UNIVERSITY OF CAPE TOWN

MASTERS DISSERTATION

Saliency Mapping in Convolutional Neural Networks to Determine Brain Age Trajectories

Author:
Daniel TAYLOR

Supervisor:
Assoc. Prof. Jonathan SHOCK
Co-Supervisor:
Assoc. Prof. Deshen MOODLEY

A dissertation submitted in fulfilment of the requirements for the degree of Master of Science
in the
DEPARMENT OF MATHEMATICS AND APPLIED MATHEMATICS

12 February 2022

Declaration of Authorship

I, Daniel Taylor, hereby declare that the work on which this thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university. I authorise the University to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: Date: 12-02-2022.

Father Zossima: *Loving humility is marvellously strong the strongest of all things and there is nothing else like it.*

– Dostoyevsky, The Brothers Karamazov

Abstract

Brain Age (BA) estimation via Deep Learning has become a strong and reliable bio-marker for brain health, but the black-box nature of Neural Networks does not easily allow insight into the causal features of brain ageing. In this work, a ResNet model was trained as a BA regressor on T1 structural brain MRI volumes from a small cross-sectional cohort of 524 individuals. Using Layer-wise Relevance Propagation (LRP) and DeepLIFT saliency mapping techniques, analyses were performed on the trained model to determine the most revealing structures over the course of brain ageing for the network, and compare these between the saliency mapping techniques. This work shows the change in attribution of relevance to different brain regions through the course of ageing. A tripartite pattern of relevance attribution to brain regions emerges. Some regions increase in relevance with age (e.g. the right Transverse Temporal Gyrus, known to be affected by healthy ageing); some decrease in relevance with age (e.g. the right Fourth Ventricle, known to dilate with age); and others remained consistently relevant across ages. This work also examines the effect of Brain Age Delta (DBA) on the distribution of relevance within the brain volume, for both older and younger individuals. It is hoped that these findings will provide clinically relevant region-wise trajectories for normal brain ageing, and a baseline against which to compare brain ageing trajectories.

Acknowledgements

Funding for this research was provided by the University of Cape Town Computer Science Department and PGFO, and ETDP-SETA over the course of the 2020 and 2021 academic years.

Model training and saliency mapping were performed on the Lambda Cloud GPU platform, using their NVIDIA RTX A6000 instances.

Preprocessing utilised FSL's [1] BET and FLIRT tools [2, 3, 4]. The randomise tool was used as well [5], for the permutation-based t-tests.

Brain regions were determined using the CerebrA 2009c atlas [6, 7], and registration used a standard MNI volume [8].

My sincerest thanks go to my supervisor, Associate Professor Jonathan Shock, who not only helped develop the ideas for directions of research in this work and guided me through the process of academic research and writing, but also helped me to think through ideas and processes deeply, give clear and concise descriptions, and to understand the limitations inherent to every person and their work – to understand that shortcomings are inevitable and that improvement comes only through deep reflection and thoughtful iteration.

I also extend my deepest thanks to my co-supervisor, Associate Professor Deshen Moodley, without whom much of the experimental rigour and structure of this work would not have been nearly of the same standard, and whose suggestions and assurances helped form this dissertation from multiple directions.

Many thanks go to Dr Jonathan Ipser for the sparing of much of his valuable time to provide great insight into brain imaging and ageing; to Dr Matthias Treder of Cardiff University for his time and effort in guiding the acquisition of our data and helping with its processing and the formation of our models; and to Dr Kamen Tsvetanov of the University of Cambridge for expediting access to our dataset.

None of this would have been possible – especially through the lock-downs and restrictions of the COVID-19 pandemic – without the love and support of my dear family and friends, who offered not only the means to keep afloat through these last two years, but the ability to live comfortably and happily, while always making themselves available for the lamenting of my most recent failures. To my dear Mother and Father in particular, I do not have the words to thank you – you have given me far more than I could ever hope for in my upbringing, and yet still allowed me to live a life of my own choosing and my own creation.

Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

LRP	**L**ayer-wise **R**elevance **P**ropagation
DeepLIFT	**Deep** **L**earning **I**mportant **F**ea**t**ures
MRI	**M**agnetic **R**esonance **I**maging
BA	**B**rain **A**ge
DBA	Brain Age Delta/**D**elta **B**rain **A**ge
ML	**M**achine **L**earning
DL	**D**eep **L**earning
CNN	**C**onvolutional **N**eural **N**etwork
ResNet	**Res**idual **Net**work
XAI	E**x**plainable **A**rtificial **I**ntelligence
Cam-CAN	**Cam**bridge **C**enter for **A**geing **N**euroscience
T1R	**T**op-**1**% **R**elevance
SOTA	**S**tate-**o**f-**t**he-**a**rt
CSF	**C**erebro**s**pinal **F**luid
MSE	**M**ean **S**quared **E**rror
MAE	**M**ean **A**bsolute **E**rror
MAPE	**M**ean **A**bsolute **P**ercentage **E**rror
SD	**S**tandard **D**eviation
CoV	**C**oefficient **o**f **V**ariation
MNI	**M**ontreal **N**eurological **I**nstitute
GCN	**G**lobal **C**ontrast Normalisation

Dedicated to my Grandpa Ian. May there come a day where ageing does not mean suffering.

Chapter 1

Introduction

1.1 Background

Deep Learning methods are a family of Machine Learning tools that create feature representations for data to extract meaningful information from it. This is useful for analysing large data structures or data that is not visualisable or too high-dimensional for humans. In such cases the powerful Deep Learning methods are able to analyse the data using learned statistical models.

One issue with modern Deep Learning tools is their inherent uninterpretability. These models commonly optimise anywhere between thousands and billions of parameters to perform their tasks successfully. This seems often to be a necessary trade-off for the analytical power brought about by the learning methods. Indeed, a trend in Deep Learning models seems to be that the larger the model, the more powerful its analytic capabilities [9]. This leaves us unable to understand how they come to their decisions, and as such they are often treated as “black-boxes”. Much research has been dedicated in recent years to solving the black-box problem, within the realm of *Explainable Artificial Intelligence* (XAI). One method of shedding light on the decision processes of these models is to create a map on the input space for a given input of the most salient features to the model’s decision. Such ‘saliency maps’ elucidate what the models deem relevant and what they are ignoring in the data.

A large amount of focus has been placed on Deep Learning (DL) in medical research. The *Medical Imaging with Deep Learning* (MIDL) conference alone had 123 conference papers in 2021. DL is an incredibly useful tool here for a number of tasks, such as lesion detections and diagnoses [10]. Many medical imaging implementations of DL utilise volumes obtained through Magnetic Resonance Imaging (MRI), which non-invasively images parts of the body and allows for the inspection of different tissue types at any location, and often at very high resolutions. We focus on the task of brain age (BA) regression from 3D MRI volumes. This involves taking MRI scans from individuals and training a DL model to predict the individuals’ ages.

We wish to study the BA regression task by analysing such DL models using saliency mapping techniques. One way in which we can learn from the BA regression task is through the analysis of individuals predicted to be significantly older than their chronological age, and which areas of the brain contribute to this disparity in which individuals. We can also analyse the saliency of specific brain regions to BA over the course of age, to determine how regions contribute to BA over time. Such information is hoped to be accessible through saliency mapping techniques applied to a BA regression model.

1.2 Problem Statement and Motivation

In the context of BA regression via DL, the black-box problem raises the question of why the model makes the decision that a given brain volume comes from a person of the predicted age. Successful application of a saliency mapping technique will show which areas of the MRI volume are being focused on for the age prediction. If a given structure is highlighted in an individual’s saliency map, it would mean that the size, shape, position and/or tissue content (e.g. white matter vs grey matter content) of that structure has drawn attention as part of the age prediction. One would expect for example that an older individual would have very large cerebral ventricles as compared to a younger individual [11]; if the regression is accurate, then we would expect a successful saliency map to highlight the ventricles.

Saliency mapping methods have been applied to many DL tasks, including in medical image analysis, in Alzheimer's Disease prediction for example, using MRI volumes [12, 13, 14, 15], and for neonate brain preterm versus term classification [16]. There are few applications thus far of saliency mapping to BA regression [17, 18].

In general, there have been few attempts to use saliency mapping techniques for regression problems. A model that can accurately predict the age of a patient's brain as compared to a healthy baseline and subsequently explain its decision would be an extremely helpful diagnostic tool for BA pathology. For example, an individual may present with an apparently older brain than their chronological age; an interpretable model would be able to highlight the areas of their brain that differ from the healthy baseline, considering that those are the areas focused on to make the decision that the brain is older than the patient's actual age. Furthermore, such models will be able to estimate population BA saliency trajectories over age. This would allow for a region-wise BA comparison on an individual basis.

While the task of BA regression does not target a specific pathology like Alzheimer's disease detection would, it does show how an individual's brain is ageing compared to a healthy baseline. With growing interest in the phenomenon of age as something like a disease state [19], such a model is a highly useful tool.

Such models are also extremely easy to use once trained. Once an MRI scan is complete, the volume is simply fed into the model, which quickly provides a prediction. The saliency mapping method is then applied post-hoc as part of the pipeline in a similar amount of time to give a region-wise explanation for the specific decision.

1.3 Research Questions

We have three research questions which we would like to address in this work:

1. *What are the differences and similarities between the explanations of BA from different saliency mapping methods?*

2. *How does accelerated brain ageing affect the distribution of BA relevance?*

3. *How does BA saliency change with age on a region-wise basis?*

1.4 Hypotheses

Based on the current literature, we form the following three hypotheses to our research questions:

1. The saliency mapping methods are mathematically similar to one-another, and so we expect structural relevance to brain ageing – although multivariate and nonlinear – to be similar among them, with some slight variations.

2. We expect that in individuals with accelerated BA, relevance is concentrated in regions relevant to BA in higher proportion than in those individuals with expected or slowed brain ageing.

3. We expect that the proportion of BA saliency increases with age in all areas that are highly relevant to BA, while in other areas this proportion (necessarily) decreases.

1.5 Aim and Objectives

The aim of this research is to create a clinically relevant and ready-to-use BA regression tool that explains its decisions graphically on the input space. The key feature of such a tool is the rich information that is provided by the saliency mapping. Although saliency mapping has been performed for the BA regression task in the past, the focus of such studies has largely been on the regions which are deemed most salient [17, 18]. We wish not only to ensure that the explanations are consistent at least with our current understanding of brain ageing, but also to utilise the information from these explanations to create population trajectories of BA saliency for specific brain regions, which has not been done before.

Specifically, the objectives of this research are as follows:

1. Create an accurate BA regression model using DL techniques.
2. Apply saliency mapping techniques to the BA regression model and compare the results to known characteristics of brain ageing.
3. Analyse the differences between saliency mapping techniques specific to the BA regression task, to determine the strengths and limitations of each.
4. Examine the link between region-specific saliency and accelerated brain ageing both in older and younger individuals.
5. Create region-wise trajectories of BA saliency over ages from a population study.

1.6 Contributions of Research

The primary contribution of this research is the development of region-specific trajectories of BA saliency over the course of age. This serves two functions:

1. Determine the saliency of a given brain structure to ageing and the change thereof over time.
2. Allow for individual comparisons to a baseline relevance trajectory on a region-specific level to assess BA.

Further contribution comes from the region-specific analysis of BA saliency in individuals exhibiting pathological (accelerated) brain ageing. This allows us to determine key contributing regions to accelerated BA. The final contribution comes from the analysis of different saliency mapping techniques applied to the BA regression task. Differences and similarities between the results of each technique offer insight not only into the strengths and limitations of each, but also into different aspects of brain ageing. Previous studies have quantified regional attribution of saliency to BA. No study has examined the differences in attribution between saliency mapping techniques, none has analysed the regional distributions of relevance associated with accelerated BA, and none has examined region-wise trajectories of saliency across age. This work will be of particular interest in clinical applications of BA analysis. Our findings have been submitted as a full paper [20] to the MIDL 2022 conference.

1.7 Dissertation Layout

The remainder of this dissertation will take the following structure:

- **Chapter 2** – The Literature Review will go over the relevant research on the subjects at hand, starting with MRI imaging and Machine Learning in the analysis of medical images, and then focusing on saliency mapping and previous work in related fields. This section will advise on how best to fulfil our aim and objectives, as well as best practices for each step of the Experimental Design.
- **Chapter 3** – The Experimental Design will discuss the acquisition and pre-processing of data, the experimental methods, and methodological issues that were encountered and how they were addressed. This will all be discussed in the context of the relevant literature.
- **Chapter 4** – The Experimental Results will show the outcome of the experiments and show the extracted data that was of greatest utility.

- **Chapter 5** – The Discussion will analyse the results and the extracted data. Here we will determine the extent to which the work has answered the research questions and made the contributions we aimed for. We will also discuss what was expected and what was unexpected in light of the literature and our domain expert's analyses.
- **Chapter 6** – The Conclusion will summarise the findings, and examine avenues for future work and improvements.

Chapter 2

Literature Review

In this chapter, we review the background of the topics of interest for this thesis, best practices for BA regression and saliency mapping, and what has and has not been done in this area of research thus far.

2.1 MRI Imaging

In this section we will look briefly at the method of MRI imaging used in our task. We use T1-weighted volumes for our experiments collected from the Cam-CAN cross-sectional dataset [21]. The dataset consists of $N = 656$ brain scans of healthy individuals ranging in age from 18 to 89 years. The subject ages are given in Figure 2.1 below, ordered by patient identification number.

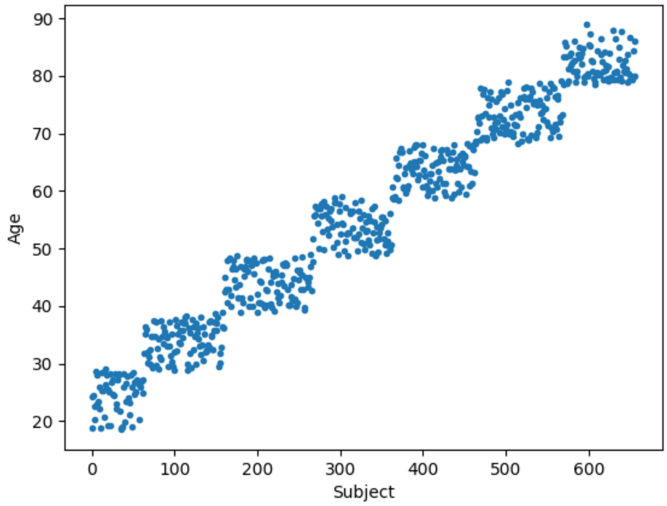


Figure 2.1: Ages of patients in the Cam-CAN cc700 dataset, ordered by patient ID.

There are many different MRI imaging sequences, including T1- and T2-weighted, diffusion-weighted, fMRI sequences like BOLD, and many more. All groups of imaging sequences though use the same basic underlying technology, which consists of a sequence of radio wave pulses onto the body of a patient placed in a very strong magnetic field (usually at about 1.5 Tesla strength in modern MRI machines, but ranging clinically from 0.2T to 7T). Under such a setting, certain types of atoms, including Hydrogen, are able to absorb radio frequency signals in such a way that their spin polarisation is altered, and as they realign the polarisations with the external magnetic field, emit radio frequency signals of their own that are detectable by a coil. In the case of human subjects, body fat and water molecules are rich in Hydrogen atoms, and the size and geometric configurations of these different molecules affects the signals by which they are detected in the radio frequency coils. This allows us to differentiate between different tissue types within the body in space when localising signals.

