on the task of brain age prediction. As the result the parameters θ of f are learned. In the second step, the DEC method learns intervals contained in the data, i.e. the intervals defined over age range Y. The DEC algorithm learns to sub-divide MRI scans into intervals based on corresponding ages by considering age-specific features, $\boldsymbol{z}_n$, of each scan.

Once the method is trained an interval assignment for a sample, $\boldsymbol{x}_n$, is determined by first calculating the Students t-distribution based on similarity between its features, $\boldsymbol{z}_n$, and an interval centroid, $\boldsymbol{\mu}_i$:

$$q_{ni} = \frac{(1 + \|\boldsymbol{z}_n - \boldsymbol{\mu}_i\|^2)^{-1}}{\sum_i (1 + \|\boldsymbol{z}_n - \boldsymbol{\mu}_i\|^2)^{-1}} \tag{6.4}$$

where $\|\cdot\|^2$ is a squared Euclidean distance. Then target distribution is calculated based on q_{ni}:

$$p_{ni} = \frac{q_{ni}^2 / \sum_n q_{ni}}{\sum_i q_{ni}^2 / \sum_n q_{ni}}. \tag{6.5}$$

Finally, interval assignment is determined by applying *argmax* to an output of Equation 6.5, i.e. finding an index of the maximum value of p_{ni}.

The DEC method does not inherently take into account data ordinality, but the brain ageing profile being extracted here is inherently ordinal in nature. In order to enforce the samples' ordinality, we therefore construct the intervals of brain ageing profile in a tree-like structure. We first train a DEC model to divide a training set into 2 intervals. For this purpose, whole dataset of N MRI scans is divided into training, validation and testing data. The wall defining the two intervals is found by assigning each sample of a test data to one of the two intervals. A subset of samples from the test set is then considered for which prediction error on age label from the model, to which the feature extractor belongs to, is smaller than a sum of the model's MAE and its standard deviation. For each age label present in the range divided into 2 intervals a mode interval label is determined. A wall is set at the first age label (starting from the lowest) at which the mode interval label changes from 1 to 2. Examples of defining intervals over age range is shown in Figure 6.4. The procedure described here for sub-dividing an interval into two is then applied recursively in a tree-like structure. Each time the procedure is applied a dataset is extracted out of the dataset used for finding first two intervals such that only the data with labels belonging to the bin to be sub-divided is used. As previously, the smallest interval is set by the MAE of the feature extractor used, rounded to the closest integer.

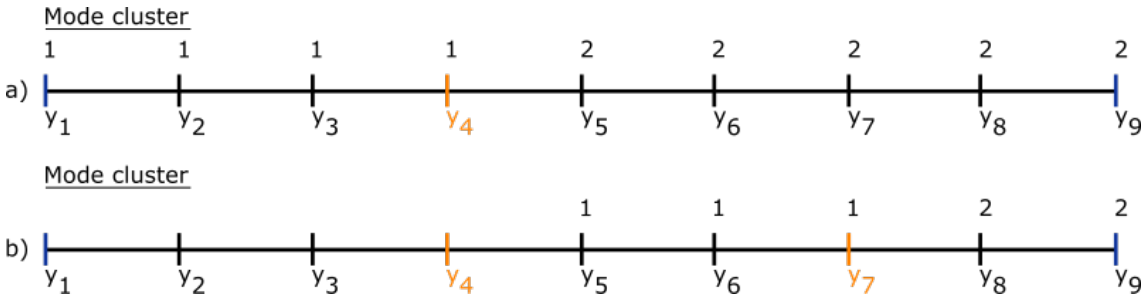

Figure 6.4: Example of dividing age range Y into three intervals using DEC. Consider a dataset of N MRI samples, $\{\boldsymbol{x}_i \in \boldsymbol{X}, y_i \in Y\}_{i=1}^{N}$, where $\boldsymbol{x}_i$ is a MRI input sample, y_i is the corresponding age, $Y \in \{y_1, y_2, ..., y_9\}$ and $y_1 < y_2 < ... < y_9$. Once DEC is trained and a cluster assignment is predicted for each sample in the test data. Mode cluster assignment is determined for each age in Y. a) First wall is set at y_4 as it is the first label at which mode cluster value changes. b) DEC is trained on data samples having age labels belonging to the second bin in the first step. Once DEC is trained on this data and mode cluster assignment is determined for each age label in this bin, the second wall is set at y_4 as mode cluster assignment value changes at it.

Implementation

The DEC model was trained using a pre-trained SENet-MR classifier, as described in Section 5.3.1 and which is used as a feature extractor, f_θ, here. This ensures that intervals are based on age-related features. The training was performed using stochastic gradient descent algorithm with initial learning rate 0.1 and momentum of 0.9. The number of epochs for the training was set to 100, with a batch size of 4. The training was stopped when less than 0.1% samples changed interval assignment between two adjacent epochs. This was measured on the validation set by running interval assignment at the end of each epoch. Training parameters were the same for all iterations as bins are searched for in a tree-like structure.

6.3.4 Ordinal DEC

As noted above, the basic DEC method does not take advantage of the additional information present in ordinal data (i.e. the fact that the data are known, a priori, to have an order associated with them). In order to tackle this issue an additional training step for DEC is proposed here. As previously, consider a dataset of N samples, $\{\boldsymbol{x}_i \in \boldsymbol{X}, y_i \in Y\}_{i=1}^{N}$, where $\boldsymbol{x}_i$ is an input sample and $y_i \in \{r_1, r_2, ..., r_K\}$ is the corresponding label (i.e. age), but where possible label values are now ordered such that $r_1 < r_2 < ... < r_{K-1} < r_K$ and where K is the number of class samples or ranks.

For the brain age prediction problem dividing the dataset into I intervals is based on two factors - the similarity of samples within an interval and the ordinality of sample labels. Defining intervals is influenced by ordinality of labels as it means that for a trained method a sample can be either assigned to an interval, where its true age labels belongs, or to one of the adjacent intervals. Therefore, the sub-division should be performed such that the order of ranks is conserved. The proposed method is to pretrain the method for grouping the samples based on their similarity using DEC, and then train the method to learn the set of the set of interval walls - $a_k, k \in \{1, ..., I-1\}$, which ensure ordinality of the bins created. The task is thus to find $I-1$ walls which divide Y into I intervals such that the first interval, I, is $i_1 = \{r_1, \ .., r_k\}$, where $r_k = a_1$ and the last interval is $i_K = \{a_{I-1}, ..., r_K\}$. Consider an example with $Y = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ and $a_k \in \{2, 5, 7\}$. For this example $i_1 = \{1, 2\}$, $i_2 = \{3, 4, 5\}$, $i_3 = \{6, 7\}$ and $i_4 = \{8, 9\}$.

In order to achieve this, the following architecture is proposed: a feature extractor pretrained with the DEC method is connected to fully-connected layers to output a vector of interval walls - $a_i, i \in \{1, ..., I-1\}$ for I intervals, i.e. in a forward pass through the proposed architecture each batch of samples is put through the pretrained DEC to predict the intervals the batch samples belong to. For training, the following loss function is proposed:

$$L = \frac{1}{N}\left(1 - \left(H\left(y_n - a^i_{m-1}\right) - H\left(y_n - a^i_m\right)\right)\right), \tag{6.6}$$

where

- a^i_{m-1} and a^i_m are the interval "walls" predicted for y_n by DEC,
- $H(x)$ is a Heaviside function, which has a value of 0.5 at x, a value of 1 for all values larger than x, and 0 otherwise (shown in Figure 6.5),
- the difference $H(x - a^i_{m-1}) - H(x - a^i_m)$ is a rectangle function Π which is

$$\Pi(x) = \begin{cases} 0 & [x] > a_m \ and \ [x] < a_{m-1} \\ \frac{1}{2} & [x] \in a_m, a_{m-1} \\ 1 & a_{m-1} < [x] < a_m \end{cases} . \tag{6.7}$$

As the Heaviside function is not continuously differentiable, it is approximated by a continuous function in the training stage:

$$H(x-a) \approx 0.5 + 0.5 \times tanh(C(x-a)), \tag{6.8}$$

where a constant C is set empirically during the experiments. Figure 6.5 shows the Heaviside function and the difference between two Heaviside functions used in Equation 6.6. The approximation to the Heaviside function given in Equation 6.8 is shown in Figure 6.6 for values of $C = 1$ and $C = 10$, showing how as C increases the approximation tends towards the instantaneous step of the Heaviside function, while still remaining smooth and therefore differentiable. C value is chosen such that a trade-off between the form of Equation 6.8 being close to the form of Heaviside function shown in Figure 5.5 and having a differentiable function.

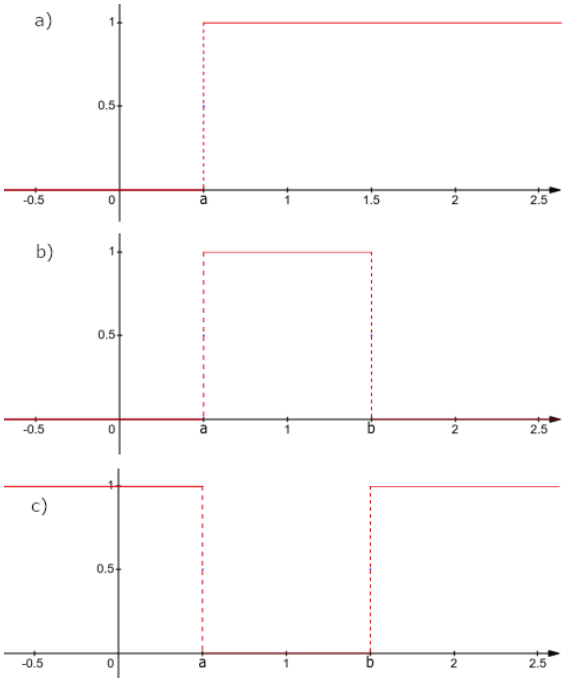

Figure 6.5: a) Heaviside function with step value of 0.5 at a denoted by $H(x-a)$. b) Graph of the difference $H(x-a) - H(x-b)$. c) Graph of the function $1 - (H(x-a) - H(x-b))$.

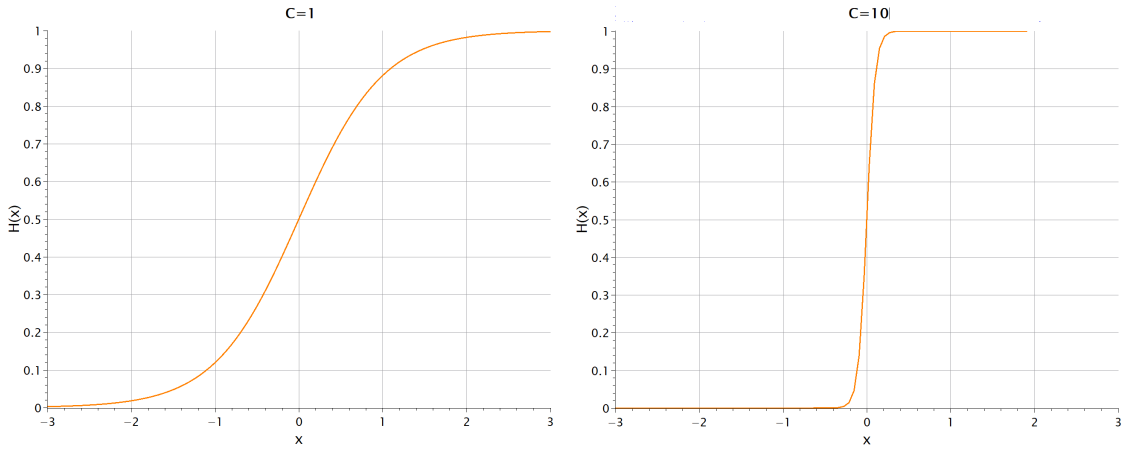

Figure 6.6: Graph of the approximation to the Heaviside function as given in Equation 5.3 for the values of C being equal to 1 and 10.

Implementation

Testing on synthetic data

The new ordinal DEC model was first trained and tested on the synthetic datasets described in Sections 4.1.2 and 4.1.3. In order to pre-train the feature extractor, a Stacked Autoencoder (SAE) (described in Section 2.3.4) was first trained on the training subset of the synthetic data. In a SAE, each layer is first initialised by pretraining as a denoising autoencoder (detailed description in Section 2.3.4), which learns to reconstruct the output of the previous layer with random noise applied. In SAE the setup described in Figure 2.17 is modified by inserting dropout layers between the corruption step and encoder, and between the activation and decoder. The activation function used in both encoder and decoder was Rectified Linear Unit (ReLU), as defined in Section 2.3.1. In SAE setup the encoder and decoder each had 5 layers. Each layer was trained using the least-squares loss function:

$$\alpha_{ls} = \|\boldsymbol{x}_i - \boldsymbol{y}_i\|^2, \tag{6.9}$$

where

- $\boldsymbol{x}_i$ denotes input to a layer,
- $\boldsymbol{y}_i$ - output of a layer,
- $\|\cdot\|^2$ – squared Euclidean distance.

The layers are pre-trained using a Stochastic Gradient Descent (SGD) algorithm with initial learning rate 0.01 and momentum of 0.9 for 300 epochs. In pre-training the layers learning rate is multiplied by 0.1 every 100 epochs and batch size is 256. After all the layers are pre-trained, the whole SAE is trained to finetune performance. The SAE's encoder is then used as a feature extractor in the ordinal DEC model. K-means clustering is applied to initialise the cluster centres [122]. In the ordinal DEC training, the initial number of clusters was varied with respect to the experimental conditions as described further in Section 6.4 as were the intial "wall" positions. The network training was performed using a SGD algorithm with initial learning rate 0.01 and momentum of 0.9. The number for epochs for the training was set to 100 with batch size of 256. The stopping condition was the same as for the standard (non-ordinal) DEC (described in Section 6.3.3).

Testing on MRI data

For the MRI data, the feature extractor used in the DEC is the convolutional part of the SENet-MR model trained on the task of brain age prediction (described in Section 6.3.3). This ensures that the similarity of the outputs is based on age-related features. This pre-trained feature extractor is then used to obtain the interval "walls" using the method proposed above.

K-means clustering is applied to initialise the cluster centres [122]. The initial number of clusters was varied with respect to the experimental conditions as described further in Section 6.4 as were the initial "wall" positions. The network training was performed using a SGD algorithm with initial learning rate 0.01 and momentum of 0.9. The number for epochs for the training was set to 100 with batch size of 4 due to GPU memory constraints. The stopping condition was the same as for the standard (non-ordinal) DEC (described in Section 6.3.3).

6.4 Results

Population-wide brain ageing profiles were obtained using the methods described, and are presented in Tables 6.2 – 6.5. In Tables 6.2 and 6.3, the intervals obtained from the distribution of predictions for the SENet-MR and SENet-OR models are presented. The model resolution was 4 for both models, as MAE achieved by the SENet with metric and ordinal regression was 3.9 and 3.6 years respectively. In Tables 6.4 and 6.5, the

intervals of the brain ageing profile obtained using the DEC and PROMETHEE II methods, respectively, are presented.

In order to facilitate the comparison with the results of Shamir and Long [6], the profiles obtained using the distribution of predictions are presented in Figures 6.7 and 6.9 as plots showing the means of the predictions for each profile interval found. Figures 6.8 and 6.10 show similar plots, but for intervals of fixed length of 5 years the same as used by Shamir and Long [6].

In Figures 6.7 – 6.10, the linear fit is shown by blue line and the lower and upper bounds on it are in green. Figures 6.7 and 6.9 show means of predicted ages plotted against the intervals of the profiles obtained using the distribution of predictions of SENet-MR and SENet-OR models defined in Table 6.2 and 6.3 respectively. A linear regression for the relationship between minimum age of an interval in the profile, $i_{min,prof}$, and predicted mean age for each interval, $y_{prof,av}$, is described by $y_{prof,av} = a_{prof} i_{min,prof} + b_{prof} = 0.91 i_{min,prof} + 6.27$. The 95% confidence interval for the b_{prof} coefficient is 1.24 to 11.39 ($p = 0.0001$). It should be noted that a line of identity falls into the 95% confidence interval of the linear fit in both plots.

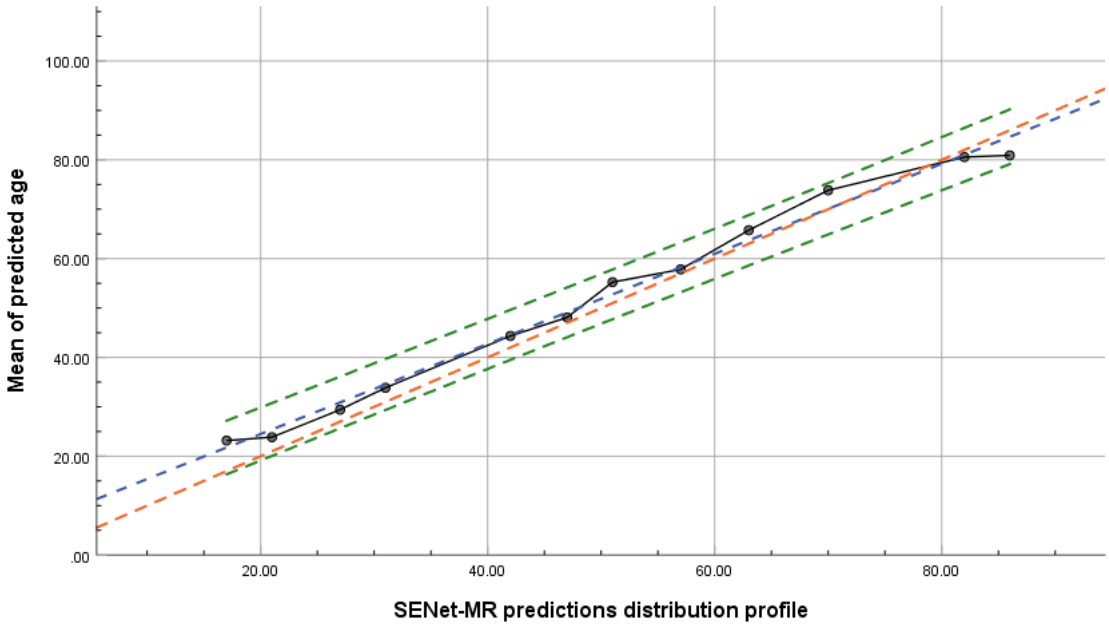

Figure 6.7: Means of predicted ages for the intervals of the profile obtained using the distribution of predictions of SENet-MR model. Orange line denotes the y=x function, blue – linear regression fit function, green – the lines representing the lower and upper bounds of the 95% confidence interval over linear regression fit parameters.

For comparison with existing literature Figure 6.8 shows means of predicted ages by the SENet-MR model plotted against 5-years-long intervals as defined in work by Shamir and Long [6]. A linear regression for the relationship between minimum age of each 5-year interval, $i_{min,preset}$, i.e. $i_{min,preset} \in \{15, 20, 25, ..., 90\}$, and predicted mean age for each interval, $y_{m,preset,av}$, is described by $y_{m,preset,av} = a_{preset} i_{min,preset} + b_{m,preset} = 0.91 i_{min,preset} + 6.17$. The 95% confidence interval for the $b_{m,preset}$ coefficient is 4.41 to 7.93 ($p = 0.0001$). It should be noted that the upper bound linear fit 95% confidence interval closely follows line of identity in Figure 6.8.

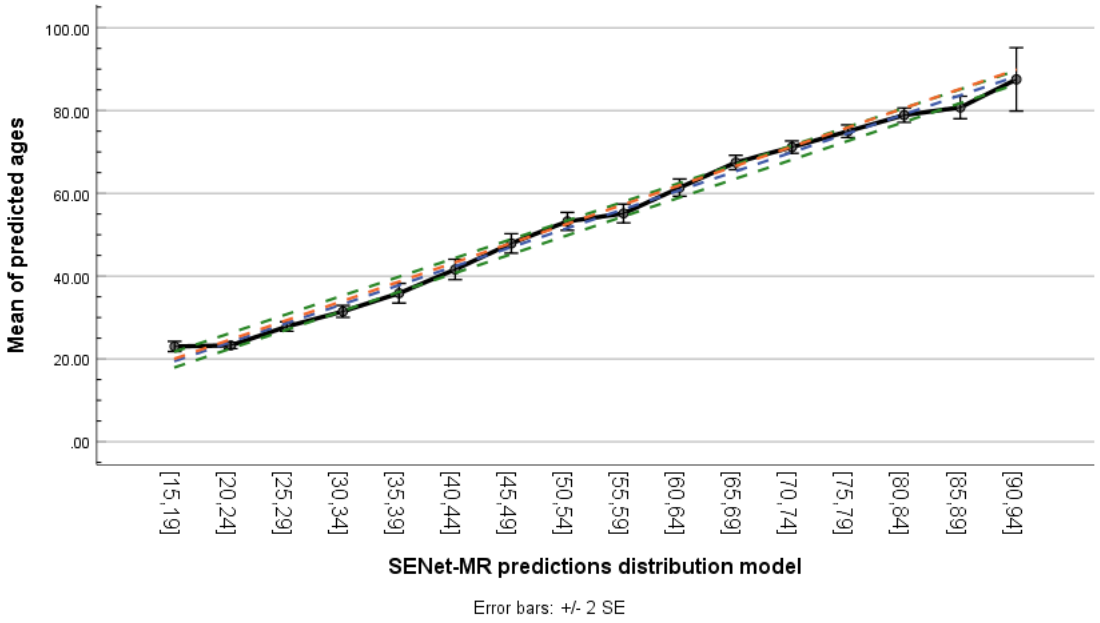

Figure 6.8: Means of predicted ages by the SENet-MR model for the intervals of fixed width of 5 years. Whiskers show 95% confidence interval over a mean. The last interval has length of 6 years to include the age of 92. Whiskers show 95% confidence interval over a mean. Orange line denotes the y=x function, blue – linear regression fit function, green – the lines representing the lower and upper bounds of the 95% confidence interval over linear regression fit parameters.

For comparison with existing literature Figure 6.10 shows means of predicted ages by the SENet-OR model plotted against 5-years-long intervals as defined in work by Shamir and Long [6]. A linear regression for the relationship between minimum age of each 5-year interval, $i_{min,preset}$, i.e. $i_{min,preset} \in \{15, 20, 25, ..., 90\}$, and predicted mean age for each interval, $y_{o,preset,av}$, is described by $y_{o,preset,av} = a_{preset} i_{min,preset} + b_{o,preset} = 0.92 i_{min,preset} + 4.04$. The 95% confidence interval for the $b_{o,preset}$ coefficient is 1.49 to 6.60

($p = 0.0001$). It should be noted that the upper bound linear fit 95% confidence interval closely follows line of identity in Figure 6.10.

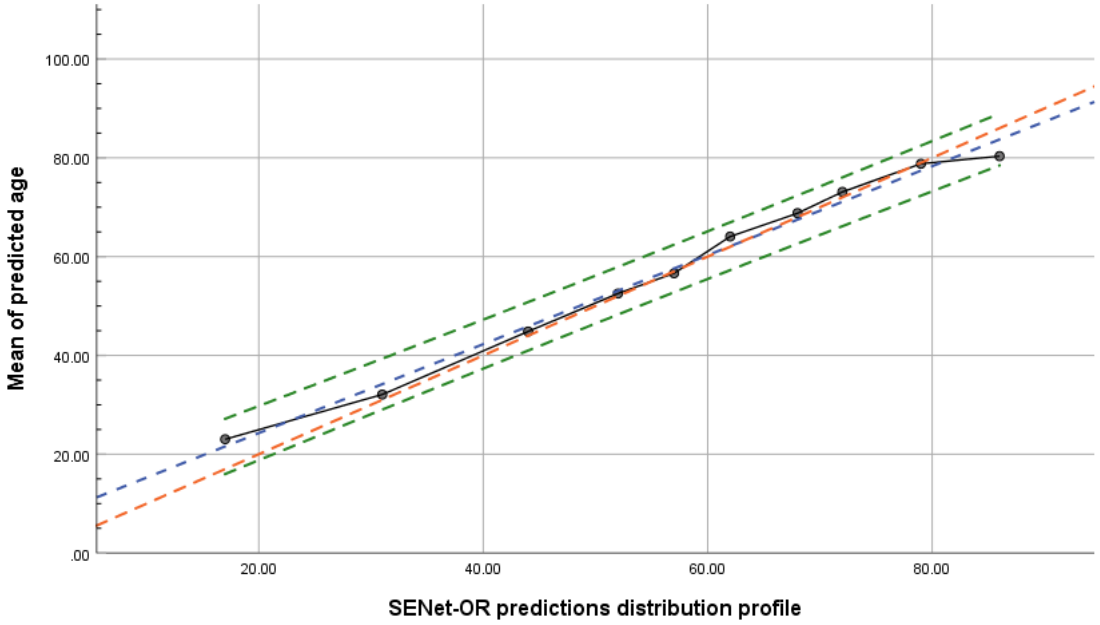

Figure 6.9: Means of predicted ages for the intervals of the profile obtained using the distribution of predictions of SENet-OR model. Orange line denotes the $y = x$ function, blue – linear regression fit function, green – the lines representing the lower and upper bounds of the 95% confidence interval over linear regression fit parameters.

The profiles represented in Tables 6.2-6.4 are visualised in Figure 6.11. In Figure 6.11 interval colouring visualises a gradient calculated as $1/(interval\ length)$ – darker the interval is coloured, steeper the gradient is. The gradient is calculated in order to facilitate the comparison with published trajectories in existing literature, i.e. darker intervals have a steeper slope and can be interpreted as the intervals corresponding to faster ageing processes.

Initialisation		Dataset	
Number of bins	**"Walls"**	**Ordinal MNIST**	**Two-variable ordinal**
3 bins	Close to the true values	[1,4]	[2,6]
	Not close to the true values	[0,9]	[0,13]
	Fractional values	[3.5, 7.8]	[5.5, 10.2]
4 bins	-	[1,4,9]	[1,6,12]
5 bins	-	[0,1,4,9]	[0,1,6,12]

Table 6.1: "Wall" initialisations in the ordinal DEC method experiments on synthetic datasets. Three different "wall" position initialisations are tried with true number of bins, i.e. 3 bins, contained in the ordinal MNIST and two-variable ordinal datasets. Then 4 and 5 bins are set in the initialisation by setting more "wall" positions.

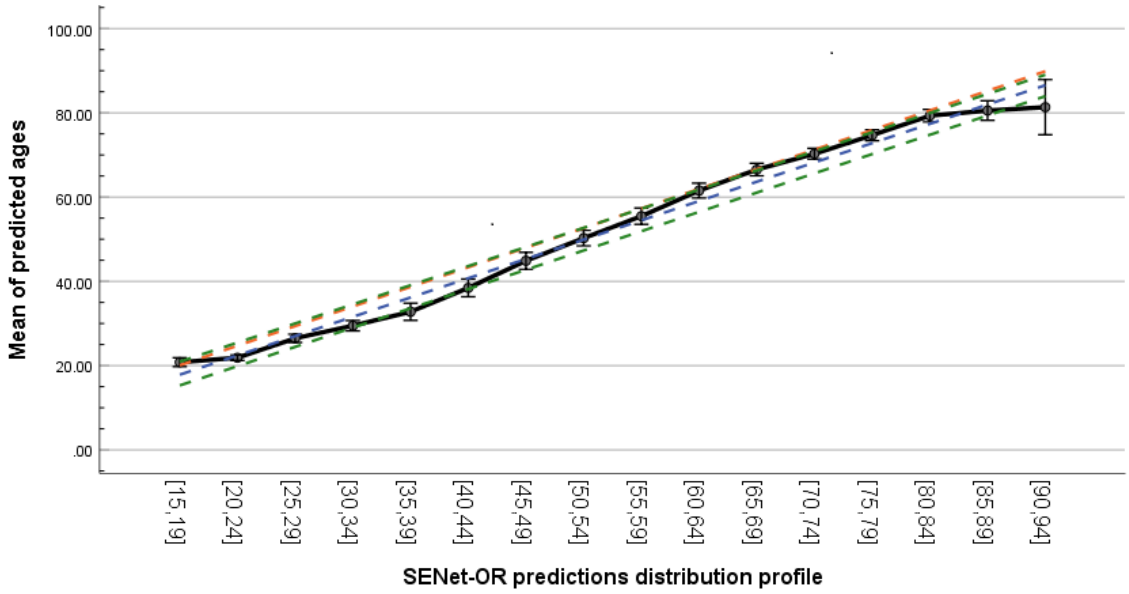

Figure 6.10: Means of predicted ages by the SENet-OR model for the intervals of fixed width of 5 years. Whiskers show 95% confidence interval over a mean. The last interval has length of 6 years to include the age of 92. Whiskers show 95% confidence interval over a mean. Orange line denotes the $y = x$ function, blue – linear regression fit function, green – the lines representing the lower and upper bounds of the 95% confidence interval over linear regression fit parameters.

The ordinal DEC method was first applied to the synthetic datasets described in Sections 4.1.2 and 4.1.3. In these datasets each sample has 2 labels - one for denoting sample class and one for a bin the sample belongs to. In order to test algorithm's performance for both datasets various "walls" position initialisations were used along with different number of clusters. These initialisations are summarised in Table 6.1. Wall search is performed

among the sample classes, not the bin classes, as shown in Figures 4.2 and 4.3. For all the initialisations the ordinal DEC method was able to recover original "wall" positions corresponding to the classes 2 and 5 in the ordinal MNIST dataset and classes 3 and 7 in the two-variable ordinal dataset. The ordinal DEC method was also able to recover original number of bins by only learning the true "wall" positions, i.e. for the number of "walls" initialised to a larger number of "walls" the learned "walls" positions are repeated to reflect true number of bins.

The ordinal DEC method was also trained on the HBAPS dataset. Training was performed with the hyperparameters tuned on synthetic data. The algorithm was, however, not able to effectively learn on this dataset as validation loss was not consistently decreasing during the training.

Bin	1	2	3	4	5	6	7	8	9	10	11	12
Age range	[17,20]	[21,26]	[27,30]	[31,41]	[42,46]	[47,50]	[51,56]	[57,62]	[63,69]	[70,81]	[82,85]	[86,92]

Table 6.2: Predicted brain ageing profile intervals from the SENet-MR classifier predictions.

Bin	1	2	3	4	5	6	7	8	9	10
Age range	[17,30]	[31,43]	[44,51]	[52,56]	[57,61]	[62,67]	[68,71]	[72,78]	[79,85]	[86,92]

Table 6.3: Predicted brain ageing profile intervals from the SENet-OR classifier predictions.

Bin	1	2	3	4	5	6	7	8	9	10
Age range	[17,22]	[23,27]	[28,39]	[40,45]	[46,51]	[52,67]	[68,71]	[72,76]	[77,85]	[86,92]

Table 6.4: Predicted brain ageing profile intervals obtained using the DEC method.

Bin		1	2	3	4	5
Number of clusters	9	[17,21]	[22,37]	[38,44]	[45,77]	[78,96]
	10	[17,21]	[22,37]	[38,44]	[44,67]	[68,96]
	12	[17,25]	[26,40]	[41,43]	[44,67]	[68,96]

Table 6.5: Predicted brain ageing profile intervals obtained using the PROMETHEE II method and convolutional part of the SENet-MR model as a feature extractor.

6.5 Discussion and conclusion

Four methods are used in this chapter for the purpose of understanding the non-linear nature of human brain ageing, i.e. identifying the periods of human adulthood when the largest structural brain changes occur. Two methods (DEC and PROMETHEE II) were based on descriptions in the existing literature and adopted for the task. The other two methods (the method considering the distribution of model predictions and ordinal DEC) are new and novel approaches proposed here for the first time.

As shown in Tables 6.2 and 6.3, the last four intervals for the oldest subjects produced by the SENet-MR and SENet-OR models are in agreement to within the model's resolution. SENet-OR grouped all the subjects younger than 30 years old into a single interval, while SENet-MR separated this range into three intervals. While not particularly clear, one paper suggested that cortical thinning, one of the structural markers of brain ageing, starts to affect the brain at the age around 30 years old [177], suggesting that this result may be of biological significance. As shown in Table 6.4, the ageing profile obtained using DEC agrees with the profile obtained using SENet-OR for the subjects older than 30 years old. Overall, all three profiles obtained agree within the accuracy achieved by the SENet-based models suggesting that they are all being driven by the same underlying brain features, and again suggesting that these features may be biologically meaningful.

The profiles identified by the PROMETHEE II method are presented in Table 6.5. The method was applied with the number of clusters pre-set to 9, 10 and 12 clusters. The method identified 5 intervals in the age range for all three initialisations. The interval walls found agree up to the SENet-MR model resolution with the walls found the method considering prediction distribution from both the SENet-MR and SENet-OR models and the DEC method. The PROMETHEE II method ranks the samples based on Euclidean distance between the features extracted from the samples whereas other methods used either consider the distribution of predictions or KL divergence based on Euclidean distance between the features. Latter difference may explain the fact that the PROMETHEE II method found 5 intervals compared to the 10 and 12 intervals found by the other methods.

The disadvantage of the PROMETHEE II method is that it uses a pre-set number of intervals. However, the method is capable of identifying whether where are less intervals than it was pre-set. The DEC method has the same feature with respect to the number of intervals in the data. On the other hand, the method considering the distribution of

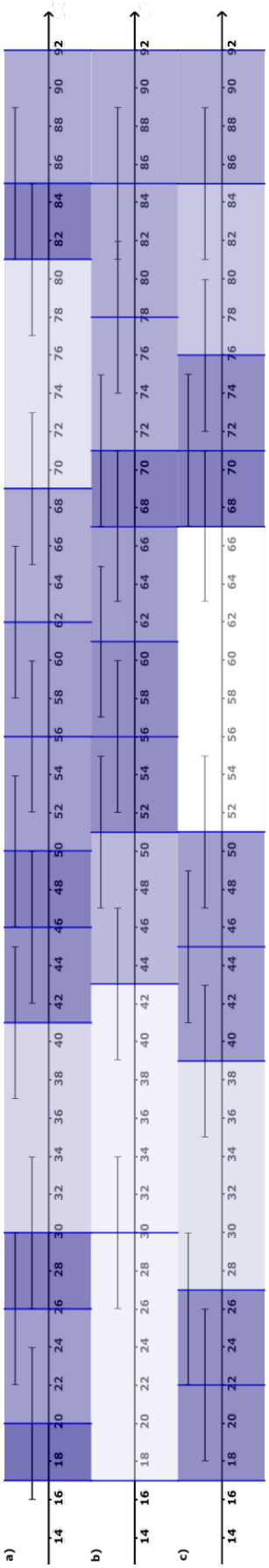

Figure 6.11: Population-wide brain ageing profiles. The profiles extracted by considering the distribution of predictions from the a) SENet-MR and b) SENet-OR models. c) The profile extracted using the DEC method. Error bars around each profile wall correspond to uncertainty around its value, i.e. 4 years as this is the SENet-MR and SENet-OR models' MAE. The colour of each bin corresponds to a slope calculated as $1/(interval\ length)$ – the darker the bin is, the more rapid the change is in the underlying age-related features.

model predictions has a disadvantage as it cannot define a number of intervals in the age range and this is only defined by a model resolution.

As there is no ground truth for either brain ageing profiles or trajectories, the profiles are compared to each other and those available in existing literature. It is, however, not possible to directly compare the results presented here to most of the existing trajectories and profiles, as existing publications do not assess whole brain ageing effects. Also, most studies to date have used relatively small cohorts (112 to 1581 subjects) which are unlikely to be comparable to the cohort described in Section 4.2.1 (a total of 8623 subjects), in terms of representation of population-wide effects. Only one study, by Vinke et al [171], reports the results of a similarly sized cohort of 10755 MRI samples. In terms of using purely cross-sectional data to define a brain ageing profile (as opposed to a trajectory), only the work by Shamir and Long [6], who investigated a cohort of 416 participants, is comparable to the current study.

Kuo et al [51] reported that degeneration of GM and increase of the CSF volume accelerated with age. A comparison of the current study's results with those by Kuo et al [51] can be performed by considering that in the profiles derived here, shorter intervals correspond to faster ageing-related changes. The trajectories for GM, WM and CSF found by Vinke et al [171] are similar to those of Kuo et al [51]. In both works the results indicate that changes accelerate with age. In this chapter the results presented in Figures 6.7 – 6.10 and Figure 6.11 indicate the rate of ageing-related changes throughout human adulthood instead of the significance of these changes. Therefore, smaller bins indicate faster changes while larger bins – slower changes. As it can be seen from Figure 6.7, the results of the current study do not match these results – the width of the bins does not consistently gets smaller as age increases. In this chapter the whole brain ageing effects were assessed and Kuo et al [51] and Vinke et al [171] assessed ageing on separate brain regions using volumetric measures. This means that regional-based methods have greater sensitivity compared to whole brain approach in measuring the rate of ageing.

There is a need for the method for extracting a brain ageing trajectory or profile to be able to take into account the fact that the difference between the subjects of 30 and 50 years of age is not equivalent to the difference between the subjects of 60 and 80 years of age [51]. Kuo et al addressed this limitation of ageing trajectories by considering linear combination of measures describing age. In the current chapter, the features are extracted using a convolutional part of a CNN pre-trained on the task of brain age prediction from

MRI scans. This ensures that the features used further in the analysis are specific to ageing, while allowing more complex trajectories to be investigated, as the feature extractor is a non-linear model.

On the other hand, there are disagreements among the authors in existing literature on the form of the trajectories. For instance, Good et al [173] found a linear decline in GM over lifetime compared to the results by Kuo et al [51] and Vinke et al [171] which identified changes accelerating with age. The authors attributed this discrepancy to the difference in segmentation and volume measurement techniques [173].

The only work in existing literature extracting a brain ageing profile using feature extraction from cross-sectional data (rather than a trajectory from longitudinal data) was performed by Shamir and Long [6]. Shamir and Long [6] plotted average predicted age for each true age label present in the dataset, and considered the gradient between the points on this graph as an indication of speed of ageing-related changes, i.e. a larger gradient corresponds to faster changes. The plot by Shamir and Long is presented in Figure 6.12. The authors found that the fastest changes occur from the age of 55 to 59 and from the age of 65 to 69. Figures 6.8 and 6.10 can be used to compare the results of this chapter to the results of Shamir and Long, as the distributions of predictions of the SENet-MR and SENet-OR models are plotted with respect to the same 5 year intervals as used by the authors. The authors used fixed intervals of 5 years and plotted means of predictions for these intervals. When the means of predictions of the SENet-MR and SENet-OR models are plotted with respect to the profiles described in Table 5.1 and 5.2 as in Figures 6.7 and 6.9 respectively, the ageing process appears more linearised and driven by a constant speed compared to the visualisation in Figure 5.12. When the means of predictions of the SENet-MR and SENet-OR models are plotted with respect to the profile as defined by Shamir and Long and shown in Figures 6.8 and 6.10 respectively, the process of brain ageing appears to be even more linearised compared to Figures 6.7 and 6.9. This can also be examined using the linear regression lines in Figures 6.7 – 6.10. For both the SENet-MR and SENet-OR models 95% confidence interval gets smaller for linear fit into the plot of mean predictions with respect to the intervals used by Shamir and Long [6] compared to the fit with respect to the intervals obtained in this chapter.