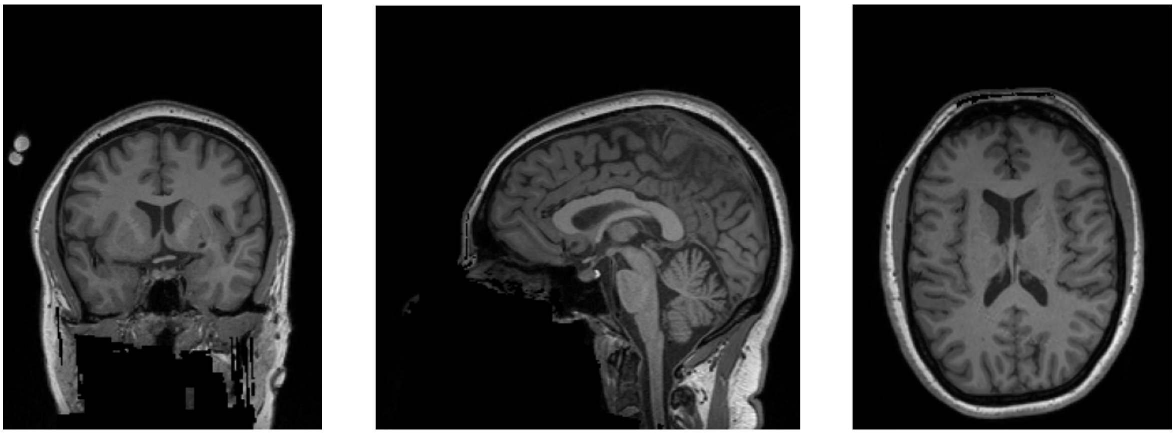
The contrast in an MRI image is determined by the different rates of return to equilibrium spin states of the Hydrogen atoms on different regions. The rate of return to equilibrium is determined (for the same radio frequency and magnetic field sequence) by two factors:

- The density of Hydrogen atom nuclei in the region,
- The proximity of other atoms.

This means that different materials – and specifically different tissue types – will be contrasted against one-another in the image.

The two most common types of structural imaging are T1- and T2-weighted imaging. T1 imaging maps a quantity across the volume that is associated with the recovery of longitudinal magnetisation – that is, the number of nuclei over time with spin in the direction of the applied magnetic field. On the other hand, T2 imaging maps a quantity across the volume associated with the decay of phase coherence of nuclei – that is, the number of nuclei in phase with one-another.

While T2-weighted imaging in the case of brain volume scanning is useful for the detection of white matter lesions, T1-weighted imaging is useful for the assessment of the cerebral cortex, and is less sensitive to white matter lesions. It is for this reason that we pay particular attention to T1-weighted images for our task. As discussed below, one of the hallmarks of brain ageing is the degradation of the cortex.

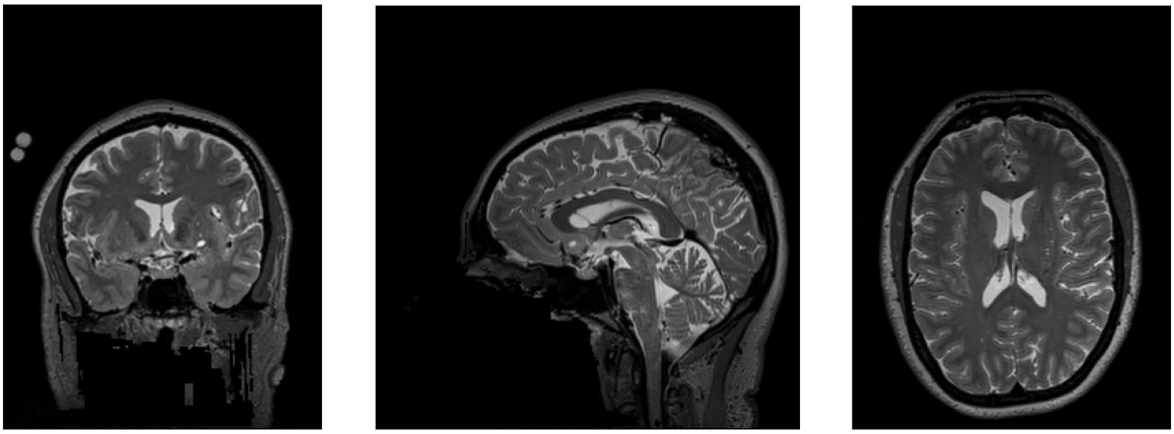


(a) Frontal section

(b) Sagittal section

(c) Transverse section

Figure 2.2: Sections of a T1-weighted MRI volume



(a) Frontal section

(b) Sagittal section

(c) Transverse section

Figure 2.3: Sections of a T2-weighted MRI volume

In Figures 2.2 and 2.3, examples are given of various sections from T1- and T2-weighted volumes of the same patient from the Cam-CAN dataset. These volumes, like all the other raw volumes in the dataset, are of shape $(256, 256, 192)$ voxels. Each voxel is approximately representative of 1mm^3. One can see in the figures that the T1-weighting quite clearly contrasts grey matter from white matter in the cortex (and other areas), and the T2-weighting clearly contrasts solid brain matter with the cerebrospinal fluid (light grey) for example.

2.2 A Broad Look at Machine Learning

In this section, we will look at Machine Learning generally in a broad sense, then focusing more clearly on Supervised Learning and its success and evolution, then examining Deep Learning with Neural Networks. This serves as background for the following section on the Convolutional Neural Network, which currently stands as the most popular Deep Learning tool for both classification and regression tasks.

ML is an extremely diverse field of study, which focuses on the broad goal of creating computational systems that can learn to perform useful tasks, and get better with more experience. The field has been in development since the early-to-mid 20th century, but has gained massive popularity and utility in recent decades, with famous milestones such as the 1997 defeat of Chess Grandmaster Garry Kasparov by IBM's Deep Blue [22]. Since then it has seen the advent of such achievements as the underpinning of modern self-driving cars [23], accurate speech-to-text software [24], realistic face generation [25], superhuman video-game performance [26] and even the ability to program [27].

Machine Learning can be divided in to three basic frameworks:

1. Supervised Learning – Iteratively teaching programs to understand a task or space with the use of labeled data, creating a function that maps inputs to their paired outputs.

2. Reinforcement Learning – Trying to use the maximisation of reward mechanisms to teach programs to use their environments, and trying to tailor the reward functions to illicit some desired behaviour.

3. Unsupervised Learning – Training a program on unlabelled data, and with minimal human input, to look for patterns in data and learn some level of representation of a dataset.

Within the realm of Supervised Learning (and often that of Unsupervised Learning) lies the powerful tool that is the Artificial Neural Network, often just called the Neural Network (NN). NNs were first put forward in their barest form in 1943 by McCulloch and Pitts [28], who reasoned that due to the "all-or-none" character of nervous activity, any network can be described in terms of propositional logic, and hence proposed a computational model for biological neural networks. Rosenblatt in 1958 created the perceptron [29], a basic model of human decision and memory pathways as a classifier with a binary output mode. Ivakhnenko and Lapa in 1965 created the first functional multi-layer perceptron under the name 'The Group Method of Data Handling' [30]. Backpropagation was first put forward by Henry Kelley in 1960 [31], and was refined by 1975 by Werbos [32] to the point that it could be implemented for the practical training of multi-layer NNs. Computational power was still insufficient for the practical use of NNs trained with backpropagation. Rumelhart et al. [33, 34] coined the term 'backpropagation' and popularised its algorithmic use. Many modern day supervised learning methods (all Deep Learning methods) use backpropagation in training. In the training phase, a model will produce outputs for labeled data, and its output will be compared with the label in a loss function, which produces a score serving as a notion of error of the model. The loss function aims to approximate the metric of error for the model accurately, but must be a differentiable function to be used in the backpropagation process. The score is used to adjust the weights and biases learned by a Neural Network according to their contribution to the decision. This can be done by any of a very large number of optimisation methods, but all successful methods use the partial derivative of the parameters with respect to the error so as to descend the error function to a local minimum. The method of using backpropagation to improve parameters iteratively using the partial derivatives of parameters to adjust them optimally is called Gradient Descent. The first detailings of Gradient Descent were made by Cauchy in 1847 [35] in application to systems of simultaneous equations. All backpropagation methods use some variation of gradient descent.

Eventually, the advances in computational power and speed allowed for the widespread use of NNs trained by backpropagation, and an increase in the networks' depths. Deep Neural Networks (DNNs) allow for the expression of highly nonlinear relationships in datasets, and have been shown to act as universal function approximators (given enough parameters) [36]. DNNs are the cornerstone of most Supervised Learning tasks, and their implementation is what is widely referred to as Deep Learning.

2.3 Convolutional Neural Networks (CNNs)

This section discusses the Convolutional Neural Network, which utilises a convolutional layer for feature detection. Many architectures, including state-of-the-art, have used and continue to use convolutional layers, and so it is that the Convolutional NN is widely regarded as the workhorse of modern Deep Learning.

We are most interested in the use of Deep Learning for image analysis. It is important to determine optimal architectures for a network to be able to pick up on visual features that determine output characteristics of interest – in our case, which features of a 3D brain MRI volume characterise age. This will inevitably lead not only to more accurate models for age regression, but also to more faithful saliency maps[1]. So we consider the currently very popular Convolutional Neural Network (CNN), which makes use of feature detectors to build up the notion of a characteristic in the training data from low- to high-level features.

In 1980, Fukushima [37] introduced the basic components of CNNs: the convolutional layer and down-sampling layers. In the original feed-forward NN designs, each neuron in a given non-input layer $L+1$ is connected to each of the neurons in the previous layer L. Its activation is given by the application of a nonlinearity f to an affine function of the input activations x_i^L:

$$\hat{x}_j^{L+1} = \sum_i \left(w_{ij} x_i^L + b_j \right) , \tag{2.1}$$

$$x_j^{L+1} = f\left(\hat{x}_j^{L+1} \right) , \tag{2.2}$$

where x_i^L is the activation of the ith neuron in layer L, w_{ij} is the weight between the neurons x_j^{L+1} and x_i^L, and b_j is a bias term. This is often visualised as in Figure 2.4, with the full connection of neurons in layer L to those in layer $L+1$.

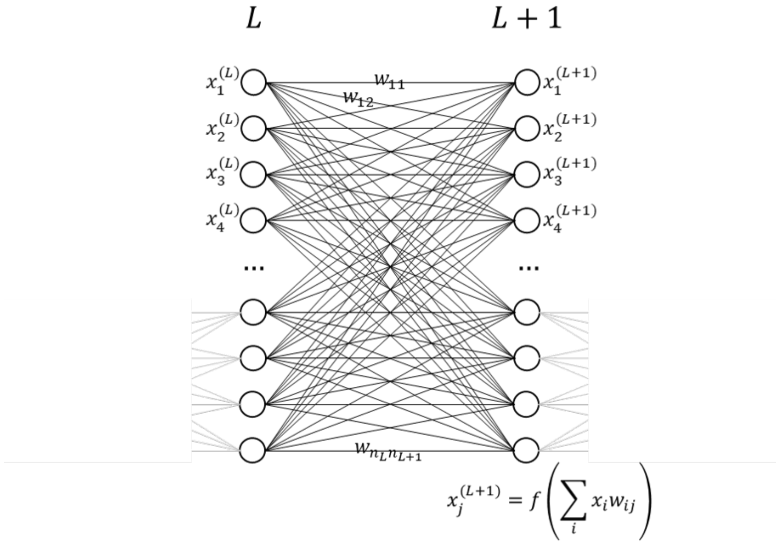


Figure 2.4: A visualisation of two adjacent layers in a fully-connected Neural Network, using nonlinear activation function f

Convolutional layers, however, make use of filters of weights between layers. These are scanned across the lower layer L with a set number of trained weights per filter. These 'convolution windows', also called kernels, scan across layers to search for features at a given level. The number of filters between two layers is then the number of feature detectors at that level of the network. The weights of convolution windows multiply against regions of the layer input, and the corresponding activations in the next layer follow Equation 2.1. Each activation in the higher layer corresponds to one scan of the convolution window of a region in the lower layer. This is shown for a simple 2-dimensional case with a single feature detector in Figure 2.5. Fukishima's

[1] We will define later exactly what is meant by 'faithfulness' of saliency maps

design was inspired by the works of Hubel and Wiesel, who identified in 1968 [38] two types of visual cells in the brain: one being a simple cell structure whose activation was maximised by straight edges with particular orientations in the visual field; and the other being a more complex cell structure with a greater field of vision, which was independent of the orientation of edges. They also proposed that a pattern recognition model could be used with combinations of these types of cells.

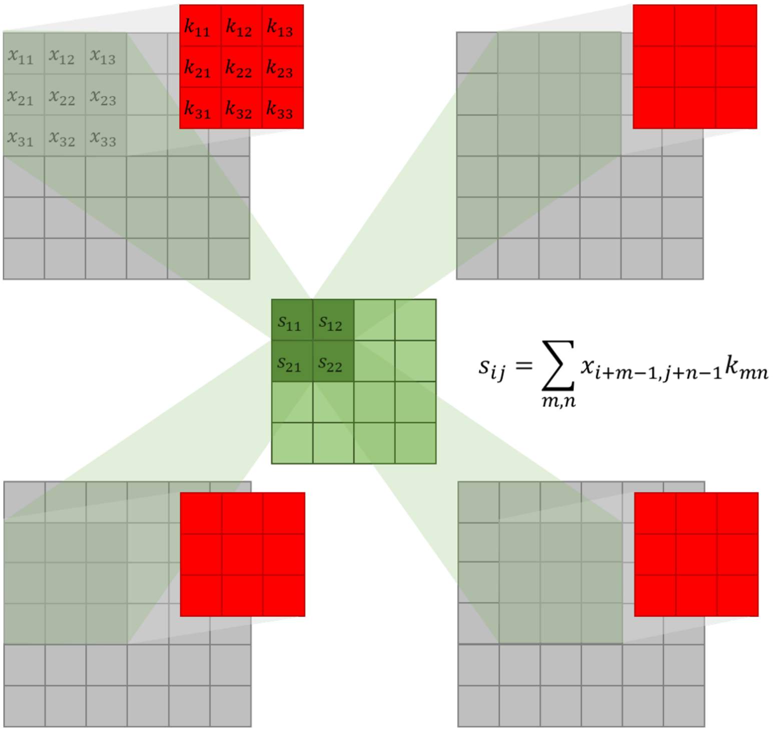

x_{11} x_{12} x_{13}
x_{21} x_{22} x_{23}
x_{31} x_{32} x_{33}
k_{11} k_{12} k_{13}
k_{21} k_{22} k_{23}
k_{31} k_{32} k_{33}
s_{11} s_{12}
s_{21} s_{22}
$s_{ij} = \sum_{m,n} x_{i+m-1,j+n-1} k_{mn}$

Figure 2.5: A simple convolutional window example, with window size 3 × 3 and stride length 1. Activations have a grid-like topology. Lower-layer activations are shown in grey, kernel values are shown in red, and higher-layer activations are shown in green.

Pooling layers down-sample the incoming data in either a linear or nonlinear way. Typically, activations are run through a convolution layer first, then the affine transformations are put through a nonlinearity as in Equation 2.2 before going through a pooling function [9]. The pooling function locally summarises nearby outputs with functions such as Max Pooling [39]. The Max Pool function takes the maximum activation of a window in the lower layer as the activation for the the corresponding position in the given layer. By doing this, the resolution of the activations is decreased by a factor of the window stride (how far the window is moved to its next position on the lower layer), but the amount of data to be processed is also decreased by the same factor, and the hope is that the most important features are retained. Average Pooling works similarly, by taking the average of values in the window in the lower layer. This has the benefit of ideally losing less information, but the trade-off is that highly contrastive activations in the window can average to a middling value which is not representative of either extreme. The Average Pooling function, and similar functions such as Sum Pooling (which simply does not divide by the window total) have the additional benefit of being differentiable, which although the layers are not trainable allows for gradient-based saliency mapping methods to be applied more easily to them than a method like Max Pool. A visualisation of Average and Max Pooling layers is given in Figure 2.6.

Later work by Waibel [40] implemented an NN with convolutional layers and down-sampling, trained with backpropagation and utilising weight sharing – that is, the assumption that the same feature detectors can be used in any position over the lower-level activations, since they are searching for the same features in different places. This allowed for shift invariance – the ability of the network to make the same predictions for one image and the same image shifted in some way. This is a hallmark of CNNs over other architecture types. Waibel's implementation is recognised as the first of a true CNN. CNNs have shown great utility in application to data with grid-like structure [9], such as time series [40, 41] and digital images [42, 43]. Not only do CNNs provide a significant performance boost over their fully-connected counterparts, but due to weight sharing, they also dramatically decrease the number of parameters that must be trained.

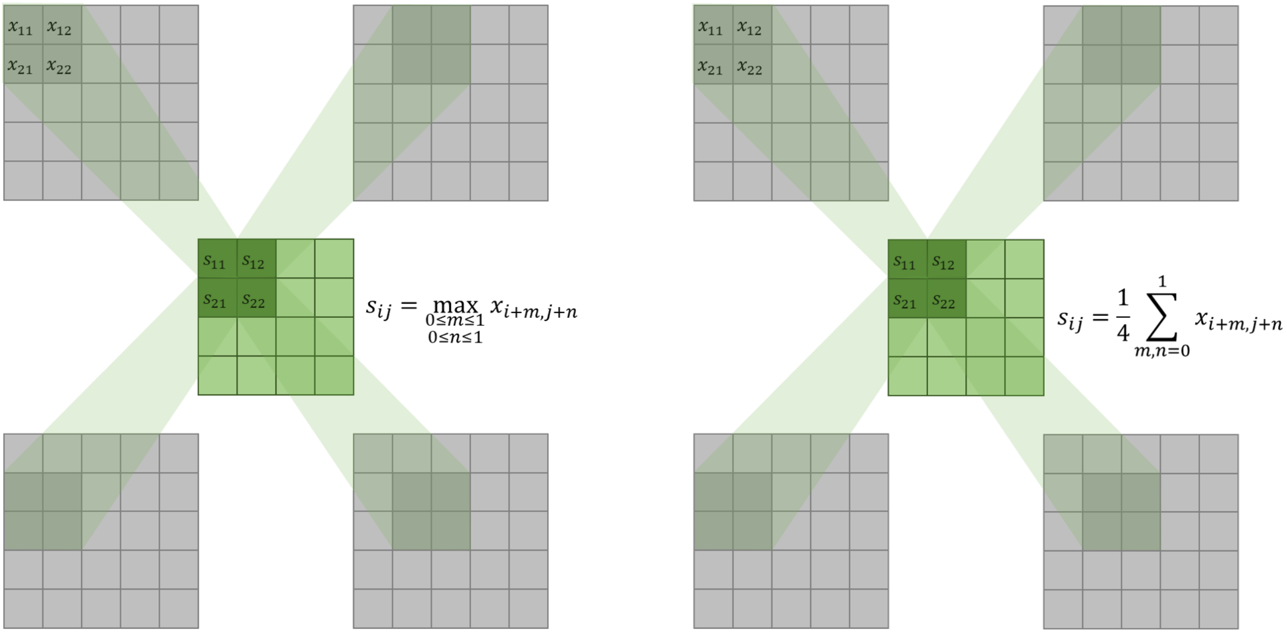


Figure 2.6: Simple examples of pooling layers with 2 × 2 pooling windows. Left: max pooling, right: average pooling.

The main motivations for the architecture of CNNs are from the work of Hubel and Wiesel, identifying the two types of neuronal cells that perform most of the activity in the mammalian primary visual cortex (V1) [9]. In a very simple-minded sense, the convolutional layers of CNNs act like the simple cell types in V1, in that they are sensitive to local activations of specific types – for example, line segments of specific orientations in lower layers. In the same vein, the pooling layers of CNNs act like the complex cell types in V1, as they are sensitive to features much like the convolutional layers, but are invariant to small shifts in the feature poses. The complex cells are also invariant to lighting changes, which has inspired pooling strategies across colour channels in some CNNs [44]. 2D imaging CNNs also loosely represent the structure of V1, in that V1 is laid out as a spatial map, with excitations in the lower part of the retina corresponding to activations in the lower parts of V1 and so forth.

It must be noted that CNNs are nowhere near a true representation of the mammalian visual system – especially not in humans, who have extraordinary visual acuity [45] (although worse night vision than many), seemingly surpassed in this regard only by some birds of prey, and by no examined mammalian species.

CNNs can be used for 2D images as well as 3D volumes (and of course 1D data, but we do not care about that for our task), for which the voxels are treated analogously to the pixels of their 2D counterparts. For 3D volumes, convolution windows become 3D boxes of learned weights to be passed over activations along three axes. Similarly, the pooling layers of a 3D CNN down-sample volumetrically. The best-performing classification and detection networks to date use some type of convolutional architecture, including Google's GoogLeNet [46], Noisy Student [47] and its successor FixEfficientNet-L2 [48] (respectively the former and current best-performing networks on the ImageNet dataset). These all take as input 2D RGB images.

2.4 Object Detection and Image Segmentation

Brain volume analysis in ML tasks is usually for the purpose of object detection and/or image segmentation. Practitioners often want to examine lesions in the brain volume to learn about specific pathology, or to isolate specific areas of the brain for atlasing tasks for example. While the task of BA regression does not directly necessitate these techniques, it does so implicitly. In order to assess accurately a subject's age, our model must be able to detect features in their MRI volume which indicate the age, and must be able to differentiate between certain structures in order to do so. This section examines the tasks of object detection and image segmentation and the success thus far of modern CNNs in, most notably, detection.

Of the objectives of visual Neural Networks, Object Detection and Image Segmentation are among those at the forefront. More than attempting to broaden the spectrum of tasks that can be performed by NNs with imagery, researchers have focused mainly on trying to replicate (and eventually surpass) human visual capabilities [9].

It is important to distinguish the tasks of:

1. **Classification:** The identification of an image or volume with a single class or category.
2. **Detection:** The reporting of the presence of an object within an image or volume.
3. **Segmentation:** The hierarchical separation and isolation of constituent parts of an image or volume.

Classification is a subset of Detection since it is all detection tasks regarding images with only one object present. Detection is therefore at least as difficult a task as classification, and is in fact generally regarded as far more difficult.

Segmentation is very closely related to Detection as well, since in order to isolate constituent parts of an image, a model must recognise their presence and location. Therefore, part of the segmentation process is feature detection. Due to their success in learning feature detectors over grid-like topology, CNNs are very useful and are widely favoured for all three of these tasks.

2.4.1 Pre-processing

Pre-processing is an important task for ensuring that input data is subject to some standardisation. Many models are able to take input of variable sizes, but if some of the data has a large range of activations and other supposedly similar data has a small range of activations (for example, similar images with very different contrasts), the model is not likely to be able to generalise very well. Pre-processing for most visual tasks is usually limited to reshaping, cropping and normalising inputs. In the case of MRI volume analysis, it is common practice to perform several other pre-processing steps as well, to isolate only brain matter, and align brain volumes to the same orientation (registration). Because of the dependence of model outputs on numerical values in many applications, it is necessary for all the samples to have activations within the same reasonable range in order to avoid issues [9]. One of the most common aims of image pre-processing is to reduce the amount of variation in the training and test data. This is especially useful for small datasets and smaller models. The most obvious method of reducing variation is to reduce image contrast – that is, the magnitude of the activation differences between pixels in an image. We usually treat the contrast in Deep Learning contexts as synonymous to the Standard Deviation [9]. That is, the contrast is given by:

$$\sigma = \sqrt{\frac{1}{3rc}\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{3}\left(x_{ijk}-\bar{x}\right)^2} \tag{2.3}$$

where i runs over the first image axis of length r, j runs over the second axis of length c, and k runs over the three colour channels. $\bar{x}$ is the image mean given by:

$$\bar{x} = \frac{1}{3rc}\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{3} x_{ijk}. \tag{2.4}$$

Global Contrast Normalisation (GCN) is a method of normalising the contrast by simply dividing through by σ after subtracting the mean:

$$x_{ijk}^{\mathrm{GCN}} = s\frac{x_{ijk}-\bar{x}}{\max\left(\varepsilon, \sqrt{\lambda + \frac{1}{3rc}\sum_{i'=1}^{r}\sum_{j'=1}^{c}\sum_{k'=1}^{3}\left(x_{i'j'k'}-\bar{x}\right)^2}\right)}. \tag{2.5}$$

Here, s is some scale factor typically chosen to be 1, ε is some threshold constant for numerical stability, and λ is a regularisation constant. λ is useful in the case that an image has very little contrast – in which case there is generally very little information – where division by σ would do little more than amplify noise.

Generally, one would use *either* ε or λ. Since the standard deviation 2.3 is simply a rescaling of the L^2 norm, GCN with $\lambda = 0$ maps images to a sphere in $\mathbb{R}^{3rc}$. The use of $\lambda > 0$ draws samples toward the origin in $\mathbb{R}^{3rc}$, but does not discard the variation in their norm. The mapping to the sphere is generally quite helpful, since NNs are typically much better at responding to directions than to exact locations – it is difficult, for example, for a network to distinguish multiple distances in the same direction.

GCN distinguishes well between brighter and dimmer regions in an image, but will not necessarily highlight edges or corners in darker parts of the image. To do this, we can use Local Contrast Normalisation (LCN), which normalises pixels with respect to the other pixels in close proximity, and not globally. The methodology is very similar to that of GCN, in that a window centered at the given pixel x_{ijk} is used to compute a mean and standard deviation, which are then used to normalise x_{ijk} according to an analogous rule to Eq. 2.5. Some variations exist, including a weighted mean and standard deviation calculation according to Gaussian wights centered on x_{ijk}.

In the BA regression literature, it is common practice to remove all non-brain matter from MRI volumes as part of pre-processing [49, 50, 51, 17]. This is called skull-stripping, and ensures that only the brain tissues are included in the BA regression input. Other tissues such as bone are not considered relevant to the task. There are several different tools that can be used to do this. We chose to use FSL's Brain Extraction Tool (BET) [2]. A fractional intensity of 0.5 is recommended by the developers of the tool for standard use.

We would like to align all of the MRI volumes to the same orientation, such that we can easily determine the location of individual brain regions using a single atlas. To do this, we register the volumes all to MNI space using a standard MNI volume [8]. The Montreal Neurological Institute created the MNI space as a standard for spatial alignment of brain MRI volumes. The standard MNI volume is an aggregate of several MNI-aligned brain scans, and it is common practice to use such a volume for registration.

2.4.2 CNNs for Detection

The early 2000s saw the rise of Support Vector Machines (SVMs) dominating over the use of CNNs in computer vision. In the early-to-mid-2010s though, advances in computational power began to turn favour again towards the use of CNNs. Girshick et al. (2014) [52] noted that at the time visual task performance was stagnating in Machine Learning, and that the leading performers were extremely complex models which gained an edge in performance with small but computationally expensive adaptations to the previously top-performing methods, which were mostly Scale-Invariant Feature Transform (SIFT) [53] and Histograms of Oriented Gradients (HOG) [54] methods. Girshick et al. proposed a multi-module ML model based on the premise that although SIFT and HOG methods roughly simulate the complex cells in V1, it was by then well-understood that there are recognition pathways deeper down the visual system, which suggests that there should be multi-stage visual processes in the human brain that are more informative for image recognition. At the time, the canonical visual recognition task was the PASCAL Visual Object Classes (VOC) object detection challenge [55]. This contains only a small amount of annotated data. The previous best performance was by SegDPM [56], which achieved a mean average precision (mAP) of 40.4%. Girshick et al.managed to achieve an mAP of 53.7% a year later. This was done by working off of two suppositions:

1. One can apply CNNs to region proposals to classify and segment objects;
2. Supervised pre-training on an auxiliary task followed by domain-specific fine-tuning will allow for significant improvement in object detection, when domain-specific training data is scarce.

The model proposed by Girshick et al. was comprised of a region extraction tool, followed by a large CNN used for computing features, and finally a set of class-specific linear SVMs to compute the final classifications. The Authors opted to develop recognition using regions, inspiring the name R-CNN. They used Selective Search [57] for region proposals so as to be able to have controlled comparisons with other detection work. The authors maintain, however, that their model is agnostic to the region proposal method. At test time, around 2000 region proposals are extracted. Feature extraction was performed by a CNN with five convolutional layers, and two fully-connected layers, with an RGB input of 227×227 pixels. The output of the CNN was a 4096-dimensional vector. This is a very low dimensionality compared to other methods, which allows for efficient detection in the SVM phase. The CNN parameters are also shared across all the categories, further aiding computational efficiency. The incoming region proposals for the CNN are of arbitrary dimensions, so they are warped in a tight bounding box around the region to the required dimensions, regardless of the region dimensions. For each feature class, the extracted feature vector is scored according to the corresponding learned SVM. Given all the scored regions in an image, the authors then applied a greedy non-maximum suppression which which rejected regions which had intersection-over-union (IoU) overlap with a higher-scoring region larger than a given threshold. The class-specific computations are limited to the multiplication of the 2000×4096 feature extraction matrix by the $4096 \times N$ SVM weight matrix (where N is the number of classes). The authors also took note of the regional operation of the model, and thus applied it to natural image segmentation, with some minor modifications. In doing so, they were able to achieve state-of-the-art results on the PASCAL VOC segmentation task, with an average segmentation accuracy of 47.9%.

The ImageNet dataset [58] is a crowd-sourced database of over 14 million labelled images, over 1 million of which are annotated with bounding boxes. Annual challenges have been put forward since 2010 with regard to ImageNet, all under the ILSVRC[2] [59] title.

Using ImageNet labelled data without bounding box labels, Girshick et al. pre-trained their CNN to high performance standards. This pre-training was a classification task; in the process of fine-tuning, the CNN was adapted to the task of detection on a new domain, warped PASCAL VOC windows. The only change made to the CNN architecture was the replacement of the ImageNet-specific 1000-way classification layer with a randomly-initialised 21-way classification layer – for the 20 PASCAL VOC classes and background. The authors treated region proposals with ≥ 0.5 IoU overlap with ground-truth boxes as positive for the box's class, and the rest as negative. If a region partially overlaps with an object to be detected in the image, it will not necessarily be easy to tell whether or not the model should then label it as true. The authors therefore adopted the strategy of using a threshold IoU with a ground truth bounding box to determine whether a region was to be labelled true or false for a given label. They tested several values for the threshold, and found that in a grid search over the values $\{0, 0.1, \ldots, 0.5\}$, the best option was 0.3. The SVMs are trained once the labels are applied to the extracted features.

The authors showed that learned features could be visualised from the network and thereby aid in its transparency. They did this by showing for individual neurons at specific places in the CNN which of the input image regions maximised its activation. They found that for each tested neuron, similar patterns arose between the most highly-activating image segments. It was found that some neurons responded to textures, others to features, and some even to high-level emergent features (such as a human, or letters and words). The authors found that removing the fully connected layers of the network still produced quite good results, even though computing the output of the network up to the layer before the fully connected layers only uses 6% of the network's learned parameters. This indicates that much of the CNN's computational power comes from the convolutional layers, and not the larger and much more computationally intensive fully-connected layers. The authors also found that the two fully connected layers were by far the most responsive to the fine-tuning regime. This would suggest that the features learned in the earlier stages of the network are nearly universal.

The authors compared their results to the previous top-performing models, Deformable Part Models (DPMs). It was found that they outperformed these older models by over 20%. Inspired by the bounding-box regression method of DPMs though, the authors tried training a linear regression model to predict a new detection window given the features computed before the fully-connected layers, for a selective search region proposal. This fixed large numbers of mislocalisations, and boosted performance by about 3%. It was found that the fine-tuning did not reduce the sensitivity of the model (the difference between the maximum and minimum AP scores), but increased both the highest and lowest performing modes. This would suggest that the fine-tuning improves the robustness of all the characteristics of the classification.

Finally, the authors applied R-CNN to the task of segmentation. This was done in three separate ways. The first method (*full*) was to compute CNN features directly on the warped window, ignoring the region shape. The downside of this is that it ignores the non-rectangular shape of the region. The second method (*fg*) was to compute CNN features using only the region's foreground mask. This allowed for non-overlapping object with similar bounding boxes to be segmented effectively. To do this, the background was replaced with the

[2] ImageNet Large Scale Visual Recognition Challenge

image mean, so that after mean subtraction, the background was zero. The third method was to concatenate the features of the *full* and *fg* features. The results showed that the two methods were in fact complementary.