The profile presented in Figure 6.7 has fairly uniform slope over whole profile and the greatest gradient is present in the last interval. This may be explained by the fact that there are only three samples present in the HBAPS test dataset and this is also reflected

by a larger standard error bar on the mean of last interval. This absence of samples in this range leads to the fact that both the SENet-MR and SENet-OR models predicted the subjects in this interval to be younger. This is also a reason for the fact that the profile shown in Figure 6.9 obtained using the distribution of predictions of the SENet-OR model levels out the intervals containing younger subjects.

From the results of Shamir and Long [6] shown in Figure 6.12 and plotting the results in this thesis with respect to their interval length (Figures 6.8 and 6.10) highlight a linear nature of ageing can be deducted. On the other hand, the profiles obtained in this chapter and presented in Tables 6.2 – 6.5 and Figure 6.11. appear to highlight non-linear nature of brain ageing as they represent the rate of changes at each interval instead of the amount of changes.

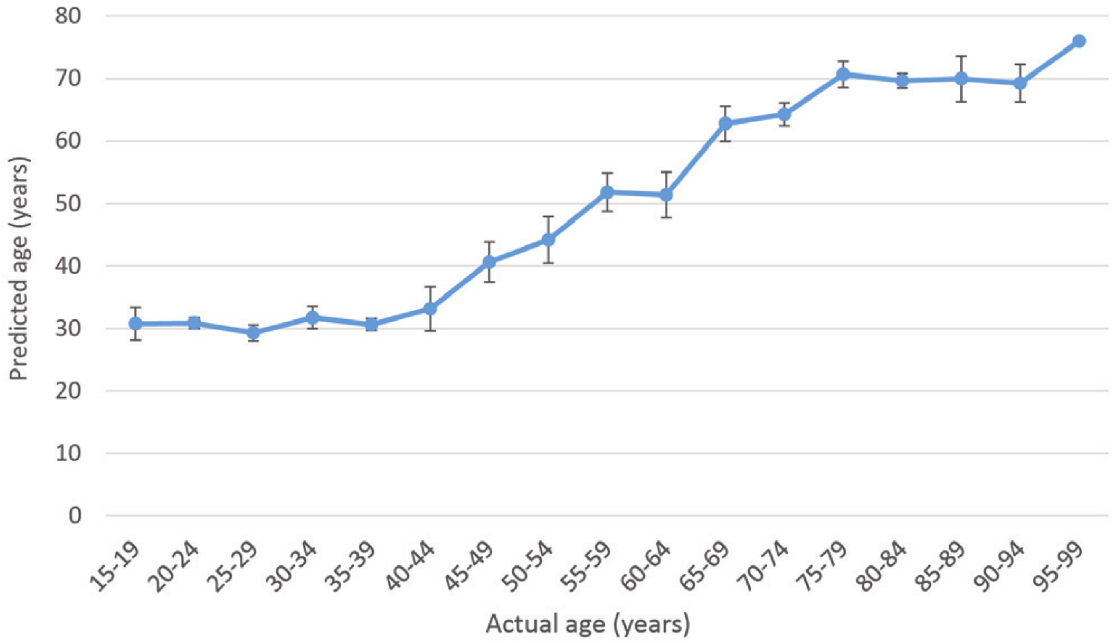

Figure 6.12: Results of Shamir and Long [6]: average predicted age with respect to 5-years bins of true age labels. Figure used with author's permission.

In the results presented in the current chapter a gradient for each ageing interval is also found by taking the inverse of the interval length. In Figure 6.11 the larger the gradient over an interval, the darker is the colouring. In the profile obtained by considering the distribution of predictions of the SENet-MR model, the interval corresponding to the greatest gradient, i.e. fastest ageing, is the first one which includes subjects from 17 to 20 years of age. In this profile, the second fastest ageing occurred in three intervals - from 27 to 30, from 47 to 50 and from 82 to 85. The profiles obtained by considering the

distribution of predictions of the SENet-OR model and of DEC have the same number of intervals, and the second fastest ageing interval corresponds to the ages from 68 to 71, while Shamir and Long [6] identified the fastest changes between the ages of 65 and 69.

The differences between the results presented in this chapter and the results of Shamir and Long [6] can be explained by two factors. Firstly, the cohort used by Shamir and Long included 416 subjects. A number of authors [223, 227, 13] have reported that performance of the brain age prediction models improves with increasing number of training samples suggesting that brain ageing process is affected by anatomical variability in the population. The cohort consisting of 416 subjects may not represent population-wide brain ageing effects due to anatomical variability of brain ageing. Secondly, the speed of brain ageing changes throughout human lifetime, i.e. the rate of ageing-related changes is larger compared to young subjects. Therefore, the length of intervals defined over whole range of ages would be varying. In this chapter rather the rate of changes is measured instead of the amount of changes. Fixing interval length as in the work by Shamir and Long [6] would affect model's ability to measure the rate of changes.

In conclusion, it should be noted that further work is needed to validate the profiles. Firstly, the DEC method can be further validated using longitudinal data. Longitudinal cohort of the ADNI dataset can be used for this purpose. This cohort has a limited number of healthy subjects and additional data collection may be needed. Data collection can, however, be replaced with training a Generative Adversarial Network (GAN) for augmenting data. The advantage of GAN is that it can be trained using a cross-sectional dataset and be used to augment longitudinal data [234]. On longitudinal data interval assignment can be performed and the samples' true labels can then be compared to the profile obtained using the HBAPS test data. Secondly, Vinke et al [171] and Kuo et al [51] have reported differences in ageing trajectories between men and women. In order to further explore population-wide effects male and female profiles could be extracted separately. Population-wide effects could also be explored further by expanding the study cohort.

Chapter 7

Extracting brain ageing related features

7.1 Main contributions

In this chapter extracting of the Magnetic Resonance Imaging (MRI) data features describing the process of brain ageing is described. The main contributions of this chapter are as follows:

1. five existing saliency mapping methods (vanilla backpropagation, guided backpropagation, the SmoothGrad method, Gradient Class Activation Mapping (Grad-CAM) and guided Grad-CAM) are applied to be compared on the task of extracting brain ageing-related featurs from MRI data as all the existing methods were adapted to work on 3D MRI data,

2. the method of ordinal saliency mapping is also proposed for applying saliency mapping methods to ordinal Convolutional Neural Network (CNN) described in Section 7.3.2,

3. all the subject-specific maps extracted for the test data of the Healthy Brain Ageing from Public Sources (HBAPS) dataset (details on the test data are provided in Section 4.2.1) are studied by averaging over different age ranges defined by the ageing profiles derived in Chapter 6 as brain ageing is a cumulative process,

4. saliency maps obtained using ordinal saliency mapping using ordinal CNN model are compared to the ones obtained using metric model,

5. all the results were interpreted for assessing biological significance by considering existing literature on brain ageing.

7.2 Introduction

In this thesis saliency mapping methods are considered for the task of brain age prediction to extract the features characterising population-wide ageing profile. For the methods to be used for this purpose, they should fulfil two conditions. Firstly, the methods should be unsupervised since there is no ground truth data available to train supervised saliency mapping methods. Secondly, the methods should be network architecture agnostic as saliency mapping needs to be applied on previously trained models for brain age prediction as described in Chapter 5.

In order to discuss the explicability of a network's output, consider a network for natural image classification (for example [74]). In the context of the natural image classification a "good" output explanation for a predicted class should have two main properties – class-discriminativeness and high-resolution. An explanation being class-discriminative means that it localises an object of the target class in an input. The explanation has high-resolution if it can capture even fine-grained details [43].

Gradient visualisation methods, which consider a gradient of an output with respect to an input, [42, 45, 46], provide high-resolution maps as they are produced in the pixel-space of the input, but experiments show that the explanations are not class-discriminative. The Layer-wise Relevance Propagation (LRP) method was proposed to tackle this issue. The LRP technique propagates relevance from an output towards the pixel-space of an input by applying pre-specified calculation rules [7]. Gu et al [126] have later shown that even though LRP improves the quality of explanations overall, its class-discriminativeness on the classification tasks still suffers due to the method of calculation in LRP. Localisation methods, which consider feature maps inside a network, were proposed to improve class-discriminativeness of the explanations. The first localisation method proposed was the Class Activation Mapping (CAM) [5], which produced low resolution saliency maps by considering a weighted sum of the last convolutional layer feature maps. As well as being low resolution, the CAM method also only allowed explaining networks which do not contain fully-connected layers. In order to make CAM applicable to any network architecture, Selavaraju et al [43] proposed the Grad-CAM method. Following this, the resolution of the

CAM and Grad-CAM explanations was improved by introducing the Guided Grad-CAM method, combining guided backpropagation [46] with Grad-CAM.

Another family of saliency mapping methods applies *occlusions*, or *perturbations*, to a network input. Zeiler and Fergus [47] proposed a method to "invert" a CNN, i.e. to create a deconvolutional network, and to propagate the information backwards through it. In the method by Zeiler and Fergus, a saliency map is computed by using the deconvolutional network and an occlusion technique, i.e. systematically masking out regions within an input in order to determine their significance in determining the output. Zhang et al [235] also used an occlusion approach to calculate saliency maps. They proposed a method of *excitation backpropagation*, which passes information from an output towards an input, i.e. top-down, using a probabilistic *Winner-Take-All process*. The main disadvantage of such an excitation backpropagation method is that it is only applicable to the classification network. In addition, a disadvantages of all occlusion-based techniques is that producing saliency maps is slower compared to gradient and feature map visualisation methods, as backward pass through a network needs to be computed a number of times equal to the number of input patches. A saliency mapping method exploiting *perturbation* was proposed by Fong and Vedaldi [236]. The method identifies the most important input regions by understanding masking of which regions prevents a network from correct classification of an input. There are a number of issues associated with using the method. Firstly, it is only applicable to classification networks. Secondly, the masks produced by the method have to undergo a quality check procedure, which involves manual setting of a threshold to choose between the good and bad masks. Thirdly, the method requires additional training and a hyperparameter search for the loss function used in this training stage. Thibeau-Sutre et al [237] used this approach for identifying the regions corresponding to the Alzheimer's disease in the inputs to the network trained to classify between healthy and diseased subjects.

Saliency mapping methods have also previously been applied to the task of brain age prediction. In the first attempts to identify the regions related to ageing, methods not involving neural networks were used, many of which have internal parameters which can be used to identify such regions. Beheshti et al [222] used a patch-based grading procedure to calculate a measure to represent cortical ageing. In the calculation, Support Vector Regression (SVR) is used and SVR weights are used to construct an *importance map*. In this chapter the terms "importance map" and "saliency map" are used interchangeably.

Kondo et al [182] and Fujimoto et al [181, 219] took similar approach and used Relevance Vector Regression (RVR) weights to produce importance maps. Aycheh et al [220] pre-processed data to extract cortical thickness, put the data through the Sparse Group Lasso (SGL) algorithm and used its outputs as an input to the Gaussian Processes (GP) model to estimate subjects' age. SGL coefficients were then used to produce importance maps. Amoroso et al [13] used a Multi-Layer Perceptron (MLP) to predict sample's brain age, and in order to identify important brain regions the Gedeon method [238] was used. This method produces importance maps by considering the weights connecting the input to the first two hidden layers of a MLP.

Huang et al [239] used a deep learning model for brain age prediction and an occlusion technique to produce saliency maps. The inputs to a network were occluded before being entered into a trained network. The importance of a region was then taken to be proportional to prediction error induced by occlusion. The authors tested two occlusion region sizes - 10×10 and 20×20 pixels in 2D slices of MRI scans, and found that occlusion size did not have an effect on the prediction error. Huang et al [239] used this information to conclude that the network used had a robust performance with respect to an input MRI scan quality. Popescu et al [38] developed a method for voxel-wise brain age prediction using deep GP and, therefore, saliency mapping was intrinsic to their methodology. Neither Huang et al [239] or Popescu et al [38] analysed the brain regions detected and their biological relevance for brain age prediction. Herent et al [52] used T1-weighted MRI data segmented into Cerebrospinal Fluid (CSF), Grey Matter (GM) and White Matter (WM), and then calculated average correlation maps with age predictions over segmented inputs. The authors used a ridge regression model along with deep learning methods. This allowed creation of weight maps from the ridge regression model.

All the above-mentioned attempts on understanding brain ageing-related features, however, suffer from representing the features over whole dataset, i.e. averaging saliency maps over whole age range. Brain ageing, however, is an cumulative process and should be studied longitudinally over the human lifetime. Herent et al [52] attempted to address this by creating two average saliency maps - for the subjects younger than 30 and older than 60 years of age, using the method by Zeiler and Fergus [47]. Bermudez et al [183] used the Grad-CAM method [43], which is described in detail in Section 2.3.6, to find most important regions of the GM maps for brain age prediction. The authors produced subject specific maps and averaged over all subject-specific maps to produce the map of regions

most important for the network to produce an output. This does not, however, describe age-specific features. Bermudez et al [183] also investigated change in averaged subject specific maps for different age groups.

The contents of this chapter address the issues of extracting brain ageing-related features from MRI data. As both metric and ordinal CNN models described in Chapter 5 were trained on the data with minimal preprocessing, in this chapter it is possible to try to understand the features of original inputs which correspond to ageing. In the above-mentioned works, importance maps were generated using only one method in each study and the results were not systematically compared to each other. All the methods used were unsupervised, as there is no existing ground truth on structural features related to brain ageing. As such, the methods can only be considered to be extracting biologically relevant ageing features if a number of them consistently detect the same features, which can be interpreted in terms of existing biological knowledge. In this thesis I have therefore compared five existing saliency mapping methods – vanilla backpropagation, guided backpropagation, the SmoothGrad method, Grad-CAM and guided Grad-CAM. All the results were also interpreted for biological significance by considering existing knowledge on brain ageing. The methods were adapted to work on 3D input MRI data, as all of them were originally developed to work on natural images. In order to understand the features in light of a fact that brain ageing is an cumulative process, subject-specific maps were averaged over a number of age ranges, to produce the maps describing the ageing profiles obtained in Chapter 6. The method of ordinal saliency mapping was also developed to apply saliency mapping methods to ordinal CNN described in Section 7.3.2. Subject specific importance maps were again obtained and averaged to describe the ageing profiles. Furthermore, the importance maps were compared to those generated using the metric model for consistency.

7.3 Methods

7.3.1 Existing saliency mapping methods

In this thesis, five existing saliency mapping methods are applied to the task of brain age prediction in order to extract MRI data features describing ageing. The methods considered are vanilla backpropagation [42], guided backpropagation [46], the SmoothGrad method [45], Grad-CAM [43] and guided Grad-CAM [43]. All the methods chosen are

widely used for saliency mapping both on the tasks handling natural images or medical data (for example, [237]).

The vanilla gradients method is used as it is a baseline method of gradient-based saliency mapping techniques. The method is described in Section 2.3.6. The method is based on exploring raw gradients with respect to an input. However, it has been previously shown that the explanations, i.e. saliency maps, produced by purely gradient-based methods are noisy. Therefore, a number of methods have been proposed to tackle the issue. A gradient with respect to an input in Equation 2.35 defining the vanilla gradient visualisation method may suffer from large fluctuations occurring due to small changes in a network's output. Therefore, it can be hypothesised that noise in resulting maps may originate from meaningless changes in derivatives with respect to an input. If the network being subject to analysis uses Rectified Linear Unit (ReLU) activation function (defined in Equation 2.6), then the gradient in Equation 2.35 will be discontinuous. The SmoothGrad technique, which is also used in this chapter, tackles this issue of noise in vanilla gradients as described in Section 2.3.6. Guided backpropagation was used instead of the deconvolutional network approach by Zeiler and Fergus [47], as it does not require any information about max-pooling to be saved on a forward pass.

From the localisation saliency mapping methods available, the Grad-CAM and guided Grad-CAM methods were chosen for use in this chapter. The CAM technique was not used, as Selvaraju et al [43] have previously shown that the Grad-CAM and guided Grad-CAM methods produce better importance maps compared to the CAM.

The LRP method was not applied for extracting brain ageing related features, because of its dependence on manually adjusted parameters. Due to the depth of the Squeeze-and-Excitation Network (SENet)-based architectures, the composite LRP rule (described in Section 2.3.6) is needed. In the calculations using the composite LRP rule at least three rules would have to be applied to produce a saliency map. For example, a LRP-0 rule is applied in the deepest layers, i.e. the ones closest to an output. Then LRP-ε is applied to the middle layers, and LRP-γ in the shallower layers, i.e. the ones closest to an input layer, and w^2-rule for the first network layer. With this setup, the γ parameter must be adjusted manually. Alternatively, the middle layers LRP-γ can be replaced by the LRP-$\alpha\beta$ and $z^{\mathcal{B}}$-rules instead of the w^2-rule, but again there are free parameters that must be adjusted manually. In tasks analysing natural images, the parameters can be adjusted for each task as objects are detectable by eye. Since ground truth is not available for the task

of brain age prediction, arbitrarily adjusting the parameters may lead to an ambiguity, i.e. information either can be lost in saliency maps or they can become oversensitive and display the features not related ageing.

Implementation and adaptation for 3D data

Local implementation of the existing saliency mapping methods is was based on the implementation provided by Ozbulak [240]. The code for the guided backpropagation and Grad-CAM methods was modified to perform value masking in the ReLU layers as defined in Equation 2.41. The modifications were required in order for the implementation to work on the SENet architecture (described in Section 2.3.2), which has residual connections including SE-blocks.

Implementing saliency mapping for natural image classification As a preliminary test of implementation, the saliency mapping methods described in Section 2.3.6 (specifically vanilla backpropagation, guided backpropagation, Grad-CAM and guided Grad-CAM were used) are first applied to the task of classifying natural images from the ImageNet dataset described in Section 4.1.3. AlexNet [74] pretrained on ImageNet (available through the Pytorch library [29] is used to predict one of the 1000 labels present in the dataset for each image. The details of this architecture can be found in the work by Krizhevsky et al [74], but are omitted here as they only relate to this initial testing and are not relevant for the main body of MRI analysis. In order to compare the methods, the two classes with the highest classification accuracy ("tiger" and "zebra") were considered.

For the SmoothGrad method, three different values were used for standard deviation in Equation 2.36 - $\sigma^2 \in \{0.05, 0.15, 1.0\}$. The LRP method was applied using the same rule throughout the model. (This is acceptable for AlexNet, as this is shallow compared to the SENet architecture used for the MRI analyses). There are 5 convolutional layers, three of which are followed by max-pooling operation, and three fully-connected layers [74]. In the MRI calculations either LRP-ε or LRP-β rules are applied. The LRP-ε and LRP-β rules are described in Table 2.1 in Section 2.3.6. The LRP-β rule is defined as the LRP-$\alpha\beta$ in Table 2.1 but only β value is varied and $\alpha = 1 - \beta$. The LRP-β rule was used using three different β values - $\beta \in \{0.05, 0.5, 0.95\}$.

Implementing saliency mapping for the task of brain age prediction Existing saliency mapping methods including the methods of vanilla backpropagation, SmoothGrad, guided backpropagation, Grad-CAM and guided Grad-CAM (described in Sections 2.3.6) were applied to the task of brain age prediction using the SENet-MR model (described in Section 5.3). The technique for saliency mapping of ordinal models is applied using the methods of vanilla backpropagation and Grad-CAM and use the SENet-OR model as described in Section 5.3.2.

7.3.2 Saliency mapping for an ordinal CNN

All the existing methods used in this chapter were originally proposed for the networks producing a single output vector. In Chapter 5, an ordinal CNN model is used to predict a sample as belonging to one of the K classes by combining the outputs of $K-1$ binary classifiers. Therefore, a technique is proposed here in order to allow saliency mapping methods to be applied to such an ordinal CNN. (The ordinal CNN itself is described in Section 5.3.2).

Assume an i^{th} input, which belongs to the input dataset as $x_i \in X$ with correponding outcome space $y_i \in Y = \{r_1, r_2, \ldots, r_K\}$, where $r_K > r_{K-1} > \ldots > r_1$ are ordered ranks. For each rank $r_k \in \{r_1, r_2, \ldots, r_{K-1}\}$ a corresponding binary classifier out of $K-1$ classifiers predicts whether for this i^{th} sample $y_i > r_k$. In order to reflect cumulative nature of ageing the rank for a sample x_i is predicted as

$$h(x_i) = r_q$$

$$q = 1 + \sum_{k=1}^{K-1} f_k(x_i), \tag{7.1}$$

where $f_k(x_i) \in \{0, 1\}$ is the classification result of the k^{th} binary classifier for a sample x_i. Here a virtual concept O is introduced, which models the features corresponding to all binary classifiers for which output is 1, i.e. for which label, y_i, of a sample x_i is larger than r_k and $f_k(x_i) = 1$. The classifiers for which output is 0 contribute to a dual virtual concept $\bar{O}$, which then models the features corresponding to all ranks for which $r_k > y_i$.

A schematic representation of the Contrastive Layer-wise Relevance Propagation (CLRP) method is shown in Figure 7.1. Consider the weights connecting the k^{th} clas-

sifier output, c_k^{max} ($c_k^{max} = max(\boldsymbol{o}_k)$, i.e. maximum value of k^{th} output vector, to an input in a network consisting of L layers - $\boldsymbol{W} = \{\boldsymbol{W}^1, \boldsymbol{W}^2, ..., \boldsymbol{W}^{L-1}, \boldsymbol{W}_k^L\}$, where

- $\boldsymbol{W}^L$ are the weights connecting the $(L-1)$-th layer and the L-th layer,
- $\boldsymbol{W}_k^L$ are the weights connecting the $(L-1)$-th layer to c_k^{max}.

The backward pass of a saliency mapping method of choice for an output binary classifier, c_k, can be defined as $f_M^k(\boldsymbol{x_i}, \boldsymbol{W}, c_k)$. The virtual concept, O, which represents the features of $\boldsymbol{x_i}$ related to its label y_i, can be then shown by a saliency map resulting from summing over all maps from the classifiers outputting 1, i.e. $f_k(x_i) = 1$:

$$\boldsymbol{R} = f_M(\boldsymbol{x_i}, \boldsymbol{W}, c_k^{max}) = \sum_{k, f_k(\boldsymbol{x}_i)=1} f_M^k(\boldsymbol{x_i}, \boldsymbol{W}, c_k^{max}). \tag{7.2}$$

Furthermore, $\bar{O}$, the dual virtual concept of O, models the features corresponding to the classifiers not contributing to predicted rank q, i.e. all the classifiers outputting 0 and $f_k(x_i) = 0$. The concept $\bar{O}$ is a concept opposite to O. For example, if there are K possible ranks in the dataset and an i^{th} sample has a predicted rank $y_i = r_k$. The concept O is formed by applying saliency mapping to all classifiers c_l with $1 <= l <= k$ and $\bar{O}$ - by applying it to all the classifiers c_l with $l > k$. The concept $\bar{O}$ represents features of $\boldsymbol{x_i}$ not related to the label y_i and is represented by a saliency map resulting from summing over all the saliency maps constructed by backpropagating from the binary classifier outputting 0, i.e. $f_k(x_i) = 0$:

$$\boldsymbol{R_{dual}} = f_{\bar{M}}(\boldsymbol{x_i}, \boldsymbol{W}, c_k^{max}) = \sum_{k, f_k(x_i)=0} f_M^k(\boldsymbol{x_i}, \boldsymbol{W}, c_k^{max}). \tag{7.3}$$

In Figure 7.1 red line corresponds to the visual concept $\bar{O}$. Then a saliency map for the ordinal technique is defined as

$$\boldsymbol{R}_{ordinal} = \boldsymbol{R} - \boldsymbol{R_{dual}}. \tag{7.4}$$

Considering the difference $\boldsymbol{R} - \boldsymbol{R}_{dual}$ in Equation 7.4 cancels the common features and ensures that only most descriptive neurons are included in resulting map.

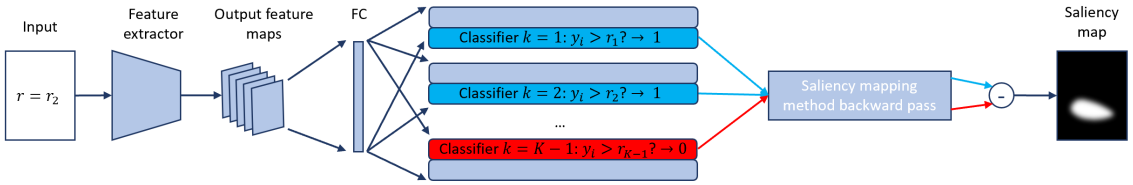

Figure 7.1: Schematic of the ordinal saliency mapping method. For the illustration of how the method produces a saliency map a synthetic input of rank r_2 out of K ranks is present. An input is put through a convolutional feature extractor and a fully-connected layer to produce an output from $K-1$ binary classifiers. Then a backward pass of a saliency mapping method of choice is performed from all classifiers. In order to construct a final saliency map, the maps resulting from the classifiers are subtracted such that the ones corresponding to 0 output are subtracted from the ones corresponding to 1.

7.3.3 Constructing saliency maps describing the population-wide ageing profiles

Here the pipeline is presented to obtain saliency maps describing the population-wide ageing profiles generated in Chapter 6. Such saliency maps describe average features relating to each interval of ageing. We consider the profile as a set of intervals, each containing a number of bins, $I = [i_1, .., i_B]$, where B is the number of bins, and each bin contains set of ages such that $i_b = \{a_1, ..., a_L\}$, where L is the bin length. First, the subject-specific maps, S_{x_i} for $x_i \in X_{test}$, are generated for all the samples in the test dataset X_{test} (described in Section 4.2.1). Secondly, to produce a saliency map describing bin i_b, the subject-specific maps are averaged over all samples labelled with ages belonging to that bin. Finally, the resulting map is multiplied by the MNI152 head mask in order to remove noise in regions outside the brain. (This is possible as data preprocessing (described in Section 4.2.3) included registering each sample onto the MNI152 template).

7.4 Results

7.4.1 Saliency mapping for natural image classification

In Figure 7.2 the results are presented for the two samples with the highest and lowest classification confidences for each label. All four input images have correctly predicted labels. The best performance is shown by the methods of guided backpropagation, Grad-CAM and LRP. On the images, which are hard to classify, i.e. for which the difference between the maximum and minimum scores in an output vector before the softmax function is small, the maps obtained by all the methods used identify the regions not related to

the object of interest are highlighted. This may be explained by the fact that for both "zebra" and "tiger" classes the network considered the stripes feature to be important. In the samples with corresponding low classification confidence the stripes is detectable in the input regions which do not correspond to the object of interest.

7.4.2 Saliency mapping for the task of brain age prediction

In Figures 7.3 – 7.5 the results of applying each saliency mapping method to the task of brain age prediction are presented. Figures 7.3 – 7.5 show saliency mapping results for samples with a range of high and low prediction errors. Figure 7.3 shows saliency maps for the sample with the lowest prediction error for the SENet-MR model. The subject is 28 years old and prediction error achieved was 0.0008 years using the SENet-MR model and 1 year using the SENet-OR model. Figure 7.4 shows saliency maps for a sample with a prediction error approximately equal to the sum of Mean Absolute Error (MAE) of the SENet-MR model and its standard deviation, 9.55 years. The subject is 52 years old, and the prediction error achieved was 9.54 years using the SENet-MR model and 7 years using the SENet-OR model. Figure 7.5 shows saliency maps for a sample with the largest prediction error by the SENet-MR model. The subject is 18 years old and prediction error achieved using the SENet-MR model is 60.2 years and 28 years using the SENet-OR model. (This extremely large error is most likely because the image was affected by preprocessing step (described in Section 4.2.3), which resulted in a size of $91 \times 109 \times 91$ voxels rather than the $196 \times 232 \times 188$ for all other samples in the HBAPS dataset).

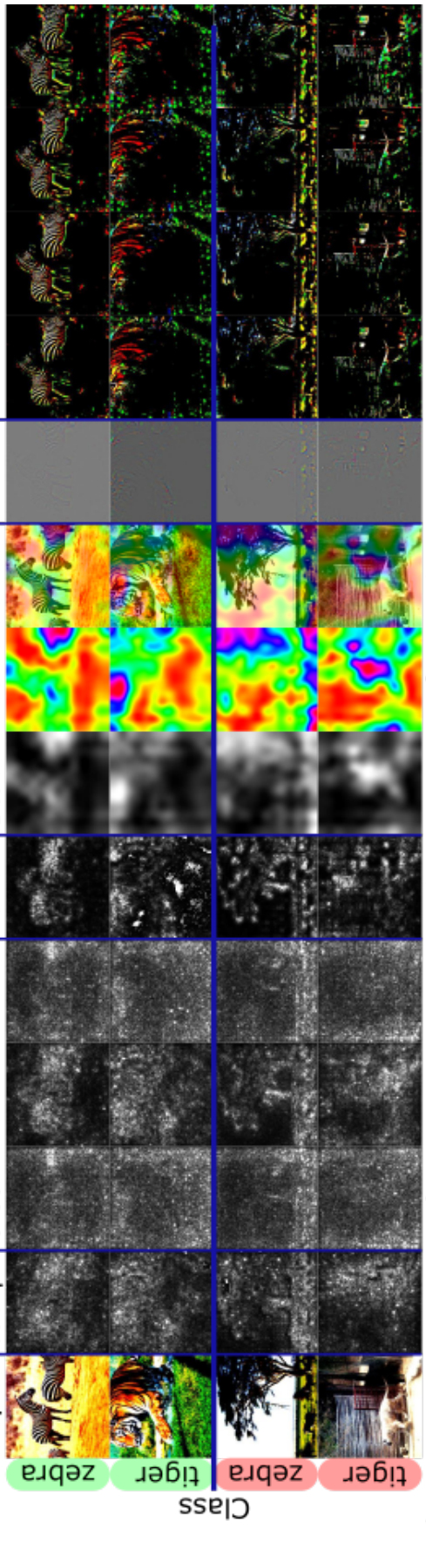

Figure 7.2: Results of applying existing saliency mapping methods to the task of classifying natural images from the ImageNet dataset. The upper two input examples denoted by green were the samples the network was the most confident about and the ones denoted by red are related to the largest uncertainty in classification. Uncertainty in sample classification can be defined as the difference between the maximum and minimum scores in an output vector before the softmax function. These values were 24.57, 22.70, 2,93 and 2.88 respectively from the top to bottom samples. Predicted labels were identified correctly for all four samples. The second, sixth and tenth columns show the saliency maps produced using vanilla backpropagation, guided backpropagation and guided Grad-CAM method respectively. The SmoothGrad method was applied using three different values of standard deviation in Equation 2.34 -$\sigma^2 \in \{0.05,\ 0.15, 1.0\}$. For the Grad-CAM method saliency maps were presented in grayscale and as heat maps (also overlaid on original image). For the LRP method both the LRP-ϵ and LRP-β rules were applied. The LRP-β rule was applied with three values of β - $\beta \in \{0.05, 0.5, 0.95\}$.

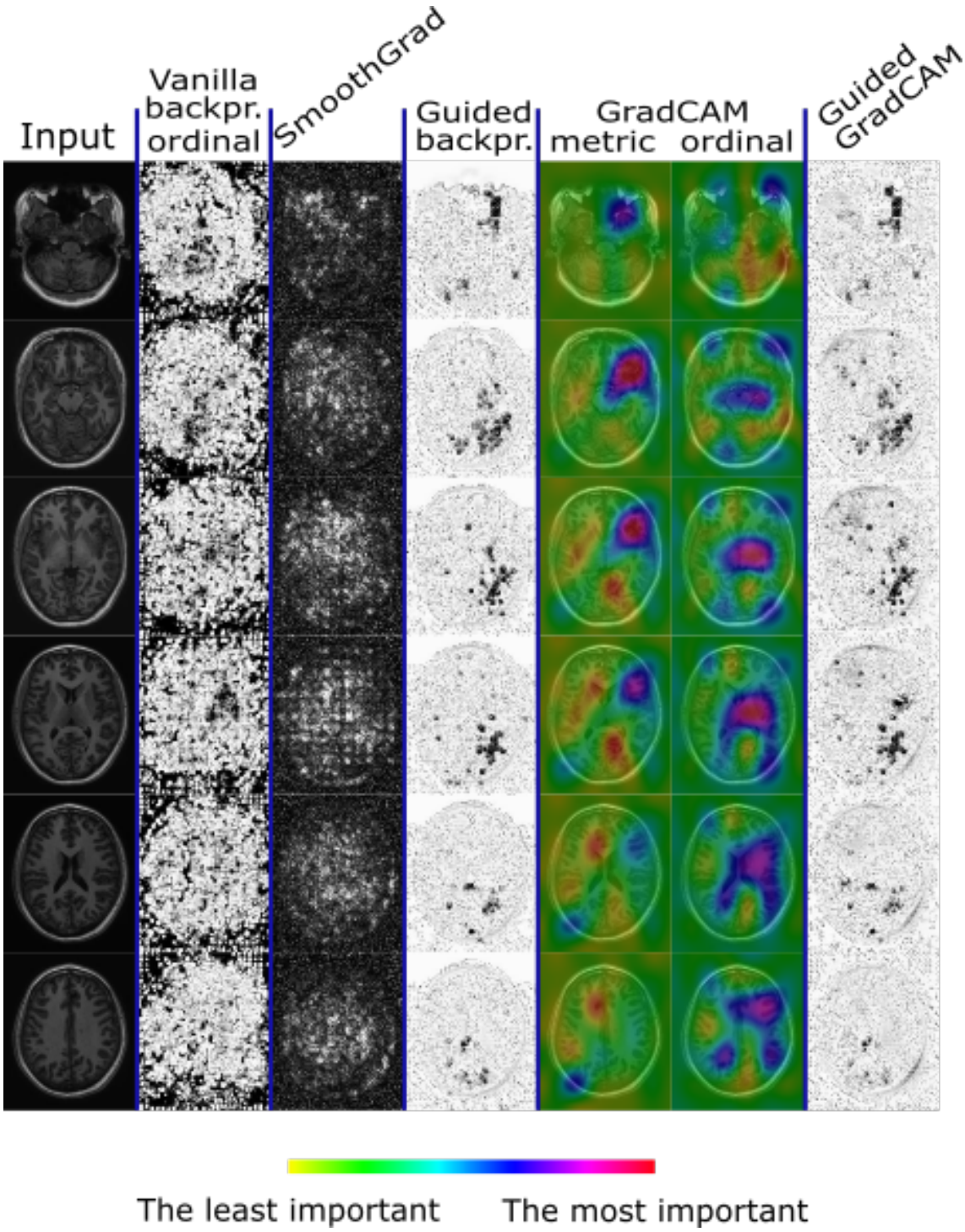

Figure 7.3: Results of applying existing saliency mapping methods to the task of brain age prediction. Saliency maps are shown for the sample with the lowest prediction error (0.0008 years) in the HBAPS testing data using the SENet-MR model. The maps named ordinal are produced by applying ordinal saliency mapping technique described in Section 7.3.2 with corresponding method. These maps are produced using the SENet-OR model and SENet-MR model is used otherwise. The leftmost column shows corresponding planes in the input sample.

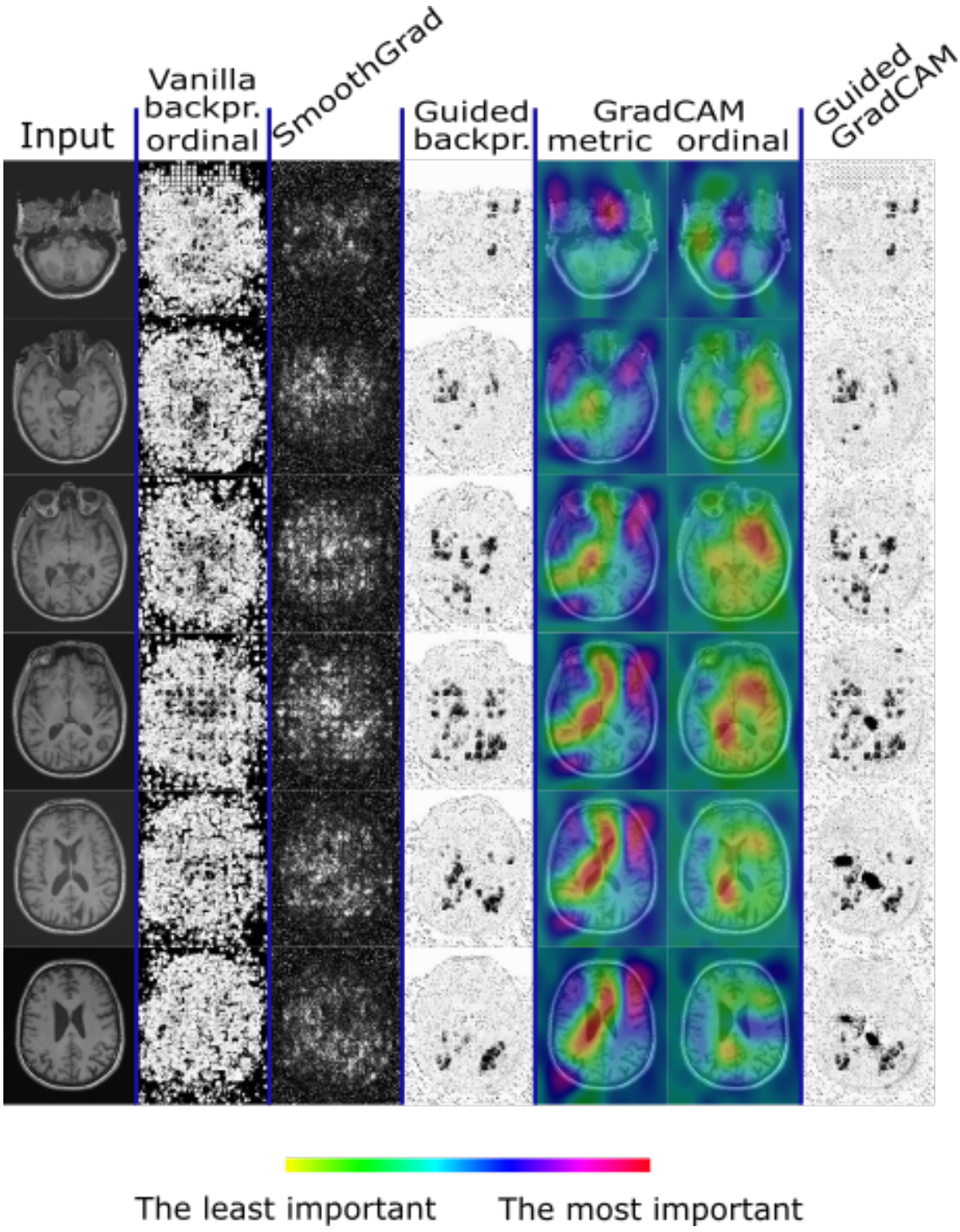

Figure 7.4: Results of applying existing saliency mapping methods to the task of brain age prediction. Saliency maps are shown for the sample with the prediction error approximately equal to the sum of MAE and standard deviation on it, 9.55 years, in the HBAPS testing data using the SENet-MR model. The maps named ordinal are produced by applying ordinal saliency mapping technique described in Section 7.3.2 with corresponding method. These maps are produced using the SENet-OR model and SENet-MR model is used otherwise. The leftmost column shows corresponding planes in the input sample.