The *full+fg* R-CNN slightly outperformed the previous leading segmentation method, O_2P (Second-order Pooling) [60]. The authors admit that under a reasonable margin of error, the performances are likely the same; but that the performance of the R-CNN model might improve with fine-tuning. It is also important to note that training the SVMs on the *full+fg* features takes about one hour on a single core, whereas training O_2P features takes over ten hours.

These models have been improved upon greatly since their inception, with architectures like Fast R-CNN [61], Faster R-CNN [62] and Mask R-CNN [63].

2.5 Explainable AI (XAI)

In this section, we discuss techniques of Explainable AI, the aim of which is to shed light on the black-box problem, and ultimately create AI systems whose actions we can understand as well as possible. It is with specific interest that we examine the DeepLIFT and LRP methods used for post-hoc explanation of DNN decisions. We discuss the emergence of these methods and their comparison thus far in literature to one-another and to other methods of saliency mapping. In this chapter we also discuss how we can evaluate the explanations provided by these XAI methods; criticisms of each of these methods are also discussed, as well as criticism of XAI in general and its executions thus far.

With the massive growth in machine learning research and adoption over the past few decades, AI is having more and more of an influence in everyday life, economics, politics, entertainment, and even justice systems. It is imperative therefore that the reasoning behind ML decisions is as clear as possible. If we cannot understand the outputs of our models, then they are leading us blindly into actions that could stray from our intended goals, and lead us by artifactual model decisions to undesirable outcomes. We must ensure that the incentives of these models are fully aligned with our own, and that they do not stray from that alignment.

There are cases of high-stakes ML models which have been implemented in the past decade that have been found to have some implicit biases [64, 65, 66, 67], skewing the outcomes of decisions in undesirable ways. This has led to a large focus on Explainable Artificial Intelligence in recent years. Model explanations also afford us insight into the models' analyses of a given task. If a model has super-human capabilities, explanation techniques could potentially provide insight into unknown relationships within a given study.

It is useful to have a model which can accurately predict a subject's age based on a brain MRI volume. Compared to a healthy baseline, this tells the subject and the present expert whether or not the subject is ageing healthily, neurophysiologically speaking. This 'healthy baseline' refers to the predictive model of healthy brain ageing represented by the predictor; we imply here the necessity for the training data of the model to be healthy individuals.

One pair of questions that arises from the prediction output is 'Why is this healthy or expected at this age?' or 'Why is this *un*healthy at this age?' If an expert were able to determine the age of a patient with a high degree of accuracy in such a way, their explanation would be comparative, contrastive and selective (as pointed out by Miller et al. [68]). The explanation would point to features in one slice or a sequence of slices of the MRI volume and compare them to the features of a mental baseline for a given age range. We would like to be able to do this as an automated process on top of or as part of our regression model, and in a consistent manner.

A regression model for a visual task like MRI volume analysis would benefit greatly from explanations which themselves are visual. Annotations or other visual cues on the input would be of great aid in showing areas of focus for the model's decision-making process. This is the aim of pixel-wise decomposition methods. These distribute some notion of relevance onto the input space for a given network decision (however relevance is defined by the creator of the method). The outputs of such process are called 'saliency maps' or 'heatmaps', due to their indication of areas of input on which the model focuses, and areas which it ignores. We hope to use saliency maps produced by pixel-wise decomposition methods to provide the same insight as (or at least similar to that of) a domain expert, and which can display the model's ability to recognise the highly nonlinear nature of brain ageing. It is convenient to approach pixel-wise decomposition for NNs via gradients. Since gradient descent via backpropagation is inherent to all modern CNNs, it is common to work with the gradients of the networks. These are readily accessible and their associated methods are generally computationally efficient, as opposed to – for example – occlusion-based methods.

2.5.1 Layer-wise Relevance Propagation (LRP)

LRP [69] is a category of gradient-based decomposition methods, which starts (like most gradient-based methods) by assigning a custom relevance score to the output of a network, given a specific input (usually just the output value or the target class activation). This relevance score is then propagated backward through the network according to some LRP assignment rule to the input layer to create a saliency map on the input space (in the case of visual data, these are pixels or voxels). We treat a regression network as a function $f : \mathbb{R}^V \to \mathbb{R}$, where $\mathbb{R}^V$ is the input space, with size and shape determined by V. Input can be a column vector, a 2-dimensional array, or something of larger dimensions.

For the sake of simplicity, the relevance assigned is commonly set as equal to the decision output of the network, $R = f(x)$. The criterion for LRP is that throughout the relevance propagation process, the total relevance is unchanged, so that between two layers L and $L+1$ with neurons i and j respectively:

$$\sum_i R_i^L = \sum_j R_j^{L+1} \tag{2.6}$$

and in particular, for input $x = x_d$ at layer 1 where the input is a column vector of length d, we have

$$R = f(x) = \sum_d R_d^{(1)}. \tag{2.7}$$

One method of implementing LRP is through a first-order Taylor approximation

$$f(x) \approx f(x_0) + \sum_d \frac{\partial f}{\partial x_{(d)}}(x_0)\left[x_{(d)} - x_{0(d)}\right]. \tag{2.8}$$

This is utilised by Sensitivity Analysis [70] to limited success. Generally, sensitivity analyses show areas of interest with low resolution [71, 72, 73].

Deep Taylor Decomposition (DTD) [72] improves dramatically upon the performance of Taylor-type LRP. Montavon et al. introduced this method separately from LRP, with their own set of criteria. The criteria are that the heatmap is *conservative*:

$$\forall x, \quad f(x) = \sum_d R_d^{(1)}(x), \tag{2.9}$$

which is simply the LRP condition; and that the heatmap is *positive*:

$$\forall x, d, \quad R_d^{(1)}(x) \geq 0. \tag{2.10}$$

Together, these conditions constitute *consistency*. If a heatmapping is consistent, then we force the condition that no relevance is assigned if a feature is not present. Although it was not strictly introduced as such DTD clearly falls under the LRP category of gradient-based approaches.

With these criteria, Montavon et al. devised rules for DTD under three types of input constraints. The first is the w^2-rule, which is used for unconstrained input spaces:

$$R_i^L = \sum_j \frac{w_{ij}^2}{\sum_{i'} w_{i'j}^2} R_j^{L+1}, \tag{2.11}$$

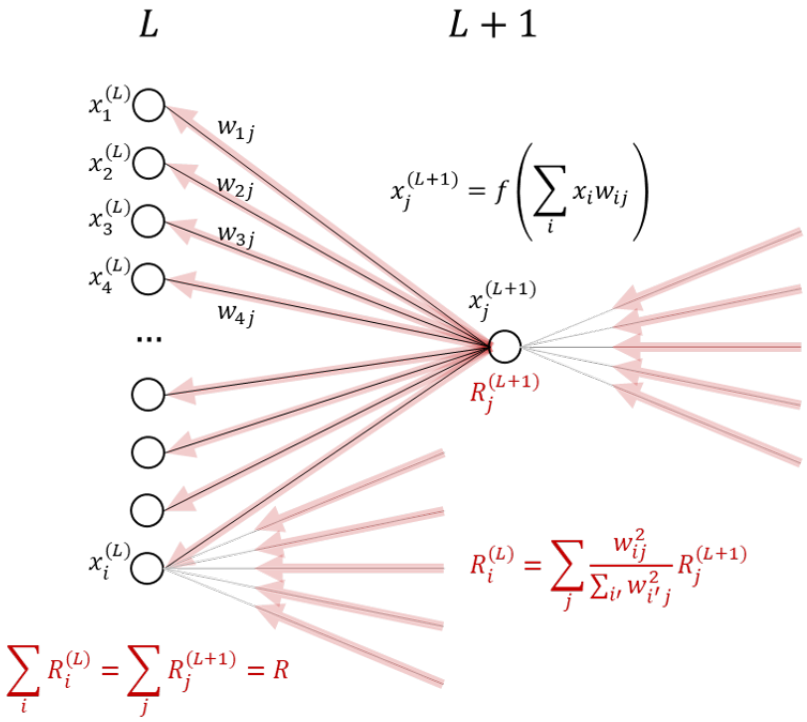
where w_{ij} is the weight connecting neuron i in layer L to neuron j in layer $L+1$. The second is the z^+-rule for non-negative activation spaces (such as follow rectified linear unit activations):

$$R_i^L = \sum_j \frac{z_{ij}^+}{\sum_{i'} z_{i'j}^+} R_j^{L+1}, \tag{2.12}$$

where $z_{ij}^+ = x_i w_{ij}^+$ and w_{ij}^+ is the non-negative part of w_{ij} and zeros elsewhere. The third rule is the $z^{\mathcal{B}}$-rule, which is used for spaces constrained from above and below:

$$R_i^L = \sum_j \frac{z_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-}{\sum_{i'} z_{i'j} - l_{i'} w_{i'j}^+ - h_{i'} w_{i'j}^-} R_j^{L+1}, \tag{2.13}$$

where l is the lower bound on the lower layer activations, and h is the upper bound. The use of these Deep Taylor Decomposition rules has proved to work well on several test datasets, including MNIST and ILSVRC. Figure 2.7 shows a diagram depicting the w^2-rule of the Deep Taylor Algorithm.



L
$L+1$
$x_1^{(L)}$
$x_2^{(L)}$
$x_3^{(L)}$
$x_4^{(L)}$
w_{1j}
w_{2j}
w_{3j}
w_{4j}
...
$x_j^{(L+1)} = f\left(\sum_i x_i w_{ij}\right)$
$x_j^{(L+1)}$
$R_j^{(L+1)}$
$x_i^{(L)}$
$R_i^{(L)} = \sum_j \frac{w_{ij}^2}{\sum_{i'} w_{i'j}^2} R_j^{(L+1)}$
$\sum_i R_i^{(L)} = \sum_j R_j^{(L+1)} = R$

Figure 2.7: Visual depiction of the flow of relevance for the w^2-rule in a simple case between fully-connected layers. The feed-forward mechanism of activations flows from left to right in the diagram, while the backward propagation of relevance, shown in red, flows from right to left.

The first method of LRP proposed by Bach et al. [69], is referred to as LRP$_z$, and decomposes relevance according to layer weights, regardless of their sign:

$$R_i^L = \sum_j \frac{z_{ij}}{\sum_{i'} z_{i'j}} R_j^{L+1}. \tag{2.14}$$

To provide numerical stability, we can add a small constant $\varepsilon \ll 1$ to the denominator to yield what is known as the LRP$_\varepsilon$ method:

$$R_i^L = \sum_j \frac{z_{ij}}{\sum_{i'} z_{i'j} + \varepsilon} R_j^{L+1}. \tag{2.15}$$

One method used successfully in many instances [73, 74, 71, 75, 16, 76] and which is one of the most commonly implemented methods of relevance propagation is the $\alpha\beta$-method [69], or LRP$_{\alpha\beta}$. Here we use the rule

$$R_i^L = \sum_j \left(\alpha \frac{(x_i w_{ij})^+}{\sum_{i'} (x_i w_{ij})^+ + b_j^+} - \beta \frac{(x_i w_{ij})^-}{\sum_{i'} (x_i w_{ij})^- + b_j^-} \right) R_j^{L+1} \tag{2.16}$$

with the condition that $\alpha - \beta = 1$ so that the relevance propagation is conservative. In some cases, the bias terms in the denominator are neglected [77]. Commonly, this method is implemented with $\alpha \in \{1, 2, 3\}$. It is easy to see that with $\alpha = 1$, $\beta = 0$, 2.16 reduces to the z^+-rule 2.12 if we assume only positive activations and do not include bias terms in the denominator. A special case of the $\alpha\beta$-rule is a DTD rule, but with cases of $\beta \geq 1$, 2.16 does not always satisfy positivity, and can produce heatmaps with both positive and negative values. A summary hierarchy of our gradient-based methods of interest is given below in Figure 2.8.

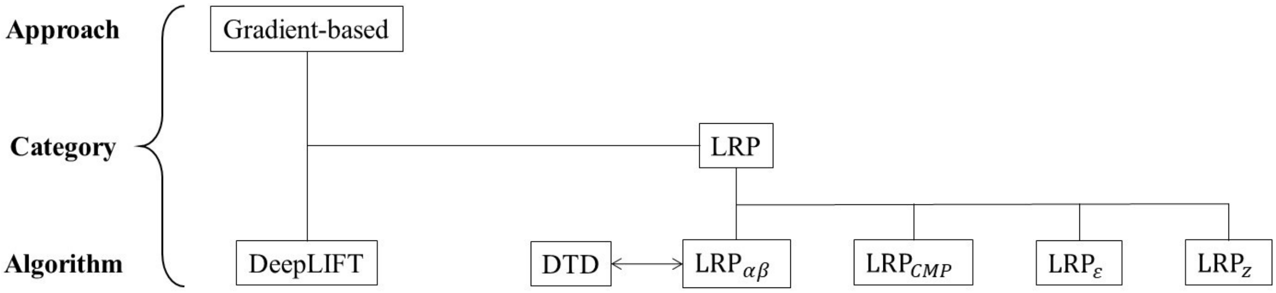


Figure 2.8: Hierarchy of gradient-based decomposition methods on which we will focus

The $\alpha\beta$-rule is shown to be robust against gradient-shattering [78], unlike LRP$_z$ and LRP$_\varepsilon$, and tends to reveal more visually appealing heatmaps.

These propagation rules work for layers L and $L+1$ with activities

$$x_j^{L+1} = g\left(\sum_i w_{ij} x_i^L + b \right) \tag{2.17}$$

where g is some nonlinearity. To tackle the problem of general renormalisation layers, Binder et al. [79] suggest the use of a first-order Taylor approximation akin to a Deep Taylor decomposition; this method proved markedly better on the CIFAR-10 dataset than an identity relevance assignment, with sequential replacement of input pixels with random noise.

To tackle the specific problem of relevance decomposition across BatchNorm layers, we can naively use an identity relevance assignment. As pointed out by Hui et al. however [80], many classifier nets have peculiarities in architecture and/or learned parameters which throw the effectiveness not only of the identity assignment, but of any relevance decomposition technique – however, it seems that in any case there is a preferable method, and so we can use the relevance decomposition

$$R_i^L = \frac{|x_i w_i|}{|x_i w_i| + |b_i| + \varepsilon} R_i^{L+1} \tag{2.18}$$

where ε is some small constant for numerical stability. The use of 2.18 defaults to the best possible assignment for a decomposition depending on the parameters at that point. This method proved effective when used on the publicly available classifier networks ResNet 50[3], MobileNet-V2[4], InceptionResNet-V2[5], and DenseNet-121[6]. The authors also note that peculiar to the ResNet architectures explored was an overlaying dot-pattern artifact, regardless of the LRP method used in its examination. This turned out to be due to the $2:1$ down-sampling in convolutional layers with 2×2 strides, and was most prominently presented by the residual components which, when down-sampling, skips entirely over some higher-level activations with its 1×1 kernel. This leads in the final analysis to a grid pattern of pixel activations whose relevance is inherently

[3] https://pytorch.org/docs/stable/torchvision/models.html#id3

[4] https://github.com/tonylins/pytorch-mobilenet-v2

[5] https://github.com/Cadene/pretrained-models.pytorch/blob/master/pretrainedmodels/models/inceptionResNetv2.py

[6] https://pytorch.org/docs/stable/torchvision/models.html#id5

higher than those of their immediately neighbouring pixels. For this reason it may be preferable to downsample using pooling functions in ResNets intended to be used for relevance propagation instead of stride lengths greater than 1.

Kohlbrenner et al. [81] provide a measured explanation of best practice for use of LRP rules. The authors note the superior visual representation of the LRP-$\alpha\beta$ method (and hence also the z^+-method) over other methods of decomposition, but also the fact of the class-insensitivity of LRP-$\alpha\beta$. Using only the $\alpha\beta$-rules, the heatmaps from different classes are identical for a given input. This stems from the constraint of positive layer activations, which leads to a the same explanations for different classes for the same image – in effect explaining that class A is or is not present for exactly the reason that class B is or is not present. In practice, most practitioners then make use of multiple methods of decomposition to have the benefits of each and reduce the deleterious consequences of any one. This is known as *Composite* LRP, denoted LRP$_{CMP}$. This is implemented in a convolutional network by making use of either the LRP$_z$ or LRP$_\varepsilon$ methods in decomposing the final fully-connected layers; using the LRP$_{\alpha\beta}$ rule for all the lower layers apart from the input – including the convolutional layers; and finally using the $z^{\mathcal{B}}$-rule for the input layer.

Not only are heatmaps from the LRP$_{CMP}$ method visually appealing, but they are also class-sensitive. The quality of different heatmaps was quantified by Kohlbrenner et al. by measuring the inside-total relevance ratios from heatmap methods for varying sizes of boundary boxes over class objects in images. The results showed that the LRP$_{CMP}$ methods reliably outperformed any other standard LRP rules on the PVOC and ImageNet datasets. The attribution methods provided by LRP$_{CMP}$ are shown both qualitatively and quantitatively to outperform other LRP and related methods.

As pointed out by Sixt et al. [82], there are potential problems with modified backpropagation explanations like LRP and Deep Taylor Decompositions. There are three main issues with such modified backpropagation explanations which call into question their 'faithfulness'. The authors do not define what is meant by faithfulness of explanations, but rather leave it as a roughly well-intuited notion. The first issue is that of class insensitivity, as discussed by Kohlbrenner et al. [81] – the explanations of different classes for the same decision are identical, scaled by the ratio of the two class outputs. This means that normalised heatmaps of one input are exactly the same for different classes. This is not a problem that faces regression networks, since there is essentially one 'class' at the output layer in this case, but the issue is still worth questioning. This same issue was noted by many, including Montavon et al. [78], who proposed a method which attempted to alleviate class insensitivity using differences between unnormalised heatmaps for different classes; but, as argued by Sixt et al., this did not fix the underlying problem, linked closely to the other two issues. The second issue is the failures of the 'sanity checks' proposed by Adebayo et al. [83]. The premise of the sanity checks is that randomising the parameters of a network layer in the process of relevance decomposition should result in drastically different saliency maps. Both LRP and the Deep Taylor techniques fail this test according to Sixt et al. Images were tested as such and measured against one-another for similarity using the SSIM metric [84], and the modified backpropagation explanations showed minimal difference compared to other methods. The third issue resulted from a test similar to the previous sanity checks, but now randomising the relevance scores at a given layer, then propagating the relevance back as before from that layer downwards. Using the cosine similarity convergence (CSC) measure (which normalises the dot product of two vectors, such that more similar vectors measure close to 1 and less similar vectors measure closer to 0), the authors compared layers' relevance scores between one true redistribution and distributions from a randomised final layer relevance. The modified backpropagation algorithms – which again included LRP and Deep Taylor – showed very little change in layer relevances, saturating quickly to a CSC measure of 1.

The concern raised by the authors it that the modified backpropagation algorithms attempt simply to recreate the input, as opposed to highlighting areas of saliency. Another point that the authors raised, pointed out initially by Geirhos et al. [85], is that CNNs trained on the ImageNet database focus on textures as opposed to shapes; but the modified backpropagation algorithms seem to focus on reconstructing the shapes in an image. Although in large part this seems to be the case, it cannot be argued that the only feature of these attribution methods is to reconstruct the input. It is well documented [71, 16, 74, 14] that saliency maps of these types are effective at least in many applications in localising relevance.

The underlying issue facing all modified backpropagation algorithms in this sense is that the necessity of positivity amounts to a sequence of non-negative matrices $\{A_i\}$. Sixt et al. prove that this converges to a rank-1 matrix $\bar{C}$ with the property that

$$\begin{aligned}\bar{C} &\equiv \prod_i A_i \\ &= \bar{\mathbf{v}}\gamma^T.\end{aligned}$$

Then for any input vector $\mathbf{v}$:

$$\begin{aligned}\bar{C}\mathbf{v} &= \bar{\mathbf{c}}\gamma^T\mathbf{v} \\ &= \lambda\bar{\mathbf{c}}, \quad \lambda \in \mathbb{R}.\end{aligned}$$

This means that for an appropriate sequence $\{A_i\}$ of matrices, the heatmap vectors will always converge to the same direction. This explains the class insensitivity problem, and why up to a certain point, the randomisation of parameters or relevance scores does little to the heatmaps of inputs for the modified backpropagation algorithms. LRP$_z$ is a gradient-based algorithm, and not a modified backpropagation algorithm, and so does not suffer this feature. LRP$_{CMP}$ alleviates the issue of class insensitivity [81] due to its inclusion of the LRP$_z$ method at fully-connected layers, and performed significantly better in the tests of [82] than other modified backpropagation methods. LRP$_z$ performs the best of all the LRP methods under the tests, but is subject to gradient-shattering and does not produce the most visually appealing heatmaps. It is concluded that the problem of convergence is slightly lessened but not completely disposed of by LRP$_{CMP}$.

For cases such as LRP$_{CMP}$ and LRP$_{\alpha\beta}$ with large α, it is unclear that the more limited similarities in heatmaps through the sanity checks are failures of faithfulness. It may be that the residual similarities of heatmaps constitute a robustness to perturbations overall as opposed to a total failure of explanation faithfulness.[7]

2.5.2 DeepLIFT

Shrikumar et al. [86] provided a novel method for producing saliency maps which remedied potential drawbacks of methods such as Gradient×Input [87] and Integrated Gradients [88, 89], which they called Deep Learning Important FeaTures (DeepLIFT). This method allowed for both positive and negative relevance scores to be propagated back through a network in such a way as to prevent the zeroing-out of relevance attributed to negative neuron activations and the entries to nonlinearities.

The DeepLIFT methodology consists of the comparison of gradients and activations from a forward pass (the same values as would be used in LRP) to reference values. The reference values can be all zeros, but usually come from passing a reference image through the network, and using the activations from this as reference. These reference activations serve as a default measure according to the given problem. A good reference input is one that is close to the original but has minimal activation of the target classes. In many cases though, it is simpler and perhaps most practical to use an image of all zeros in the forward pass as reference (this of course does not mean that all the reference activations will be zero), as the authors used for the MNIST dataset.

It will be interesting and of great insight to examine in our application what will constitute a good reference input. It may be that for the regression model, a middling age will be a ‘happy medium’ against which to compare other inputs – in which case we can use an accurately predicted middle-aged volume – or we might prefer to use a blurred image, or even all zeros (the background MRI activation) as reference.

DeepLIFT assigns contribution scores from higher layers to layers before. If we let t be some target output neuron, and $x_0, x_1, \ldots, x_{n-1}$ some set of neurons in an intermediate layer which are necessary and sufficient to compute t; and t^0 be the reference value of the output neuron such that $\Delta t = t - t^0$, and similarly $\Delta x_i = x_i - x_i^0$ for each of the aforementioned x_i, then in the DeepLIFT paradigm the contributions from each of the x_i to t, $C_{\Delta x_i \Delta t}$, must satisfy:

[7] We shall provide later in this section our own definition of faithfulness of explanations

$$\sum_{i=1}^{n} C_{\Delta x_i \Delta t} = \Delta t. \tag{2.19}$$

This is called the *Summation-To-Delta property*, and is similar in sentiment to the Deep Taylor property of conservativeness, Eq. 2.9, and the general LRP condition, Eq. 2.6. $C_{\Delta x_i \Delta t}$ can be thought of as the amount of relevance attributed from Δt to Δx_i.

For a given input neuron x with difference from reference Δx and target neuron t with difference from reference Δt, the authors define the multiplier $m_{\Delta x \Delta t}$ as:

$$m_{\Delta x \Delta t} = \frac{C_{\Delta x \Delta t}}{\Delta x} \tag{2.20}$$

such that as $\Delta x \to 0$, we have necessarily that $\Delta t \to 0$ and so $m_{\Delta x \Delta t}$ acts very similarly to the partial derivative $\dfrac{\partial t}{\partial x}$. Of course, we always have finite differences though.

If we assume some input layer $x_0, x_1, \ldots, x_{n-1}$, some hidden layer $y_0, y_1, \ldots, y_{n-1}$ and some target output neuron t, then DeepLIFT enforces the *chain rule for multipliers*:

$$m_{\Delta x_i \Delta t} = \sum_j m_{\Delta x_i \Delta y_j} m_{\Delta y_j \Delta t}. \tag{2.21}$$

This is consistent with Eq. 2.19, the Summation-To-Delta property. This is of course identical to the chain rule in differential calculus. Given the multiplier from every neuron to its immediate successors, one can efficiently work out the contribution of any one neuron to any later neuron in the network using the multipliers and the difference from reference value. Thusly we are able to make heatmaps.

The authors introduced positive and negative components of difference-from-reference values in order to allow the separate attribution of relevance contributions. If y is a neuron in the network that is not in the output layer, then we define Δy^+ and Δy^- as the positive and negative components respectively of Δy such that:

$$\Delta y^+ + \Delta y^- = \Delta y \tag{2.22}$$
$$C_{\Delta y \Delta t} = C_{\Delta y^+ \Delta t} + C_{\Delta y^- \Delta t} \tag{2.23}$$

The separation of positive and negative terms is used only in the context of the Reveal-Cancel rule. In this case we may find that $m_{\Delta y^+ \Delta t}$ and $m_{\Delta y^- \Delta t}$ differ.

There are three rules used for assigning relevance scores using DeepLIFT. The first is the Linear Rule, which is to be applied to dense layers and convolutional layers. If y is a linear function of its inputs x_i, $y = \sum_i w_i x_i + b$ then we define the positive and negative parts of Δy as:

$$\begin{aligned}\Delta y^+ &= \sum_i 1\left\{w_i \Delta x_i > 0\right\} w_i \Delta x_i \\ &= \sum_i 1\left\{w_i \Delta x_i > 0\right\} w_i (\Delta x_i^+ + \Delta x_i^-),\end{aligned}$$

$$\begin{aligned}\Delta y^- &= \sum_i 1\left\{w_i \Delta x_i < 0\right\} w_i \Delta x_i \\ &= \sum_i 1\left\{w_i \Delta x_i < 0\right\} w_i (\Delta x_i^+ + \Delta x_i^-).\end{aligned}$$

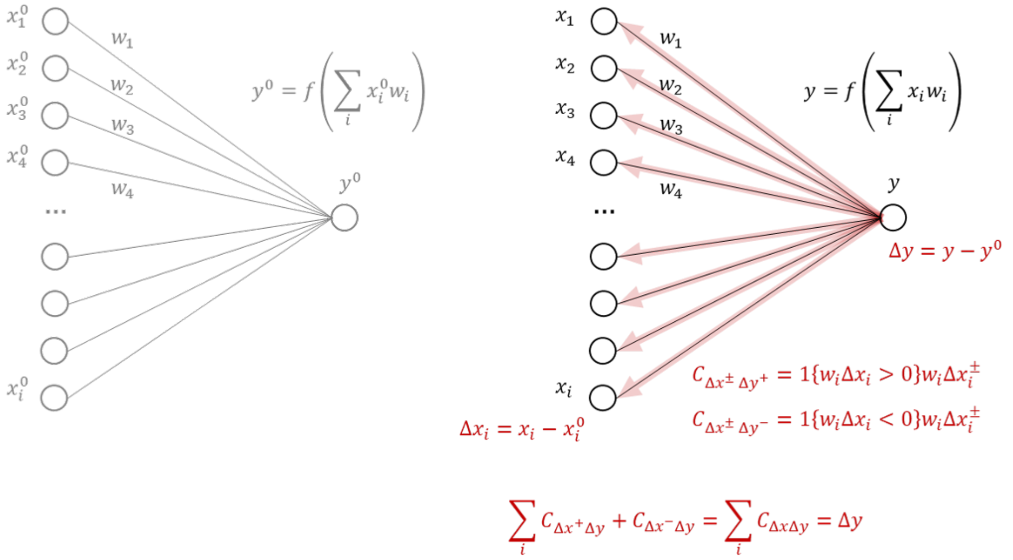
Thus for the contribution score, we choose:

$$
\begin{aligned}
C_{\Delta x_i^+ \Delta y^+} &= 1\left\{w_i \Delta x_i > 0\right\} w_i \Delta x_i^+, \\
C_{\Delta x_i^- \Delta y^+} &= 1\left\{w_i \Delta x_i > 0\right\} w_i \Delta x_i^-, \\
C_{\Delta x_i^+ \Delta y^-} &= 1\left\{w_i \Delta x_i < 0\right\} w_i \Delta x_i^+, \\
C_{\Delta x_i^- \Delta y^-} &= 1\left\{w_i \Delta x_i < 0\right\} w_i \Delta x_i^-.
\end{aligned}
$$

Then from Eq. 2.20, we get the multipliers:

$$
\begin{aligned}
m_{\Delta x_i^+ \Delta y^+} = m_{\Delta x_i^- \Delta y^+} &= 1\left\{w_i \Delta x_i > 0\right\} w_i, \\
m_{\Delta x_i^+ \Delta y^-} = m_{\Delta x_i^- \Delta y^-} &= 1\left\{w_i \Delta x_i < 0\right\} w_i.
\end{aligned}
$$

In the case that $\Delta x_i = 0$, it would be consistent with Eq. 2.19 to assign contributions of 0 to both positive and negative components, but it is possible that each of Δx_i^+ and Δx_i^- are non-zero. If this is the case then of course it is incorrect to assign their multipliers to 0. Instead, a compromise is struck, and convention is to set $m_{\Delta x_i^+ \Delta y^+} = m_{\Delta x_i^- \Delta y^+} = m_{\Delta x_i^+ \Delta y^-} = m_{\Delta x_i^- \Delta y^-} = 0.5 w_i$ when $\Delta x_i = 0$. The Linear Rule is illustrated in Figure 2.9. This also shows the necessity for a forward pass on the network using the reference input as well.



x_1^0
x_2^0
x_3^0
x_4^0
w_1
w_2
w_3
w_4
$y^0 = f\left(\sum_i x_i^0 w_i\right)$
y^0
...
x_i^0
x_1
x_2
x_3
x_4
w_1
w_2
w_3
w_4
$y = f\left(\sum_i x_i w_i\right)$
y
$\Delta y = y - y^0$
...
x_i
$\Delta x_i = x_i - x_i^0$
$C_{\Delta x^\pm \Delta y^+} = 1\{w_i \Delta x_i > 0\} w_i \Delta x_i^\pm$
$C_{\Delta x^\pm \Delta y^-} = 1\{w_i \Delta x_i < 0\} w_i \Delta x_i^\pm$
$\sum_i C_{\Delta x^+ \Delta y} + C_{\Delta x^- \Delta y} = \sum_i C_{\Delta x \Delta y} = \Delta y$

Figure 2.9: Visual depiction of the Linear Rule contribution score distribution for the simple case between fully-connected layers. Again the feed-forward flow of the network is from left to right in the diagram, while the contribution score redistribution, shown in red, flows from right to left. On the left is shown the forward propagation to the same neuron in the reference input case.