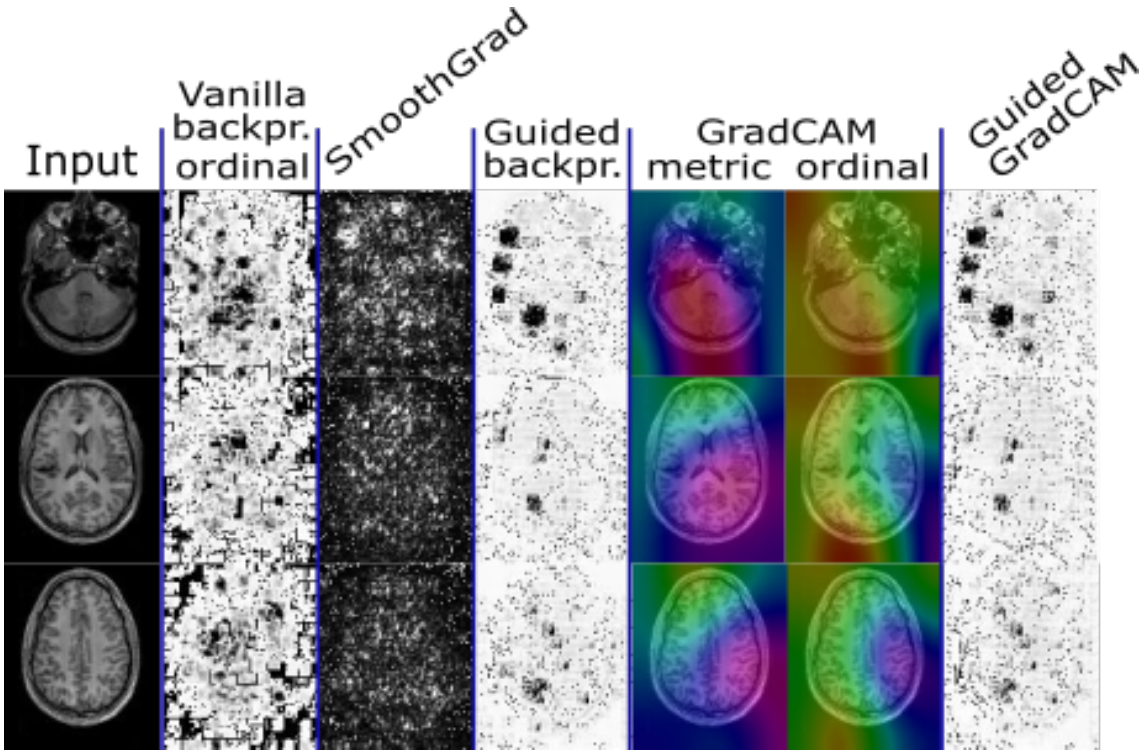

Figure 7.5: Results of applying existing saliency mapping methods to the task of brain age prediction. Saliency maps are shown for the sample with the largest prediction error (60.2 years) in the HBAPS testing data using the SENet-MR model. The maps named ordinal are produced by applying ordinal saliency mapping technique described in Section 7.3.2 with corresponding method. These maps are produced using the SENet-OR model and SENet-MR model is used otherwise. The leftmost column shows corresponding planes in the input sample. Three planes are presented from the samples as resulting saliency maps are less informative compared to the samples with smaller assciated prediction error.

Figures 7.6 - 7.16 show average saliency maps for population-wide brain ageing profiles obtained in Chapter 6. Figures 7.6 and 7.7 show the results of using five saliency mappings methods (vanilla and guided backpropagation, SmoothGrad, Grad-CAM and guided Grad-CAM) used to generate average saliency maps for profiles produced using the Deep Embedded Clustering (DEC) method and the method considering the distribution of predictions of the SENet-MR model. Average maps were produced as described in Section 7.3.3 and are shown with respect to a single representative plane in the MNI152 template. Figures 7.8 - 7.10 and 7.11 - 7.13 show the transverse, saggital and coronal views of the saliency maps produced using the Grad-CAM method for the profiles produced using the DEC method and the method considering the distribution of predictions of the SENet-MR model. These two profiles highlight the importance of ventricles in younger subjects and

the importance of subarachnoid space in older ones. The profiles extracted using these two methods are very similar as the same feature extractor is used – convolutional part of the SENet-MR model. The maps also highlight the importance of the frontal lobe region in older subjects.

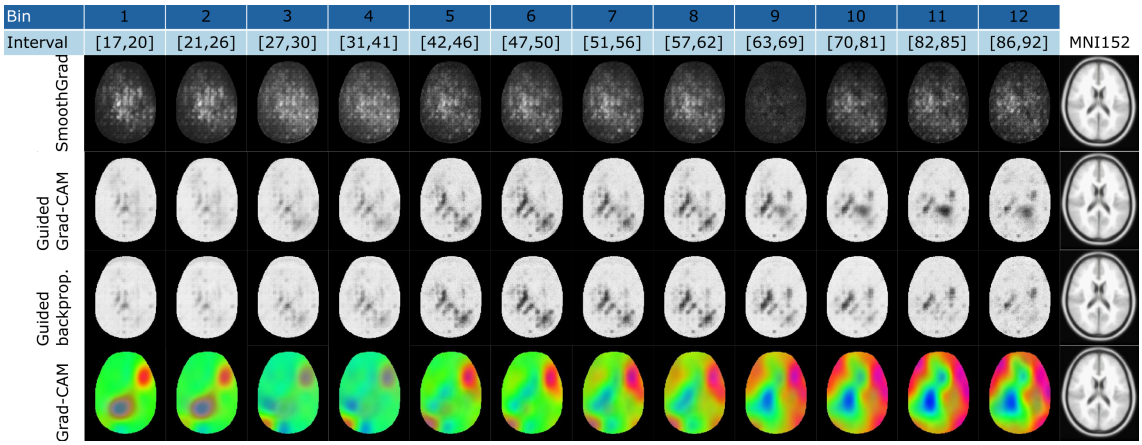

Figure 7.6: Average saliency maps obtained as described in Section 7.3.3 for the profile constructed using the distribution of predictions of the SENet-MR model (details in Section 6.3.1). The maps are constructed using four saliency mapping methods – guided backpropagation, SmoothGrad, Grad-CAM and guided Grad-CAM. The rightmost column shows corresponding plane in the MNI152 template.

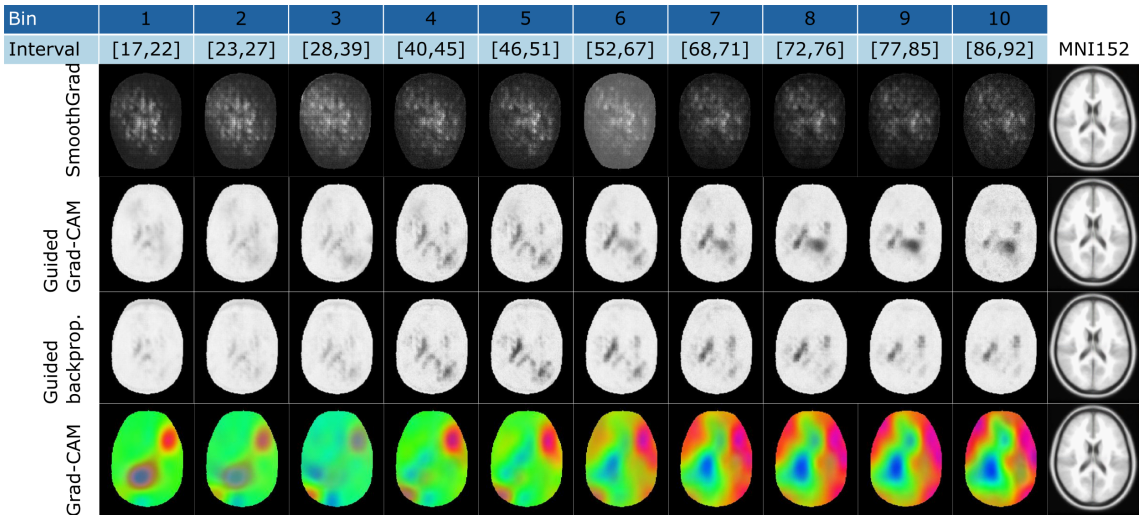

Figure 7.7: Average saliency maps obtained as described in Section 7.3.3 for the profile constructed using the DEC method (details in Section 6.3.3). The maps are constructed using four saliency mapping methods - guided backpropagation, SmoothGrad, Grad-CAM and guided Grad-CAM. The rightmost column shows corresponding plane in the MNI152 template.

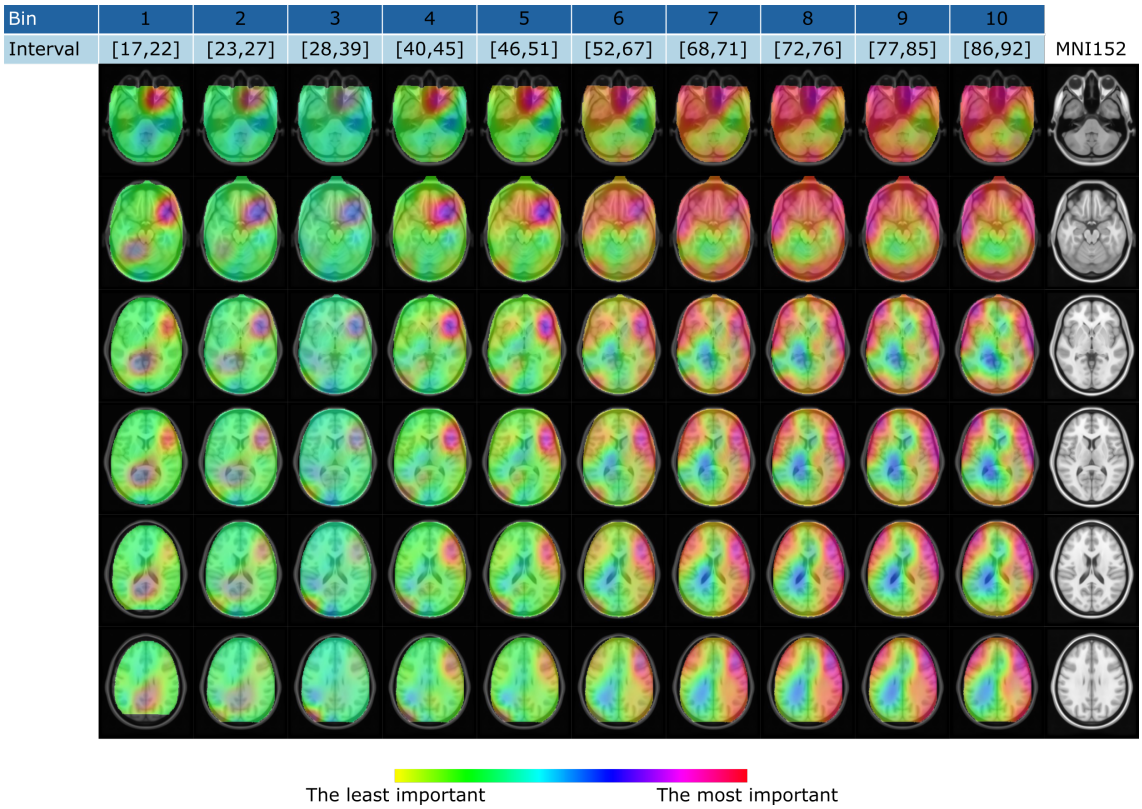

Figure 7.8: Average saliency maps in transverse plane obtained as described in Section 7.3.3 for the profile constructed using the DEC method (details in Section 6.3.3). The maps are constructed using the Grad-CAM method. The rightmost column shows corresponding plane in the MNI152 template.

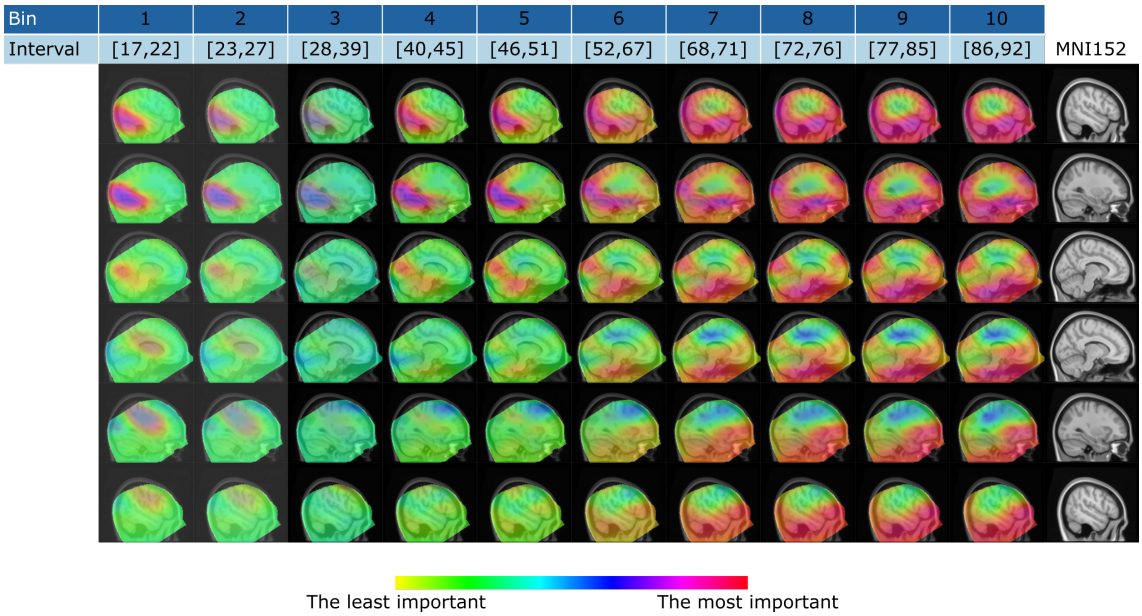

Figure 7.9: Average saliency maps in saggital plane obtained as described in Section 7.3.3 for the profile constructed using the DEC method (details in Section 6.3.3). The maps are constructed using the Grad-CAM method. The rightmost column shows corresponding plane in the MNI152 template.

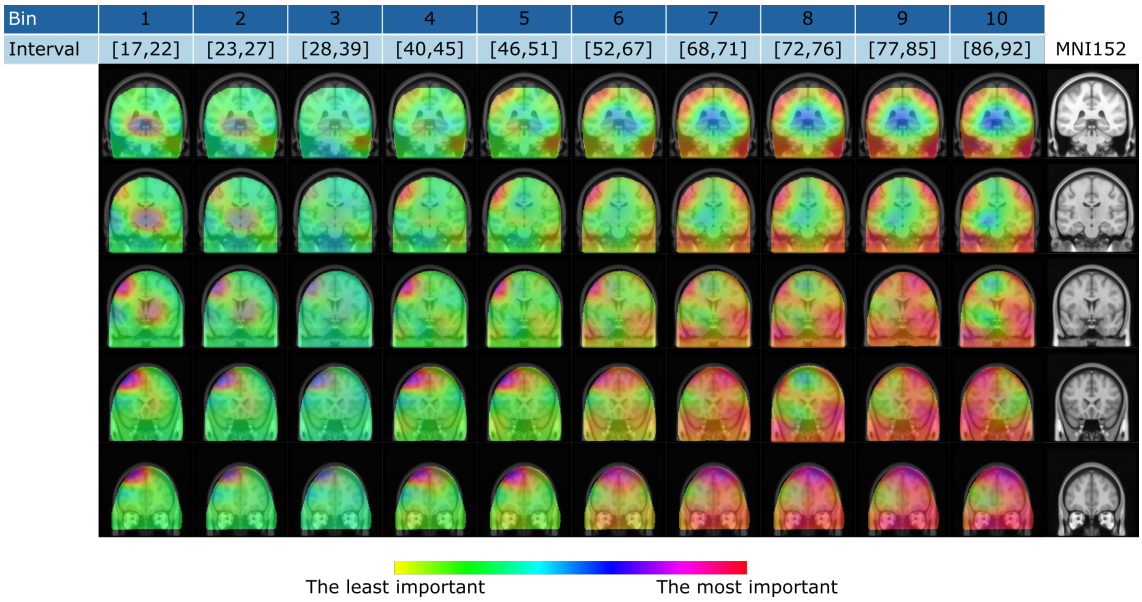

Figure 7.10: Average saliency maps in coronal plane obtained as described in Section 7.3.3 for the profile constructed using the DEC method (details in Section 6.3.3). The maps are constructed using the Grad-CAM method. The rightmost column shows corresponding plane in the MNI152 template.

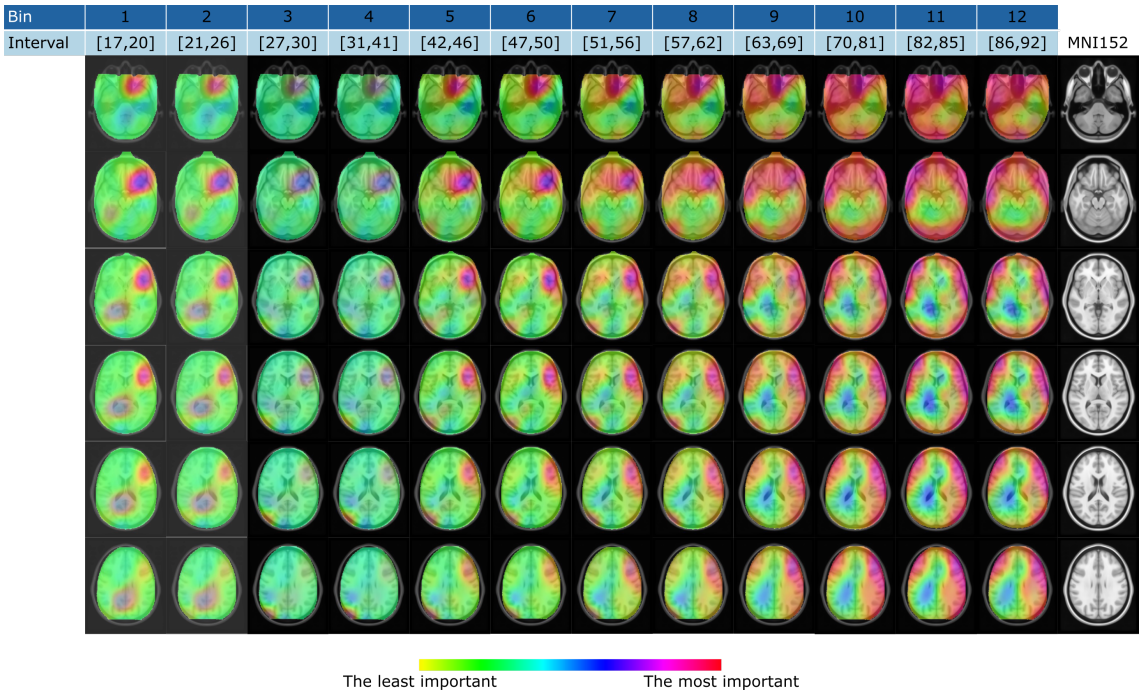

Figure 7.11: Average saliency maps in transverse plane obtained as described in Section 7.3.3 for the profile constructed using the distribution of predictions of the SENet-MR model (details in Section 6.3.1). The maps are constructed using the Grad-CAM method. The rightmost column shows corresponding plane in the MNI152 template.

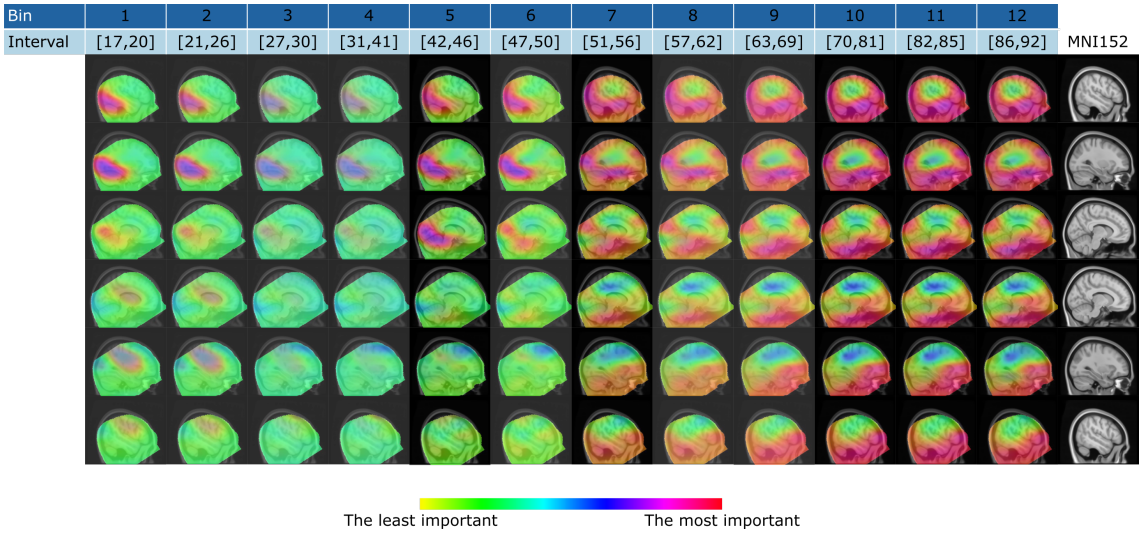

Figure 7.12: Average saliency maps in saggital plane obtained as described in Section 7.3.3 for the profile constructed using the distribution of predictions of the SENet-MR model (details in Section 6.3.1). The maps are constructed using the Grad-CAM method. The rightmost column shows corresponding plane in the MNI152 template.

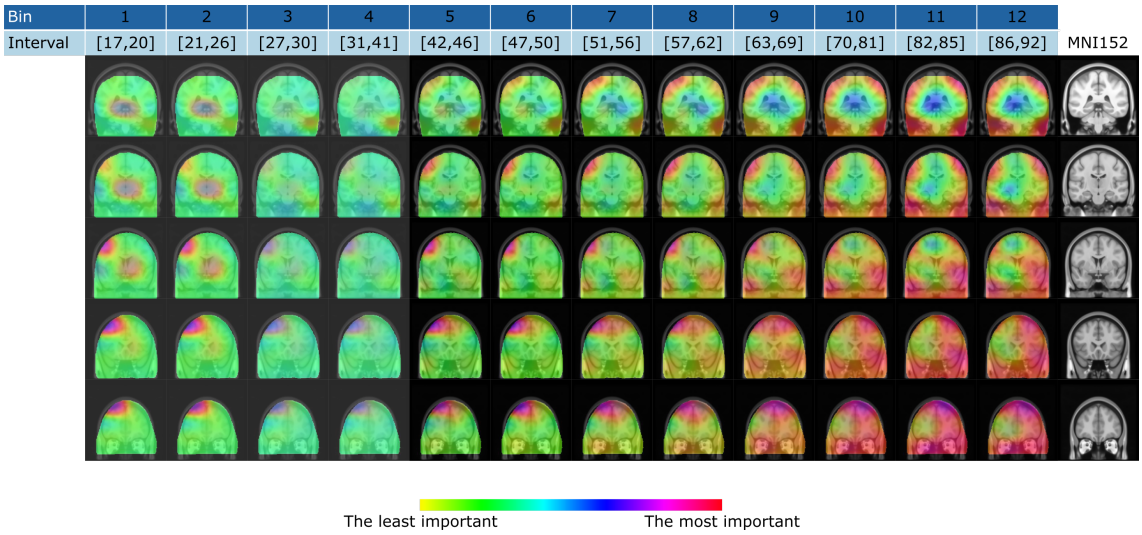

Figure 7.13: Average saliency maps in coronal plane obtained as described in Section 7.3.3 for the profile constructed using the distribution of predictions of the SENet-MR model (details in Section 6.3.1). The maps are constructed using the Grad-CAM method. The rightmost column shows corresponding plane in the MNI152 template.

Figures 7.14 - 7.16 show average maps for the profile obtained by considering prediction distribution of the SENet-OR model (details given in Section 6.3.1). The averaged maps were produced using the technique for saliency mapping of ordinal models proposed in Section 7.3.2. This profile also identifies ventricle importance in younger subjects, but also finds the frontal lobe regions to be of importance in younger ages.

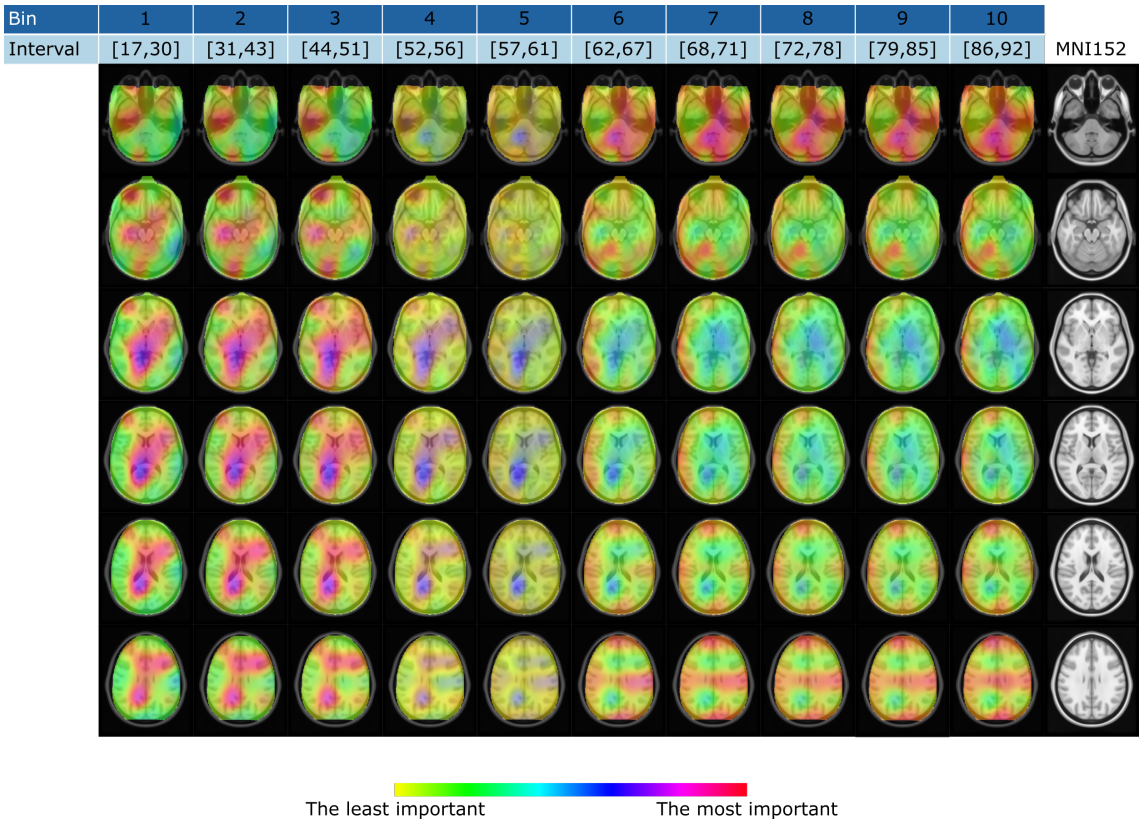

Figure 7.14: Average saliency maps in transverse plane obtained as described in Section 7.3.3 for the profile constructed using the distribution of predictions of the SENet-OR model (details in Section 6.3.1). The maps are constructed using the Grad-CAM method. The rightmost column shows corresponding plane in the MNI152 template.

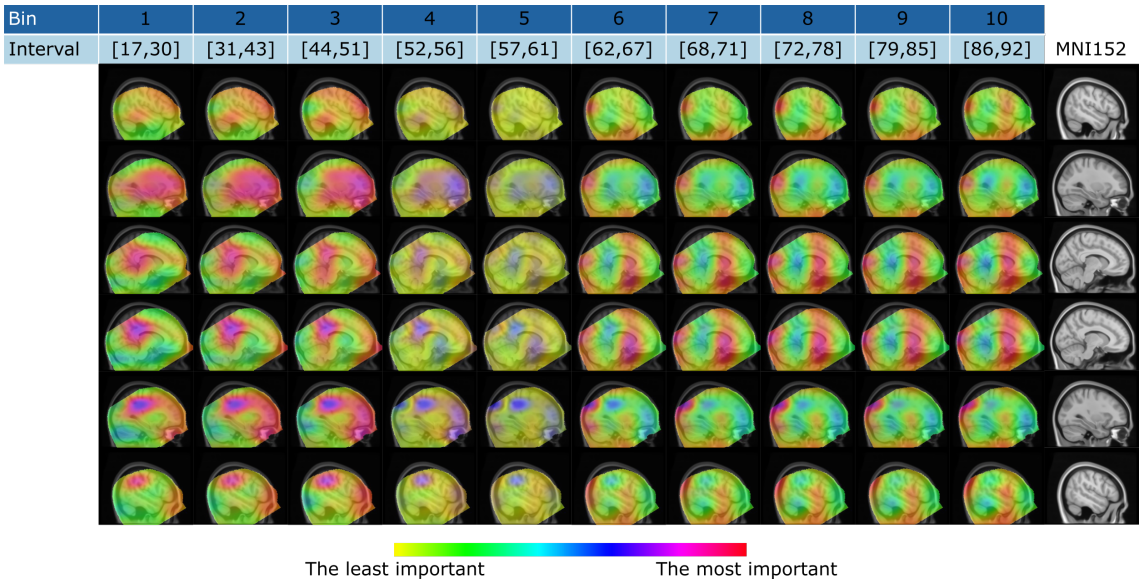

Figure 7.15: Average saliency maps in saggital plane obtained as described in Section 7.3.3 for the profile constructed using the distribution of predictions of the SENet-OR model (details in Section 6.3.1). The maps are constructed using the Grad-CAM method. The rightmost column shows corresponding plane in the MNI152 template.

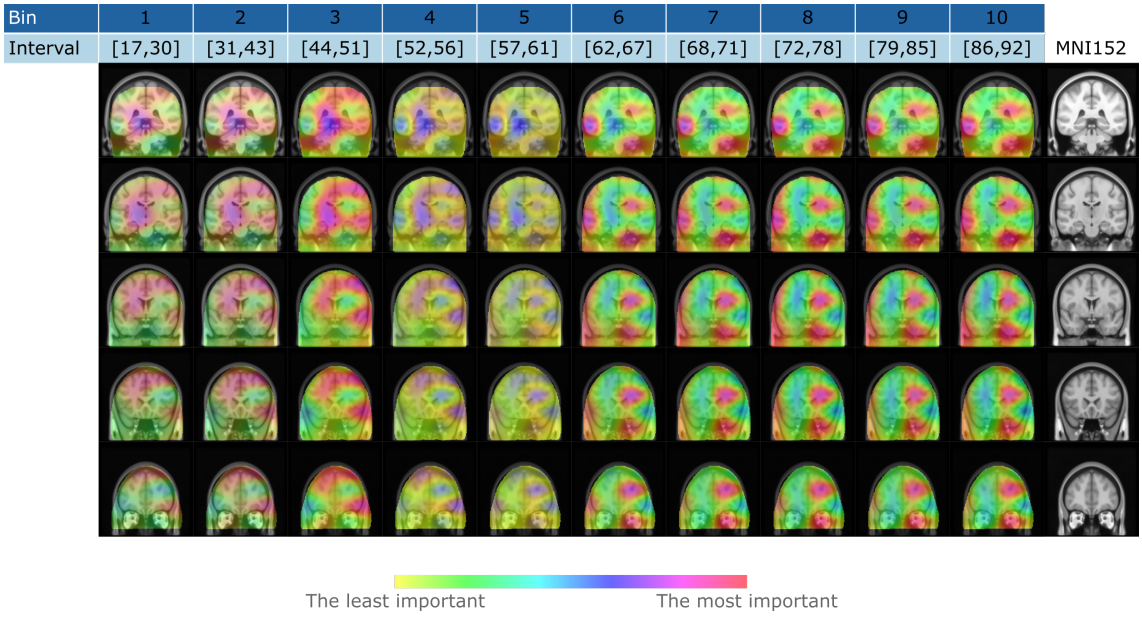

Figure 7.16: Average saliency maps in coronal plane obtained as described in Section 7.3.3 for the profile constructed using the distribution of predictions of the SENet-OR model (details in Section 6.3.1). The maps are constructed using the Grad-CAM method. The rightmost column shows corresponding plane in the MNI152 template.

Figures 7.17 – 7.19 show example subject-specific saliency maps presented for the three age groups – young (up to 30 years of age), middle aged (31 – 50 years of age) and older

		Predicted Age		
		Young	**Middle Aged**	**Older Adults**
True Age	**Young**	28.0, 28.0	28.0, 42.0	-
	Middle Aged	40.0, 28.5	40.0, 40.0	42.0, 57.4
	Older Adults	-	59.0, 26.2	82.0, 82.0

Table 7.1: True ages and predicted ages using the SENet-MR model for the samples for which saliency maps are shown in Figure 7.17 – 7.18. For each sample true age and predicted ages are separated by a comma.

		Predicted Age		
		Young	**Middle Aged**	**Older Adults**
True Age	**Young**	21.0, 21.0	23.0, 33.0	-
	Middle Aged	42.0, 27.0	41.0, 41.0	44.0, 53.0
	Older Adults	-	58.0, 43.0	68.0, 68.0

Table 7.2: True ages and predicted ages using the SENet-OR model for the samples for which saliency maps are shown in Figure 7.19. For each sample true age and predicted ages are separated by a comma.

adults (older than 51 years of age). Saliency maps are visualised based on the samples' true ages and predicted ages. Figures 7.17 and 7.18 show saliency maps produced using the Grad-CAM and guided backpropagation methods respectively, based on the predictions of the SENet-MR model. Figure 7.19 shows saliency maps produced using the ordinal Grad-CAM method based on the predictions of the SENet-OR model. In Figures 7.17 - 7.19 no examples are presented of the samples with true age in the young group and predicted age in the old group and or vice versa, as it is shown in Figures 7.3 – 7.5 that the samples with such large prediction error would almost certainly relate only to samples affected at the preprocessing stage (such as the example shown in Figure 7.5). Tables 7.1 and 7.2 show true and predicted ages for each sample used in Figures 7.17 - 7.18 and Figure 7.19 respectively. In Tables 7.1 and 7.2. the true and predicted ages of samples are shown in the same order as they are presented in corresponding Figures 7.17 - 7.18 and 7.19.

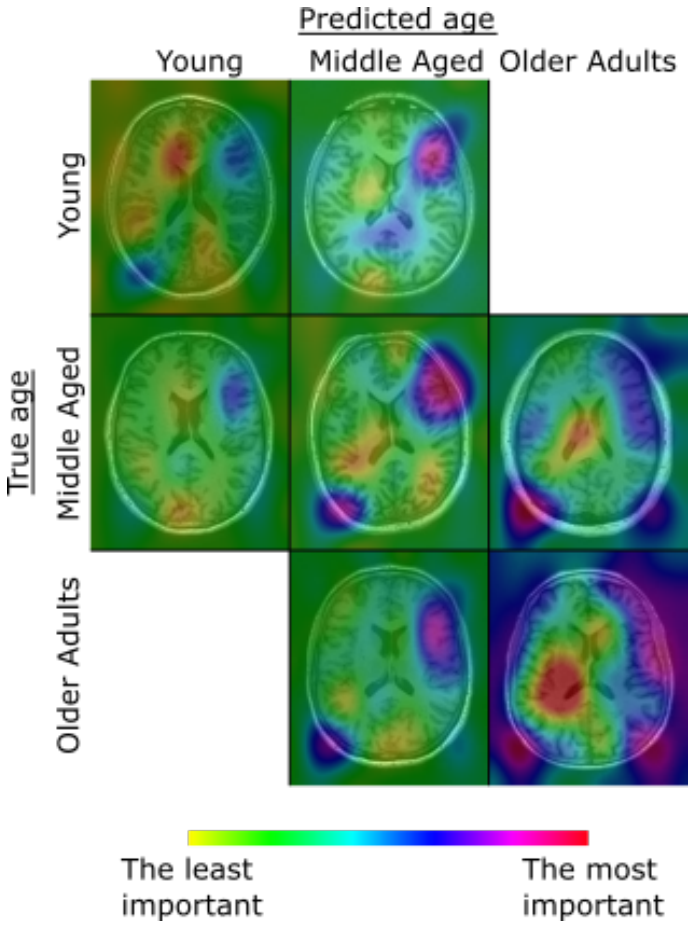

Figure 7.17: Example saliency maps produced using the Grad-CAM method represented to visualise the subjects from three groups – young (up to 30 years of age), middle aged (31 – 50 years of age) and older adults (51 and older). Subject-specific saliency maps are presented based on their true and predicted ages based on the predictions by the SENet-MR model. Saliency maps are shown overlaid over corresponding plane in an input sample after preprocessing.