The second assignment method is the Rescale Rule, which is used on nonlinearity layers which take one input and have a single output (such as ReLU, tanh or sigmoid functions). To this end, let y be a nonlinear transformation of its input neuron x, $y = f(x)$. By the summation-to-delta property 2.19, we must have that $C_{\Delta x \Delta y} = \Delta y$. Thus of course we must also have $m_{\Delta x \Delta y} = \dfrac{\Delta y}{\Delta x}$. For this rule, we set the differences-from-reference Δy^+ and Δy^- proportional to Δx^+ and Δx^- respectively, according to:

$$
\begin{aligned}
\Delta y^+ &= \frac{\Delta y}{\Delta x} \Delta x^+ = C_{\Delta x^+ \Delta y^+}, \\
\Delta y^- &= \frac{\Delta y}{\Delta x} \Delta x^- = C_{\Delta x^- \Delta y^-}.
\end{aligned}
$$

The corresponding multipliers are then given by

$$m_{\Delta x^+ \Delta y^+} = m_{\Delta x^- \Delta y^-} = m_{\Delta x \Delta y} = \frac{\Delta y}{\Delta x}.$$

Now we see that as $x^0 \to x$, we will have $\Delta x \to 0$ and $\Delta y \to 0$ such that the multipliers again become the corresponding partial derivatives evaluated at $x = x^0$.

The third assignment method is called the Reveal-Cancel Rule, and is also applied to nonlinearities. This aims to alleviate potential issues of cancelling relevance attributions by positive and negative terms not being considered separately. Consider again a nonlinear neuron $y = f(x)$. This time, as opposed to setting Δy^+ and Δy^- proportional to Δx^+ and Δx^- respectively, we consider the impact of positive terms in the absence of negative terms, and the impact of negative terms in the absence of positive terms:

$$\begin{aligned}
\Delta y^+ =& \frac{1}{2}\left(f(x^0 + \Delta x^+) - f(x^0)\right) \\
&+ \frac{1}{2}\left(f(x^0 + \Delta x^- + \Delta x^+) - f(x^0 + \Delta x^-)\right) \\
\Delta y^- =& \frac{1}{2}\left(f(x^0 + \Delta x^-) - f(x^0)\right) \\
&+ \frac{1}{2}\left(f(x^0 + \Delta x^+ + \Delta x^-) - f(x^0 + \Delta x^+)\right).
\end{aligned}$$

And the multipliers are given by:

$$m_{\Delta x^+ \Delta y^+} = \frac{C_{\Delta x^+ \Delta y^+}}{\Delta x^+} = \frac{\Delta y^+}{\Delta x^+}, m_{\Delta x^- \Delta y^-} = \frac{\Delta y^-}{\Delta x^-}.$$

The authors note that in some cases, such as the use of a ReLU layer, it may be preferable to use the Rescale Rule for the nonlinearity, as relevance may be unduly assigned to noise terms by the Reveal-Cancel Rule, where it considers the positive and negative terms separately.

To test the proficiency of DeepLIFT at assigning relevance scores, the authors trained a CNN on the MNIST dataset to 99.2% test accuracy, and used reference inputs of all zeros, which is the background of the MNIST digits. After computing the heatmaps of given samples for the target class c_0 and a separate class c_t, the authors found the difference of the relevance scores for each pixel, $S_{x_i\text{diff}} = S_{x_i c_o} - S_{x_i c_t}$. Using this as a ranking mechanism for saliency for the one class over the other, the top 20% of pixels (in the original images) according to the ranking were then erased. The resulting images were passed through the network again to determine the new classifications according to c_0 and c_t. The difference of the log-odds scores were then computed. This test was performed for DeepLIFT, Absolute Gradients [90], Guided Backpropagation [91], Gradient×Input [87], and Integrated Gradients [88, 89]. Tests were run for DeepLIFT with the Rescale Rule as well as with the Reveal-Cancel Rule, and the best overall performer was DeepLIFT with Reveal-Cancel. The final nonlinearity of the network was a softmax output.

The authors also experimented on a genomic classification task. The task was set up using background ACGT sequences with the expected probabilities 0.3, 0.2, 0.2, and 0.3 respectively for each base molecule, and motif models based on well-known Position Weight Matrices (PWMs) for the GATA1 and TAL1 genes. The background sequences would be created and 0 to 3 instances of each of the motifs overlaid at non-overlapping positions. There were four possibilities for the outcome of the detection: in Case 1, both a GATA1 motif and a TAL1 motif are detected; in Case 2, only a GATA1 motif is detected; Case 3, only a TAL1 motif is detected; and Case 4, neither motif is detected. Each case can occur with probability $\frac{1}{4}$.

The authors trained a CNN with two hidden layers, a global average pooling layer, and a single fully-connected layer, which achieved > 98 auROC on all tasks (cases) of the synthesised test set. They then identified the top 5 matches in each sequences to a given motif according to the log-odds score, and plotted the log-odds score against the relevance assignment for DeepLIFT and other relevance attribution methods. The references used were background ACGT sequences.

The experiment was performed comparing the outputs from Guided Backpropagation, Gradient × Input [87], Integrated Gradients, DeepLIFT with only the Rescale Rule, DeepLIFT with only the Reveal-Cancel Rule, and DeepLIFT with the Rescale Rule on convolutional layers and the Reveal-Cancel Rule on fully-connected layers. Both DeepLIFT methods using the Reveal-Cancel Rule tended to assign greater relevance to TAL1 motifs if they were present with GATA1 motifs than if they were detected alone, and the same was true of GATA1 motifs in the presence of TAL1 motifs versus in their absence. With the exception only of Guided Backpropagation, all of the other tested attribution methods also delivered false negative values for the presence of both motifs at times. The authors believe that this may be due to the need to learn an *and*-like relation, which the authors previously argued cannot attribute relevance accurately with models like the Rescale Rule. Guided Backpropagation did not show the false negatives, but did have a tendency to show false positives.

Using the Reveal-Cancel Rule on all nonlinearities assigned undue positive and negative relevance to many irrelevant areas of the sequences. These were not as large as the assignments to the most relevant areas, but were still noticeable in comparison to the blended use of Reveal-Cancel and the Rescale Rule. This happens because the Reveal-Cancel Rule can assign relevance to neurons with activations of zero due to its splitting of positive and negative assignments. As discussed before, for nonlinearities such as ReLU units, the noise that is cancelled out by the nonlinearity is preferably ignored, which can be done with the use of the Rescale Rule.

The attribution method described by the Linear Rule – without considering positive and negative contributions separately – can be rewritten in terms of layer-to-layer relevance attribution [92] like LRP's *Epsilon-Rule*:

$$R_i^L = \sum_j \frac{z_{ij} - \bar{z}_{ij}}{\sum_{i'} \left(z_{i'j} - \bar{z}_{i'j}\right)} R_j^{L+1}. \tag{2.24}$$

Here, $z_{ij} = w_{ij}^{(L,L+1)} x_i^L$ and $\bar{z}_{ij} = w_{ij}^{(L,L+1)} \bar{z}_i^L$, with $w_{ij}^{(L,L+1)}$ the weight between the ith neuron in layer L and the jth neuron in layer $L+1$, x_i^L the activation of the ith neuron in layer L, and $\bar{x}_i^L$ the reference activation of the ith neuron in layer L. Eq. 2.24 comes from setting the total relevance R_i^L equal to the sum of all contributions of that neuron to the neurons in the layer above, $R_i^L = \sum_j C_{\Delta x_i \Delta y_j}$, and stating that the relevance associated with any neuron in layer $L+1$ must have total relevance equal to the sum of contributions from neurons in the previous layer: $R_j^{(L+1)} = \sum_i C_{\Delta x_i \Delta y_j}$.

Pianpanit et al. [93] examined different interpretation techniques on modified 3D CNN models which were used to diagnose Parkinson's Disease (PD) from 3D Single Photon Emission Tomography (SPECT) images. The metric used in determining evaluation performance was the Dice coefficient, which was used on the four models. The Dice coefficient D is a measure for comparison between an image's predicted segmentation P and a ground truth segmentation G. It is defined as

$$D = \frac{2|P \cap G|}{|P| + |G|}. \tag{2.25}$$

Clearly, D lies in the range $[0, 1]$ with $D = 1$ showing identical segmentation. The attribution methods examined were Gradients [90], Guided Backpropagation [91], Grad-CAM [94], DeepLIFT and SHAP [95]. The dataset used was from the Parkinson's Progression Markers Initiative (PPMI) database. The first architecture used in the assessment of the methods was the PD Net [96] architecture, which was developed for use on the PPMI dataset. The second architecture used was a modified version of PD Net, which was lengthened by adding more convolutional layers; this was referred to as Deep PD Net. The PD Net types were tested with and without the inclusion of BatchNorm layers, and were compared with an SVM. It was found that the Deep PD Net with the inclusion of BatchNorm layers provided the highest specificity of the tested models, and the second highest sensitivity, leading to its having the highest accuracy. Guided backprop showed the highest mean Dice Coefficient in its relevance propagation, with DeepLIFT showing middling performance on this benchmark, below Guided Grad-CAM [94] and Gradients too, but outperforming both Grad-CAM and SHAP.

Chatterjee et al. [97] used DeepLIFT and other saliency map techniques to analyse the predictions of networks designed to detect the presence of COVID-19 or pneumonia in chest X-ray images. The first of the two datasets that the authors used was the COVID-19 image collection [98]. This consists of 236 COVID-19 patients, 12 simultaneous COVID-19 and ARDS patients, 4 ARDS patients, 1 Chlamydophilia patient, 1 Klebsiella patient, 2 Legionella patients, 12 Pneumocystis patients, 16 SARS patients, 13 Streptococcus patients, and 5 healthy control patients. The second dataset was the Chest X-ray dataset [99], which contained 1583 healthy patients (of whom 500 were randomly chosen to be used), 1493 Viral Pneumonia patients (of whom 250 were randomly chosen to be used), and 2780 Bacterial pneumonia patients (of whom 250 were randomly chosen to be used).

Four different architecture types were explored in their use for the experiments. The first was ResNet [100], which overcomes the issues faced by many deep networks, of gradient explosion and degradation (whereby accuracy saturates quickly in shallow layers of a deep network). This is done by adding the output of the previous layer of a network to the following layer – utilising what is called a 'skip layer'. The authors opted to try two versions of this – ResNet18 and ResNet34. The second type of architecture was InceptionNet [46], which efficiently examines both local and global input features by using kernels of various sizes at the same level of the network, effectively making the network 'wider' as opposed to deeper. The authors opted to try InceptionV3 for analysis. The third type was InceptionResNetV2 [101], which utilises the skip layers of ResNet in the architecture of InceptionNet to increase efficiency. The fourth type was DenseNet [102], which connects layers directly to one-another by matching up feature map sizes throughout the network. The authors opted for the use of DenseNet161 in he experiments. It was found that DenseNet161 had the highest precision of the five proposed architectures, but the authors used all five to create an ensemble architecture which well outperformed any one of the five proposed architectures. All methods apart from occlusion failed to be run on DenseNet161 due to memory limitations on the GPU. DeepLIFT also encountered problems with the ResNet models, apparently due to the in-layer ReLU operations. The relevance methods all highlighted similar areas in the lungs for the respective pathologies, notably focusing on different areas for COVID-19 as for bacterial and viral pneumonia.

2.5.3 Other Saliency Mapping Methods

Having discussed LRP and DeepLIFT, as well as some of their past implementations, we look at some other saliency mapping methods briefly, and the comparisons that have been made to LRP and DeepLIFT.

Chang et al. [103] introduced Variational Dropout Saliency Maps (VDSM) as another saliency attribution technique using the comparison between input activations and reference activations. Their methods consist of a Smallest Deletion Region (SDR) objective, and a Smallest Supporting Region (SSR) objective, which are respectively the smallest region of the original input to mask, and the smallest region of the input that can be substituted into a reference image to maximise the class probability.

The VDSM method was compared to several other saliency methods which make use of reference values, including DeepLIFT. The evaluation was broken down into two parts, both examining how the log-odds scores were affected on CNN trained on MNIST. The first procedure considers the removal of pixels from the image according to each specific saliency method from most to least salient pixels, whereas the second method removed pixels from least to most salient. The saliency methods were evaluated according to the greatest log-odds change compared to the initial classification. The VDSM method performed best on both tasks, with the SDR objective outperforming all others in the first experimental procedure, and the SSR objective outperforming all others in the second procedure. DeepLIFT seemed to display middling performance in both tasks. An important note is that the tasks were performed over all areas of the image for which the saliency assignments reported relevance.

The experiments were performed both for the choice of background reference inputs (the same as used in [86]), and reference inputs generated by a VAE. The use of the VAE is a novel approach proposed by the authors, and helped alleviate the affect of network artifacts creating unrealistic saliency maps, at least in the VDSM method performing the SSR objective. In the experiments, the VDSM methods tended to highlight more of the digit in question than methods like DeepLIFT and PDA [104]. This may be due to the fact that the VDSM methods were designed specifically for these tasks, whereas the application of gradient-based methods such as DeepLIFT are far more general. Whereas DeepLIFT asks the broad question 'Why did the network make this decision for this example?' VDSM asks one of two questions: 'Which parts of this are the most characteristic of the class of concern?' or 'Which parts of this should change to be more characteristic of another class?' Generally one of these questions is what is being asked in any occlusion-based saliency task. This may be a more desirable characteristic in specific cases, but we do not know whether or not it will be applicable to many saliency tasks. On the other hand, it is useful to have a saliency method tailored to the specific needs of a task, such that the outcome of the task is specific in its explanation and our interpretation of it can be as concrete as possible – this is more difficult with more general saliency methods such as DeepLIFT and LRP.

We note that while VDSM outperformed DeepLIFT on this task, several features limit its generalisability. The metrics used to determine the saliency mapping performances were foundationally identical to the VDSM protocols themselves, and so these performance comparisons may not be fair. The method for use of the VDSM method in regression task settings is not entirely clear, as the masking of regions within the input can yield input samples that are far outside the model's domain of inference. The VDSM method is also very computationally expensive, requiring many inference calls per saliency map to determine SDRs and SSRs.

Sundararajan et al. [89] propose the Integrated Gradients method of relevance attribution, arguing its formulation from the standpoint of two axioms that ought to be satisfied by an attribution framework. The method, like DeepLIFT, uses a reference comparison for each input; and as the name suggests, is a gradient-based method. The first of these is called *Sensitivity(a)*, which is satisfied if for every input and reference pair that differ in one feature and have different predictions, have a non-zero attribution of relevance to that feature. The second axiom is that of *Implementation Invariance*, which is satisfied if attributions are always identical for the same input into functionally equivalent networks. Networks are functionally equivalent if for every input, they produce exactly the same output. The authors show that due to the use of discrete gradients in LRP and DeepLIFT, these methods and ones like them do not satisfy Implementation Invariance. This is generally to do with how DeepLIFT and LRP treat nonlinearities like ReLU, and the fact that generally it is not true that $\frac{f(x_1 - x_0)}{g(x_1 - x_0)} = \frac{f(x_1 - x_0)}{h(x_1 - x_0)} \cdot \frac{h(x_1 - x_0)}{g(x_1 - x_0)}$. This makes sense because these methods explicitly ask 'Why does this network come to this decision?' as opposed to 'Why does the function represented by this network come to this decision?' The use of discrete gradients separates these into two different questions, whereas to a method using partial derivatives, they are the exact same question.

The Integrated Gradients technique combines the Sensitivity(a) property satisfied by LRP, DeepLIFT and the like (but not by gradients), with the Implementation Invariance of gradients. This is done by performing the path integral along the straight line in $\mathbb{R}^n$ from the reference x' to the input x:

$$\text{IntegratedGradients}_i(x) = (x_i - x_i') \int_0^1 \left(\frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} \right) d\alpha \tag{2.26}$$

where $\frac{\partial F(x)}{\partial x_i}$ is the gradient of the function F (representing the network) along the ith dimension. In networks composed of ReLUs, Sigmoids and pooling operations, 2.26 is Conservative (the authors refer to it as Completeness) in the sense of Eq. 2.9:

$$\sum_i \text{IntegratedGradients}_i(x) = F(x) - F(x').$$

Of course, the integral itself must be approximated, and for greater accuracy, the method sacrifices computational load. Generally an approximation is performed using the Riemann Sum technique. Finding a reasonable trade-off between computational intensity and accuracy (and therefore the satisfaction of the axioms) is the biggest problem faced by Integrated Gradients. It is not clear from the experiments performed by the authors whether Integrated Gradients perform better than DeepLIFT or LRP for any given task, but these methods seem to be more highly used in practice than Integrated Gradients. This is likely due to the necessary trade-off that was earlier discussed.