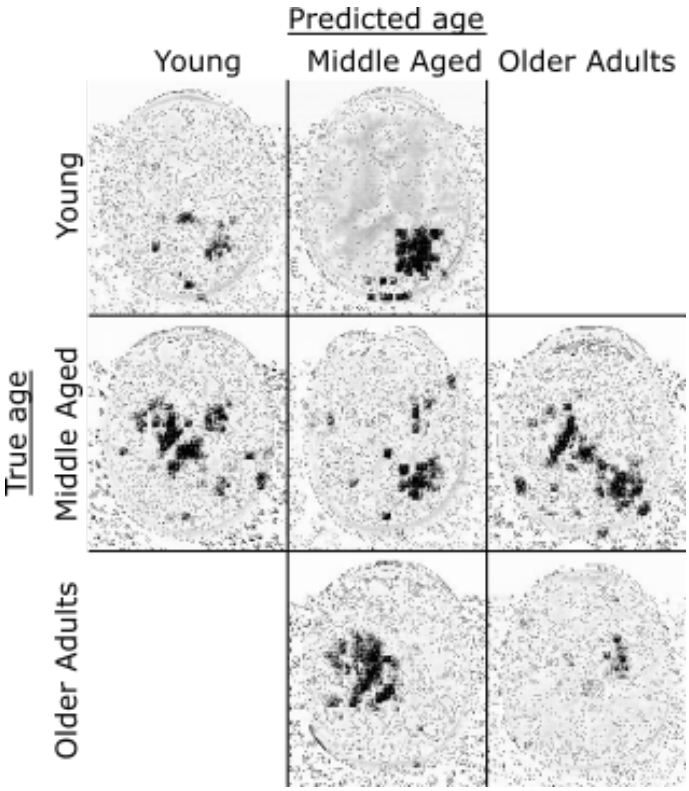

Figure 7.18: Example saliency maps produced using the guided backpropagation method represented to visualise the subjects from three groups – young (up to 30 years of age), middle aged (31 – 50 years of age) and older adults (51 and older). Subject-specific saliency maps are presented based on their true and predicted ages based on the predictions by the SENet-MR model.

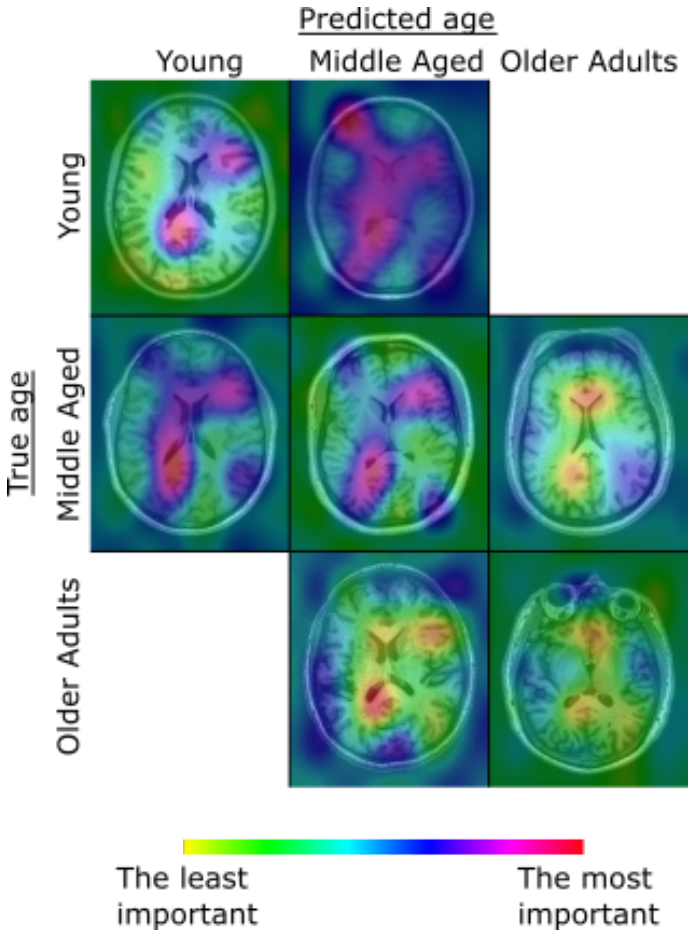

Figure 7.19: Example saliency maps produced using ordinal Grad-CAM method represented to visualise the subjects from three groups – young (up to 30 years of age), middle aged (31 – 50 years of age) and older adults (51 and older). Subject-specific saliency maps are presented based on their true and predicted ages based on the predictions by the SENet-OR model. Saliency maps are shown overlaid over corresponding plane in an input sample after preprocessing.

7.5 Discussion and conclusion

7.5.1 Saliency mapping for natural image classification

Saliency mapping methods were first applied to the task of natural image classification. Varying the value of standard deviation in Equation 2.36 governing the SmoothGrad method affects the amount of noise applied to an input. As it can be seen in Figure 7.2, for too small or too large values of standard deviation ($\sigma^2 = 0.05$ or $\sigma^2 = 1.0$) the noise applied dominates and saliency map's sensitivity suffers. Conversely, it can be seen that the SmoothGrad approach with standard deviation adjusted appropriately can signif-

icantly improve sensitivity of maps produced using purely gradient-based methods such as vanilla gradients method [45]. The method of guided backpropagation has, however, been shown to produce sharper maps compared to the SmoothGrad method. From Figure 7.3 it is also evident that combining guided backpropagation with the Grad-CAM method leads to saliency maps which are dominated by the regions identified by guided backpropagation.

In Figure 7.2 it can also be seen that using the LRP-ε rule β excludes certain features from resulting maps compared to the LRP-β rule. This can be explained by noting that the LRP-ε rule uses all model parameters with equal weight in backpropagating importance scores (details in Table 2.1, Section 2.3.6). The LRP-β rule on the other hand is applied on the product of layer input and its parameters, with the operators defined as $(\cdot)^+ = max(0, \cdot)$ and $(\cdot)^- = min(0, \cdot)$. Results of these two operations are called positive and negative saliency respectively. Using positive and negative saliency, two importance maps are produced and combined using $\alpha = 1 - \beta$ and β weights respectively. Since β regulates the flow of negative saliency in importance backpropagation, a higher β value leads to more features in the resulting map which appear to simply represent noise. This can also be observed in Figure 7.2. The difference between LRP maps produced using the two different rules, and different β values, was less evident for samples with lower certainty in the underlying classification decision.

Overall, the best quality maps for the task of natural image classification were produced using the LRP method. It can also be observed that the quality of saliency maps falls for the inputs with less confident classification decision. It is also evident that stripes in samples of classes "tiger" and "zebra" were important for classification. In input images with smaller classification confidence saliency maps detect stripes on objects not related to actual image class.

7.5.2 Saliency mapping for the task of brain age prediction

Figures 7.3 – 7.5 show the results of applying existing saliency mapping methods and applying the technique for ordinal saliency mapping introduced in Section 7.3.2. The vanilla backpropagation method is applied both to the SENet-MR and SENet-OR models. As it was observed for the task of natural images classification for saliency mapping the models for brain age prediction vanilla backpropagation lacks sensitivity, i.e. it identified part of an input corresponding to head, but not specific brain regions. Vanilla backpropagation applied to the ordinal model is more affected by noise compared to applying it on the

metric model. The SmoothGrad method has better sensitivity and is less affected by noise compared to pure vanilla backpropagation, but produced worse results compared to the method of guided backpropagation. All gradient-based saliency mapping methods are affected by a dot-pattern artifact. The artifacts appear with a fixed distance between them which suggests that the models for prediction do not use all input voxels equally efficiently. Hui and Binder have demonstrated this effect experimentally by measuring relevance layer-wise using the LRP method on ResNet-50 architecture (detailed description in Section 2.3.2). In ResNet architecture residual connections also involve downsampling. They have shown that the dots in resulting maps originate from the downsampling operation as it limits the propagation of relevance to the regions affected by sampling [241]. The Grad-CAM method uses last convolutional feature maps in a model to produce a saliency map and does not suffer from dot artefacts. Feature maps are also less affected by noise, i.e. the features not of interest, compared to the gradients and resulting maps show more distinct regions of interest compared to the gradient-based methods. It can also be observed that in the guided Grad-CAM method the contribution from guided backpropagation dominates and, therefore, the maps produced using the methods of guided backpropagation and guided Grad-CAM are similar. It should be noted, however, that the method of guided backpropagation ignored the features describing whole brain atrophy occurring with age, i.e. increasing subarachnoid space (see Figure 7.20). Therefore, these features are detected by the Grad-CAM method, but not by the guided Grad-CAM method.

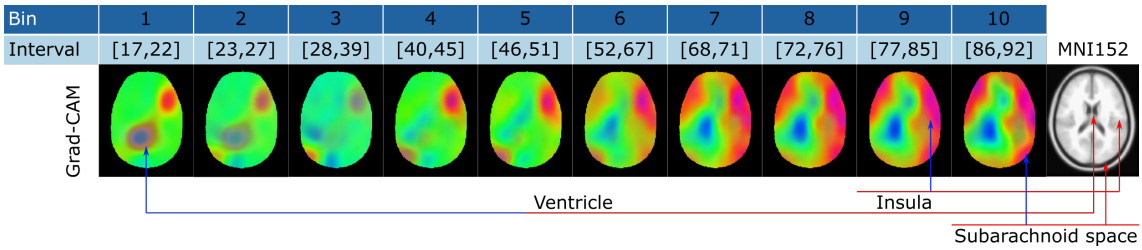

Figure 7.20: Average saliency maps obtained as described in Section 7.3.3 for the profile constructed using the DEC method (details in Section 6.3.3) and the Grad-CAM method for saliency mapping. The rightmost column shows corresponding plane in the MNI152 template. Brain regions affected by the age-related changes are highlighted by arrows: blue arrows point to corresponding regions in saliency maps and red – to the regions in the MNI152 template respectively.

Figure 7.17 shows sample examples grouped by their true and predicted ages. In Figure 7.17 saliency maps are constructed by considering the predictions by the SENet-

MR model using the Grad-CAM method. The figure demonstrates that older predicted age corresponds to more features in an input sample being used by the model to make a prediction. Figure 7.18 shows saliency maps constructed using the guided backpropagation method for the same bins and model as in Figure 7.17. From Figure 7.18 it can be concluded that the SENet-MR model uses less features of an input sample to predict a sample's age to be in the same group as its true label lies, than for samples predicted to be in a different, younger or older, group. Therefore, a sample may be predicted to belong to a group, different from where its true label belongs, due to the SENet-MR model extracting the features not related to the sample's true label. Figure 7.19 shows saliency maps examples obtained using the SENet-OR model and ordinal Grad-CAM method. Comparing this to Figure 7.18, it can be seen that the SENet-OR model used more input sample's features in subjects who are or are predicted to be young and middle aged compared to the SENet-MR model. On the other hand, the SENet-OR model uses much less input features compared to the SENet-MR model in the old subjects who are predicted to be old.

The variability over the population of the brain ageing process cannot be investigated using subject-specific saliency maps. Therefore, in Figures 7.6 – 7.7, average saliency maps for the brain ageing profiles extracted in Chapter 6 are shown. In Figure 7.6 average saliency maps for the profile obtained by considering the distribution of predictions of the SENet-MR model are presented. Figure 7.7 shows average saliency maps produced for the DEC profile which used SENet-MR model's feature extractor. For these two profiles, out of the gradient-based saliency mapping methods, the method of guided backpropagation was the least affected by noise, i.e. detected the least amount of features not related to brain age prediction such as background or parts of head instead of brain regions. In both profiles, the method of guided backpropagation found ventricular features to be the most important for the model to make predictions. Since combining the methods of Grad-CAM and guided backpropagation results in saliency maps with features dominated by guided backpropagation, average saliency obtained using the guided Grad-CAM method for both profiles again found the ventricles to be the most important feature. While it would be interesting to directly and quantitative compare the maps, preliminary investigations showed that direct approaches to voxel-wise spatial correlations did not give interpretable results due to the high levels of background noise in the maps produced using the method of guided backpropagation. Further work is needed to determine whether other metrics might be able to directly compare the methods. For the methods of guided backpropagation and

Grad-CAM, in these profiles the intensity of important features increased with age [171, 6]. Saliency maps constructed using the Grad-CAM method for both profiles highlighted the importance of ventricle features in younger subjects (up to 30 years of age) and increasing importance of whole brain atrophy with age, i.e. increasing subarachnoid space. In Figure 7.20 the regions affected by the age-related changes are highlighted. These results are consistent with previously published volumetric trajectories for various brain regions, as it has previously been found that rate of volumetric changes increases with age and that brain volume does not change significantly in young subjects [171].

Below, the results are compared to existing literature. However, comparison with all existing works is not possible due to fundamental differences in methods or datasets used. Comparison with the works by Fujimoto et al [181, 219] and Kondo et al [182] is not possible due to the way in which their results are represented, for example: in these works the regions of interest were presented by assessing the importance of areas of brain cortex rather than regions of the whole brain parenchyma. The maps were also averaged over the whole dataset. While both Kondo et al [182] and Fujimoto et al [181] collected data from subjects covering a similar age range to those presented here, they included much smaller subject numbers, again potentially making comparison problematic. Comparison with maps presented by Beheshti et al [222] is also not meaningful, due to their measurements of importance again being performed only on cortical areas, and them being averaged over 78 subjects aged from 21 to 60 years of age. For the same reasons a work by Aycheh et al [220] is also not used for comparison. Popescu et al [38] constructed Deep Gaussian Processes (DGP) model for voxel-wise brain age prediction, which allowed saliency maps to be produced directly out of the model. Direct comparison with their results, however, is not possible as the model used data segmented into GM and WM as an input. Therefore, only whether insula region, but not ventricles and subarachnoid space, is affected can be assessed. The authors also only presented saliency map for a single subject. Similarly, comparison with the results by Huang et al [239] cannot be performed as in this work only GM importance maps are presented, and these are averaged over 318 subjects.

Amoroso et al [13] used the Gedeon method to identify the most important features and produced average saliency map over 484 subjects aged 7 – 80 years. The features detected relate to whole brain atrophy, but do not identify changes in ventricular size. This may be related to the fact that the saliency maps presented were averaged over the range of samples' ages spanning whole lifetime, whereas my results show clear differences in the

degree of ventricular involvement with age. Amoroso et al found the inferior frontal gyrus to be involved in structural ageing process. The features describing the profiles obtained using the DEC method (features are shown in Figures 7.8 - 7.10) also highlight this region in subjects older than 68 years of age. In the profile obtained by considering the distirbution of SENet-MR model's predictions the region of inferior frontal gyrus becomes of importance in the intervals containing the subjects older than 63 years of age (see Figures 7.11 - 7.13). Heschl's gyrus was found to be important in the DEC method profiled after 52 years of age, while in the profile derived from the SENet-MR model's error distirbution it is highlighted for the subjects over 57 years of age. The importance of the regions of inferior frontal gyrus and Heschl's gyrus are not, however, highlighted in the profile extracted by considering the distribution of errors of the SENet-OR model (Figures 7.14 - 7.16). Any of the three profiles identifies the regions of anterior cingulate parahippocampal gyrus, cingulate and precentral gyri to be related to ageing for any of the intervals, while Amoroso et al described them as age-related. Amoroso et al also note that teporal lobe plays the role in ageing. In the DEC profile and the profile derived from the distribution of errors of the SENet-MR model the temporal region is of importance for the subjects older than 68 and 51 years of age respectively. The differences between the features shown in Figures 7.8 – 7.16 and the features identified by Amoroso et al, may also be explained by the fact that the saliency mapping methods presented in this chapter either use backpropagation from an output to an input, i.e. gradient-based methods, or from an output to the last convolutional feature maps, i.e. Grad-CAM, while the Gedeon method used by Amoroso et al used the considers only two last fully-connected layers in a MLP model for brain age prediction [13].

Feng et al presented average maps separately for each life decade from 20s to 80s. Feng et al found that frontal lobes are particularly affected by ageing. The profiles obtained using the DEC method and by considering the distribution of errors of the SENet-MR model in Figures 7.6 – 7.13 also highlight the importance of frontal lobe features and its importance increases with age [39]. Frontal lobe was not, however, identified to be important by the profile based on the distribution of predictions of the SENet-OR model.

Bermudez et al [183] presented the test data in three groups based on their age labels: young (0 – 30 years of age), middle age (30 – 50 years of age) and older adult subjects (50 – 96 years of age). The authors presented saliency maps based on samples' true and predicted ages in the same way as I have in Figures 7.17 – 7.19. The authors randomly chose 10 samples for each category and presented average saliency maps. Bermudez et

al used Grad-CAM method and observed that the most important features for the model were ventricular size, cerebral cortex and head size, and subarachnoid space. Head size measure is not relevant for comparison to the current dataset as the youngest subject in the HBAPS dataset is 17 years old and after that age head circumference does not change significantly [242]; the other features highlighted by Bermudez et al are largely consistent with those I found, however. Changes in the subarachnoid space are significant during human lifetime, as it has previously been shown than whole brain volume as well as GM volume [171] decrease with age. Bermudez et al found that intensity and amount of important features increases with age. This can is consistent with the fact that ageing accelerates during lifetime [171, 6]. As can be seen in both subject-specific and average saliency maps for each profile represented in Figures 7.6 – 7.16, the maps generated using the Grad-CAM and ordinal Grad-CAM methods identified subarachnoid space and ventricular size to be important features. For the subject-specific maps and average saliency maps obtained using the SENet-MR model, the intensity of important features also increased with age, again consistent with the results of Bermudez et al. For the average saliency maps constructed using ordinal Grad-CAM method for the profile considering the distribution of predictions of the SENet-OR model, the intensity of the features does not increase with age (Figures 7.14 - 7.16), but the maps nevertheless highlight the importance of ventricular size and subarachnoid space. The reason why in the current dataset intensity decreases rather than increases in the ordinal profile shown in Figures 7.14 - 7.16 needs to be further investigated.

Herent et al [52] used three approaches to assess the features important for brain ageing in MRI data. First, the authors constructed correlation maps of GM and CSF with predicted brain age averaged over test data, produced using GM and CSF segmentation masks. Secondly, the authors represented importance map visualising the weights of the ridge regression model trained on the task of brain age prediction. Thirdly, Herent et al constructed saliency maps using the occlusion method of Zeiler and Fergus [47] on a CNN model for brain age prediction trained on 2D inputs . Two maps were presented – average maps for the subjects younger than 30 years of age and the subjects older than 60 years of age. Correlation maps of GM and CSF with predicted brain age constructed by the authors indicated positive correlation with age of the subarachnoid space and ventricular features. Importance map visualising weights of the ridge regression model highlighted the importance of the ventricles and grey matter atrophy with age. The results of this chapter are in agreement with all of these findings, as discussed above. Results of Herent et al also

showed that average occlusion maps for the subjects reveal the importance of insula as a feature for brain age prediction. Saliency maps for the profiles presented in the Figure 7.6 – 7.13 also show importance of insula features. The profile constructed by considering the distribution of predictions of the SENet-MR model and using the DEC method also identified the importance of insula features whose importance increased with age, which is also consistent in the results by Herent et al. Popescu et al [38] have also detected the importance of insula features in relation to brain ageing. In the profile constructed by considering the distribution of predictions of the SENet-OR model shown in Figures 7.14 - 7.16, the importance of insula features is also evident, but the importance decreases with age.

In conclusion, I have shown that the methods of guided backpropagation, Grad-CAM and guided Grad-CAM are usable for studying brain ageing features from MRI data with minimal preprocessing. Average maps derived for the whole brain ageing profiles extracted in Chapter 6 appear to be of biological significance, as features are in accordance with previous research and existing knowledge on the biology of brain ageing. Saliency maps produced using proposed methodology for saliency mapping for ordinal models also highlighted biologically interpretable features.

Further research is, however, needed in order to extend current knowledge on human brain ageing even more. Firstly, further investigation is needed to understand why the profile extracted using the ordinal saliency mapping method and averaging over the intervals obtained from the SENet-OR model did not show the important features to be increasing in intensity with age. This may be explained either by the properties of the ordinal saliency mapping methodology, which need to be studied, or the properties of the SENet-OR model. Secondly, it can be observed in Figures 7.6 – 7.7 that the features related to left ventricles appear more important than the ones related to right ventricles and the features describing whole brain atrophy lie more in the right hemisphere compared to the left one. A number of studies have previously reported structural hemispheric asymmetries and their relation to the process of brain ageing (for example, [243]), but this is still an open discussion whether any of the hemispheres ages faster [244]. For example, Esteves et al [244] used functional MRI in order to study hemispheric asymmetries and found that the subjects of older age with present hemispheric asymmetries performed better in the presented tasks [244]. This may be explained either by the fact that most data samples are collected from left-handed individuals, but could also indicate biases in the underlying MR

images caused by data acquisition or preprocessing issues. This could be investigated by computing separate average maps for left- and right-handed subjects, by comparing data from different sources, or by investigating alternative pre-processing steps (for example a different template for the registration). Investigating separate average maps for male and female subjects could also provide additional insights into the process of brain ageing. Overall, however, it appears that appropriately created saliency maps can provide us with information that may be of interest at both the individual and group level, and which may reveal biologically meaningful changes over the human lifespan.

Chapter 8

Examining brain age prediction models' generalisability

8.1 Main contributions

In this chapter generalisability of the brain age prediction models presented in Chapter 5 is examined. In order to examine models' generalisability a dataset more representative of data collected in clinical setting is used along with modified Healthy Brain Ageing from Public Sources (HBAPS) test data. Generalisability of the brain age prediction models is studied in light of following contributions:

- the models' accuracy comparable to the results achieved in Chapter 5 is also measured on test data from the HBAPS dataset with resolution of its samples reduced;
- test data more representative of the clinical setting, using an independent, locally collected, dataset (the Institute of Psychiatry, Psychology & Neuroscience (IoPPN) dataset described in Section 4.2.2) is collated and preprocessed in order to analyse the performance of brain age prediction methods on this type of data; two subsets of the IoPPN dataset are used consisting of Magnetic Resonance Imaging (MRI) data acquired using T1-weighted sequences - one consisting of healthy subjects and one containing the samples acquired from diseased subjects;
- saliency maps are extracted for each prediction on the IoPPN test data;
- HBAPS test data is also used to investigate saliency mapping methods for the purpose of outlier detection.

8.2 Introduction

Using brain age prediction models as a biomarker for various brain diseases has been studied by academic researchers. Developing a putative biomarker from a brain age prediction model can be achieved by training the model on healthy subjects and using the model to predict the age of clinical patients. Then if predicted brain age is greater than patient's chronological age by a number of years greater than model's Mean Absolute Error (MAE), this can be interpreted as occurrence of accelerated ageing in the patient's brain. Authors have investigated such differences between healthy subjects and patients with a number of diseases [245] – Alzheimer's disease [224], traumatic brain injury [246], HIV [226], epilepsy [225], Down's syndrome [228] and diabetes [247]. Brain ageing process is influenced by many complex factors, however, including both heritable factors [35] and ones that change during an individual's lifetime [248]. While clearly an oversimplification, representing this process by a single measure, as a difference between predicted and chronological age, allows development of a straightforward biomarker [35].

However, clinical usage of brain age prediction-based biomarker should be studied further as its reliability and ability to capture population-wide variation in the ageing process needs to be further validated. In Chapter 5 I have discussed the task of brain age prediction from MRI data – existing research in this field is presented along with results of this thesis by discussing both Deep Learning (DL)-based and non-DL-based approaches. However, all the works discussed there, including the results of this thesis, use either datasets collected purely for research purposes or limited clinical datasets, and do not fully represent population-wide effects nor variations in the quality of clinical data.

In this chapter generalisability of the models presented in Chapter 5 is first examined by measuring their performance on test data from the HBAPS dataset, after deliberately reducing its resolution. Such experiments explore the brain age predicting models from two perspectives. Firstly, usability of these models in clinical environment, where MRI data is acquired with varying resolution, not always with the highest technically possible, due to temporal and financial constraints is studied. Such experiments also allow understanding of how much information is needed in an MRI scan to make a brain age prediction with accuracy comparable to that from the whole, high resolution, scan.

In this chapter the approaches presented in Chapters 5 and 7 are also tested on data more representative of the clinical setting, using an independent, locally collected, dataset, in which both healthy controls and patients are present – the IoPPN dataset described in

Section 4.2.2. The aim of this section of the chapter is to apply the methods presented in this thesis to a more clinically-representative dataset and analyse the performance of such methods on this type of data. First, the Squeeze-and-Excitation Network (SENet)-MR and SENet-OR models (described in Chapter 5) are used on the dataset for brain age prediction. Then saliency maps are extracted for each prediction, using the Gradient Class Activation Mapping (Grad-CAM) method as it had the best performance among existing methods described in Chapter 7. Saliency mapping is also investigated for the purpose of outlier detection.

8.3 Methods

8.3.1 Studying the performance of the brain age prediction models on the HBAPS test data

In order to study robustness of the brain age prediction models with respect to input data resolution, i.e. whether model accuracy measured by MAE is affected by test data resolution, the HBAPS test data resolution was needed to be reduced. Then brain age can be predicted for each sample using pre-trained SENet-MR model as described in Chapter 5.

Shi et al [249] have proposed a methodology for simulating low resolution images from higher resolution ones. The authors described a two-stage procedure for lowering resolution. First, the samples are blurred using a Gaussian kernel parametrised by a standard deviation of 1 voxel. Secondly, blurred samples are then down sampled to imitate the partial volume effect [249].

The HBAPS data described in Section 4.2.1 was preprocessed as detailed in Section 4.2.3. As part of this preprocessing stage, the MRI scans are resampled and registered onto the MNI152 template and resulting samples have voxel size of $1 \times 1 \times 1$ mm, i.e. a digital resolution of 1 mm. In this chapter, low resolution data is produced from the HBAPS test data in order to examine brain age prediction model robustness with respect to resolution. Here, a simplified approach for lowering sample resolution is considered – the planes are repeated along the head-to-foot direction, simulating the anisotropic voxel dimensions typical of 2D MRI acquisitions, rather than the higher, and near isotropic, resolution typical of 3D acquisitions. As the result, the HBAPS test data is used with two voxel sizes in this chapter: $1 \times 1 \times 4$ mm and $1 \times 1 \times 5$ mm. For example, for achieving voxel size of $1 \times 1 \times 4$ mm every forth plane is extracted from the original data sample and then

used to create four identical adjacent planes in the head-to-foot direction in resulting data sample. The approach taken in this thesis for simulating low resolution MRI data is similar to the approach described by Savio et al [250]. Savio et al averaged the intensities of three consecutive planes in a data sample originally having $1 \times 1 \times 1$ mm voxel size to simulate the samples with 3 mm resolution along the head-to-foot direction. In this Chapter the robustness of the brain age prediction models with respect to input data resolution is examined on the $1 \times 1 \times 4$ mm and $1 \times 1 \times 5$ mm voxel sizes as these resolutions are recommended for clinical trial protocols in brain tumour studies by U.S. Food and Drug Administration (FDA) and National Institutes of Health of the United States Department of Health and Human Services (NCI). As one of the aims of this chapter is to test the performance of brain age prediction algorithms on data more representative of the clinical setting, the resolutions recommended for clinical studies are used in this experiment on lower resolution data.

8.3.2 Studying the performance of brain age prediction models and saliency mapping on the IoPPN test data

For brain age prediction on the IoPPN dataset, pretrained SENet-MR and SENet-OR models are used as described in Section 5.3. The SENet-MR and SENet-OR models are pretrained on T1-weighted MRI data collected from healthy subjects and in this chapter they are tested on T1-weighted data from healthy and diseased subjects separately. For each prediction a saliency map is then produced using the Grad-CAM method as detailed in Section 7.3.1.

8.4 Results

8.4.1 Performance of the Se-MR model on reduced resolution data

Robustness of the SENet-MR model (details in Section 5.3) is tested with respect to low resolution data using whole HBAPS test data with lowered resolution as described in Section 8.3.1. The distributions of predicted ages for the HBAPS testing data with voxel size of $1 \times 1 \times 4$ mm and $1 \times 1 \times 5$ mm are shown in Figure 8.1. MAE and the Pearson's correlation coefficient, R, of chronological and predicted ages at both resolutions are presented in Table 8.2 along with the SENet-MR model's performance on the original HBAPS test data (i.e. $1 \times 1 \times 1$ mm).

The linear regression fit for the SENet-MR model's predictions on data with different voxel sizes is shown by the dashed dark blue line in Figure 8.1, along with the 95% intervals (dashed light blue-green lines) and the line of identity (in orange). The parameters into linear fits into the plots in Figure 8.1 are presented in Table 8.1. From Table 8.1 it also can be noted that for linear fits into the predictions on HBAPS data with two different resolutions ($1 \times 1 \times 4$ mm and $1 \times 1 \times 5$ mm) intercepts differ by less than 1.

Voxel size	**a**	**95% confidence interval bounds on a**		**b**	**95% confidence interval bounds on b**	
		Lower	**Upper**		**Lower**	**Upper**
$1 \times 1 \times 4$ mm	0.928	0.911	0.944	3.493	2.687	4.298
$1 \times 1 \times 5$ mm	0.916	0.898	0.933	4.237	3.392	5.082

Table 8.1: Parameters of the linear regression fit into the plot of the SENet-MR predictions with respect to true age labels for both whole HBAPS test dataset with different voxel size. Linear regression fit into the plot is represented by a function $y_p = a \times y_t - b$, where y_t and y_p are true age labels and the SENet-MR model predictions respectively, while a and b correspond to the slope and intercept of the function. For both a and b values 95% confidence interval lower and upper bounds are stated. $p = 0.0005$ for both fits and for all slope and bias values.

The SENet-MR model's performance on the test data with voxel size of $1 \times 1 \times 4$ mm is the same as on the data with original resolution, as it can be seen in Table 8.2. On the test data with voxel size of $1 \times 1 \times 5$ mm the performance on the SENet-MR model is worse, but only slightly, i.e. MAE is higher by 0.27 years and Pearson's correlation coefficient, R, is lower by 0.01.

Voxel size [mm]	**MAE [years]**	**R**	**RMSE [years$^{0.5}$]**	**Error variance [years]**	**σ [years$^{0.5}$]**
$1 \times 1 \times 1$	3.87	0.96	6.88	32.36	5.69
$1 \times 1 \times 4$	3.87	0.96	6.88	32.36	5.69
$1 \times 1 \times 5$	4.14	0.95	7.28	35.94	5.99

Table 8.2: MAE and the Pearson's correlation coefficient, R, of chronological and predicted ages, Root Mean Square Error (RMSE), error variance and standard deviation on MAE, σ, for the SENet-MR model predictions on the HBAPS test data with three different resolutions.

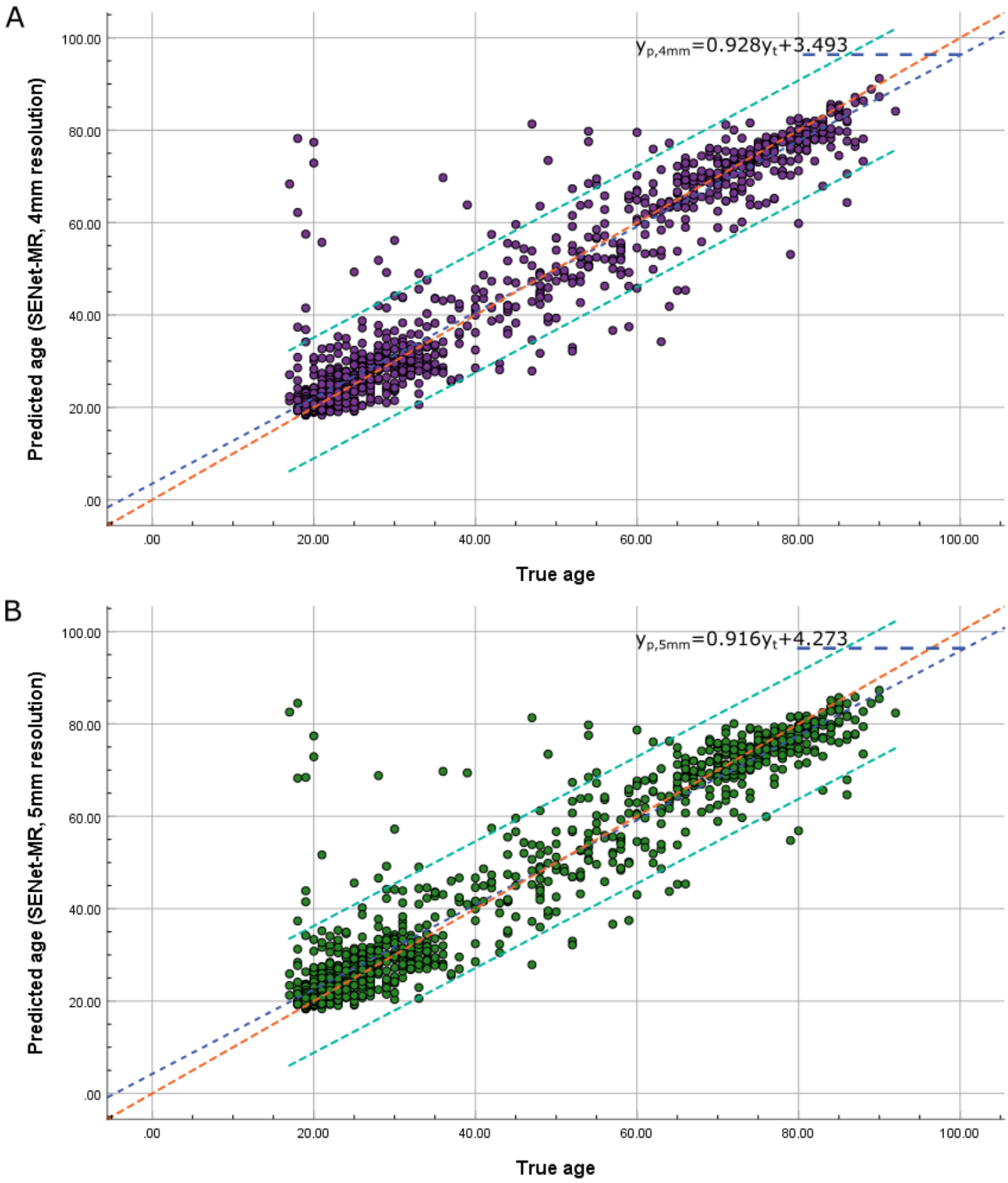

Figure 8.1: Distribution of predictions of the SENet-MR model on whole HBAPS test data with voxel size of $1 \times 1 \times 4$ mm (A) and $1 \times 1 \times 5$ mm (B). The brain age predicted by the models on the testing data (y-axis) is plotted against chronological age label (x-axis). Orange line denotes the y=x function, blue – linear regression fit function, green – the lines representing the lower and upper bounds of the 95% confidence interval over linear regression fit parameters.

8.4.2 Detecting outliers on the HBAPS dataset

In Section 7.4.2 an example was given of a sample being affected by incorrect preprocessing, resulting in a smaller size ($91 \times 109 \times 91$ voxels instead of $196 \times 232 \times 188$). The brain age prediction models described in Chapter 5 failed to extract important features and this resulted in large prediction error. Therefore, gross outliers can be detected by identifying the samples with an unexpected size after preprocessing. Hence the distributions of predicted ages for the HBAPS testing data excluding the preprocessed samples of size $91 \times 109 \times 91$ voxels (16 samples) are shown in Figure 8.2. MAE and the Pearson's correlation coefficient, R, of chronological and predicted ages for the HBAPS test data with and without the affected samples are presented in Table 8.4.

The linear regression fit for the SENet-MR and SENet-OR models' predictions on data without the samples affected by preprocessing is shown by the dashed dark blue line in Figure 8.2, along with the 95% intervals (dashed light blue-green lines) and the line of identity (in orange). Parameters of the linear regression fit into the plots presented in Figure 8.2 are shown in Table 8.3. From Table 8.3 it can be observed that the difference between 0 and the constant term of linear fit into the SENet-OR model's predictions is not statistically significant, while the constant of linear fit into the SENet-MR model's predictions does not statistically significantly deviate from 0.

Model	a	p_a	95% confidence interval bounds on a		b	p_b	95% confidence interval bounds on b	
			Lower	Upper			Lower	Upper
SENet-MR	0.928	0.0005	0.911	0.944	3.493	0.0005	2.687	4.298
SENet-OR	0.916	0.0005	0.898	0.933	4.237	0.145	3.392	5.082

Table 8.3: Parameters of the linear regression fits into the plot of the SENet-MR and SENet-OR models' predictions with respect to true age labels for whole HBAPS test dataset excluding the samples affected at the preprocessing stage. Linear regression fits into the plots are represented by the function $y_{MR,wap} = a \times y_t - b$ and $y_{OR,wap} = a \times y_t - b$ for the SENet-MR and SENet-OR models' predictions respectively. $y_{MR,wap}$ and $y_{MR,wap}$ are corresponding predictions by the SENet-MR and SENet-OR models, while y_t is a true age label. a and b correspond to the slope and intercept of the function. For both a and b values $p-$values are given corresponding to p_a and p_b along with 95% confidence interval lower and upper bounds.

HBAPS samples	Model	MAE [years]	R	RMSE [years$^{0.5}$]	Error variance [years]	σ [years$^{0.5}$]
All	SENet-MR	3.87	0.96	6.88	32.36	5.69
	SENet-OR	3.62	0.97	5.91	21.79	4.67
Without preprocessing outliers	SENet-MR	3.43	0.97	5.30	16.32	4.04
	SENet-OR	3.37	0.98	5.06	14.27	3.78

Table 8.4: MAE and the Pearson's correlation coefficient, R, of chronological and predicted ages, Root Mean Square Error (RMSE), error variance and standard deviation on MAE, σ, for the SENet-MR and SENet-OR model predictions on the whole HBAPS test dataset and without the samples affected by preprocessing.

As expected, once the samples affected at the preprocessing stage were removed the MAE was reduced and Pearson's correlation coefficient, R, increased for both prediction models. Here in Figure 8.3 a sample with large prediction error (60.2 years) is presented along with corresponding saliency map which is different from a HBAPS test data sample with 0 prediction error presented in Figure 8.4.