Lundberg and Lee [95] proposed the SHapley Additive exPlanation (SHAP) Values as a measure of feature importance. The Shapley Values are classically defined as the importance of a feature to the output of a model as quantified in a very computationally intensive manner in which the effect of any one feature is quantified relative to the rest of a feature subset. In a feature subset $S \subseteq F$ of the set of all features, importance of the inclusion of a feature to input is quantified by measuring the output of the model being trained with a feature i present relative to being trained without it. One instance of the model $f_{S\cup\{i\}}$ trained in the presence of i, the other instance f_S in its absence, the difference of the predictions $f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)$ forms a vote toward the importance of the feature. This is computed for all possible subsets $S \subseteq F - \{i\}$. The Shapley Values are then computed as a weighted sum of these differences:

$$\phi_i = \sum_{S \subseteq F - \{i\}} \left[\frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left(f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right) \right]. \tag{2.27}$$

Needless to say, the computational overhead involved in computing Shapley Values is immense, so several methods of approximating them have been put forward. SHAP estimates the Shapley Values with a conditional expectation function of the original model. The SHAP method, as explained by the authors, shares some similarity with LRP and DeepLIFT. The revelation of this fact was the inspiration of the Reveal-Cancel Rule [86], which is proposed as a good approximation of the Shapley Values. Lundberg and Lee also proposed a method of combining the SHAP method and DeepLIFT, to form what they called Deep SHAP. In this regard, they defined the DeepLIFT multipliers in terms of the SHAP values.

Ancona et al. [92] evaluated and compared four gradient-based attribution methods on a theoretical basis – ε-LRP (that is, LRP$_\varepsilon$, (2.15)), DeepLIFT, Gradient×Input [87] and Integrated Gradients [89]. The authors also presented a new metric, *Sensitivity-n*, which they claim generalises the notions of Completeness [89] and Summation-to-Delta [86], and used it as an evaluation tool in comparing the four gradient-based methods, as well as the occlusion-based method proposed by Zeiler and Fergus [105]. The occlusion-based method, known as *Occlusion-1*, is used by the authors as a benchmark for perturbation-based attribution methods. The authors show that both DeepLIFT and LRP$_\varepsilon$ can be reformulated to be applied by use of the chain rule for derivatives, as long as the derivative evaluated on a nonlinearity is replaced with some function which is dependent on the method. In this manner, all four of the gradient-based methods can be computed from the partial derivatives of the network layers. The formulation of the *Sensitivity-n* metric is based on the notion that several attribution methods actually explain slightly different things. For example, the authors argue that Occlusion-1 better identifies the importance of single features, while Integrated Gradients better quantifies the importance of several features which are present simultaneously. The authors define the metric as follows:

Definition 1. Sensitivity-n An attribution satisfies *Sensitivity-n* if the sum of the attributions of any subset of features with cardinality n is equal to the change in the output S_c caused by removing the features in the subset. In other words, if for all subsets $x_S = [x_1, \ldots, x_n] \subseteq x$ of features, it holds that $\sum_{i=1}^{n} R_i^c(x) = S_c(x) - S_c(x_{[x_S=0]})$.

Here, $S_c(x)$ is the target neuron output for input x, and $x_{[x_S=0]}$ is the input x altered such that all the subset of components x_S is set to zero (which is often the background activation). In the case that $n = N$, the total number of input neurons, the condition becomes $\sum_{i=1}^{n} R_i^c(x) = S_c(x) - S_c(\bar{x})$, where $\bar{x}$ is all zeros (no features). This special case is equivalent to the *Completeness* of [89] and the *Summation-to-Delta* of [86], and it is in this regard that *Sensitivity-n* is a generalisation of the two. Under certain outlined conditions, the authors note that LRP$_\varepsilon$ also satisfies *Sensitivity-N*. By construction, the *Occlusion-1* method satisfies *Sensitivity*-1. The authors show, however, that no one of the five methods examined satisfies *Sensitivity-n* for all n, unless applied to a linear model (or a model which behaves linearly on the given data), in which case all the methods are functionally equivalent. We can use different values of n with *Sensitivity-n* to measure the correlation between the sum of attributions $\sum_i R_i^c(x)$ and the target output variation $S_c(x) - S_c(x_{[x_S=0]})$ for different attribution methods. To this end, the authors estimated the correlations for given values of n for 100 randomly sampled subsets of features from a single input x. The authors calculated the Pearson Correlation Coefficients (PCCs) between the relevance sums and the target output variations for 1000 samples from each dataset to obtain an average at each value of n, which was ranged from 1 to $\sim 0.8N$ for each dataset. The datasets used were MNIST, CIFAR10, ImageNet, and IMDB. MNIST was used to test two architectures with four different activation functions. Using a DNN for sentiment analysis, the authors showed the fact that the linear behaviour of the network led to the equivalence of the PCCs across the board for all the tested attribution methods.

The authors concluded that *Occlusion-1* better identified the few most important features, while the gradient-based methods better measured global, nonlinear effects on relevance. They also note the much slower implementation of *Occlusion-1*. Integrated Gradients and DeepLIFT were found to be highly correlated, which suggests that DeepLIFT (being easier, faster and less intensive to implement) is a more economical way to approximate Integrated Gradients. LRP$_\varepsilon$ proved to be equivalent to Gradient×Input when subject to ReLU nonlinearities, but failed under other nonlinearities, having $f(0) \neq 0$.

2.5.4 Evaluating Visual Explanations

It is imperative that we have a measure of quality for visual explanations, such that their utility does not remain vague and unquantifiable. In order to do this we must first determine what makes a good explanation for humans (since it is humans to whom we wish to explain).

Thus far we have spoken about LRP and DeepLIFT in their applications of assigning relevance to input pixels, implicitly identifying 2D images as input. It is definitely not the case that pixels are the only space onto which pixel-wise decomposition methods can map. Indeed, Shrikumar et al. used strings of DNA sequences as their input for a test example in the original DeepLIFT paper to great success. The rules we have explored for our decomposition methods are all applicable to input of any dimensionality. In particular, for brain MRI volumes, these methods all have 3D analogues to the 2D applications, using the same rules. These will in turn create 3-dimensional saliency maps.

It is important to have a concrete notion of each of explainability and interpretability in application to ML. Mittelstadt et al. [106] compare and contrast the concept of an explanation from the perspectives of Philosophy, Sociology and Neuroscience against current explanation techniques used in ML models. They make the following definitions:

Definition 2. Interpretability is the degree of human comprehensibility of a given black-box model or decision.

Definition 3. Explanation refers to the ways of conveying information about a phenomenon.

With specific regard to ML models, explanations can be used in a variety of different ways, most notably for the tasks of improving the model, allowing users and researchers to learn from the model, and to create trust for the model by stakeholders. The authors explain that full scientific explanations are not necessarily the aim of XAI, but rather its aim is to provide human-understandable reasoning for causal relationships between the variables in a model. It is likely that for models applied to complex decision-making tasks, a full scientific explanation would not be understandable to humans.

Two of the broad aims of XAI are to create models that are transparent to users, and to create post-hoc explanations. The transparency of a model refers to the understanding of its inner workings, while the post-hoc explanations describe the model behaviour. Transparency can be specified by an understanding of the model's function ('simulatability'), individual component analysis ('decomposability'), or the algorithm ('algorithmic transparency'). Post-hoc explanations on the other hand attempt to explain how a model behaves and why by analysing the trained parameters and outputs. The authors note that despite these available explanation techniques, many researchers have opted to use methods of retro-fitting local or approximate models over more complex ones. This is an attempt to create more interpretable models that approximate the decision making abilities of the more complex models. They explain how there is necessarily a trade-off between the performance of an approximated model, the simplicity of the approximation, and the size of the domain in which it remains valid.

The authors give an analogy for the bulk of current XAI methods as being akin to scientific models, in that many current methods lie in the building of approximate models which are highly interpretable. The main issue with such models is that they can only describe the system in question over a restricted domain, and so can be misleading if provided as an explanation, especially for lay users. They state though, that it is useful to keep in mind Box's Maxim, "All models are wrong, but some are useful" [107]. It is important to note that on the XAI side of this analogy though, what we are simplifying is itself a model – since we use NNs as universal function approximators – and so we are simplifying a model which itself simplifies a true function; this becomes then an approximation of an approximation. The authors also note that in its current state, XAI research efforts generally do not expend significant effort in characterising the domains of the simplified models. This provides little insight into the decision-making processes of the original model, and what insight it does provide can be misleading.

The authors make the case for the use of contrastive explanations in XAI. These are explanations which include or are based on counterfactual data. For example, a model could be devised to tell a user that some output x_i was not maximised and some other output x_j was maximised by an input data-point X_1 with reference to another data-point X_2 which evaluates better on x_i and not as highly on x_j[8]. The authors note that work, primarily by Miller et al. [68] showed that human explanations are characterised by three features:

1. Human explanations are contrastive, in that humans offer and ask for explanations for events relative to other possible events having happened. For example, "Why did event X occur as a result of this action, and not event Y?"
2. Human explanations are selective, in that a person offering an explanation selects only the information about an event (about which there are innumerable explanations) that he believes is relevant. It is infeasible to provide all the information about an event every time an explanation is needed.
3. Human explanations are social, in that an one tends to provide an explanation in terms that the recipient will understand. In relation to the last point, not only does the recipient have a preferred mode of information reception to maximise understanding, but he also cannot expect to have a full scientific explanation for every event for which he needs explanation. Not only would it be temporally infeasible; he could not hope to understand all the information for any of the events.

These features lead to the important notion of discourse in explanation. A person explaining does not necessarily select the optimal information to convey understanding to the person receiving the explanation. The recipient then asks further questions, since he does not fully understand, and the resulting discourse theoretically completes an explanation, unless there is some information lacking. The authors note that since the people offering explanations may have different goals to the people receiving them, not only does discussion ensue, but there may be risk of malicious explanation in an attempt to garner trust from the recipient. It is for this reason that the authors urge caution in the use of post-hoc explanations. They point out that not only is it important to have information exchange in this process of explanation, but also argumentation on the justafiability of the explanation. Thus the explanations for ML algorithms should ideally provide room for debate over their justifications.

In relation to this concern, Dombrowski et al. [108] found that not only can the explanations of image detection networks (provided by methods including but not limited to DeepLIFT and LRP methods such as Deep Taylor) be sensitive to small perturbations, as pointed out initially by Ghorbani et al. [109], but they can be manipulated easily with small perturbations such that all final class probabilities are roughly the same as for the original image, leaving the image identical to the human eye, and having *any* explanation which the creator likes. The authors show that this is due to the fact that on the explanation manifold of the network, similar images with similar network outputs can have vastly distant explanations, since the explanation distance is determined by geodesics on its manifold. These geodesic distances are bounded from above proportionally to the principal curvatures of the manifold. Furthermore, the principal curvatures are bounded by the parameter β of the Softplus function

$$\mathrm{Softplus}_\beta(x) = \frac{1}{\beta}\ln\left(1 + e^{\beta x}\right).$$

The ReLU function

$$\mathrm{ReLU}(x) = \max(0, x)$$

is the limit of the Softplus function as $\beta \to \infty$. Therefore with the use of ReLU nonlinearities in a network, the curvature of the explanation manifold is unbounded, and the network explanations are generally much more susceptible to manipulation. The authors' experiments show clearly the drastic difference made by the use of Softplus nonlinearities with low values of β over the use of high values, or ReLU nonlinearities. They note that β is a hyper-parameter of the network, but that generally a value close to 1 works well. They refer to the provision of robustness by low-β Softplus functions as β-smoothing. A basic visual explanation of β-smoothing is shown in Figure 2.10. They found that the SmoothGrad method of explanation roughly mimics

[8] Note the conceptual similarity to the difference-from-reference in DeepLIFT.

the buffering effects of β-smoothing, but not quite to the same degree and at much greater computational cost.

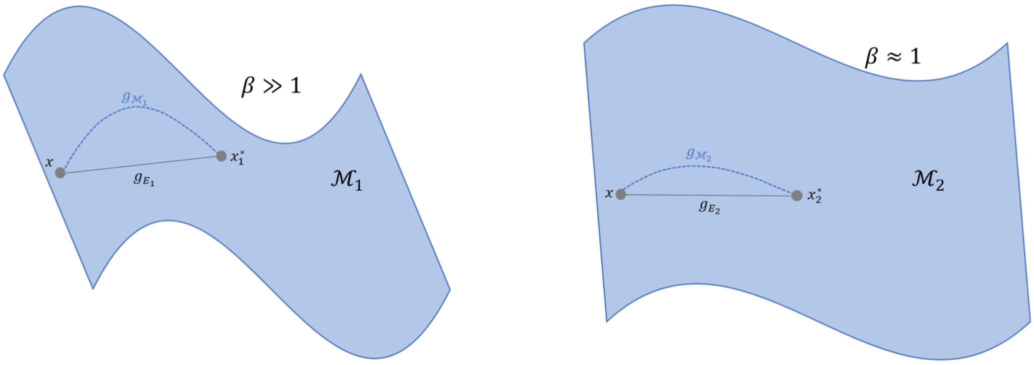


Figure 2.10: β-smoothing decreases the difference between decision manifold geodesic distance $g_{\mathcal{M}}$ and the Euclidean distance g_E between an input x and its altered counterpart x^*

The issues of the inability for discourse in visual explanations, and unwanted explanations from adversarially generated input beg the question of what exactly we want in a 'faithful' explanation. Many authors refer to explanation faithfulness without a definition, but at some point use a metric of comparison between two methods that they intuit to represent this faithfulness [82, 83, 110]. The definition we follow closely resembles that of Ribeiro et al. [111].

Definition 4. Explanation faithfulness is the similarity of explanations for similar inputs to a model.

Definition 4 looks similar in nature to function continuity. This is different from the implicit meaning of faithfulness measured, for example, by Sixt et al. Their notion of faithfulness stems solely from the sanity checks of Adebayo et al., and it is not entirely clear what the corresponding measurements mean, as discussed earlier.

The issue of how to validate the explanations we produce in our work is of great importance. Bach et al. [69] proposed the pixel-flipping method for evaluating the heatmaps produced via LRP. This method is carried out by organising the pixels of the heatmap by activation value, and setting a certain proportion of the pixels in the original image to the average image activation if they are in that proportion of the highest-activated pixels in the heatmap. Subsequently the change in prediction of the image is evaluated. This was done with specific application to the MNIST dataset, which consists of 2D images; but the procedure would be analogous with a 3D volume. This method has its merits in respect to use for the MNIST dataset, but for the case of BA regression, it is not clear at all that it would be useful, or even how a useful analogous validation would work. The pixel-flipping method in application to 3D MRI would hypothetically set activations in key brain ageing areas to the average activation value. In the case of a young individual for example, that would likely remove great portions of the cortex, if not all of it, and enlarge the ventricles. This would in the best case lead to an increase in the predicted age of the edited volume (which would be useful), and in the worst case produce an edited volume which the network does not understand, and therefore cannot regress accurately. A worse situation befalls an older subject, since the ageing pathology is characterised by loss of grey matter and dilation of the ventricles, and therefore by decreased MRI volume activations. The setting of highlighted areas from the heatmaps to zero in the original volumes is sure to create an image well outside the scope of the model. The fundamental issue with this method of evaluation is the production of an edited volume which the model does not understand. This is a particularly destructive issue for a regression network, since a classification network would show a lack of confidence in categorisations if it did not understand its input (in an ideal situation in which it is not being fooled) by low probability scores. A regression model can only output a single score, which is its decision and not an indication of its certainty in that decision.

Samek et al. [73] proposed an evaluation framework for pixel-wise explanations which generalises the pixel-flipping method of [69] by replacing $m \times m$ regions around the most salient pixels of an input image for chosen values of m, and not with the mean value of the image activations, but with uniformly distributed (random) values. Again, this suffers from the conceptual issue that what is being manipulated is a class probability, and that this is not an available metric for a regression network.