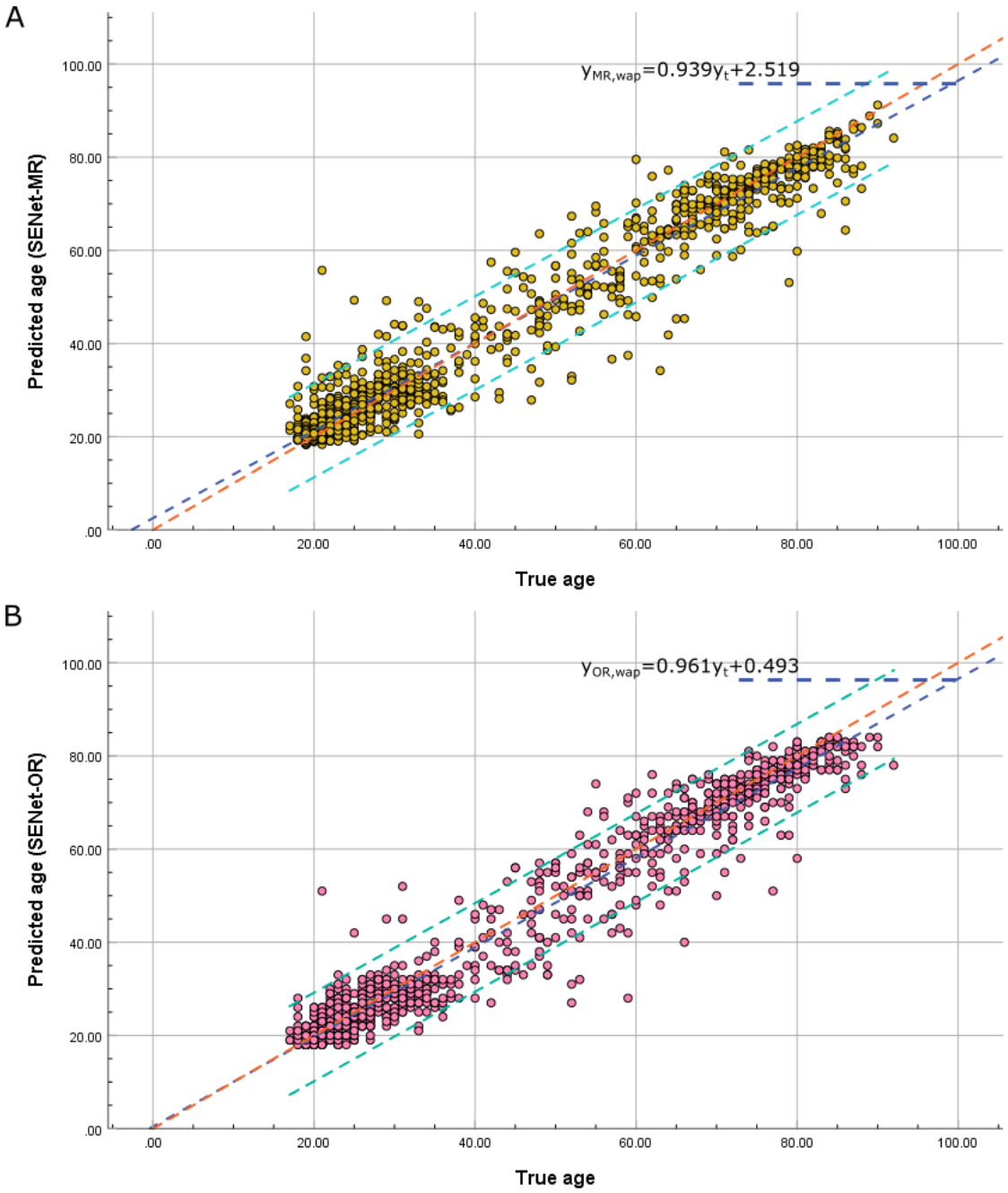

Figure 8.2: Distribution of predictions of the SENet-MR (A) and SENet-OR (B) models on the HBAPS test data excluding the samples affected by preprocessing. The brain age predicted by the models on the testing data (y-axis) is plotted against chronological age label (x-axis). Orange line denotes the $y = x$ function, blue – linear regression fit function, green – the lines representing the lower and upper bounds of the 95% confidence interval over linear regression fit parameters.

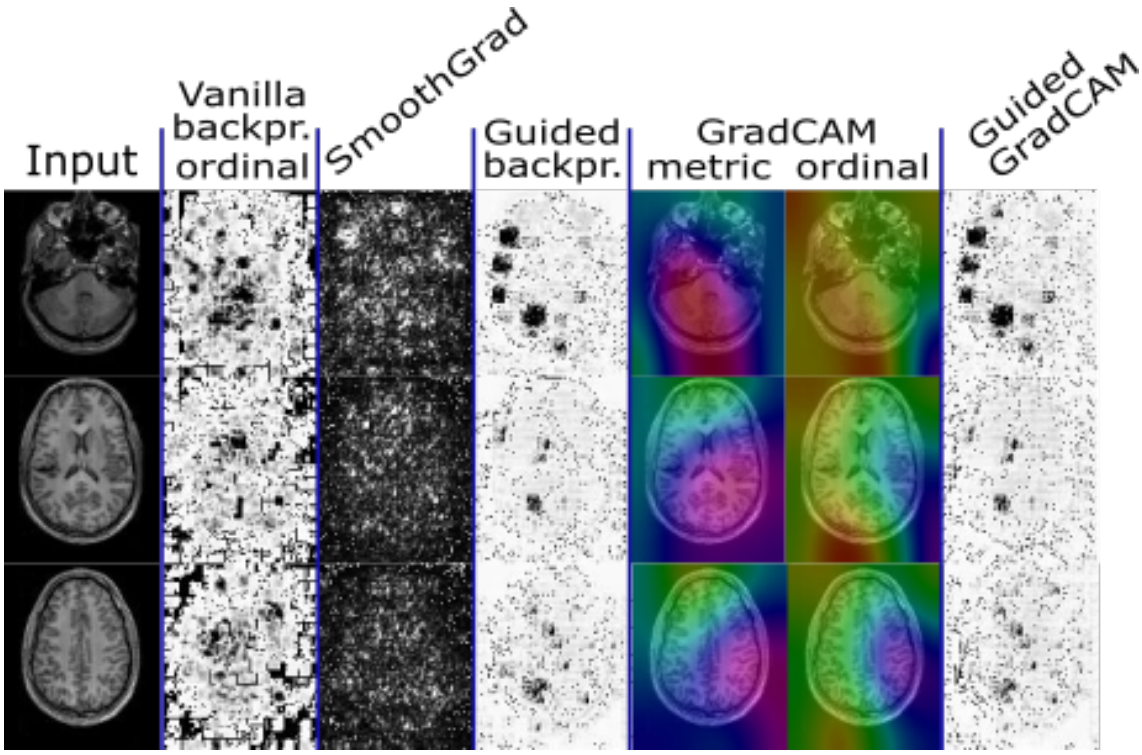

Figure 8.3: Example of the HBAPS test data sample affected by the preprocessing stage. Saliency maps are shown for the sample with the largest prediction error (60.2 years) in the HBAPS testing data using the SENet-MR model. The maps named ordinal are produced by applying ordinal saliency mapping technique described in Section 7.3.2 with corresponding method. These maps are produced using the SENet-OR model and SENet-MR model is used otherwise. The leftmost column shows corresponding planes in the input sample. Three planes are presented from the samples as resulting saliency maps are less informative compared to the samples with smaller associated prediction error.

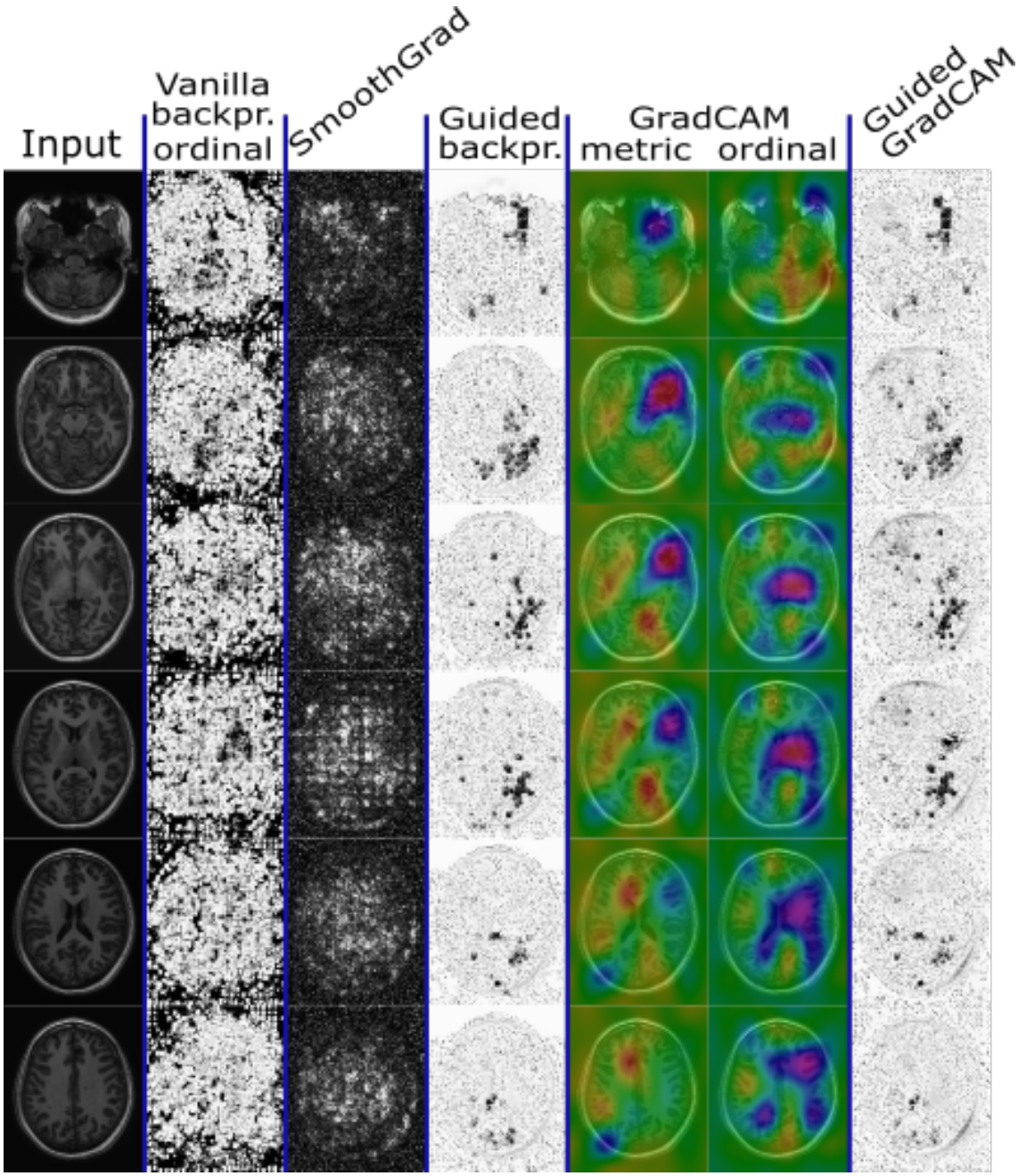

Figure 8.4: Results of applying existing saliency mapping methods to the task of brain age prediction. Saliency maps are shown for the sample with the lowest prediction error (0.0008 years) in the HBAPS testing data using the SENet-MR model. The maps named ordinal are produced by applying ordinal saliency mapping technique described in Section 7.3.2 with corresponding method. These maps are produced using the SENet-OR model and SENet-MR model is used otherwise. The leftmost column shows corresponding planes in the input sample.

8.4.3 Brain age prediction on the IoPPN dataset

In order to further analyse the performance of brain age prediction models described in Chapter 5 they are applied to the subsets of IoPPN dataset - healthy and diseased subjects imaged using T1-weighted sequence (described in Section 4.2.2). MAE and the Pearson's correlation coefficient, R, of chronological and predicted ages from the IoPPN dataset are presented in Table 8.5. In Figure 8.5, the average prediction error for each age in the IoPPN dataset is plotted for the SENet-MR and SENet-OR models and for the healthy and diseased subjects separately.

IoPPN data	Model	MAE [years]	R	Error variance [years]	σ [years$^{0.5}$]
Healthy	SENet-MR	21.41	0.37	147.13	12.13
	SENet-OR	31.26	0.29	175.21	13.24
Diseased	SENet-MR	14.89	0.57	126.46	11.25
	SENet-OR	20.53	0.53	191.75	13.85

Table 8.5: MAE and the Pearson's correlation coefficient, R, of chronological and predicted ages, error variance and standard deviation on MAE, σ, for the SENet-MR and SENet-OR models predictions on healthy and diseased subjects of the IoPPN dataset.

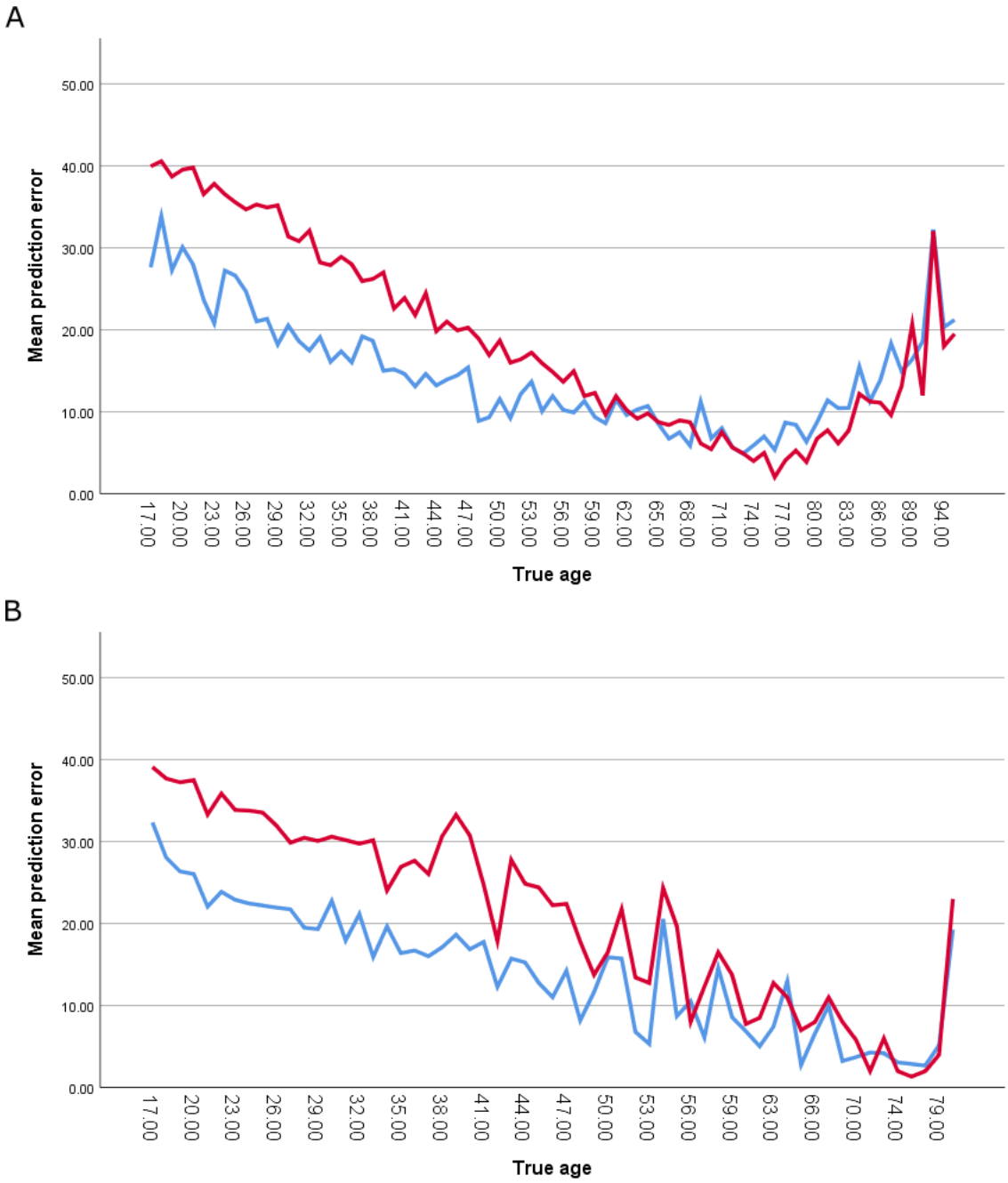

Figure 8.5: Average prediction error for each true chronological age in the IoPPN dataset by the SENet-MR (blue line) and SENet-OR (red line) models on healthy (A) and diseased (B) subjects separately.

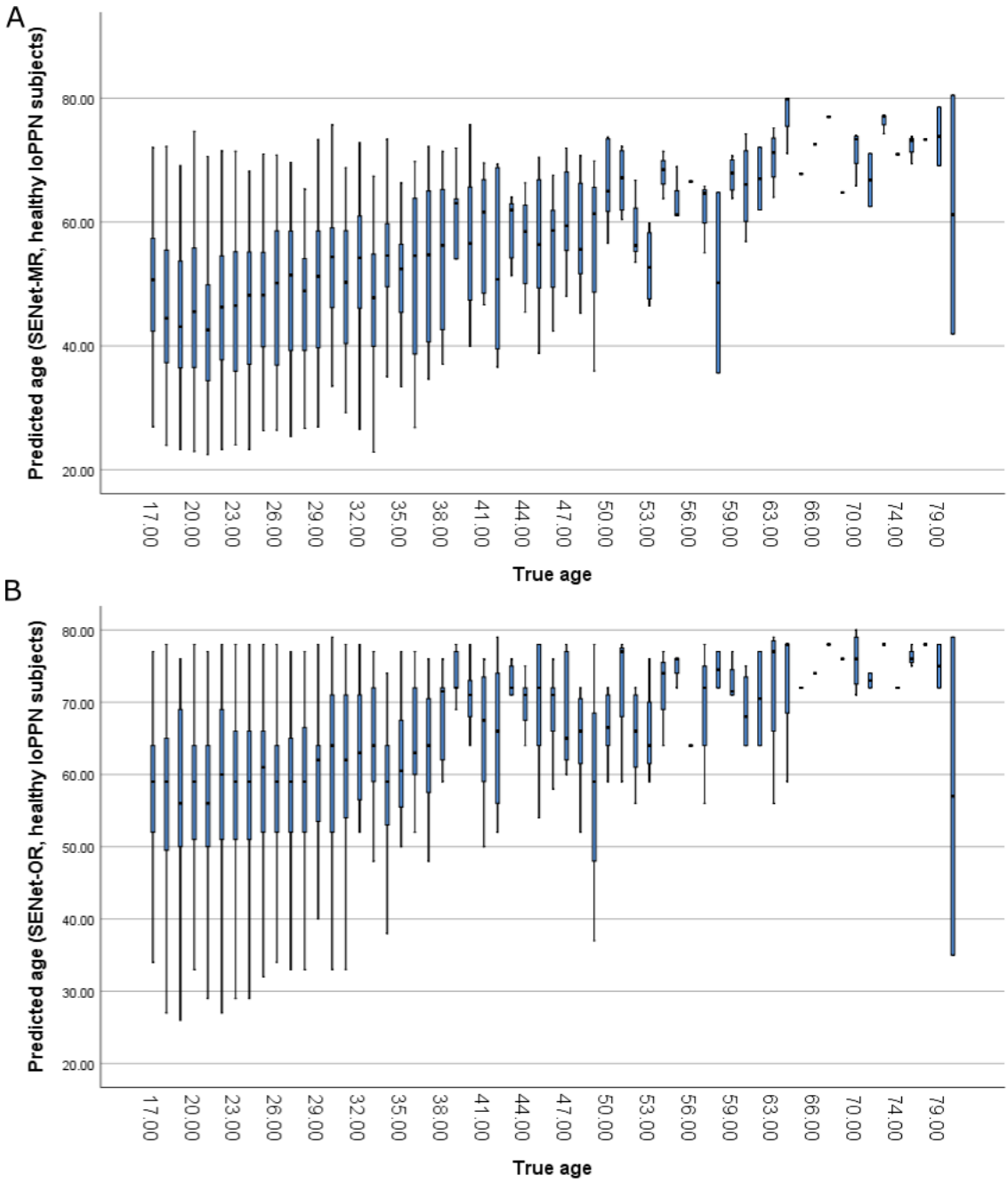

Figure 8.6: Box plots of the predictions on the IoPPN healthy subjects imaged using T1-weighted sequence by the SENet-MR (A) and SENet-OR (B) models with respect to the true age labels.

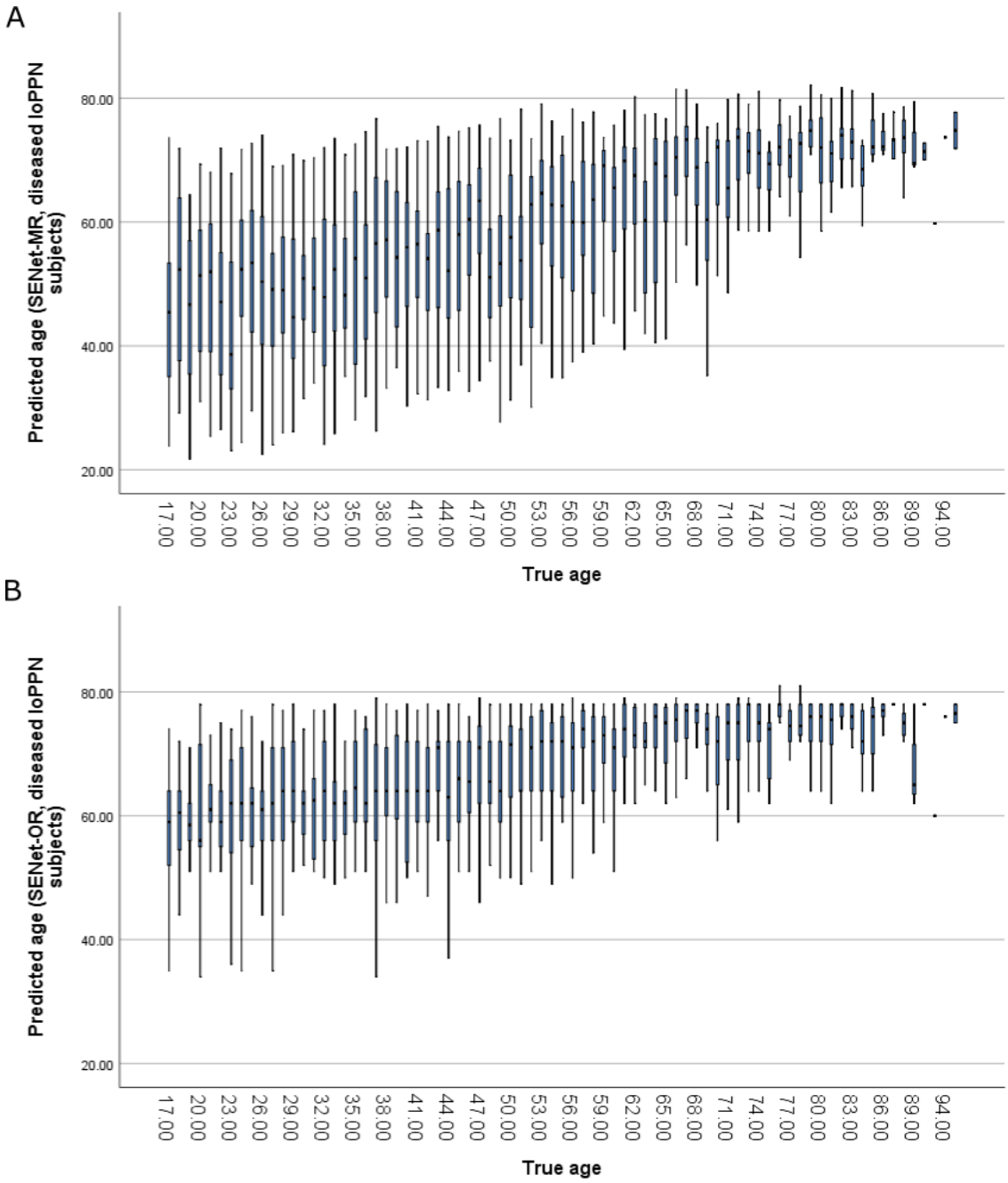

Figure 8.7: Box plots of the predictions on the IoPPN diseased subjects imaged using T1-weighted sequence by the SENet-MR (A) and SENet-OR (B) models with respect to the true age labels.

In Figures 8.6 and 8.7 box plots representing the predictions distributions for the SENet-MR and SENet-OR models on the IoPPN dataset healthy and diseased subjects are shown. The SENet-MR model's predictions are more concentrated around mean compared to the SENet-OR models' predictions on the samples from both healthy and diseased subjects. The SENet-OR model's MAE was lower for both the healthy and diseased subjects in

the IoPPN dataset compared to the SENet-MR model's MAE. The SENet-OR model's predictions were also more tended towards the largest true age label available in the HBAPS training data used for training the SENet-OR model.

No "gross failures" of preprocessing were found in the IoPPN test data. Figures 8.8 – 8.13 show saliency mapping results for samples with a range of high and low prediction errors:

- Figure 8.8 shows saliency maps for the samples from the IoPPN and HBAPS datasets with 0 prediction error using the SENet-MR model. Both samples are acquired from the healthy subjects who were 28 years old.
- Figure 8.9 shows saliency maps for the samples with associated prediction error equal to the sum of MAE of the SENet-MR model and its standard deviation, 9.55 years. The subjects are healthy and 22 years of age.
- Figure 8.10 shows saliency maps for the samples with prediction error of 20 years by the SENet-MR model. The subjects are healthy and 29 years old.
- Figure 8.11 shows saliency maps for the sample from the IoPPN test data diseased subject and for the sample from HBAPS test data healthy subject with 0 prediction error using the SENet-MR model. Both samples are acquired from the healthy subjects who were 69 years old.
- Figure 8.12 shows saliency maps for the samples with associated prediction error equal to the sum of MAE of the SENet-MR model and its standard deviation, 9.55 years. The subjects are 52 years of age. IoPPN test data sample is diseased, while HBAPS test data sample is healthy.
- Figure 8.13 shows saliency maps for the samples with prediction error of 20 years by the SENet-MR model. The subjects are 69 years old. IoPPN test data sample is diseased, while HBAPS test data sample is healthy.

From Figures 8.8 – 8.10 it can be observed that the features extracted from the samples belonging to the IoPPN dataset's healthy subjects are similar to the ones extracted from the HBAPS test data samples with the same associated prediction error. The features extracted are also in agreement with the average features describing whole brain population-wide profile extracted using the Deep Embedded Clustering (DEC) method.

In the samples associated with the prediction error of 20 years the SENet-MR model used more features from original sample for both healthy subject from HBAPS and IoPPN test data. From Figures 8.11 - 8.13 it can be seen that for the diseased subjects from the IoPPN dataset are associated with features different from those extracted from the healthy test subjects from the HBAPS dataset.

In order to further examine performance of the brain age prediction models on the IoPPN dataset in Figure 8.14 saliency maps for a T2-weighted sample from the IoPPN dataset are presented for comparison. The subject presented is 30 years old but predicted age by the SENet-MR model is 65.4 years. It should be noted that saliency maps for a T2-weighted sample presented in Figure 7.8 are different from the saliency maps for T1-weighted samples shown in Figures 8.8 – 8.10. The model used more input sample features from the T2-weighted sample compared to the T1-weighted samples.

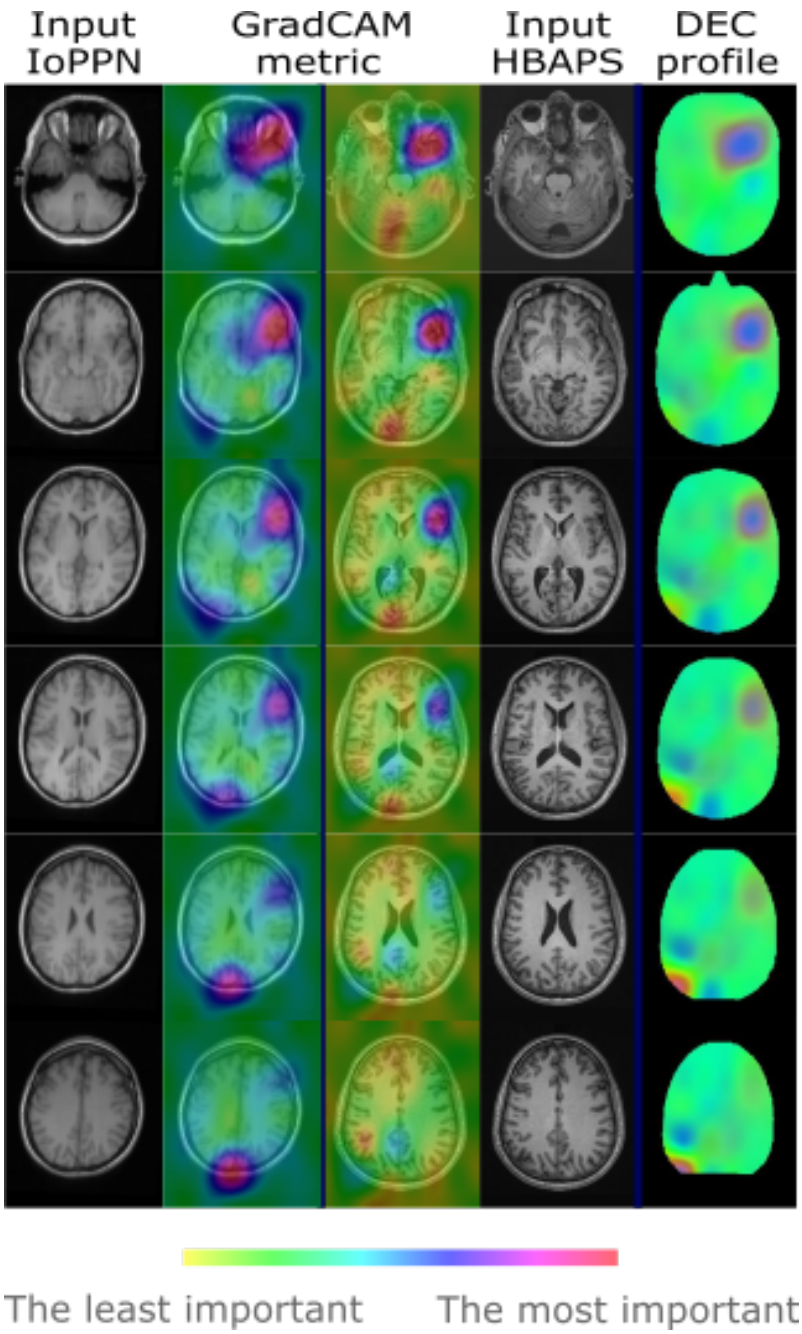

Figure 8.8: Comparison of the Grad-CAM saliency maps obtained for the samples from the IoPPN test data consisting of healthy subjects and HBAPS test data with 0 prediction error using the SENet-MR model. Both samples belong to the subjects who are 28 years old. The two leftmost columns show planes of a sample belonging to the IoPPN dataset and corresponding saliency map, the third and fourth columns – for the HBAPS test data sample. The rightmost column shows corresponding planes in the features describing whole brain ageing profile obtained using the DEC method (described in Section 6.3.3). The maps describing the profile are multiplied by the MNI152 head mask for clarity.

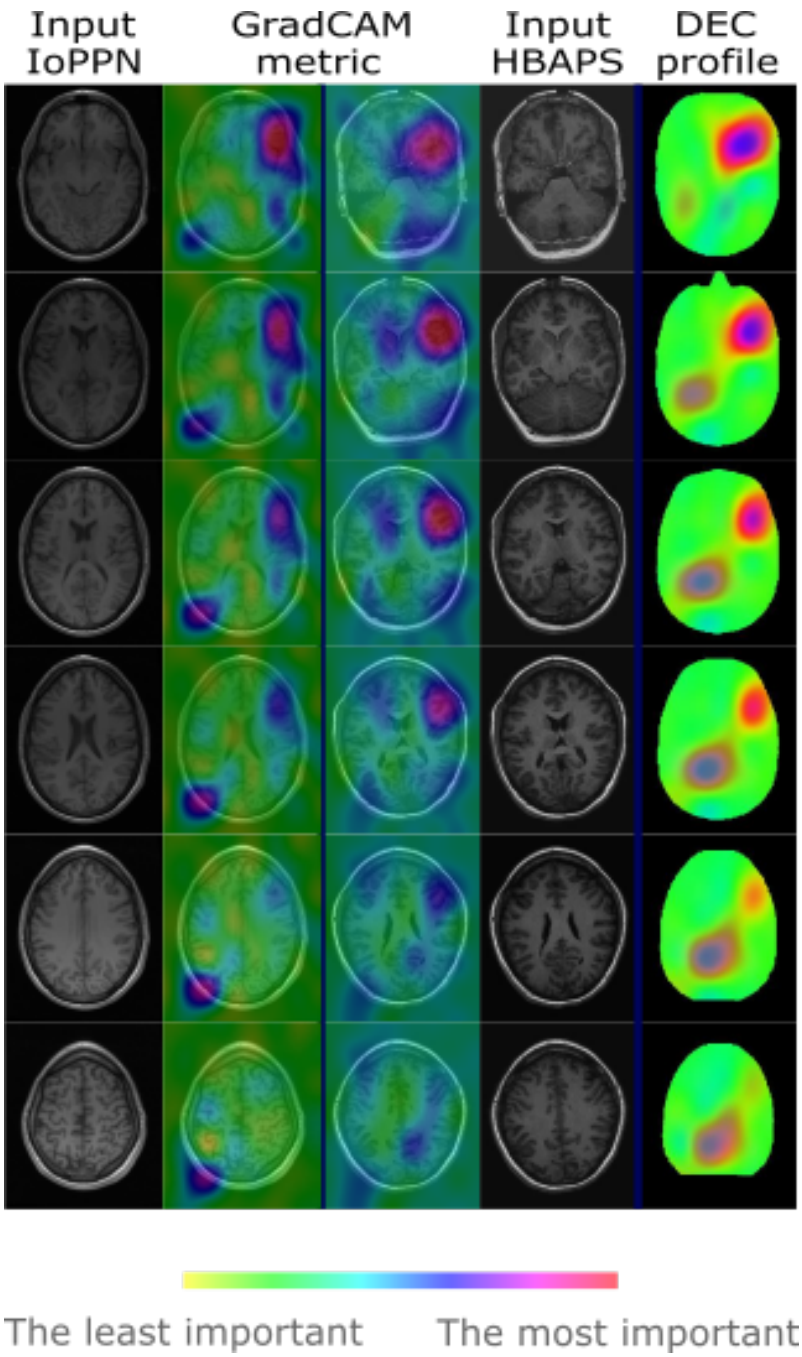

Figure 8.9: Comparison of the Grad-CAM saliency maps obtained for the samples from the IoPPN test data consisting of healthy subjects and HBAPS test data with the prediction error equal to the sum of MAE and standard deviation on it, 9.55 years, in the HBAPS testing data using the SENet-MR model. Both samples belong to the subjects who are 22 years old. The two leftmost columns show planes of a sample belonging to the IoPPN dataset and corresponding saliency map, the third and fourth columns – for the HBAPS test data sample. The rightmost column shows corresponding planes in the features describing whole brain ageing profile obtained using the DEC method (described in Section 6.3.3). The maps describing the profile are multiplied by the MNI152 head mask for clarity.

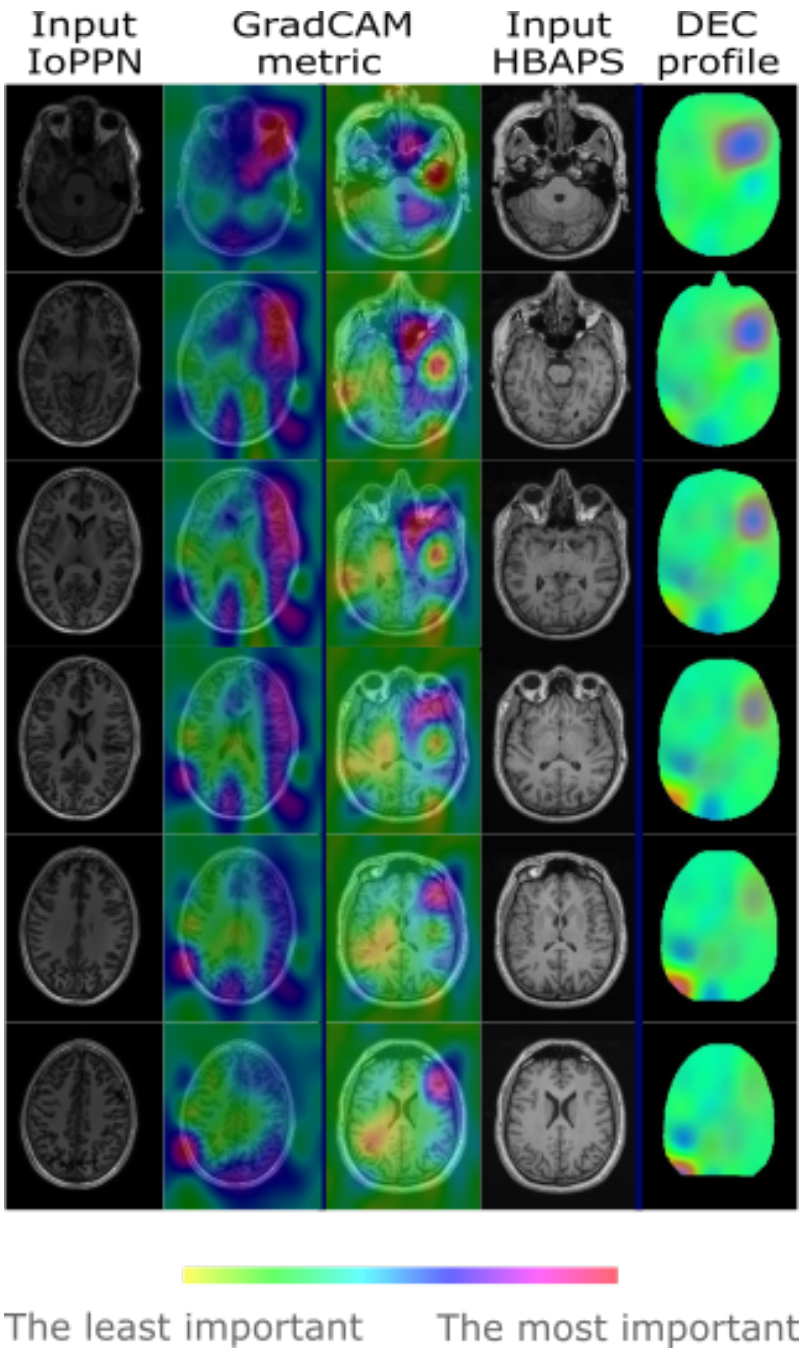

Figure 8.10: Comparison of the Grad-CAM saliency maps obtained for the samples from the IoPPN test data consisting of healthy subjects and HBAPS test data with prediction error of 20 years using the SENet-MR model. Both samples belong to the subjects who are 29 years old. The two leftmost columns show planes of a sample belonging to the IoPPN dataset and corresponding saliency map, the third and fourth columns – for the HBAPS test data sample. The rightmost column shows corresponding planes in the features describing whole brain ageing profile obtained using the DEC method (described in Section 6.3.3). The maps describing the profile are multiplied by the MNI152 head mask for clarity.

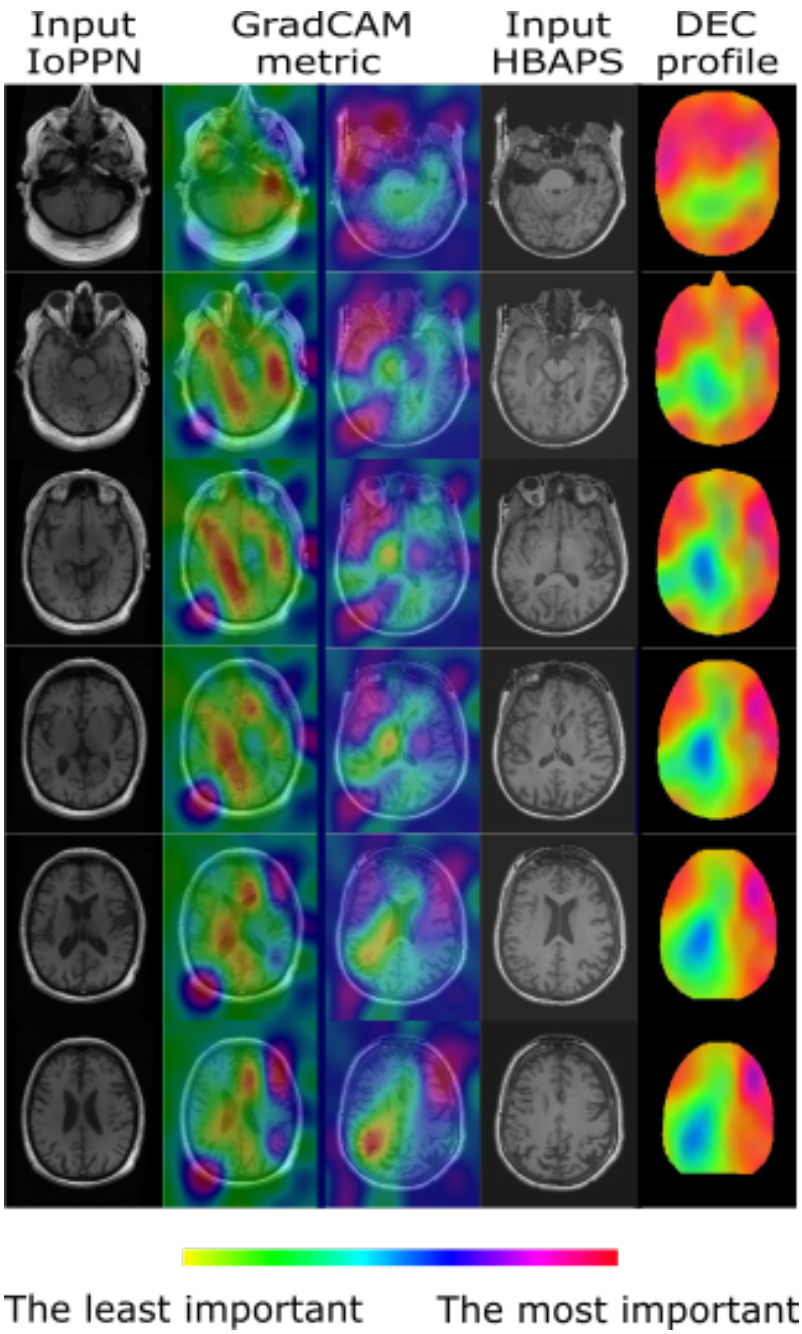

Figure 8.11: Comparison of the Grad-CAM saliency maps obtained for the samples from the IoPPN test data consisting of diseased subjects and HBAPS test data with 0 prediction error using the SENet-MR model. Both samples belong to the subjects who are 69 years old. The two leftmost columns show planes of a sample belonging to the IoPPN dataset and corresponding saliency map, the third and fourth columns – for the HBAPS test data sample. The rightmost column shows corresponding planes in the features describing whole brain ageing profile obtained using the DEC method (described in Section 6.3.3). The maps describing the profile are multiplied by the MNI152 head mask for clarity.

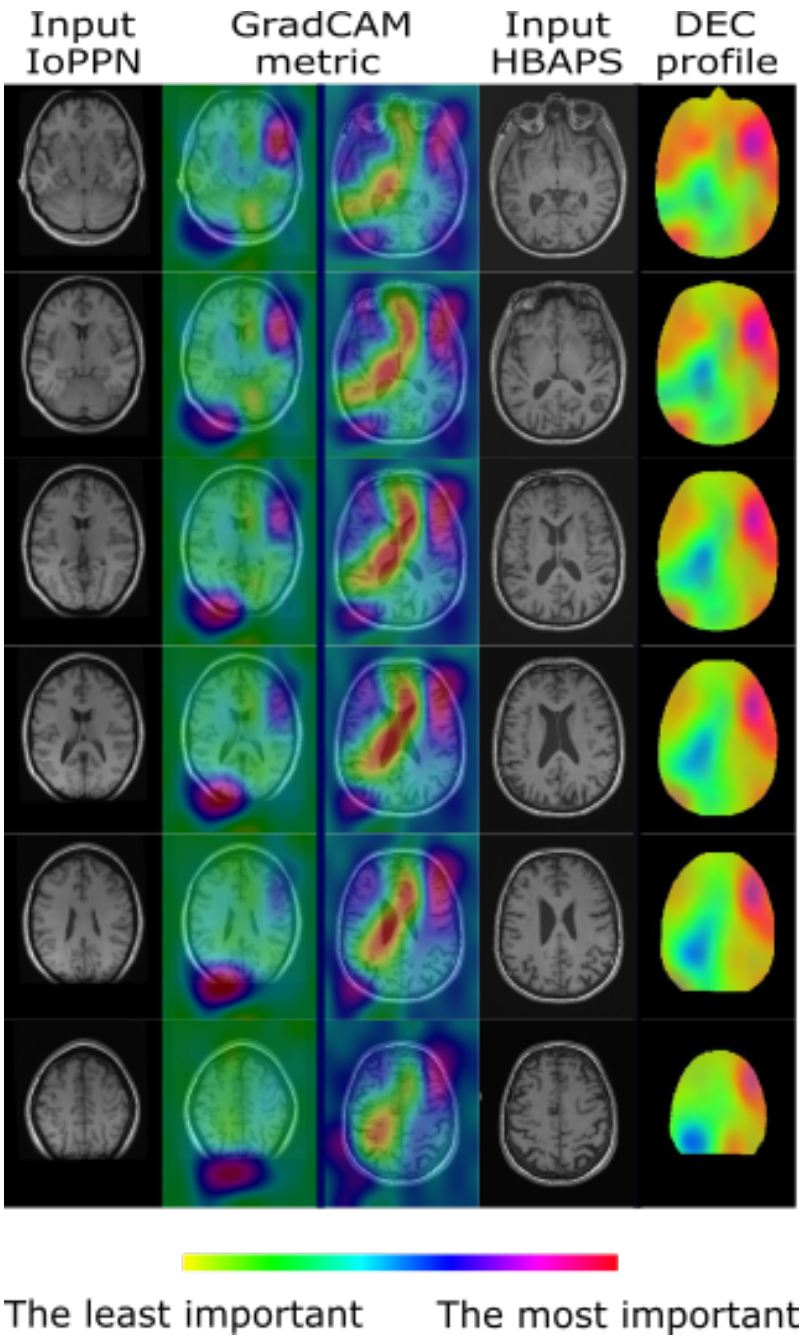

Figure 8.12: Comparison of the Grad-CAM saliency maps obtained for the samples from the IoPPN test data consisting of diseased subjects and HBAPS test data with the prediction error equal to the sum of MAE and standard deviation on it, 9.55 years, in the HBAPS testing data using the SENet-MR model. Both samples belong to the subjects who are 52 years old. The two leftmost columns show planes of a sample belonging to the IoPPN dataset and corresponding saliency map, the third and fourth columns – for the HBAPS test data sample. The rightmost column shows corresponding planes in the features describing whole brain ageing profile obtained using the DEC method (described in Section 6.3.3). The maps describing the profile are multiplied by the MNI152 head mask for clarity.

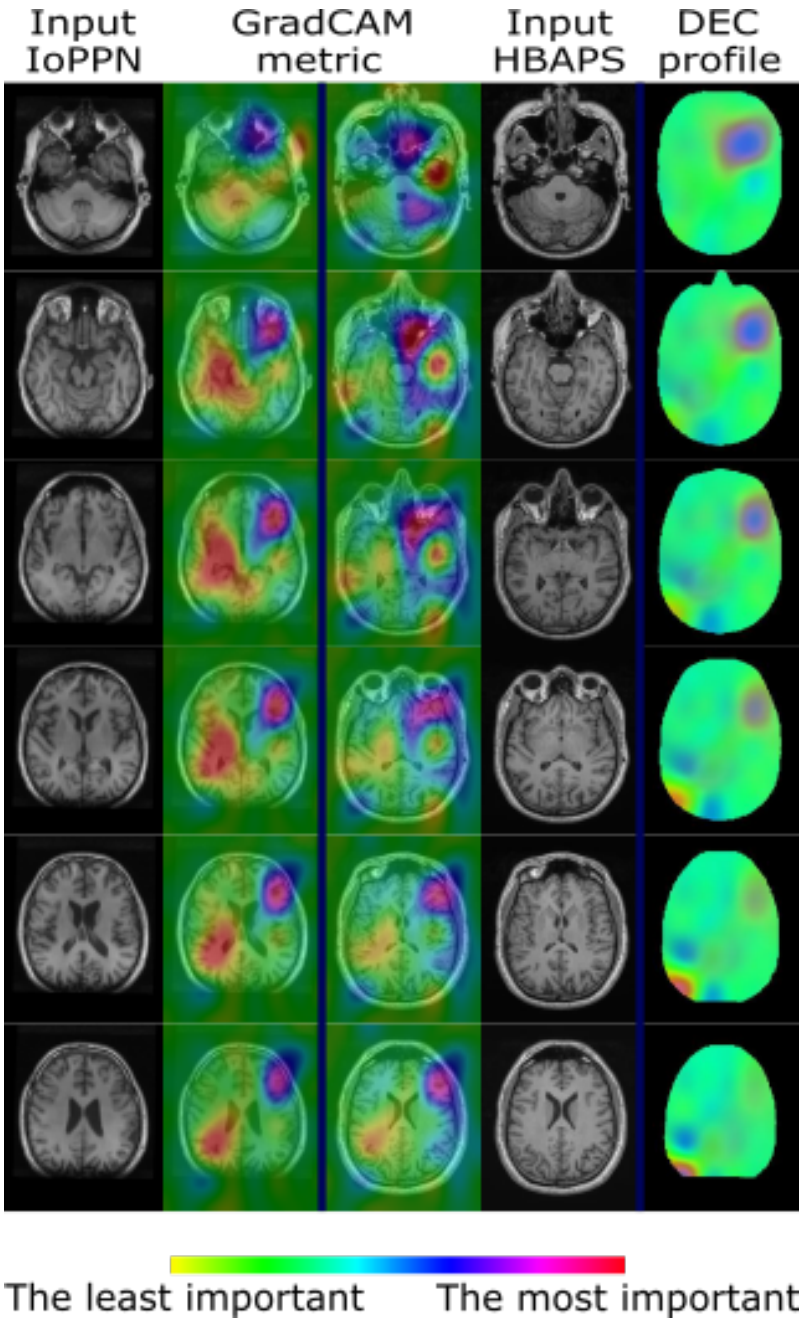

Figure 8.13: Comparison of the Grad-CAM saliency maps obtained for the samples from the IoPPN diseased subjects and HBAPS dataset sample with prediction error of 20 years using the SENet-MR model. Both samples belong to the subjects who are 29 years old. The two leftmost columns show planes of a sample belonging to the IoPPN dataset's diseased sample and corresponding saliency map, the third and fourth columns – for the HBAPS test data sample. The rightmost column shows corresponding planes in the features describing whole brain ageing profile obtained using the DEC method (described in Section 6.3.3). The maps describing the profile are multiplied by the MNI152 head mask for clarity.

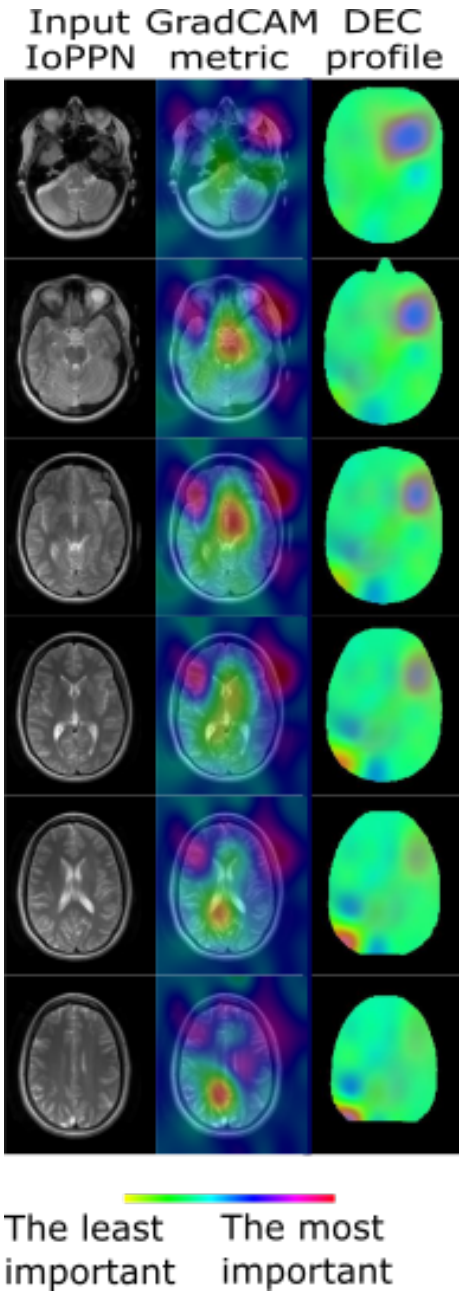

Figure 8.14: Example of the saliency maps obtained for a non-T1-weighted sample from the IoPPN dataset, with prediction error of 35.4 years using the SENet-MR model. The subject is 30 years old. The method of saliency mapping is Grad-CAM. The rightmost column shows corresponding planes in the features describing the whole brain ageing profile obtained using the DEC method (described in Section 6.3.3). The maps describing the profile are multiplied by the MNI152 head mask for clarity.

8.5 Discussion and conclusions

In this Chapter the methods for brain age prediction and extracting age-related features presented in Chapter 5 and 7 were first also examined on robustness with respect to input resolution. The brain age prediction models were also examined on robustness with respect to input resolution. The models were not affected by input resolution, but their performance was affected by variable quality of images, i.e. variability of acquisition sequences (and potentially by the disease status of subjects).

In Figure 8.1 and Table 8.2 the performance of the SENet-MR model on the HBAPS test data with lowered resolution is presented. It can be observed that the data resolution did not affect model's performance up to voxel size of $1 \times 1 \times 4$ mm. With the voxel size of $1 \times 1 \times 5$ mm, the model's MAE increased by 7%. Accurate comparison with existing literature is not possible. For instance, Feng et al [39] have demonstrated that a single plane from a 3D MRI data can be used to achieve comparable performance to using whole MRI scan. Feng et al trained a network equivalent to the network used for analysing 3D MRI data, but taking 2D slices as an input. The MAE achieved by considering a single MRI slice in sagittal plane from the test samples was 5.5 years, compared to 4.06 years achieved on whole MRI scans [39]. The experiment by Feng et al is different from the one in this chapter. In the experiment here the brain age prediction model uses an input an array of data with lowered resolution, while in the experiments by Feng et al a single plane of higher resolution data is used.

Further to this, the brain age prediction models' performance is studied with respect to identifying prediction outliers. Firstly, as was shown in Chapter 7, the HBAPS test data contained the samples affected at the preprocessing stage. As a results of preprocessing, a number of samples have smaller overall data size compared to most of the samples. These samples also have much higher associated prediction error and the features identified in saliency maps related to predictions are not highlighting any biologically relevant regions. Example of the HBAPS test data sample affected by the preprocessing stage is repeated in Figure 8.15 for convenience. Table 8.4 and Figure 8.2 present the performance of the SENet-MR and SENet-OR models on whole HBAPS test data with the samples corrupted by preprocessing being excluded. As expected, the models' measured performance improved - MAE decreased by 12.8% and 7.4% for the SENet-MR and SENet-OR models respectively. After removing the corrupted samples, it can also be observed that the predictions of

SENet-OR model are more concentrated within the 95% confidence interval of the linear fit than the SENet-MR model predictions.

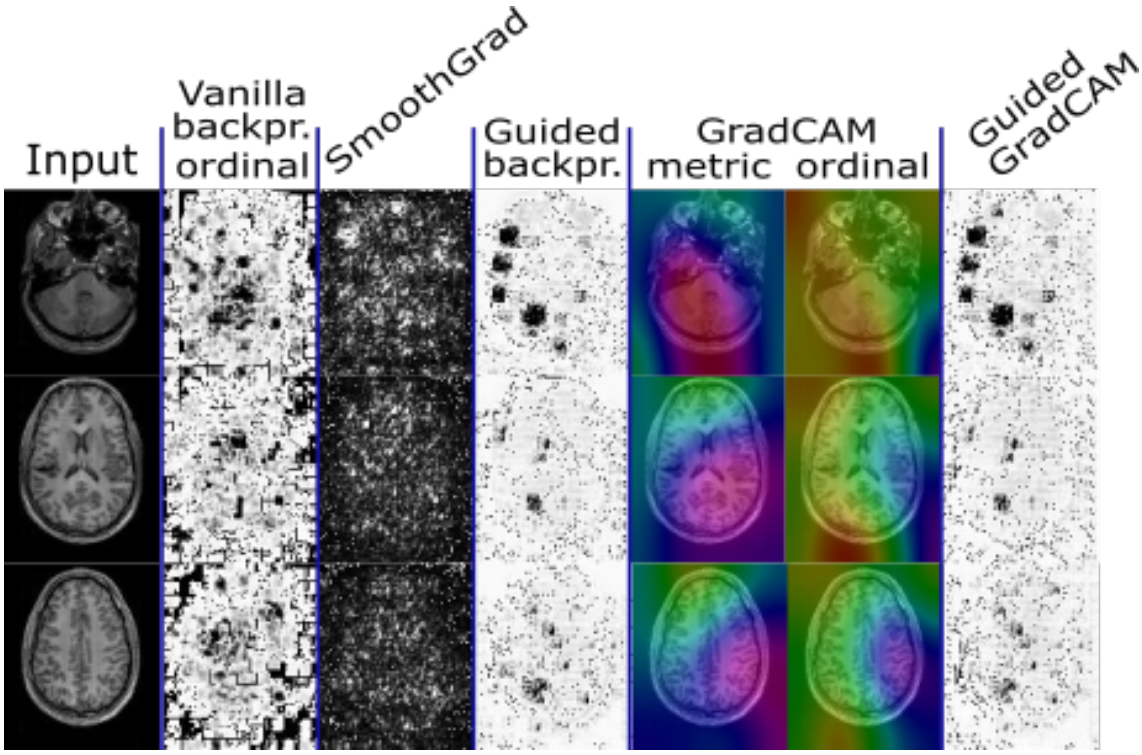

Figure 8.15: Example of the HBAPS test data sample affected by the preprocessing stage. Saliency maps are shown for the sample with the largest prediction error (60.2 years) in the HBAPS testing data using the SENet-MR model. The maps named ordinal are produced by applying ordinal saliency mapping technique described in Section 7.3.2 with corresponding method. These maps are produced using the SENet-OR model and SENet-MR model is used otherwise. The leftmost column shows corresponding planes in the input sample. Three planes are presented from the samples as resulting saliency maps are less informative compared to the samples with smaller associated prediction error.

Finally, the robustness of the SENet-MR and SENet-OR models' predictions was examined using the independent IoPPN dataset, which is more representative of the variability expected in a true clinical dataset. For the testing purposes two subsets of the IoPPN dataset consisting of only the samples acquired using T1-weighted sequences are used. One subset consists of only healthy and one consisting of only diseased subjects as detailed in Section 4.2.2. The MAE of the SENet-OR model was larger compared to the SENet-MR model on both healthy and diseased subjects from the IoPPN dataset by 46.0% and 37.9%. The increase in MAE for both brain age prediction models can be explained by the fact that the samples in the IoPPN dataset are of varying quality and were acquired us-

ing different imaging protocols while the HBAPS training data consisted of data acquired using research study protocols which are mostly based on the Alzheimer's Disease Neuroimaging Initiative (ADNI) protocol. It should be noted that for both the SENet-MR and SENet-OR models the MAE was lower on diseased subjects compared to the healthy ones. This is due to the fact that for many samples acquired from diseased subjects prediction error was as low as for healthy subjects. This highlights the need of collecting longitudinal datasets as this would allow to benchmark the changes of individual subjects with respect to their previous scans taken in younger age and/or before the disease. This also supports the importance of collecting longitudinal samples in clinical practice for using brain age assessment as a biomarker for neurodegenerative conditions. MAE was lower on diseased subjects also possibly due to the fact that there are more diseased subjects in the IoPPN dataset compared to healthy ones.

In Figure 8.5 average prediction error is presented for each age label in the IoPPN dataset for the SENet-MR and SENet-OR models and both healthy and diseased subjects. In Chapter 5 it is shown that the brain age prediction models underperformed on the youngest and the oldest subjects in both whole HBAPS test data and HBAPS test data containing only independent subjects. As can be observed in Figure 8.5 on a larger but less "clean" dataset such as the IoPPN dataset these effects are more apparent for the healthy and diseased subjects. The influence of variability in a dataset may be also observed in the box plot of the SENet-MR and SENet-OR predictions in Figures 8.6 and 8.7. From Figures 8.6 and 8.7 it can be noted that the SENet-OR model's predictions were also more tended towards the largest true age label available in the HBAPS training data used for training the SENet-OR model. This may indicate that the SENet-OR model may be biased towards older subjects and further investigation of the model is needed.

Performance of the SENet-MR and SENet-OR models on the IoPPN dataset could also be influenced by the biases introduced by the HBAPS dataset used for training of the models. HBAPS dataset consists of healthy controls determined mostly by their self-assessment and self-reporting of their conditions. Such datasets are also acquired by collecting data from volunteers. As the result such datasets are not representative of whole population. Among older subjects healthier subjects tend to volunteer in the research studies as this requires physical and mental effort. This also may introduce a bias. For instance, older subjects with mobility difficulties are less likely to participate even they are cognitively

not affected by ageing. As the result subjects contained in the research datasets contain more homogeneous subjects compared to population.

Comparison of the features extracted from healthy IoPPN subjects' samples and the HBAPS test data samples with the same associated prediction error shows that these highlight the same regions, as shown in Figures 8.8 – 8.13. Therefore, the brain age prediction models presented in this thesis appear to be capable of extracting consistent (and therefore likely relevant) age-related features from the samples of variable quality.

For comparison in Figure 8.14 saliency maps from a T2-weighted sample from the IoPPN dataset are presented. The features are presented in comparison to those describing the corresponding interval in the whole brain population-wide profile extracted using the DEC method. The clear differences in the features extracted from a T2-weighted sample show that variability of acquisition sequences has affected brain age prediction models' performance, but also that such differences can potentially be detected and used as part of a quality control process to exclude outliers, or highlight them for further manual investigation.

In conclusion, it should be noted that further work is needed to understand the reasons for models' performance on clinical datasets. Firstly, it would be needed to investigate the tendency of ordinal model's prediction towards the largest true age label present in the data used for training. The SENet-OR model uses $K-1$ binary classifiers (K is a number of true age labels in training data) after the feature extraction step in order to predict brain age as described in Section 5.3.2. Saliency mapping can be used in order to identify the bias by constructing saliency maps separately from each binary classifier and comparing the differences. Possible existence of the bias in ordinal model may also be investigated by deconvolutional network approach introduced by Zeiler and Fergus [47]. Deconvolutional network approach can be used to identify which of the binary classifiers contributed the most to accurate predictions. As the result of this investigation, a modification to the loss function could be introduced to reduced the bias towards older age labels. For example, in Section 5.3.2 it is described that all binary classifiers were equally contributing to the loss during the training. Investigation of the binary classifier could allow to introduce non-equal weighting for reducing the bias. Secondly, information about the diseased subjects could be obtained in order to understand the brain age prediction models' performance on the samples acquired from these subjects. This could allow to understand why for many diseased subjects prediction error was as low as 0. Thirdly, testing the models described in

this thesis on clinical dataset containing longitudinal data would also benefit the analysis of such models' ability to be used as a biomarker in clinical setting. In such analysis the data used for a single subject could be collected also at younger ages and/or before the disease onset.

Nevertheless, I have shown that it is possible to use the models described in this thesis on clinical data if data quality control measures are in place, i.e. the acquisition sequences are the same as the one used to acquire the data for training. Therefore, there is also need for further research into the methods for outlier detection. Saliency maps produced for brain age predictions represent useful candidate for outlier detection method.

Chapter 9

Conclusion

This thesis has presented novel methods for the analysis structural Magnetic Resonance Imaging (MRI) data with the aim of understanding human brain ageing. A deep learning model is developed for the purpose of brain age prediction. The methods are proposed to extract whole brain population-wide ageing profiles and understand the features describing them. In this thesis clinical applicability of brain age prediction algorithms is also discussed. In the final chapter, the key contributions of this thesis are summarised and their limitations are discussed along with possible further developments.

9.1 Key contributions

Chapter 4 describes first contribution of this thesis - compiling a dataset of structural MRI scans from publicly available data. The Healthy Brain Ageing from Public Sources (HBAPS) dataset consists of 10,878 MRI T1-weighted scans acquired in healthy subjects was constructed from 34 publicly available datasets with the age range from 17 to 96 years of age.

The second contribution is developing deep ordinal regression model for predicting brain age from MRI scans in order to take into account accumulative nature of brain ageing. Ordinal regression model was compared to the metric regression model also implemented and trained. Ordinal regression model achieved the lowest Mean Absolute Error (MAE) on MRI data with minimal preprocessing data up-to-date in comparison to existing literature. The model's performance was also not affected by the fact that the training data used was compiled from 34 data sources.

Further in Chapter 6 the models for brain age prediction are used for extracting a population-wide brain ageing profiles. For this purpose two existing methods (the methods of Deep Embedded Clustering (DEC) and Preference Ranking Organization METHod for Enrichment of Evaluations (PROMETHEE) II) were applied and two new method were developed (the method considering the distribution of predictions of the brain age prediction model and ordinal DEC). All the profiles obtained agree within the accuracy achieved by the brain age predictions models suggesting that they are all being driven by the same underlying brain features, and again suggesting that these features may be biologically meaningful.

In Chapter 7 age-related features describing the whole brain population-wide ageing profiles are obtained and this another contribution of this thesis. For this purpose, the age-related features were obtained using five existing saliency mapping methods and in order to take into account accumulative nature of ageing the methodology for saliency mapping ordinal models was developed. Among existing saliency mapping methods the methods of guided backpropagation, Gradient Class Activation Mapping (Grad-CAM) and guided Grad-CAM were shown to be usable for studying brain ageing features from MRI data with minimal preprocessing. Average maps derived for the whole brain ageing profiles extracted in Chapter 6 are of biological significance as features are in accordance with previous research and existing knowledge on biology of brain ageing. Saliency maps produced using proposed methodology for saliency mapping for ordinal models also highlighted biologically interpretable features.

In Chapter 8 final contribution of this thesis is discussed, i.e. examining generalisability of the models for brain age prediction presented in Chapter 5. This is studied by measuring their performance on test data of the HBAPS dataset with reduced resolution. Further, the methods discussed above are applied to a clinical dataset. For this purpose independent, locally collected, dataset, in which two subsets of Institute of Psychiatry, Psychology & Neuroscience (IoPPN) test data are used - one consisting of healthy subjects and one from diseased ones.

9.2 Limitations of this work and further developments

First, limitations of this work originate from the limitations of the HBAPS dataset. The HBAPS dataset is highly unbalanced as 50% of the samples lie in the range from 17 to 30 years of age and there are few samples available for the samples labelled to be 17 years old

and for the samples labelled to be older than 82 years old. The samples labelled to be of age 36 to 47 years are also underrepresented. In order to achieve more uniform data distribution more data should be added to the dataset for the age labels which are underrepresented. As this may be problematic to use MRI data collected from real subjects, especially in older subjects, data augmentation should be used. Recently, Generative Adversarial Network (GAN) was proposed for MRI data augumentation and head MRI data in particular (for example the work by Kazuhiro et al [251]).

Improving the HBAPS dataset would improve the quality of brain age prediction models. The brain age prediction model can also be further developed by improving their architecture. For this purpose the deconvolutional network approach by Zeiler and Fergus [47] can be used to understand which network parts contributed the most to accurate predictions and what have they detected at each step during a forward pass. Then more educated choice of improved architecture would be possible. This would allow to remove unnecessary layers and training would be less time-consuming. Then more extensive network hyperparameter optimization would be possible for further results improvement.

Data augmentation would also allow to validate and improve the whole brain population-wide ageing profiles as in this thesis a cross-sectional dataset is used. Longitudinal data samples can be first assigned to the ageing intervals using a pretrained method and the samples' true labels can then be compared to the profiles obtained using the HBAPS test data. In order to further explore population-wide effects male and female profiles could be extracted separately. Population-wide effects could also be explored further by expanding the study cohort. Main methodological contribution needed further to develop extracting brain ageing profiles is fine-tuning training the ordinal DEC method proposed in this thesis to be trained on MRI data after successfully implementing it on synthetic data. Advantage of DEC-based method is that once the method is trained it can be used to analyse new samples to study in which profile interval they belong. The ordinal DEC method proposed in this thesis allows to also take into ordinality of samples in problems such as brain age prediction.

Extracting age-related features for the ageing profiles extracted also had a number limitations. The features extracted resulted in asymmetrical maps. In order to understand this property of the features extracted computing separate average maps for male and female subjects and left- and right-handed subjects would be needed. Another important limitation is that the profile features extracted using the method for ordinal saliency mapping

proposed in this thesis did not reveal important features to be increasing in intensity with age. This is not in accordance with the feature profiles extracted using existing saliency mapping methods and known biological course of structural brain changes. Understanding the cause of this discrepancy between the features related to ageing profiles is out of scope of this thesis. Therefore, further work is needed to understand whether these differences are explained by the properties of the ordinal saliency mapping methodology, which need to be studied, or the properties of the ordinal brain age prediction model.

Further work is also needed to make the brain age prediction models described in this thesis usable in clinical environment. This would need further research into the methods of outlier detection, i.e. identifying input samples which are affected either at the acquisition or preprocessing stages. In this thesis it was shown that saliency mapping methods are good candidate for developing outlier detection methods.

Bibliography

[1] Petar Velickovic. *The resurgence of structure in deep neural networks*. PhD thesis, University of Cambridge, January 2019. → pages 9, 41, 42, 44, 48, 49

[2] Sumit Saha. A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way, December 2018. URL https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53. → pages 9, 47

[3] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-Excitation Networks. *eprint arXiv:1709.01507*, September 2017. URL https://ui.adsabs.harvard.edu/abs/2017arXiv170901507H. → pages 9, 11, 54, 104, 106

[4] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal Regression with Multiple Output CNN for Age Estimation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4920–4928, 2016. → pages 10, 31, 56, 57, 58

[5] B. Zhou, A. Khosla, Lapedriza A, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *CVPR*, 2016. → pages 10, 30, 33, 63, 65, 162

[6] Lior Shamir and Joe Long. Quantitative Machine Learning Analysis of Brain MRI Morphology throughout Aging. *Current aging science*, 9(4):310–317, 2016. doi:10.2174/1874609809666160413113711. Place: United Arab Emirates. → pages 14, 137, 149, 150, 157, 158, 159, 160, 190, 192

[7] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-Wise Relevance Propagation: An Overview. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Muller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep*

Learning, pages 193–209. Springer International Publishing, Cham, 2019. ISBN 978-3-030-28954-6. doi:10.1007/978-3-030-28954-6_10. URL https://doi.org/10.1007/978-3-030-28954-6_10. → pages 20, 30, 33, 63, 70, 71, 72, 162

[8] Brian K. Atchinson and Daniel M. Fox. From The Field: The Politics Of The Health Insurance Portability And Accountability Act. *Health Affairs*, 16(3):146–150, May 1997. doi:10.1377/hlthaff.16.3.146. URL https://www.healthaffairs.org/doi/10.1377/hlthaff.16.3.146. Publisher: Health Affairs. → pages 20, 95

[9] Ulrike Kübler. Structured Clinical Interview for DSM-IV (SCID). In Marc D. Gellman and J. Rick Turner, editors, *Encyclopedia of Behavioral Medicine*, pages 1919–1920. Springer, New York, NY, 2013. ISBN 978-1-4419-1005-9. doi:10.1007/978-1-4419-1005-9_66. URL https://doi.org/10.1007/978-1-4419-1005-9_66. → pages 20, 95

[10] Origins of DSM-I: a study in appearance and reality. *American Journal of Psychiatry*, 148(4):421–431, April 1991. doi:10.1176/ajp.148.4.421. URL https://ajp.psychiatryonline.org/doi/abs/10.1176/ajp.148.4.421. Publisher: American Psychiatric Publishing. → pages 20, 95

[11] Verna C. Pangman, Jeff Sloan, and Lorna Guse. An examination of psychometric properties of the Mini-Mental State Examination and the Standardized Mini-Mental State Examination: Implications for clinical practice. *Applied Nursing Research*, 13(4):209–213, November 2000. doi:10.1053/apnr.2000.9231. URL https://www.sciencedirect.com/science/article/pii/S0897189700254529. → pages 20, 95

[12] *Handbook of clinical neurology.* Elsevier, Amsterdam, 2003. OCLC: 698911345. → pages 20, 95

[13] Nicola Amoroso, Marianna La Rocca, Loredana Bellantuono, Domenico Diacono, Annarita Fanizzi, Eufemia Lella, Angela Lombardi, Tommaso Maggipinto, Alfonso Monaco, Sabina Tangaro, and Roberto Bellotti. Deep Learning and Multiplex Networks for Accurate Modeling of Brain Age. *Frontiers in aging neuroscience*, 11: 115–115, May 2019. doi:10.3389/fnagi.2019.00115. URL https://pubmed.ncbi.nlm.nih.gov/31178715. Publisher: Frontiers Media S.A. → pages 21, 82, 101, 102, 105, 160, 164, 190, 191

[14] Geneviève Richard, Knut Kolskår, Kristine M. Ulrichsen, Tobias Kaufmann, Dag Alnæs, Anne-Marthe Sanders, Erlend S. Dørum, Jennifer Monereo Sánchez, Anders Petersen, Hege Ihle-Hansen, Jan Egil Nordvik, and Lars T. Westlye. Brain age prediction in stroke patients: Highly reliable but limited sensitivity to cognitive performance and response to cognitive training. *NeuroImage: Clinical*, 25:102159, January 2020. doi:10.1016/j.nicl.2019.102159. URL http://www.sciencedirect.com/science/article/pii/S2213158219305054. → pages 21, 101, 102, 104, 105

[15] 2018 Ageing Report: Policy challenges for ageing societies. Technical report, European Commision, May 2018. → pages 28

[16] Dementia: a public health priority. Technical report, World Health Organization, Geneva, 2012. → pages 28

[17] The epidimiology and impact of dementia: current state and future trends. Technical report, World Health Organization, Geneva, 2015. → pages 28

[18] Carlos López-Otín, Maria A. Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. The Hallmarks of Aging. *Cell*, 153(6):1194–1217, June 2013. doi:10.1016/j.cell.2013.05.039. URL https://doi.org/10.1016/j.cell.2013.05.039. Publisher: Elsevier. → pages 29, 77

[19] Trey Hedden and John D. E. Gabrieli. Insights into the ageing mind: a view from cognitive neuroscience. *Nature Reviews Neuroscience*, 5(2):87–96, February 2004. doi:10.1038/nrn1323. URL https://doi.org/10.1038/nrn1323. → pages 29

[20] Katherine L. Roberts and Harriet A. Allen. Perception and Cognition in the Ageing Brain: A Brief Review of the Short- and Long-Term Links between Perceptual and Cognitive Decline. *Frontiers in aging neuroscience*, 8:39–39, March 2016. doi:10.3389/fnagi.2016.00039. URL https://pubmed.ncbi.nlm.nih.gov/26973514. Publisher: Frontiers Media S.A. → pages 29

[21] Andras Bilkei-Gorzo. Genetic mouse models of brain ageing and Alzheimer's disease. *Pharmacology & Therapeutics*, 142(2):244–257, May 2014. doi:10.1016/j.pharmthera.2013.12.009. URL http://www.sciencedirect.com/science/article/pii/S0163725813002519. → pages 29

[22] Ake T. Lu, Eilis Hannon, Morgan E. Levine, Eileen M. Crimmins, Katie Lunnon, Jonathan Mill, Daniel H Geschwind, and Steve Horvath. Genetic architecture of epigenetic and neuronal ageing rates in human brain regions. *Nature communications*, 8:15353–15353, May 2017. doi:10.1038/ncomms15353. URL https://pubmed.ncbi.nlm.nih.gov/28516910. Publisher: Nature Publishing Group. → pages 29

[23] Tobias Kaufmann, Dennis van der Meer, Nhat Trung Doan, Emanuel Schwarz, Martina J. Lund, Ingrid Agartz, Dag Alnæs, Deanna M. Barch, Ramona Baur-Streubel, Alessandro Bertolino, Francesco Bettella, Mona K. Beyer, Erlend Bøen, Stefan Borgwardt, Christine L. Brandt, Jan Buitelaar, Elisabeth G. Celius, Simon Cervenka, Annette Conzelmann, Aldo Córdova-Palomera, Anders M. Dale, Dominique J.-F de Quervain, Pasquale Di Carlo, Srdjan Djurovic, Erlend S. Dørum, Sarah Eisenacher, Torbjørn Elvsåshagen, Thomas Espeseth, Helena Fatouros-Bergman, Lena Flyckt, Barbara Franke, Oleksandr Frei, Beathe Haatveit, Asta K. Håberg, Hanne F. Harbo, Catharina A. Hartman, Dirk Heslenfeld, Pieter J. Hoekstra, Einar A. Høgestøl, Terry Jernigan, Rune Jonassen, Erik G. Jönsson, Peter Kirsch, Iwona Kłoszewska, Knut-Kristian Kolskår, Nils Inge Landrø, Stephanie Le Hellard, Klaus-Peter Lesch, Simon Lovestone, Arvid Lundervold, Astri J. Lundervold, Luigi A. Maglanoc, Ulrik F. Malt, Patrizia Mecocci, Ingrid Melle, Andreas Meyer-Lindenberg, Torgeir Moberget, Linn B. Norbom, Jan Egil Nordvik, Lars Nyberg, Jaap Oosterlaan, Marco Papalino, Andreas Papassotiropoulos, Paul Pauli, Giulio Pergola, Karin Persson, Geneviève Richard, Jaroslav Rokicki, Anne-Marthe Sanders, Geir Selbæk, Alexey A. Shadrin, Olav B. Smeland, Hilkka Soininen, Piotr Sowa, Vidar M. Steen, Magda Tsolaki, Kristine M. Ulrichsen, Bruno Vellas, Lei Wang, Eric Westman, Georg C. Ziegler, Mathias Zink, Ole A. Andreassen, and Lars T. Westlye. Genetics of brain age suggest an overlap with common brain disorders. *bioRxiv*, page 303164, January 2018. doi:10.1101/303164. URL http://biorxiv.org/content/early/2018/04/17/303164.abstract. → pages 29, 101

[24] Katja Franke, Gabriel Ziegler, Stefan Klöppel, and Christian Gaser. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. *NeuroImage*, 50(3):883–892, April 2010. doi:10.1016/j.neuroimage.2010.01.005. URL http://www.sciencedirect.com/science/article/pii/S1053811910000108. → pages 29, 101, 102, 103

[25] Nico U. F. Dosenbach, Binyam Nardos, Alexander L. Cohen, Damien A. Fair, Jonathan D. Power, Jessica A. Church, Steven M. Nelson, Gagan S. Wig, Alecia C. Vogel, Christina N. Lessov-Schlaggar, Kelly Anne Barnes, Joseph W. Dubis, Eric Feczko, Rebecca S. Coalson, John R. Pruett, Jr, Deanna M. Barch, Steven E. Petersen, and Bradley L. Schlaggar. Prediction of individual brain maturity using fMRI. *Science (New York, N.Y.)*, 329(5997):1358–1361, September 2010. doi:10.1126/science.1194144. URL https://pubmed.ncbi.nlm.nih.gov/20829489. → pages 29

[26] Rajat Raina, Anand Madhavan, and Andrew Y. Ng. Large-Scale Deep Unsupervised Learning Using Graphics Processors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 873–880, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 978-1-60558-516-1. doi:10.1145/1553374.1553486. URL https://doi.org/10.1145/1553374.1553486. event-place: Montreal, Quebec, Canada. → pages 29

[27] D. C. Cireşan, U. Meier, L. M. Gambardella, and J. Schmidhuber. Deep, Big, Simple Neural Nets for Handwritten Digit Recognition. *Neural Computation*, 22(12):3207–3220, December 2010. doi:10.1162/NECO_a_00052. → pages 29

[28] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A Matlab-like Environment for Machine Learning. In *BigLearn, NIPS Workshop*, 2011. → pages 29

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf. → pages 29, 40, 86, 167

[30] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghe-

mawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. URL https://www.tensorflow.org/. → pages 29, 40

[31] Human Connectome Project | Mapping the human brain connectivity, . URL http://www.humanconnectomeproject.org/. Library Catalog: www.humanconnectomeproject.org. → pages 29

[32] ADNI | Alzheimer's Disease Neuroimaging Initiative, . URL http://adni.loni.usc.edu/. Library Catalog: adni.loni.usc.edu. → pages 29, 137

[33] LONI Image Data Archive (IDA), . URL https://ida.loni.usc.edu/login.jsp. → pages 29

[34] NITRC: Welcome, . URL https://www.nitrc.org/. → pages 29

[35] James H. Cole, Rudra P.K. Poudel, Dimosthenis Tsagkrasoulis, Matthan W.A. Caan, Claire Steves, Tim D. Spector, and Giovanni Montana. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, 163:115–124, December 2017. doi:10.1016/j.neuroimage.2017.07.059. URL http://www.sciencedirect.com/science/article/pii/S1053811917306407. → pages 30, 98, 101, 129, 131, 196

[36] Qi Qi, Baolin Du, Mingyong Zhuang, Yue Huang, and Xinghao Ding. Age Estimation from MR Images via 3D Convolutional Neural Network and Densely Connect. In Long Cheng, Andrew Chi Sing Leung, and Seiichi Ozawa, editors, *Neural Information Processing*, pages 410–419, Cham, 2018. Springer International Publishing. ISBN 978-3-030-04239-4. → pages 30, 101, 102, 105

[37] Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz,

Michael M. Hoffman, Wei Xie, Gail L. Rosen, Benjamin J. Lengerich, Johnny Israeli, Jack Lanchantin, Stephen Woloszynek, Anne E. Carpenter, Avanti Shrikumar, Jinbo Xu, Evan M. Cofer, Christopher A. Lavender, Srinivas C. Turaga, Amr M. Alexandari, Zhiyong Lu, David J. Harris, Dave DeCaprio, Yanjun Qi, Anshul Kundaje, Yifan Peng, Laura K. Wiley, Marwin H. S. Segler, Simina M. Boca, S. Joshua Swamidass, Austin Huang, Anthony Gitter, and Casey S. Greene. Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society, Interface*, 15(141):20170387, April 2018. doi:10.1098/rsif.2017.0387. URL https://pubmed.ncbi.nlm.nih.gov/29618526. Publisher: The Royal Society. → pages 30

[38] Sebastian G. Popescu, James H. Cole, Ben Glocker, and David J. Sharp. Deep learning methods for estimating brain age from structural MRI scans. Amesterdam, 2018. → pages 30, 101, 104, 105, 164, 190, 193

[39] Xinyang Feng, Zachary C. Lipton, Jie Yang, Scott A. Small, and Frank A. Provenzano. Estimating brain age based on a uniform healthy population with deep learning and structural MRI. *Neurobiology of Aging*, March 2020. doi:10.1016/j.neurobiolaging.2020.02.009. URL http://www.sciencedirect.com/science/article/pii/S0197458020300361. → pages 30, 82, 98, 101, 131, 191, 219

[40] T. Huang, H. Chen, R. Fujimoto, K. Ito, K. Wu, K. Sato, Y. Taki, H. Fukuda, and T. Aoki. Age estimation from brain MRI images using deep learning. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 849–852, April 2017. ISBN 1945-8452. Journal Abbreviation: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). → pages 30, 101, 102, 105

[41] Quin Xie, Kevin Faust, Randy Van Ommeren, Adeel Sheikh, Ugljesa Djuric, and Phedias Diamandis. Deep learning for image analysis: Personalizing medicine closer to the point of care. *Critical Reviews in Clinical Laboratory Sciences*, 56(1):61–73, January 2019. doi:10.1080/10408363.2018.1536111. URL https://doi.org/10.1080/10408363.2018.1536111. Publisher: Taylor & Francis. → pages 30

[42] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps.

eprint arXiv:1312.6034, December 2013. URL https://ui.adsabs.harvard.edu/abs/2013arXiv1312.6034S. → pages 30, 33, 63, 162, 165

[43] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *eprint arXiv:1610.02391*, October 2016. URL https://ui.adsabs.harvard.edu/abs/2016arXiv161002391S. → pages 30, 63, 66, 68, 162, 164, 165, 166

[44] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. *arXiv e-prints*, May 2016. URL https://ui.adsabs.harvard.edu/abs/2016arXiv160501713S. → pages 30, 62, 63, 70

[45] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise. *arXiv e-prints*, page arXiv:1706.03825, June 2017. URL https://ui.adsabs.harvard.edu/abs/2017arXiv170603825S. → pages 30, 33, 63, 64, 162, 165, 187

[46] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for Simplicity: The All Convolutional Net. *eprint arXiv:1412.6806*, December 2014. URL https://ui.adsabs.harvard.edu/abs/2014arXiv1412.6806S. → pages 30, 33, 63, 67, 68, 162, 163, 165

[47] Matthew D Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. *eprint arXiv:1311.2901*, November 2013. URL https://ui.adsabs.harvard.edu/abs/2013arXiv1311.2901Z. → pages 30, 63, 67, 131, 163, 164, 166, 192, 222, 226

[48] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436, June 2015. ISBN 1063-6919. Journal Abbreviation: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). → pages 30

[49] S-M Moosavi-Dezfooli, A Fawzi, and P Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, 2016. → pages 30

[50] J. Su, D. V. Vargas, and K. Sakurai. One Pixel Attack for Fooling Deep Neural Networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, October 2019. doi:10.1109/TEVC.2019.2890858. → pages 30

[51] P.-L. Kuo, J. A. Schrack, M. D. Shardell, M. Levine, A. Z. Moore, Y. An, P. Elango, A. Karikkineth, T. Tanaka, R. de Cabo, L. M. Zukley, M. AlGhatrif, C. W. Chia, E. M. Simonsick, J. M. Egan, S. M. Resnick, and L. Ferrucci. A roadmap to build a phenotypic metric of ageing: insights from the Baltimore Longitudinal Study of Aging. *Journal of Internal Medicine*, 287(4):373–394, April 2020. doi:10.1111/joim.13024. URL https://doi.org/10.1111/joim.13024. Publisher: John Wiley & Sons, Ltd. → pages 32, 81, 82, 134, 135, 157, 158, 160

[52] Paul Herent, Simon Jegou, Gilles Wainrib, and Thomas Clozel. Brain age prediction of healthy subjects on anatomic MRI with deep learning : going beyond with an "explainable AI" mindset. *bioRxiv*, January 2018. doi:10.1101/413302. URL http://biorxiv.org/content/early/2018/09/10/413302.abstract. → pages 32, 164, 192

[53] Ryuichiro Hataya. moskomule/senet.pytorch, April 2020. URL https://github.com/moskomule/senet.pytorch. original-date: 2017-09-12T12:26:27Z. → pages 33

[54] Antoine Dejonckheere. antoinedejonckheere/promethee, April 2019. URL https://github.com/antoinedejonckheere/promethee. original-date: 2016-05-23T14:02:13Z. → pages 33

[55] Vladimir Lukiyanov. vlukiyanov/pt-dec, April 2020. URL https://github.com/vlukiyanov/pt-dec. original-date: 2018-07-02T22:53:51Z. → pages 33

[56] GitHub - utkuozbulak/pytorch-cnn-visualizations: Pytorch implementation of convolutional neural network visualization techniques, . URL https://github.com/utkuozbulak/pytorch-cnn-visualizations. → pages 33

[57] Woo-Jeoung Nam. wjNam/Relative_attributing_propagation, March 2020. URL https://github.com/wjNam/Relative_Attributing_Propagation. original-date: 2019-11-11T05:11:52Z. → pages 33

[58] *Cambridge international dictionary of English*. Cambridge University Press, Cambridge, 1995. → pages 34

[59] Robert Hirsch. *Exploring Color Photography: From Film to Pixels*. New York, 6 edition, 2014. → pages 35

[60] Stephen Johnson. *Stephen Johnson on Digital Photography*. O'Reilly Media, Inc., 2006. ISBN 0-596-52370-X. → pages 35

[61] R. Mezrich. A perspective on K-space. *Radiology*, 195(2):297–315, May 1995. doi: 10.1148/radiology.195.2.7724743. Place: United States. → pages 35

[62] Seeing is believing: introduction to image contrast. In Donald W. McRobbie, Elizabeth A. Moore, Martin J. Graves, and Martin R. Prince, editors, *MRI from Picture to Proton*, pages 30–46. Cambridge University Press, Cambridge, 2 edition, 2006. ISBN 978-0-511-54540-5. doi:10.1017/CBO9780511545405.003. URL https://www.cambridge.org/core/books/mri-from-picture-to-proton/seeing-is-believing-introduction-to-image-contrast/91914073E6CDB68A8A8C4B2DDA28F95C. → pages 36

[63] R. Bakshi, S. Ariyaratana, R. H. Benedict, and L. Jacobs. Fluid-attenuated inversion recovery magnetic resonance imaging detects cortical and juxtacortical multiple sclerosis lesions. *Archives of neurology*, 58(5):742–748, May 2001. doi: 10.1001/archneur.58.5.742. Place: United States. → pages 36

[64] E. M. Haacke, S. Mittal, Z. Wu, J. Neelavalli, and Y.-C. N. Cheng. Susceptibility-weighted imaging: technical aspects and clinical applications, part 1. *AJNR. American journal of neuroradiology*, 30(1):19–30, January 2009. doi:10.3174/ajnr.A1400. → pages 36

[65] V. J. Wedeen, R. P. Wang, J. D. Schmahmann, T. Benner, W. Y. I. Tseng, G. Dai, D. N. Pandya, P. Hagmann, H. D'Arceuil, and A. J. de Crespigny. Diffusion spectrum magnetic resonance imaging (DSI) tractography of crossing fibers. *NeuroImage*, 41(4):1267–1277, July 2008. doi:10.1016/j.neuroimage.2008.03.036. Place: United States. → pages 36

[66] Peter J. Kostelec and Senthil Periaswamy. Image Registration for MRI. 2002. → pages 36

[67] John Ashburner and Karl J. Friston. CHAPTER 4 – Rigid Body Registration. 2007. → pages 38

[68] John C. Mazziotta, Arthur W. Toga, Alan Evans, Peter Fox, and Jack Lancaster. A Probabilistic Atlas of the Human Brain: Theory and Rationale for Its Development: The International Consortium for Brain Mapping (ICBM). *NeuroImage*, 2(2, Part A):89–101, June 1995. doi:10.1006/nimg.1995.1012. URL http://www.sciencedirect.com/science/article/pii/S1053811985710129. → pages 38

[69] Pravat K Mandal, Rashima Mahajan, and Ivo D Dinov. Structural brain atlases: design, rationale, and applications in normal and pathological cohorts. *Journal of Alzheimer's disease : JAD*, 31 Suppl 3(0 3):S169–S188, 2012. doi:10.3233/JAD-2012-120412. URL https://pubmed.ncbi.nlm.nih.gov/22647262. → pages 39

[70] Luis Bote-Curiel, Sergio Muñoz-Romero, Alicia Gerrero-Curieses, and L. José Rojo-Álvarez. Deep Learning and Big Data in Healthcare: A Double Review for Critical Beginners. *Applied Sciences*, 9(11), 2019. doi:10.3390/app9112331. → pages 39, 40

[71] Simon Haykin. CHAPTER 4 - Neural Networks: A Guided Tour. In Naresh K. Sinha and Madan M. Gupta, editors, *Soft Computing and Intelligent Systems*, pages 71–80. Academic Press, San Diego, January 2000. doi:10.1016/B978-012646490-0/50007-X. URL http://www.sciencedirect.com/science/article/pii/B978012646490050007X. → pages 39

[72] Georgios Z. Papadakis, Apostolos H. Karantanas, Manolis Tsiknakis, Aristidis Tsatsakis, Demetrios A. Spandidos, and Kostas Marias. Deep learning opens new horizons in personalized medicine. *Biomedical reports*, 10(4):215–217, April 2019. doi:10.3892/br.2019.1199. URL https://www.ncbi.nlm.nih.gov/pubmed/30988951. → pages 39

[73] Guangming Zhu, Bin Jiang, Liz Tong, Yuan Xie, Greg Zaharchuk, and Max Wintermark. Applications of Deep Learning to Neuro-Imaging Techniques. *Frontiers in Neurology*, 10:869, 2019. doi:10.3389/fneur.2019.00869. URL https://www.frontiersin.org/article/10.3389/fneur.2019.00869. → pages 39, 40

[74] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2999134.2999257. → pages 39, 43, 86, 162, 167

[75] Luca Saba, Mainak Biswas, Venkatanareshbabu Kuppili, Elisa Cuadrado Godia, Harman S. Suri, Damodar Reddy Edla, Tomaž Omerzu, John R. Laird, Narendra N. Khanna, Sophie Mavrogeni, Athanasios Protogerou, Petros P. Sfikakis, Vijay Viswanathan, George D. Kitas, Andrew Nicolaides, Ajay Gupta, and Jasjit S. Suri. The present and future of deep learning in radiology. *European Journal of Radiology*, 114:14–24, May 2019. doi:10.1016/j.ejrad.2019.02.038. URL http://www.sciencedirect.com/science/article/pii/S0720048X19300919. → pages 39

[76] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016. URL http://arxiv.org/abs/1605.02688. → pages 40

[77] Francois Chollet and others. Keras, 2015. URL https://github.com/fchollet/keras. Publisher: GitHub. → pages 40

[78] Luca Boldrini, Jean-Emmanuel Bibault, Carlotta Masciocchi, Yanting Shen, and Martin-Immanuel Bittner. Deep Learning: A Review for the Radiation Oncologist. *Frontiers in Oncology*, 9:977, 2019. doi:10.3389/fonc.2019.00977. URL https://www.frontiersin.org/article/10.3389/fonc.2019.00977. → pages 40

[79] Cian M. Scannell, Mitko Veta, Adriana D.M. Villa, Eva C. Sammut, Jack Lee, Marcel Breeuwer, and Amedeo Chiribiri. Deep-Learning-Based Preprocessing for Quantitative Myocardial Perfusion MRI. *Journal of Magnetic Resonance Imaging*, 51(6):1689–1696, June 2020. doi:10.1002/jmri.26983. URL https://doi.org/10.1002/jmri.26983. Publisher: John Wiley & Sons, Ltd. → pages 40

[80] Muhammad Irfan Sharif, Jian Ping Li, Javeria Naz, and Iqra Rashid. A comprehensive review on multi-organs tumor detection based on machine learning. *Pattern Recognition Letters*, 131:30–37, March 2020. doi:10.1016/j.patrec.2019.12.006. URL http://www.sciencedirect.com/science/article/pii/S0167865519303691. → pages 40

[81] Kirti Raj Bhatele and Sarita Singh Bhadauria. Brain structural disorders detection and classification approaches: a review. *Artificial Intelligence Review*, 53(5):3349–3401, June 2020. doi:10.1007/s10462-019-09766-9. URL https://doi.org/10.1007/s10462-019-09766-9. → pages 40

[82] G. S. Cover, W. G. Herrera, M. P. Bento, S. Appenzeller, and L. Rittner. Computational methods for corpus callosum segmentation on MRI: A systematic literature review. *Computer methods and programs in biomedicine*, 154:25–35, February 2018. doi:10.1016/j.cmpb.2017.10.025. → pages 40

[83] J. Dolz, L. Massoptier, and M. Vermandel. Segmentation algorithms of subcortical brain structures on MRI for radiotherapy and radiosurgery: A survey. *IRBM*, 36(4):200–212, August 2015. doi:10.1016/j.irbm.2015.06.001. URL http://www.sciencedirect.com/science/article/pii/S1959031815000688. → pages 40

[84] Frank F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65 6:386–408, 1958. → pages 41

[85] Koichiro Yasaka, Hiroyuki Akai, Akira Kunimatsu, Shigeru Kiryu, and Osamu Abe. Deep learning with convolutional neural network in radiology. *Japanese Journal of Radiology*, 36(4):257–272, April 2018. doi:10.1007/s11604-018-0726-3. URL https://doi.org/10.1007/s11604-018-0726-3. → pages 43, 46

[86] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pages 249–256, 2010. → pages 43

[87] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv e-prints*, February 2015. → pages 44, 108

[88] Jose Bernal, Kaisar Kushibar, Daniel S. Asfaw, Sergi Valverde, Arnau Oliver, Robert Martí, and Xavier Lladó. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *arXiv e-prints*, December 2017. → pages 44

[89] CS231n Convolutional Neural Networks for Visual Recognition, . URL http://cs231n.github.io/neural-networks-3/. → pages 45

[90] Michael A. Nielsen. Neural Networks and Deep Learning. 2015. URL http://neuralnetworksanddeeplearning.com. → pages 45

[91] Kunihiko Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1(2):119–130, January 1988. doi:

10.1016/0893-6080(88)90014-7. URL http://www.sciencedirect.com/science/article/pii/0893608088900147. → pages 45

[92] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. doi:10.1109/5.726791. → pages 45

[93] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A Fast Learning Algorithm for Deep Belief Nets. *Neural Comput.*, 18(7):1527–1554, July 2006. doi:10.1162/neco.2006.18.7.1527. URL http://dx.doi.org/10.1162/neco.2006.18.7.1527. → pages 46

[94] G. E. Hinton and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504, July 2006. doi:10.1126/science.1127647. URL http://science.sciencemag.org/content/313/5786/504.abstract. → pages 46

[95] Aditya Khamparia and Karan Mehtab Singh. A systematic review on deep learning architectures and applications. *Expert Systems*, 36(3):e12400, June 2019. doi:10.1111/exsy.12400. URL https://doi.org/10.1111/exsy.12400. → pages 46

[96] Bo Zhu, Jeremiah Z. Liu, Stephen F. Cauley, Bruce R. Rosen, and Matthew S. Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555:487, March 2018. URL https://doi.org/10.1038/nature25988. → pages 46

[97] S. Wang, Z. Su, L. Ying, X. Peng, S. Zhu, F. Liang, D. Feng, and D. Liang. Accelerating magnetic resonance imaging via deep learning. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 514–517, April 2016. doi:10.1109/ISBI.2016.7493320. → pages 46

[98] Raghav Mehta, Aabhas Majumdar, and Jayanthi Sivaswamy. BrainSegNet: a convolutional neural network architecture for automated segmentation of human brain structures. *Journal of Medical Imaging*, 4(2):1–11, April 2017. URL https://doi.org/10.1117/1.JMI.4.2.024003. → pages 46

[99] Christian Wachinger, Martin Reuter, and Tassilo Klein. DeepNAT: Deep convolutional neural network for segmenting neuroanatomy. *Segmenting the Brain*, 170:434–445, April 2018. doi:10.1016/j.neuroimage.2017.02.035. URL http://www.sciencedirect.com/science/article/pii/S1053811917301465. → pages 46

[100] Hao Chen, Qi Dou, Lequan Yu, Jing Qin, and Pheng-Ann Heng. VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. *Segmenting the Brain*, 170:446–455, April 2018. doi:10.1016/j.neuroimage.2017.04.041. URL http://www.sciencedirect.com/science/article/pii/S1053811917303348. → pages 46

[101] Xiaohong W. Gao, Rui Hui, and Zengmin Tian. Classification of CT brain images based on deep learning networks. *Computer Methods and Programs in Biomedicine*, 138:49–56, January 2017. doi:10.1016/j.cmpb.2016.10.007. URL http://www.sciencedirect.com/science/article/pii/S0169260716305296. → pages 46

[102] Tong Duc Phong, Hieu N. Duong, Hien T. Nguyen, Nguyen Thanh Trong, Vu H. Nguyen, Tran Van Hoa, and Vaclav Snasel. Brain Hemorrhage Diagnosis by Using Deep Learning. In *Proceedings of the 2017 International Conference on Machine Learning and Soft Computing*, ICMLSC '17, pages 34–39, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4828-7. doi:10.1145/3036290.3036326. URL http://doi.acm.org/10.1145/3036290.3036326. → pages 46

[103] Sergey M. Plis, Devon R. Hjelm, Ruslan Salakhutdinov, Elena A. Allen, Henry J. Bockholt, Jeffrey D. Long, Hans J. Johnson, Jane S. Paulsen, Jessica A. Turner, and Vince D. Calhoun. Deep learning for neuroimaging: a validation study. *Frontiers in Neuroscience*, 8:229, 2014. doi:10.3389/fnins.2014.00229. URL https://www.frontiersin.org/article/10.3389/fnins.2014.00229. → pages 46

[104] Z. Md Alom, M. Tarek Taha, Chris Yakopcic, Stefan Westberg, Paheding Sidike, S. Mst Nasrin, Mahmudul Hasan, C. Brian Van Essen, A. S. Abdul Awwal, and K. Vijayan Asari. A State-of-the-Art Survey on Deep Learning Theory and Architectures. *Electronics*, 8(3), 2019. doi:10.3390/electronics8030292. → pages 49

[105] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *eprint arXiv:1502.03167*, February 2015. URL https://ui.adsabs.harvard.edu/abs/2015arXiv150203167I. → pages 49

[106] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0-387-31073-8. → pages 51

[107] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. → pages 51

[108] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. ISBN 1063-6919. doi:10.1109/CVPR.2016.90. Journal Abbreviation: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). → pages 51, 129

[109] Kaiming He and Jian Sun. Convolutional Neural Networks at Constrained Time Cost. 2015. → pages 51

[110] R. Herbrich, T. Graepel, and K. Obermayer. Support vector learning for ordinal regression. In *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*, volume 1, pages 97–102 vol.1, September 1999. doi:10.1049/cp:19991091. → pages 56

[111] Amnon Shashua and Anat Levin. Ranking with Large Margin Principle: Two Approaches. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 961–968. MIT Press, 2003. URL http://papers.nips.cc/paper/2269-ranking-with-large-margin-principle-two-approaches.pdf. → pages 56

[112] Koby Crammer and Yoram Singer. Pranking with Ranking. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 641–647. MIT Press, 2002. URL http://papers.nips.cc/paper/2023-pranking-with-ranking.pdf. → pages 56

[113] Eibe Frank and Mark Hall. A Simple Approach to Ordinal Classification. In Luc De Raedt and Peter Flach, editors, *Machine Learning: ECML 2001*, pages 145–156. Springer Berlin Heidelberg, 2001. ISBN 978-3-540-44795-5. → pages 56

[114] Ling Li and Hsuan-tien Lin. Ordinal Regression by Extended Binary Classification. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 865–872. MIT Press, 2007. URL http://papers.nips.cc/paper/3125-ordinal-regression-by-extended-binary-classification.pdf. → pages 56

[115] G. Dong, G. Liao, H. Liu, and G. Kuang. A Review of the Autoencoder and Its Variants: A Comparative Perspective from Target Recognition in Synthetic-Aperture Radar Images. *IEEE Geoscience and Remote Sensing Magazine*, 6(3):44–68, September 2018. doi:10.1109/MGRS.2018.2853555. → pages 60

[116] James B. MacQueen. Some methods for classification and analysis of multivariate observations. 1967. → pages 61

[117] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. → pages 61

[118] Sohil Atul Shah and Vladlen Koltun. Deep Continuous Clustering. *eprint arXiv:1803.01449*, March 2018. URL https://ui.adsabs.harvard.edu/abs/2018arXiv180301449A. → pages 61

[119] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep Clustering for Unsupervised Learning of Visual Features. *eprint arXiv:1807.05520*, July 2018. URL https://ui.adsabs.harvard.edu/abs/2018arXiv180705520C. → pages 61

[120] Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. Improved Deep Embedded Clustering with Local Structure Preservation. In *IJCAI*, 2017. → pages 61

[121] Vaibhav Krishna, Tian Guo, and Nino Antulov-Fantulin. Is Simple Better? Revisiting Non-linear Matrix Factorization for Learning Incomplete Ratings. *eprint arXiv:1710.05613*, October 2017. URL https://ui.adsabs.harvard.edu/abs/2017arXiv171005613K. → pages 61

[122] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised Deep Embedding for Clustering Analysis. *eprint arXiv:1511.06335*, November 2015. URL https://ui.adsabs.harvard.edu/abs/2015arXiv151106335X. → pages 61, 62, 148

[123] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *arXiv e-prints*, page arXiv:1511.05942, November 2015. URL https://ui.adsabs.harvard.edu/abs/2015arXiv151105942C. → pages 62

[124] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv e-prints*, February 2017. _eprint: 1702.08608. → pages 63

[125] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to Explain Individual Classification Decisions. *J. Mach. Learn. Res.*, 11:1803–1831, August 2010. Publisher: JMLR.org. → pages 63

[126] Jindong Gu, Yinchong Yang, and Volker Tresp. Understanding Individual Decisions of CNNs via Contrastive Backpropagation. In C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision – ACCV 2018*, pages 119–134, Cham, 2019. Springer International Publishing. ISBN 978-3-030-20893-6. → pages 63, 72, 162

[127] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 3319–3328. JMLR.org, 2017. → pages 63

[128] Luisa M. Zintgraf, Taco S. Cohen, and Max Welling. A New Method to Visualize Deep Neural Networks. *arXiv e-prints*, March 2016. _eprint: 1603.02518. → pages 63

[129] Jean-Pierre Brans and Bertrand Mareschal. Promethee Methods. In JosÉ Figueira, Salvatore Greco, and Matthias Ehrogott, editors, *Multiple Criteria Decision Analysis: State of the Art Surveys*, pages 163–186. Springer New York, New York, NY, 2005. ISBN 978-0-387-23081-8. doi:10.1007/0-387-23081-5_5. URL https://doi.org/10.1007/0-387-23081-5_5. → pages 75, 76

[130] R. Peters. Ageing and the brain. *Postgraduate medical journal*, 82(964):84–88, February 2006. doi:10.1136/pgmj.2005.036665. URL https://pubmed.ncbi.nlm.nih.gov/16461469. Publisher: BMJ Group. → pages 77

[131] Lars Nyberg, Carl-Johan Boraxbekk, Daniel Eriksson Sörman, Patrik Hansson, Agneta Herlitz, Karolina Kauppi, Jessica K. Ljungberg, Hugo Lövheim, Anders Lundquist, Annelie Nordin Adolfsson, Anna Oudin, Sara Pudas, Michael Rönnlund, Mikael Stiernstedt, Anna Sundström, and Rolf Adolfsson. Biological and environmental predictors of heterogeneity in neurocognitive ageing: Evidence from Betula and other longitudinal studies. *Ageing Research Reviews*, 64:101184, December 2020. doi:10.1016/j.arr.2020.101184. URL https://www.sciencedirect.com/science/article/pii/S1568163720303196. → pages 77

[132] G Sophia Borgeest, Richard N Henson, Meredith Shafto, David Samu, Cam-CAN, and Rogier A Kievit. Greater lifestyle engagement is associated with better age-adjusted cognitive abilities. *PloS one*, 15(5):e0230077–e0230077, May 2020. doi:10.1371/journal.pone.0230077. URL https://pubmed.ncbi.nlm.nih.gov/32437448. Publisher: Public Library of Science. → pages 77, 78

[133] Ian J. Deary, Simon R. Cox, and W. David Hill. Genetic variation, brain, and intelligence differences. *Molecular Psychiatry*, February 2021. doi:10.1038/s41380-021-01027-y. URL https://doi.org/10.1038/s41380-021-01027-y. → pages 78

[134] Adejoke Yetunde Onaolapo, Adebimpe Yemisi Obelawo, and Olakunle James Onaolapo. Brain Ageing, Cognition and Diet: A Review of the Emerging Roles of Food-Based Nootropics in Mitigating Age-related Memory Decline. *Current aging science*, 12(1):2–14, 2019. doi:10.2174/1874609812666190311160754. URL https://pubmed.ncbi.nlm.nih.gov/30864515. Publisher: Bentham Science Publishers. → pages 78

[135] Lars Nyberg, Carl-Johan Boraxbekk, Daniel Eriksson Sörman, Patrik Hansson, Agneta Herlitz, Karolina Kauppi, Jessica K. Ljungberg, Hugo Lövheim, Anders Lundquist, Annelie Nordin Adolfsson, Anna Oudin, Sara Pudas, Michael Rönnlund, Mikael Stiernstedt, Anna Sundström, and Rolf Adolfsson. Biological and environmental predictors of heterogeneity in neurocognitive ageing: Evidence from Betula and other longitudinal studies. *Ageing Research Reviews*, 64:101184, December 2020. doi:10.1016/j.arr.2020.101184. URL https://www.sciencedirect.com/science/article/pii/S1568163720303196. → pages 78

[136] Michael Rönnlund, Lars Nyberg, Lars Bäckman, and Lars-Göran Nilsson. Stability, growth, and decline in adult life span development of declarative memory: cross-sectional and longitudinal data from a population-based study. *Psychology and aging*, 20(1):3–18, March 2005. doi:10.1037/0882-7974.20.1.3. Place: United States. → pages 78

[137] Tamara A. Rahhal, Cynthia P. May, and Lynn Hasher. Truth and character: sources that older adults can remember. *Psychological science*, 13(2):101–105, March 2002. doi:10.1111/1467-9280.00419. URL https://pubmed.ncbi.nlm.nih.gov/11933991. → pages 78

[138] H. E. Garrett. A developmental theory of intelligence. *American Psychologist*, 1(9): 372–378, 1946. doi:10.1037/h0056380. Place: US Publisher: American Psychological Association. → pages 78

[139] B. Balinsky. An analysis of the mental factors of various age groups from nine to sixty. *Genetic Psychology Monographs*, 23:191–234, 1941. Place: US Publisher: Heldref Publications. → pages 78

[140] Elliot M. Tucker-Drob and Timothy A. Salthouse. Adult age trends in the relations among cognitive abilities. *Psychology and aging*, 23(2):453–460, June 2008. doi:10.1037/0882-7974.23.2.453. URL https://pubmed.ncbi.nlm.nih.gov/18573019. → pages 78

[141] Zimu Wu, Aung Zaw Zaw Phyo, Tagrid Al-Harbi, Robyn L. Woods, and Joanne Ryan. Distinct Cognitive Trajectories in Late Life and Associated Predictors and Outcomes: A Systematic Review. *Journal of Alzheimer's disease reports*, 4(1):459–478, October 2020. doi:10.3233/ADR-200232. URL https://pubmed.ncbi.nlm.nih.gov/33283167. Publisher: IOS Press. → pages 78

[142] Cheryl Grady. The cognitive neuroscience of ageing. *Nature Reviews Neuroscience*, 13 (7):491–505, July 2012. doi:10.1038/nrn3256. URL https://doi.org/10.1038/nrn3256. → pages 79

[143] Malgorzata Urbanska, Magdalena Blazejczyk, and Jacek Jaworski. Molecular basis of dendritic arborization. *Acta neurobiologiae experimentalis*, 68(2):264–288, 2008. Place: Poland. → pages 79

[144] K. Henze and W. Martin. Evolutionary biology: Essence of mitochondria. *Nature*, 426:127–128, 2003. → pages 79

[145] Douglas R. Green. *Means to an End: Apoptosis and Other Cell Death Mechanisms*. Cold Spring Harbor Laboratory Presss, 2011. → pages 79

[146] Ai-Ling Lin, Daniel Coman, Lihong Jiang, Douglas L. Rothman, and Fahmeed Hyder. Caloric restriction impedes age-related decline of mitochondrial function and neuronal activity. *Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism*, 34(9):1440–1443, September 2014. doi:10.1038/jcbfm.2014.114. → pages 79

[147] Debolina Ghosh, Kelsey R. LeVault, Aaron J. Barnett, and Gregory J. Brewer. A reversible early oxidized redox state that precedes macromolecular ROS damage in aging nontransgenic and 3xTg-AD mouse neurons. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 32(17):5821–5832, April 2012. doi:10.1523/JNEUROSCI.6192-11.2012. → pages 79

[148] Steven H. Graham and Hao Liu. Life and death in the trash heap: The ubiquitin proteasome pathway and UCHL1 in brain aging, neurodegenerative disease and cerebral Ischemia. *Ageing research reviews*, 34:30–38, March 2017. doi:10.1016/j.arr.2016.09.011. → pages 79

[149] Jesse S. Kerr, Bryan A. Adriaanse, Nigel H. Greig, Mark P. Mattson, M. Zameel Cader, Vilhelm A. Bohr, and Evandro F. Fang. Mitophagy and Alzheimer's Disease: Cellular and Molecular Mechanisms. *Trends in neurosciences*, 40(3):151–166, March 2017. doi:10.1016/j.tins.2017.01.002. → pages 79

[150] Ralph A. Nixon. The role of autophagy in neurodegenerative disease. *Nature Medicine*, 19(8):983–997, August 2013. doi:10.1038/nm.3232. URL https://doi.org/10.1038/nm.3232. → pages 79

[151] Mark P. Mattson and Thiruma V. Arumugam. Hallmarks of Brain Aging: Adaptive and Pathological Modification by Metabolic States. *Cell Metabolism*, 27(6):1176–1199, June 2018. doi:10.1016/j.cmet.2018.05.011. URL https://www.sciencedirect.com/science/article/pii/S1550413118303188. → pages 79, 80

[152] Manu S. Goyal, Andrei G. Vlassenko, Tyler M. Blazey, Yi Su, Lars E. Couture, Tony J. Durbin, Randall J. Bateman, Tammie L-S. Benzinger, John C. Morris, and Marcus E. Raichle. Loss of Brain Aerobic Glycolysis in Normal Human Aging. *Cell metabolism*, 26(2):353–360.e3, August 2017. doi:10.1016/j.cmet.2017.07.010. URL https://pubmed.ncbi.nlm.nih.gov/28768174. → pages 79

[153] Manoj Kumar Jaiswal, Wolf-Dieter Zech, Miriam Goos, Christine Leutbecher, Alberto Ferri, Annette Zippelius, Maria Teresa Carrì, Roland Nau, and Bernhard U Keller. Impairment of mitochondrial calcium handling in a mtSOD1 cell culture model of motoneuron disease. *BMC neuroscience*, 10:64–64, June 2009. doi:10.1186/1471-2202-10-64. URL https://pubmed.ncbi.nlm.nih.gov/19545440. Publisher: BioMed Central. → pages 79

[154] H. Lodish and S. L. Zipursky. *Molecular Cell Biology.* W. H. Freeman, New York, 4 edition, 2000. → pages 79

[155] Harris Eyre, Bernhard Baune, and Helen Lavretsky. Clinical Advances in Geriatric Psychiatry: A Focus on Prevention of Mood and Cognitive Disorders. *The Psychiatric clinics of North America*, 38(3):495–514, September 2015. doi:10.1016/j.psc.2015.05.002. URL https://pubmed.ncbi.nlm.nih.gov/26300035. Edition: 2015/06/26. → pages 79

[156] Julian N. Trollor and Michael J. Valenzuela. Brain ageing in the new millennium. *Australian and New Zealand Journal of Psychiatry*, 35(6):788–805, December 2001. doi:10.1046/j.1440-1614.2001.00969.x. URL https://doi.org/10.1046/j.1440-1614.2001.00969.x. Publisher: John Wiley & Sons, Ltd. → pages 80, 81, 83

[157] Kalpita Banerjee Ishita Guha Thakurta Maitrayee Sinha Maria Bindu Bagh Sasanka Chakrabarti, Soumyabrata Munshi. Mitochondrial Dysfunction during Brain Aging: Role of Oxidative Stress and Modulation by Antioxidant Supplementation. *Aging and disease*, 2(3):242, 2011. URL http://www.aginganddisease.org/EN/abstract/article_147281.shtml. Publisher: Aging and disease. → pages 80

[158] Philip W. Landfield, Eric M. Blalock, Kuey-Chu Chen, and Nada M. Porter. A new glucocorticoid hypothesis of brain aging: implications for Alzheimer's disease. *Current Alzheimer research*, 4(2):205–212, April 2007. doi:10.2174/156720507780362083. URL https://pubmed.ncbi.nlm.nih.gov/17430248. → pages 80

[159] Adith Mohan, Karen A. Mather, Anbupalam Thalamuthu, Bernhard T. Baune, and Perminder S. Sachdev. Gene expression in the aging human brain: an overview. *Current opinion in psychiatry*, 29(2):159–167, March 2016. doi:10.1097/YCO.0000000000000238. Place: United States. → pages 80

[160] Elaine Nicpon Marieb. *Essentials of human anatomy and physiology.* Benjamin Cummings, San Francisco, 6 edition, 2000. → pages 80

[161] Luísa T. Ferreira, Ana C. Figueiredo, Bernardo Orr, Danilo Lopes, and Helder Maiato. Chapter 3 - Dissecting the role of the tubulin code in mitosis. In Helder Maiato and Melina Schuh, editors, *Methods in Cell Biology*, volume 144, pages 33–74. Aca-

demic Press, January 2018. ISBN 0091-679X. doi:10.1016/bs.mcb.2018.03.040. URL https://www.sciencedirect.com/science/article/pii/S0091679X18300402. → pages 80

[162] Patrícia Ramos, Agostinho Santos, Nair Rosas Pinto, Ricardo Mendes, Teresa Magalhães, and Agostinho Almeida. Iron levels in the human brain: a post-mortem study of anatomical region differences and age-related changes. *Journal of trace elements in medicine and biology : organ of the Society for Minerals and Trace Elements (GMS)*, 28(1):13–17, January 2014. doi:10.1016/j.jtemb.2013.08.001. Place: Germany. → pages 81

[163] P. Cras, M. Kawai, D. Lowery, P. Gonzalez-DeWhitt, B. Greenberg, and G. Perry. Senile plaque neurites in Alzheimer disease accumulate amyloid precursor protein. *Proceedings of the National Academy of Sciences of the United States of America*, 88 (17):7552–7556, September 1991. doi:10.1073/pnas.88.17.7552. URL https://pubmed.ncbi.nlm.nih.gov/1652752. → pages 81

[164] Thomas L. Kemper. Neuroanatomical and neuropathological changes during aging and dementia. In *Clinical neurology of aging, 2nd ed.*, pages 3–67. Oxford University Press, New York, NY, US, 1994. ISBN 0-19-507167-0 (Hardcover). → pages 81

[165] P. Delaère, Y. He, G. Fayet, C. Duyckaerts, and J. J. Hauw. Beta A4 deposits are constant in the brain of the oldest old: an immunocytochemical study of 20 French centenarians. *Neurobiology of aging*, 14(2):191–194, April 1993. doi:10.1016/0197-4580(93)90096-t. Place: United States. → pages 81

[166] D. W. Dickson, G. Singer, P. Davies, S. H. Yen, and H. A. Crystal. Regional immunocytochemical studies of brains of prospectively studied demented and monondemented elderly humans. *Neurobiology of Aging*, 13, 1992. → pages 81

[167] Ronald S. Petralia, Mark P. Mattson, and Pamela J. Yao. Communication breakdown: the impact of ageing on synapse structure. *Ageing research reviews*, 14:31–42, March 2014. doi:10.1016/j.arr.2014.01.003. URL https://pubmed.ncbi.nlm.nih.gov/24495392. Edition: 2014/02/02. → pages 81

[168] M. Jucker, K. Bättig, and W. Meier-Ruge. Effects of aging and vincamine derivatives on pericapillary microenvironment: stereological characterization of the cere-

bral capillary network. *Neurobiology of aging*, 11(1):39–46, February 1990. doi: 10.1016/0197-4580(90)90060-d. Place: United States. → pages 81

[169] H. C. Fang. Observations on aging characteristics of cerebral blood vessels macroscopic and microscopic features. 1976. → pages 81

[170] O. Piguet, K.L. Double, J.J. Kril, J. Harasty, V. Macdonald, D.A. McRitchie, and G.M. Halliday. White matter loss in healthy ageing: A postmortem analysis. *Neurobiology of Aging*, 30(8):1288–1295, August 2009. doi:10.1016/j.neurobiolaging.2007.10.015. URL http://www.sciencedirect.com/science/article/pii/S0197458007004174. → pages 81

[171] Elisabeth J. Vinke, Marius de Groot, Vikram Venkatraghavan, Stefan Klein, Wiro J. Niessen, M. Arfan Ikram, and Meike W. Vernooij. Trajectories of imaging markers in brain aging: the Rotterdam Study. *Neurobiology of Aging*, 71:32–40, November 2018. doi:10.1016/j.neurobiolaging.2018.07.001. URL http://www.sciencedirect.com/science/article/pii/S0197458018302513. → pages 81, 82, 134, 135, 136, 157, 158, 160, 190, 192

[172] Christopher M Henstridge, Rosemary J Jackson, JeeSoo M Kim, Abigail G Hermann, Ann K Wright, Sarah E Harris, Mark E Bastin, John M Starr, Joanna Wardlaw, Thomas H Gillingwater, Colin Smith, Chris-Anne McKenzie, Simon R Cox, Ian J Deary, and Tara L Spires-Jones. Post-mortem brain analyses of the Lothian Birth Cohort 1936: extending lifetime cognitive and brain phenotyping to the level of the synapse. *Acta neuropathologica communications*, 3:53–53, September 2015. doi: 10.1186/s40478-015-0232-0. URL https://pubmed.ncbi.nlm.nih.gov/26335101. Publisher: BioMed Central. → pages 81

[173] Catriona D. Good, Ingrid S. Johnsrude, John Ashburner, Richard N.A. Henson, Karl J. Friston, and Richard S.J. Frackowiak. A Voxel-Based Morphometric Study of Ageing in 465 Normal Adult Human Brains. *NeuroImage*, 14(1):21–36, July 2001. doi:10.1006/nimg.2001.0786. URL http://www.sciencedirect.com/science/article/pii/S1053811901907864. → pages 82, 134, 135, 158

[174] A. Pfefferbaum, D. H. Mathalon, E. V. Sullivan, J. M. Rawles, R. B. Zipursky, and K. O. Lim. A quantitative magnetic resonance imaging study of changes in brain morphology from infancy to late adulthood. *Archives of neurology*, 51(9):874–887,

September 1994. doi:10.1001/archneur.1994.00540210046012. Place: United States. → pages 82, 136

[175] Kelvin O. Lim, Robert B. Zipursky, Mark C. Watts, and Adolf Pfefferbaum. Decreased Gray Matter in Normal Aging: An in Vivo Magnetic Resonance Study. *Journal of Gerontology*, 47(1):B26–B30, January 1992. doi:10.1093/geronj/47.1.B26. URL https://doi.org/10.1093/geronj/47.1.B26. → pages 82, 136

[176] R. C. Gur, P. D. Mozley, S. M. Resnick, G. L. Gottlieb, M. Kohn, R. Zimmerman, G. Herman, S. Atlas, R. Grossman, D. Berretta, and et al. Gender differences in age effect on brain atrophy measured by magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America*, 88(7):2845–2849, April 1991. doi:10.1073/pnas.88.7.2845. URL https://pubmed.ncbi.nlm.nih.gov/2011592. → pages 82, 136

[177] N. Raz, F. M. Gunning-Dixon, D. Head, J. H. Dupuis, and J. D. Acker. Neuroanatomical correlates of cognitive aging: evidence from structural magnetic resonance imaging. *Neuropsychology*, 12(1):95–114, January 1998. doi:10.1037//0894-4105.12.1.95. Place: United States. → pages 82, 136, 155

[178] Rosemary Fama and Edith V. Sullivan. Thalamic structures and associated cognitive functions: Relations with age and aging. *Neuroscience and biobehavioral reviews*, 54:29–37, July 2015. doi:10.1016/j.neubiorev.2015.03.008. URL https://pubmed.ncbi.nlm.nih.gov/25862940. Edition: 2015/04/09. → pages 82, 134, 135

[179] David H. Salat, Randy L. Buckner, Abraham Z. Snyder, Douglas N. Greve, Rahul S.R. Desikan, Evelina Busa, John C. Morris, Anders M. Dale, and Bruce Fischl. Thinning of the Cerebral Cortex in Aging. *Cerebral Cortex*, 14(7):721–730, July 2004. doi:10.1093/cercor/bhh032. URL https://doi.org/10.1093/cercor/bhh032. → pages 82, 134, 135

[180] Dongtao Wei, Kangcheng Wang, Jie Meng, Kaixiang Zhuang, Qunlin Chen, Wenjing Yan, Peng Xie, and Jiang Qiu. The reductions in the subcallosal region cortical volume and surface area in major depressive disorder across the adult life span. *Psychological medicine*, 50(3):422–430, February 2020. doi:10.1017/S0033291719000230. Place: England. → pages 82, 134, 135

[181] Ryuichi Fujimoto, Chihiro Kondo, Koichi Ito, Kai Wu, Kazunori Sato, Yasuyuki Taki, Hiroshi Fukuda, and Takafumi Aoki. Age estimation using effective brain local features from T1-weighted images. *Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2016:5941–5944, August 2016. doi:10.1109/EMBC.2016.7592081. Place: United States. → pages 82, 103, 164, 190

[182] Chihiro Kondo, Koichi Ito, Kai Wu, Kazunori Sato, Yasuyuki Taki, Hiroshi Fukuda, and Takafumi Aoki. An age estimation method using brain local features for T1-weighted images. *Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2015:666–669, August 2015. doi:10.1109/EMBC.2015.7318450. Place: United States. → pages 82, 164, 190

[183] Camilo Bermudez, Andrew J. Plassard, Shikha Chaganti, Yuankai Huo, Katherine S. Aboud, Laurie E. Cutting, Susan M. Resnick, and Bennett A. Landman. Anatomical context improves deep learning on the brain age estimation task. *Magnetic Resonance Imaging*, 62:70–77, October 2019. doi:10.1016/j.mri.2019.06.018. URL http://www.sciencedirect.com/science/article/pii/S0730725X1830609X. → pages 82, 101, 104, 105, 129, 130, 164, 165, 191

[184] Daniel G. Amen, Sachit Egan, Somayeh Meysami, Cyrus A. Raji, and Noble George. Patterns of Regional Cerebral Blood Flow as a Function of Age Throughout the Lifespan. *Journal of Alzheimer's disease : JAD*, 65(4):1087–1092, 2018. doi:10.3233/JAD-180598. Place: Netherlands. → pages 82

[185] Stijn Bonte, Pieter Vandemaele, Stijn Verleden, Kurt Audenaert, Karel Deblaere, Ingeborg Goethals, and Roel Van Holen. Healthy brain ageing assessed with 18F-FDG PET and age-dependent recovery factors after partial volume effect correction. *European Journal of Nuclear Medicine and Molecular Imaging*, 44(5):838–849, May 2017. doi:10.1007/s00259-016-3569-0. URL https://doi.org/10.1007/s00259-016-3569-0. → pages 83

[186] Fatemeh Seifar, Farid Hajibonabi, Neda Parnianfard, Mahnaz Ranjkesh, Ebrahim Valikhani, and Mohammad Yazdchi. Age-Related Glucose Metabolism in FDG-PET Brain Studies. 1(1), 2018. → pages 83

[187] O. Almkvist. Functional brain imaging as a looking-glass into the degraded brain: reviewing evidence from Alzheimer disease in relation to normal aging. *Acta psychologica*, 105(2-3):255–277, December 2000. doi:10.1016/s0001-6918(00)00064-0. Place: Netherlands. → pages 83

[188] Mark D'Esposito, Leon Y. Deouell, and Adam Gazzaley. Alterations in the BOLD fMRI signal with ageing and disease: a challenge for neuroimaging. *Nature Reviews Neuroscience*, 4(11):863–872, November 2003. doi:10.1038/nrn1246. URL https://doi.org/10.1038/nrn1246. → pages 83

[189] Michael Cliff, Dan W. Joyce, Melissa Lamar, Thomas Dannhauser, Derek K. Tracy, and Sukhwinder S. Shergill. Aging effects on functional auditory and visual processing using fMRI with variable sensory loading. *Cortex*, 49(5):1304–1313, May 2013. doi:10.1016/j.cortex.2012.04.003. URL https://www.sciencedirect.com/science/article/pii/S0010945212001098. → pages 83

[190] Randy L Buckner. The brain's default network: origins and implications for the study of psychosis. *Dialogues in clinical neuroscience*, 15(3):351–358, September 2013. doi:10.31887/DCNS.2013.15.3/rbuckner. URL https://pubmed.ncbi.nlm.nih.gov/24174906. Publisher: Les Laboratoires Servier. → pages 83

[191] Hui-Jie Li, Xiao-Hui Hou, Han-Hui Liu, Chun-Lin Yue, Guang-Ming Lu, and Xi-Nian Zuo. Putting age-related task activation into large-scale brain networks: A meta-analysis of 114 fMRI studies on healthy aging. *Neuroscience & Biobehavioral Reviews*, 57:156–174, October 2015. doi:10.1016/j.neubiorev.2015.08.013. URL https://www.sciencedirect.com/science/article/pii/S0149763415002316. → pages 83

[192] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. doi:10.1109/5.726791. → pages 84

[193] FCP Classic Data Table, . URL http://fcon_1000.projects.nitrc.org/fcpClassic/FcpTable.html. → pages 93

[194] ABIDE, . URL http://fcon_1000.projects.nitrc.org/indi/abide/abide_I.html. → pages 93

[195] Christa Dang, Nawaf Yassi, Karra D. Harrington, Ying Xia, Yen Ying Lim, David Ames, Simon M. Laws, Martha Hickey, Stephanie Rainey-Smith, Hamid R. Sohrabi, James D. Doecke, Jurgen Fripp, Olivier Salvado, Peter J. Snyder, Michael Weinborn, Victor L. Villemagne, Christopher C. Rowe, Colin L. Masters, and Paul Maruff. Rates of age- and amyloid -associated cortical atrophy in older adults with superior memory performance. *Alzheimer's & Dementia : Diagnosis, Assessment & Disease Monitoring*, 11:566–575, August 2019. doi:10.1016/j.dadm.2019.05.005. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6939054/. → pages 93

[196] BNU 1, . URL http://fcon_1000.projects.nitrc.org/indi/CoRR/html/bnu_1.html. → pages 93

[197] BNU 2, . URL http://fcon_1000.projects.nitrc.org/indi/CoRR/html/bnu_2.html. → pages 93

[198] Data Description — Consortium for Reliability and Reproducibility (CoRR) documentation, . URL http://fcon_1000.projects.nitrc.org/indi/CoRR/html/samples.html. → pages 93, 94, 95

[199] Chaogan Yan, Dongqiang Liu, Yong He, Qihong Zou, Chaozhe Zhu, Xinian Zuo, Xiangyu Long, and Yufeng Zang. Spontaneous Brain Activity in the Default Mode Network Is Sensitive to Different Resting-State Conditions with Limited Cognitive Load. *PLOS ONE*, 4(5):e5743, May 2009. doi:10.1371/journal.pone.0005743. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0005743. Publisher: Public Library of Science. → pages 93

[200] Jason R. Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A. Shafto, Marie Dixon, Lorraine K. Tyler, Cam-CAN, and Richard N. Henson. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *NeuroImage*, 144:262–269, January 2017. doi:10.1016/j.neuroimage.2015.09.018. URL https://www.sciencedirect.com/science/article/pii/S1053811915008150. → pages 93

[201] COBRE, . URL http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html. → pages 94

[202] Vince D. Calhoun, Jing Sui, Kent Kiehl, Jessica A. Turner, Elena A. Allen, and Godfrey Pearlson. Exploring the Psychosis Functional Connectome: Aberrant Intrinsic Networks in Schizophrenia and Bipolar Disorder. *Frontiers in Psychiatry*, 2, 2012. doi:10.3389/fpsyt.2011.00075. URL https://www.frontiersin.org/articles/10.3389/fpsyt.2011.00075/full. Publisher: Frontiers. → pages 94

[203] Denise C. Park and Sara B. Festini. *The Middle-Aged Brain: A Cognitive Neuroscience Perspective*. Oxford University Press. ISBN 978-0-19-066226-4. URL https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199372935.001.0001/acprof-9780199372935-chapter-15. Publication Title: Cognitive Neuroscience of Aging Section: Cognitive Neuroscience of Aging. → pages 94

[204] Avram J. Holmes, Marisa O. Hollinshead, Timothy M. O'Keefe, Victor I. Petrov, Gabriele R. Fariello, Lawrence L. Wald, Bruce Fischl, Bruce R. Rosen, Ross W. Mair, Joshua L. Roffman, Jordan W. Smoller, and Randy L. Buckner. Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures. *Scientific Data*, 2(1):150031, July 2015. doi:10.1038/sdata.2015.31. URL https://doi.org/10.1038/sdata.2015.31. → pages 94

[205] David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E.J. Behrens, Essa Yacoub, and Kamil Ugurbil. The WU-Minn Human Connectome Project: An Overview. *NeuroImage*, 80:62–79, October 2013. doi:10.1016/j.neuroimage.2013.05.041. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3724347/. → pages 94

[206] Ke Zhao, Wen-Jing Yan, Yu-Hsin Chen, Xi-Nian Zuo, and Xiaolan Fu. Amygdala Volume Predicts Inter-Individual Differences in Fearful Face Recognition. *PLOS ONE*, 8(8):e74096, August 2013. doi:10.1371/journal.pone.0074096. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0074096. Publisher: Public Library of Science. → pages 94

[207] IXI Dataset – Brain Development, . URL https://brain-development.org/ixi-dataset/. → pages 94

[208] Kate Brody Nooner, Stanley J. Colcombe, Russell H. Tobe, Maarten Mennes, Melissa M. Benedict, Alexis L. Moreno, Laura J. Panek, Shaquanna Brown, Stephen T. Zavitz, Qingyang Li, Sharad Sikka, David Gutman, Saroja Bangaru, Rochelle Tziona Schlachter, Stephanie M. Kamiel, Ayesha R. Anwar, Caitlin M. Hinz, Michelle S. Kaplan, Anna B. Rachlin, Samantha Adelsberg, Brian Cheung, Ranjit Khanuja, Chaogan Yan, Cameron C. Craddock, Vincent Calhoun, William Courtney, Margaret King, Dylan Wood, Christine L. Cox, A. M. Clare Kelly, Adriana Di Martino, Eva Petkova, Philip T. Reiss, Nancy Duan, Dawn Thomsen, Bharat Biswal, Barbara Coffey, Matthew J. Hoptman, Daniel C. Javitt, Nunzio Pomara, John J. Sidtis, Harold S. Koplewicz, Francisco Xavier Castellanos, Bennett L. Leventhal, and Michael P. Milham. The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry. *Frontiers in Neuroscience*, 6, October 2012. doi:10.3389/fnins.2012.00152. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3472598/. → pages 94

[209] Emily Flagg. The Parkinson's Progression Markers Initiative (PPMI) Protocol. page 126. → pages 95

[210] Jie Lu, Hesheng Liu, Miao Zhang, Danhong Wang, Yanxiang Cao, Qingfeng Ma, Dongdong Rong, Xiaoyi Wang, Randy L. Buckner, and Kuncheng Li. Focal Pontine Lesions Provide Evidence That Intrinsic Functional Connectivity Reflects Polysynaptic Anatomical Pathways. *The Journal of Neuroscience*, 31(42):15065–15071, October 2011. doi:10.1523/JNEUROSCI.2364-11.2011. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3397237/. → pages 95

[211] Yun Xu and Royston Goodacre. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing*, 2(3):249–262, 2018. doi:10.1007/s41664-018-0068-2. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6373628/. → pages 98

[212] Yanhao Ren, Qiang Luo, Weikang Gong, and Wenlian Lu. *Transfer Learning Models on Brain Age Prediction*. August 2019. ISBN 978-1-4503-7262-6. doi:10.1145/3364836.3364893. → pages 98, 101, 102, 105

[213] Douglas N. Greve and Bruce Fischl. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48(1):63–72, October 2009. doi:10.1016/j.neuroimage.2009.06.060. URL http://www.sciencedirect.com/science/article/pii/S1053811909006752. → pages 100

[214] Adriana M Azor, James H Cole, Anthony J Holland, Maureen Dumba, Maneesh C Patel, Angelique Sadlon, Anthony P. Goldstone, and Katherine E. Manning. Increased brain age in adults with Prader-Willi syndrome. *NeuroImage. Clinical*, 21: 101664–101664, 2019. doi:10.1016/j.nicl.2019.101664. URL https://pubmed.ncbi.nlm.nih.gov/30658944. → pages 101

[215] Maria Ly, Gary Z. Yu, Helmet T. Karim, Nishita R. Muppidi, Akiko Mizuno, William E. Klunk, and Howard J. Aizenstein. Improving brain age prediction models: incorporation of amyloid status in Alzheimer's disease. *Neurobiology of Aging*, 87:44–48, March 2020. doi:10.1016/j.neurobiolaging.2019.11.005. URL http://www.sciencedirect.com/science/article/pii/S019745801930394X. → pages 101

[216] Stephen M. Smith, Diego Vidaurre, Fidel Alfaro-Almagro, Thomas E. Nichols, and Karla L. Miller. Estimation of brain age delta from brain imaging. *NeuroImage*, 200:528–539, October 2019. doi:10.1016/j.neuroimage.2019.06.017. URL http://www.sciencedirect.com/science/article/pii/S1053811919305026. → pages 101

[217] Maxwell L. Elliott, Daniel W. Belsky, Annchen R. Knodt, David Ireland, Tracy R. Melzer, Richie Poulton, Sandhya Ramrakha, Avshalom Caspi, Terrie E. Moffitt, and Ahmad R. Hariri. Brain-age in midlife is associated with accelerated biological aging and cognitive decline in a longitudinal birth cohort. *Molecular Psychiatry*, December 2019. doi:10.1038/s41380-019-0626-7. URL https://doi.org/10.1038/s41380-019-0626-7. → pages 101

[218] B. A. Jonsson, G. Bjornsdottir, T. E. Thorgeirsson, L. M. Ellingsen, G. Bragi Walters, D. F. Gudbjartsson, H. Stefansson, K. Stefansson, and M. O. Ulfarsson. Brain age prediction using deep learning uncovers associated sequence variants. *Nature Communications*, 10(1):5409, November 2019. doi:10.1038/s41467-019-13163-9. URL https://doi.org/10.1038/s41467-019-13163-9. → pages 101, 131

[219] R. Fujimoto, K. Ito, K. Wu, K. Sato, Y. Taki, H. Fukuda, and T. Aoki. Brain age estimation from T1-weighted images using effective local features. In *2017 39th*

Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 3028–3031, July 2017. ISBN 1558-4615. → pages 101, 102, 164, 190

[220] Habtamu M. Aycheh, Joon-Kyung Seong, Jeong-Hyeon Shin, Duk L. Na, Byungkon Kang, Sang W. Seo, and Kyung-Ah Sohn. Biological Brain Age Prediction Using Cortical Thickness Data: A Large Scale Cohort Study. *Frontiers in Aging Neuroscience*, 10:252, 2018. doi:10.3389/fnagi.2018.00252. URL https://www.frontiersin.org/article/10.3389/fnagi.2018.00252. → pages 101, 102, 103, 164, 190

[221] Jenessa Lancaster, Romy Lorenz, Rob Leech, and James H Cole. Bayesian Optimization for Neuroimaging Pre-processing in Brain Age Classification and Prediction. *Frontiers in aging neuroscience*, 10:28–28, February 2018. doi:10.3389/fnagi.2018.00028. URL https://pubmed.ncbi.nlm.nih.gov/29483870. Publisher: Frontiers Media S.A. → pages 101, 102, 103

[222] Iman Beheshti, Pierre Gravel, Olivier Potvin, Louis Dieumegarde, and Simon Duchesne. A novel patch-based procedure for estimating brain age across adulthood. *NeuroImage*, 197:618–624, August 2019. doi:10.1016/j.neuroimage.2019.05.025. URL http://www.sciencedirect.com/science/article/pii/S1053811919304173. → pages 101, 102, 103, 163, 190

[223] Pascal Sturmfels, Saige Rutherford, Mike Angstadt, Mark Peterson, Chandra Sripada, and Jenna Wiens. A Domain Guided CNN Architecture for Predicting Age from Structural Brain Images. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 3rd Machine Learning for Healthcare Conference*, volume 85 of *Proceedings of Machine Learning Research*, pages 295–311, Palo Alto, California, August 2018. PMLR. URL http://proceedings.mlr.press/v85/sturmfels18a.html. → pages 101, 102, 105, 160

[224] Katja Franke and Christian Gaser. Longitudinal changes in individual BrainAGE in healthy aging, mild cognitive impairment, and Alzheimer's disease. *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry*, 25(4):235–245, 2012. doi:10.1024/1662-9647/a000074. Place: Germany Publisher: Hogrefe Publishing. → pages 102, 196

[225] Heath R. Pardoe, James H. Cole, Karen Blackmon, Thomas Thesen, and Ruben Kuzniecky. Structural brain changes in medically refractory focal epilepsy resemble premature brain aging. *Epilepsy Research*, 133:28–32, July 2017. doi: 10.1016/j.eplepsyres.2017.03.007. URL http://www.sciencedirect.com/science/article/pii/S0920121116302650. → pages 102, 196

[226] James H. Cole, Jonathan Underwood, Matthan W.A. Caan, Davide De Francesco, Rosan A. van Zoest, Robert Leech, Ferdinand W.N.M. Wit, Peter Portegies, Gert J. Geurtsen, Ben A. Schmand, Maarten F. Schim van der Loeff, Claudio Franceschi, Caroline A. Sabin, Charles B.L.M. Majoie, Alan Winston, Peter Reiss, and David J. Sharp. Increased brain-predicted aging in treated HIV disease. *Neurology*, 88(14): 1349, April 2017. doi:10.1212/WNL.0000000000003790. URL http://n.neurology.org/content/88/14/1349.abstract. → pages 102, 196

[227] Han Peng, Weikang Gong, Christian F. Beckmann, Andrea Vedaldi, and Stephen M. Smith. Accurate brain age prediction with lightweight deep neural networks. *bioRxiv*, page 2019.12.17.879346, January 2019. doi:10.1101/2019.12.17.879346. URL http://biorxiv.org/content/early/2019/12/18/2019.12.17.879346.abstract. → pages 104, 105, 129, 130, 131, 160

[228] James H Cole, Tiina Annus, Liam R Wilson, Ridhaa Remtulla, Young T Hong, Tim D Fryer, Julio Acosta-Cabronero, Arturo Cardenas-Blanco, Robert Smith, David K Menon, Shahid H Zaman, Peter J Nestor, and Anthony J Holland. Brain-predicted age in Down syndrome is associated with beta amyloid deposition and cognitive decline. *Neurobiology of aging*, 56:41–49, August 2017. doi:10.1016/j.neurobiolaging.2017.04.006. URL https://pubmed.ncbi.nlm.nih.gov/28482213. Edition: 2017/04/18 Publisher: Elsevier. → pages 105, 196

[229] James H. Cole and Katja Franke. Predicting Age Using Neuroimaging: Innovative Brain Ageing Biomarkers. *Trends in Neurosciences*, 40(12):681–690, December 2017. doi:10.1016/j.tins.2017.10.001. URL https://doi.org/10.1016/j.tins.2017.10.001. Publisher: Elsevier. → pages 129

[230] Naftali Raz, Ulman Lindenberger, Paolo Ghisletta, Karen M Rodrigue, Kristen M Kennedy, and James D Acker. Neuroanatomical correlates of fluid intelligence in healthy adults and persons with vascular risk factors. *Cerebral cortex (New York,*

N.Y. : 1991), 18(3):718–726, March 2008. doi:10.1093/cercor/bhm108. URL https://pubmed.ncbi.nlm.nih.gov/17615248. Edition: 2007/07/05. → pages 134

[231] Dongtao Wei, Kaixiang Zhuang, Lei Ai, Qunlin Chen, Wenjing Yang, Wei Liu, Kangcheng Wang, Jiangzhou Sun, and Jiang Qiu. Structural and functional brain scans from the cross-sectional Southwest University adult lifespan dataset. *Scientific Data*, 5(1):180134, July 2018. doi:10.1038/sdata.2018.134. URL https://doi.org/10.1038/sdata.2018.134. → pages 134, 135

[232] Xinke Shen, Tao Liu, Dacheng Tao, Yubo Fan, Jicong Zhang, Shuyu Li, Jiyang Jiang, Wanlin Zhu, Yilong Wang, Yongjun Wang, Henry Brodaty, Perminder Sachdev, and Wei Wen. Variation in longitudinal trajectories of cortical sulci in normal elderly. *NeuroImage*, 166:1–9, February 2018. doi:10.1016/j.neuroimage.2017.10.010. URL http://www.sciencedirect.com/science/article/pii/S1053811917308224. → pages 134

[233] Lior Shamir, Nikita Orlov, D. Mark Eckley, Tomasz Macura, Josiah Johnston, and Ilya G. Goldberg. Wndchrm – an open source utility for biological image analysis. *Source Code for Biology and Medicine*, 3(1):13, July 2008. doi:10.1186/1751-0473-3-13. URL https://doi.org/10.1186/1751-0473-3-13. → pages 137

[234] Lan Lan, Lei You, Zeyang Zhang, Zhiwei Fan, Weiling Zhao, Nianyin Zeng, Yidong Chen, and Xiaobo Zhou. Generative Adversarial Networks and Its Applications in Biomedical Informatics. *Frontiers in Public Health*, 8:164, 2020. doi:10.3389/fpubh.2020.00164. URL https://www.frontiersin.org/article/10.3389/fpubh.2020.00164. → pages 160

[235] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down Neural Attention by Excitation Backprop. *arXiv e-prints*, page arXiv:1608.00507, August 2016. URL https://ui.adsabs.harvard.edu/abs/2016arXiv160800507Z. → pages 163

[236] Ruth Fong and Andrea Vedaldi. Interpretable Explanations of Black Boxes by Meaningful Perturbation. *arXiv e-prints*, page arXiv:1704.03296, April 2017. URL https://ui.adsabs.harvard.edu/abs/2017arXiv170403296F. → pages 163

[237] Elina Thibeau-Sutre, Olivier Colliot, Didier Dormont, and Ninon Burgos. Visualization approach to assess the robustness of neural networks for medical im-

age classification. volume 11313, March 2020. doi:10.1117/12.2548952. URL https://doi.org/10.1117/12.2548952. → pages 163, 166

[238] T. D. Gedeon. Data mining of inputs: analysing magnitude and functional measures. *International journal of neural systems*, 8(2):209–218, April 1997. doi: 10.1142/s0129065797000227. Place: Singapore. → pages 164

[239] Chu-Chung Huang, Mu-En Liu, Hung-Wen Kao, Kun-Hsien Chou, Albert C. Yang, Ying-Hsiu Wang, Tong-Ru Chen, Shih-Jen Tsai, and Ching-Po Lin. Effect of Alzheimer's Disease Risk Variant rs3824968 at SORL1 on Regional Gray Matter Volume and Age-Related Interaction in Adult Lifespan. *Scientific reports*, 6:23362, March 2016. doi:10.1038/srep23362. → pages 164, 190

[240] GitHub - utkuozbulak/pytorch-cnn-visualizations: Pytorch implementation of convolutional neural network visualization techniques, . URL https://github.com/utkuozbulak/pytorch-cnn-visualizations. → pages 167

[241] Lucas Y. W. Hui and Alexander Binder. BatchNorm Decomposition for Deep Neural Network Interpretation. In Ignacio Rojas, Gonzalo Joya, and Andreu Catala, editors, *Advances in Computational Intelligence*, pages 280–291, Cham, 2019. Springer International Publishing. ISBN 978-3-030-20518-8. → pages 188

[242] Jonathan D. Rollins, Julianne S. Collins, and Kenton R. Holden. United States Head Circumference Growth Reference Charts: Birth to 21 Years. *The Journal of Pediatrics*, 156(6):907–913.e2, June 2010. doi:10.1016/j.jpeds.2010.01.009. URL http://www.sciencedirect.com/science/article/pii/S002234761000020X. → pages 192

[243] Vassili A Kovalev, Frithjof Kruggel, and D. Yves von Cramon. Gender and age effects in structural brain asymmetry as measured by MRI texture analysis. *NeuroImage*, 19(3):895–905, July 2003. doi:10.1016/S1053-8119(03)00140-X. URL https://www.sciencedirect.com/science/article/pii/S105381190300140X. → pages 193

[244] Madalena Esteves, Ricardo Magalhães, Paulo Marques, Teresa . Castanho, Carlos Portugal-Nunes, José M. Soares, Armando Almeida, Nadine C. Santos, Nuno Sousa, and Hugo Leite-Almeida. Functional Hemispheric (A)symmetries in the Aged Brain-Relevance for Working Memory. *Frontiers in aging neuroscience*, 10:58–58, March

2018. doi:10.3389/fnagi.2018.00058. URL https://pubmed.ncbi.nlm.nih.gov/29593523. → pages 193

[245] Tobias Kaufmann, Dennis van der Meer, Nhat Trung Doan, Emanuel Schwarz, Martina J. Lund, Ingrid Agartz, Dag Alnæs, Deanna M. Barch, Ramona Baur-Streubel, Alessandro Bertolino, Francesco Bettella, Mona K. Beyer, Erlend Bøen, Stefan Borgwardt, Christine L. Brandt, Jan Buitelaar, Elisabeth G. Celius, Simon Cervenka, Annette Conzelmann, Aldo Córdova-Palomera, Anders M. Dale, Dominique J. F. de Quervain, Pasquale Di Carlo, Srdjan Djurovic, Erlend S. Dørum, Sarah Eisenacher, Torbjørn Elvsåshagen, Thomas Espeseth, Helena Fatouros-Bergman, Lena Flyckt, Barbara Franke, Oleksandr Frei, Beathe Haatveit, Asta K. Håberg, Hanne F. Harbo, Catharina A. Hartman, Dirk Heslenfeld, Pieter J. Hoekstra, Einar A. Høgestøl, Terry L. Jernigan, Rune Jonassen, Erik G. Jönsson, Lars Farde, Lena Flyckt, Göran Engberg, Sophie Erhardt, Helena Fatouros-Bergman, Simon Cervenka, Lilly Schwieler, Fredrik Piehl, Ingrid Agartz, Karin Collste, Pauliina Victorsson, Anna Malmqvist, Mikael Hedberg, Funda Orhan, Peter Kirsch, Iwona Kłoszewska, Knut K. Kolskår, Nils Inge Landrø, Stephanie Le Hellard, Klaus-Peter Lesch, Simon Lovestone, Arvid Lundervold, Astri J. Lundervold, Luigi A. Maglanoc, Ulrik F. Malt, Patrizia Mecocci, Ingrid Melle, Andreas Meyer-Lindenberg, Torgeir Moberget, Linn B. Norbom, Jan Egil Nordvik, Lars Nyberg, Jaap Oosterlaan, Marco Papalino, Andreas Papassotiropoulos, Paul Pauli, Giulio Pergola, Karin Persson, Geneviève Richard, Jaroslav Rokicki, Anne-Marthe Sanders, Geir Selbæk, Alexey A. Shadrin, Olav B. Smeland, Hilkka Soininen, Piotr Sowa, Vidar M. Steen, Magda Tsolaki, Kristine M. Ulrichsen, Bruno Vellas, Lei Wang, Eric Westman, Georg C. Ziegler, Mathias Zink, Ole A. Andreassen, Lars T. Westlye, and Karolinska Schizophrenia Project (KaSP). Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nature Neuroscience*, 22(10):1617–1623, October 2019. doi:10.1038/s41593-019-0471-7. URL https://doi.org/10.1038/s41593-019-0471-7. → pages 196

[246] James H. Cole, Robert Leech, and David J. Sharp. Prediction of brain age suggests accelerated atrophy after traumatic brain injury. *Annals of neurology*, 77(4):571–581, April 2015. doi:10.1002/ana.24367. → pages 196

[247] Katja Franke, Christian Gaser, Brad Manor, and Vera Novak. Advanced BrainAGE in older adults with type 2 diabetes mellitus. *Frontiers in aging neuroscience*, 5:90, December 2013. doi:10.3389/fnagi.2013.00090. → pages 196

[248] Tiia Ngandu, Jenni Lehtisalo, Alina Solomon, Esko Levälahti, Satu Ahtiluoto, Riitta Antikainen, Lars Bäckman, Tuomo Hänninen, Antti Jula, Tiina Laatikainen, Jaana Lindström, Francesca Mangialasche, Teemu Paajanen, Satu Pajala, Markku Peltonen, Rainer Rauramaa, Anna Stigsdotter-Neely, Timo Strandberg, Jaakko Tuomilehto, Hilkka Soininen, and Miia Kivipelto. A 2 year multidomain intervention of diet, exercise, cognitive training, and vascular risk monitoring versus control to prevent cognitive decline in at-risk elderly people (FINGER): a randomised controlled trial. *Lancet (London, England)*, 385(9984):2255–2263, June 2015. doi:10.1016/S0140-6736(15)60461-5. Place: England. → pages 196

[249] Feng Shi, Jian Cheng, Li Wang, Pew-Thian Yap, and Dinggang Shen. LRTV: MR Image Super-Resolution With Low-Rank and Total Variation Regularizations. *IEEE transactions on medical imaging*, 34(12):2459–2466, December 2015. doi:10.1109/TMI.2015.2437894. URL https://pubmed.ncbi.nlm.nih.gov/26641727. → pages 197

[250] Sami J. Savio, Lara CV Harrison, Tiina Luukkaala, Tomi Heinonen, Prasun Dastidar, Seppo Soimakallio, and Hannu J. Eskola. Effect of slice thickness on brain magnetic resonance image texture analysis. *BioMedical Engineering OnLine*, 9(1):60, October 2010. doi:10.1186/1475-925X-9-60. URL https://doi.org/10.1186/1475-925X-9-60. → pages

[251] Koshino Kazuhiro, Rudolf A Werner, Fujio Toriumi, Mehrbod S Javadi, Martin G Pomper, Lilja B Solnes, Franco Verde, Takahiro Higuchi, and Steven P Rowe. Generative Adversarial Networks for the Creation of Realistic Artificial Brain Magnetic Resonance Images. *Tomography (Ann Arbor, Mich.)*, 4(4):159–163, December 2018. doi:10.18383/j.tom.2018.00042. URL https://pubmed.ncbi.nlm.nih.gov/30588501. Publisher: Grapho Publications, LLC. → pages 226