Dabkowski and Gal [112] proposed a more general metric of image saliency, which was put to use by Chang et al. [103] in their proposition of the VDSM method. They provide two working definitions for saliency:

Definition 5. Smallest Sufficient Region (SSR) is the smallest region of the image which alone allows for confident classification of the image.

Definition 6. Smallest Destroying Region (SDR) is the smallest region of the image which when removed prevents a confident classification.

Chang et al. referred to the SSR as the 'Smallest Supporting Region' and to the SDR as the 'Smallest Deletion Region', but of course were referring to the same respective concepts as those defined above. The issue again comes up of measuring confident classification. The authors present a metric for saliency maps based off of the SSR objective. They note that saliency is different from localisation, although localisation is an important requirement of saliency. Their metric therefore assumes that the most salient parts of an image should not only lead to confident classification on their own, but should have as small an area as possible. As opposed to masking, which may produce adversarial artifacts, the authors choose to crop the salient regions to the tightest bounding box enclosing the entire salient region. The saliency metric is then simply

$$s(a, p) = \log \tilde{a} - \log p \tag{2.28}$$

where a is the area of the rectangular cropped image size, $\tilde{a} = \max(0.05, a)$ to prevent numerical instability from small regions, and p is the probability of the class in question returned by the cropped region. 2.28 is clearly just the logarithm of $\frac{\tilde{a}}{p}$, the fractional area of the cropped region divided by the corresponding class probability; since the intuition is that a good saliency detector should concentrate relevance in as small a region as possible which alone will produce confident classification, low values of s should indicate an effective saliency map. p is evaluated by feeding only the cropped area a into the network and obtaining the class probability. To do this, the cropped region is resized to the network input size, disregarding the aspect ratio. The authors note that of course this works best with networks that are largely invariant to scale and aspect ratio.

In application to the BA regression problem, the cropping of images to salient regions, and the later resizing of the cropped regions raise two problems:

1. If the cortex is highlighted by in the heatmap – which is very likely, as it is a key marker of BA – then since the cortex encloses the rest of the cerebrum, the cropped region a cannot be a small fraction of the total MRI volume, and so the first term of 2.28 will penalise our heatmaps.
2. If the cortex is not highlighted and relevance is concentrated on a smaller region, such as the ventricles, the method of Dabkowski and Gal will produce a cropped region which does not resemble a brain volume, and this will likely confuse the network into producing an output of whose validity we cannot be certain. Therefore, even if we can produce a measure p_0 of the confidence of the initial prediction of the uncropped input (perhaps by comparison of the true and predicted ages), the value of p given by a highly cropped region a cannot be trusted and the second term in 2.28 will be unreliable.

It is clear that a different method of evaluation is necessary for saliency mapping with regression tasks. Very little has been done in the way of saliency mapping for regression. One of the earliest implementations was by Millan and Achard [113]. The authors apply saliency mapping to augmented data which has a ground truth location of injected noise. A DNN is trained on the augmented data to determine the extent of augmentation (the degradation in 'quality' of a signal). The labels on which the model is trained is the extent of the added noise, but the model never sees the ground truth alterations. The authors apply their own post-hoc explanation method AGRA, alongside other saliency mapping techniques. They then compare the produced saliency maps to the ground truth alterations to evaluate the methods' explanations. These comparisons are performed using the PCC and MSE metrics, and their tailored AGRA method outperforms the other methods, GRAD, GRAD×Input, SmoothGRAD, and Integrated Gradients by these metrics on the given task.

The key insight of the use of augmented data and the associated possession of ground truth alterations is that the alterations are exactly what we expect the explanations to capture. We would expect in the case of noise added to a signal that the explanation of the signal quality would point out where the signal deviates from a high-quality one, and therefore we would expect a perfect explanation to highlight exactly the areas of augmentation. This is why the comparison metrics are used between the ground truths and the saliency maps. This necessarily provides a quantitative metric for the faithfulness of a regression model explanation, according to Definition 4. Explanations for similar inputs will necessarily have similar ground-truth explanations.

There is no ground truth explanation for a BA regression task. However, if we are able to devise a similar task with augmented brain data and ground truth comparisons of augmentation, then we can at least show whether or not the saliency mapping techniques are able to explain brain structural regression outputs.

We cannot quantitatively evaluate the saliency maps produced for the BA regression task in this way since there is no ground truth comparison. We can however look to qualitative measures. Since we wish to achieve clinical relevance, the quality of explanations must be affirmed by domain experts. If domain experts disagree with the explanation quality, then no matter the performance on any other metrics, the saliency maps have not captured even our current understanding of brain ageing. It is in this regard that we consider domain expert analysis as the most useful metric for the utility of the BA saliency maps.

2.5.5 Criticism of XAI

Here we discuss work that has criticised modern XAI methods, and some of the recommendations for a path forward and better practice. It is vitally important that we understand the limitations of the tools that we are using, in order that we do not place undue faith in their capabilities.

There are many different methods of attempting to understand how a model makes its decisions. One method is to see how the parts of the model respond to input features [114, 90, 71, 115] and which input features maximise certain outcomes. Other methods are built with interpretability measures as a feature [116, 117, 118]. We have of course already explored some of the many methods of saliency maps. These methods all implicitly assume that the task is to provide a contextual, selective explanation that users can understand, and that the specifics of the inner workings of the models do not all necessarily matter. The risk in this framework of explanation is exactly that – we do not know what is going on at every step of the way. We are hoping to base important decisions on these models, and they seem to be inherently uninterpretable from a structural standpoint, and that has understandably worried many.

Rudin [64] strongly suggests against the use of post-hoc explanations and model explanations, and advocates the widespread use of more interpretable ML models. Rudin points out that saliency maps are good at pointing out what a model is looking at and what it is omitting, but that they fail to explain what the models are doing with that information. She also notes the important fact that for some methods, saliency maps can be exactly the same for multiple classes. While these are valid concerns, it is important to note what was talked about previously by Mittelstadt et al. [106]. For a sufficiently complex model, one cannot hope to understand all the goings-on of the processing of information that leads to a decision. One must necessarily choose a subset of information to convey best an understanding of the model's decision.

Rudin argues that for most tasks, there should exist a sufficiently simple model (such that its workings can be fully understood by humans) which performs reasonably well. She states as well that the commonly held belief that more complex models perform better and less complex models perform worse is not entirely true. She gives the example of the proprietary COMPAS recidivism model, which was famously shown in a ProPublica analysis to have racial biases in determining the allowance of bail for incarcerated individuals [66, 67, 65]. It had also been shown that the proprietary algorithm was well-approximated by simple, transparent, and more interpretable algorithms [119]. Rudin shows a simple, fully interpretable decision tree algorithm which closely approximated the COMPAS model, which shows no inherent bias.

The example of the COMPAS recidivism algorithm is eye-opening; and there are surely more cases like it. We must not forget, however, that for a task as complicated as medical image analysis, there is likely no model that performs well on a given imaging task whose action is fully interpretable by humans. Rudin argues the existence of interpretable models by a Rashomon Set argument. This begins with the notion that the set of all models which perform reasonably well on a given task for a specific dataset is large. The largeness of this set makes it likely in turn that there exists within it at least one interpretable model. This may be true for some tasks, but for any given task, there must be a lower bound on the interpretability (however that is measured) of models with good performances. It is very likely that for at least some tasks, the lower bound on interpretability is higher than the ability for humans to comprehend. Further, even if there do exist interpretable models which perform well on a complex task, we do not know that these fall within existing methods, and so we do not know whether or not they are computationally feasible. Rudin also points out that the largeness of the Rashomon Set is largely due to uncertainty in the data. This means that more diverse models can perform well on the task, but also means that as we increase the size of the dataset (which is a necessary action to optimise model performance if we can access more data) the Rashomon Set decreases in size, decreasing the probability of its containing interpretable models. This also contradicts Rudin's claim that there is no necessary trade-off between the interpretability of a model and its performance, since models necessarily perform at least as well (usually better) with more training data.

It would seem that the interpretability of a model does indeed have a trade-off mechanism with model performance. Although it is clear that the visual explanations provided by saliency maps are unable to make a model fully interpretable, we must consider what would be necessary to create such models, and with that in mind consider our best options. We have to decide whether our use of ML in our lives would preferably be more accurate or more interpretable. The primary purpose for the integration of ML technology into the working world is as a tool of great power and precision, and so it would make most sense to forgo the prospect of interpretability in favour of of better-performing models whose decisions we can explain.

2.6 Medical Imaging with CNNs

The use of ML techniques in medical image analysis has taken off substantially in the past decade. In particular, analysis with the use of DNNs – of which CNNs are heavily favoured – has led to levels of diagnosis and insight on par with experts or even at super-human levels [120, 121].

Litjens et al. [120] analyse the use of ML techniques in medical image analysis, surveying over 300 contributions, and focusing primarily on DNN implementations. The main applications of ML to the field are in classification and detection of anomalies, lesions and diseases, and the segmentation of images such as organs and substructures, or lesions. Another growing area of interest is that of image enhancement and image generation, using networks such as auto-encoders or adversarial networks.

The use of DNNs has extended to the analysis of many areas of the human body, including skin lesions (classifying cancerous moles, for example), images of the eye, X-ray and CT scans of the chest and abdomen, CT angiographic data, cardiac and pulmonary scanning, breast tissue images, musculoskeletal images, and in significant focus, brain imagery. In particular, brain MRI and fMRI scanning is utilised.

2.6.1 Brain Imaging

There are of course many different uses for DNNs in analysing brain imagery, such as cancer detection, or risk or presence of neurodegenerative disease. We focus on the task of imaging the brain in this section, and narrow our focus later on the task of brain ageing.

Prince et al. [10] note the difficulty in using MRI volumes in machine learning due to the inherent lack of quantativity in the technique, as compared for example to CT scanning. The differences in scan intensities are vast not only between scanners, but also within single scanners at different implementation times – although to a lower degree [122]. It was found that even with strict control over the scan parameters, inter-scanner variations were still significant [123]. This has led to difficulty especially in segmentation protocols. Some successful methods to combat these issues are dataset harmonisation techniques based in machine learning used to standardise contrasts and intensities [124].

Landman et al. [10] examine the obstacles in the way of creating accurate brain atlases. They point out that while cranial and subcortical structures of the brain are highly regular, there is a high degree of variability in the structure of cortical, vascular, soft tissue and peripheral nerve courses in the brain. While organ and soft tissue structure in the body is much less regular than the skeletal structures, it is well known that the volumes and shapes of cerebral structures are correlated with pathology and function [125, 126]. The central objectives necessary to create useful and reliable brain atlases are to establish homologies between perspectives of imaging, between positions within the same individual, between localised images and the whole structure, between the same anatomical features of different individuals, between in vivo and postmortem perspectives, and between extremes of imaging conditions. Many of these challenges are faced by our task too. Cortical fold structure, obtained by the analysis of the sulci and gyri, is key to the functions and organisation of functional regions. They play critical roles in understanding the development and degeneration of the brain, and variability of its structure. To this end it is extremely useful to be able to label areas of the brain delineated by the cortical folds. There are many methods used for regional segmentation of the cortex, including the mapping of a pre-labeled atlas to individual brain volumes [127], and the automatic segmentation of regions to be labeled later by an expert [128, 129]. Graph-based approaches are useful in determining relationships between sulcal areas of interest, with nodes representing individual sulci, and edges representing their relations.

Erus et al. [10] discuss the challenges of MRI neuroimaging harmonisation. They comment that the low sample sizes of many studies may be a key contributor to the low reproducibility of findings. Many attempts have been made to gather big data for the sake of neuroimaging research, but this inevitably leads to the challenge of heterogeneous imaging. The ENIGMA project [130, 131] looked to examine associations between genome variations and changes to cerebral structure. The iSTAGING [132] project has pooled over 20000 participants between the ages of 45 and 89 to form one of the largest existing MRI ageing databases. The aim of the study is to analyse the diverse anatomical changes that occur due to ageing, cerebrovascular disease, and Alzheimer's Disease. A particularly useful tool has come from semi-supervised learning techniques which reduce the dimensionality of the complex structures down to a manageable number of dimensions, each corresponding to a specific pattern of brain changes due to one or more sources of brain structure change. Data harmonisation however has still been an issue, even in studies with consistent standardisation of field strengths and protocols. One harmonisation approach which decreased scanner-related differences and improved longitudinal consistency is focused on the image processing stage, whereby regional volumes are computed that are consistent both between sites and longitudinally. This allowed the creation of scanner-specific atlases. One limitation of this method is that only a small number of individuals were able to be scanned at both sites.

Erus et al. discuss the development of the Brain Development Index (BDI)[133], which summarises the physiological changes to the brain in maturation from childhood through adolescence to early adulthood. This model, trained on 621 subjects between the ages of 8 and 22, was based on a support vector regression model. It has achieved a correlation coefficient of $r = 0.89$ and a mean absolute error of 1.22 years. The authors note that deviations from the trajectory were correlated to cognitive performance changes, with higher BDIs than their chronological age corresponding to significantly superior cognitive speed compared to those with lower BDIs than their chronological age.

The authors also note the application of machine learning to brain ageing in later life. It has been shown that specific patterns of grey matter loss have been attributed to ageing, even in individuals without concurrent pathology[134, 135]. They note however that advanced brain ageing is particularly difficult to model, but can lead to significant insights into pathologies such as Alzheimer's Disease, and the similarities and dissimilarities between such pathologies and ageing. Multivariate pattern analysis was applied to a large dataset of individuals ranging in ages from 20 to 90 years, with 2705 participants from the Study of Health In Pomerania (SHIP) cohort [136], in order to quantify atrophy patterns associated with both ageing (SPARE-BA) and Alzheimer's Disease (SPARE-AD) [137] in relation to the risk factors of smoking, anti-hypertensive and anti-diabetic drugs and waist circumference (in males). A study of 1472 individuals analysing the SPARE-AD and SPARE-BA indices, as well as the presence of the APOE-4 allele (the strongest sporadic genetic risk factor for Alzheimer's Disease) in individuals, looked to quantify the imaging differences brought upon subjects by the APOE-4 allele [138], but found that there was no significant association between imaging bio-markers and the presence of the gene.

It is well understood that there is vast heterogeneity in the pathology of brain ageing, and a study of 400 participants [139] attempted to characterise the various pathways by which these emerge. The method first used multivariate pattern analyses to develop BA models, then a mixture of classifiers and unsupervised distribution mixture models to isolate 5 distinct phenotypes of advanced brain ageing.

Yang et al. [12] and Rieke et al. [13] in the same year used sensitivity maps and backpropagative methods to create heatmaps as visualisation methods pertaining to 3D CNNs and the diagnosis of Alzheimer's Disease (AD). Both studies produce satisfactory heatmaps, focusing on known areas of saliency with respect to AD presence. The different heatmapping methods had differing areas of concentration of saliency in both cases, and both studies concluded that a mixture of explanatory techniques is most efficacious in the goal of identifying areas of saliency.

Böhle et al. [14] used LRP to create heatmaps which serve to explain the diagnosis of Alzheimer's Disease in 344 individuals, 193 of whom were Alzheimer's patients, and the other 151 healthy controls. Heatmaps were also created using Guided Backpropagation (GP) [91]. The authors noted that as opposed to GP, LRP serves as a better individual marker of AD relevance, and performed better on the authors' metrics designed to compare the heatmaps quantitatively. The authors also noted however that there are several limitations of LRP in the context of the classification task. The first is that there is no ground truth comparison – this was aided in the text by their use of a brain atlas. The second limitation is the sensitivity of LRP to the algorithm by which it is applied, although they also stated that their parameter of concern – the β of $LRP_{\alpha\beta}$ – did not destabilise the heatmaps. The third limitation, inherent to all heatmaps, is that the voxel-wise saliency highlighting nature of the heatmaps does not allow us necessarily to understand the underlying reasons for which a given voxel is highlighted – that is, we do not know if a part of the brain is important in the decision because of, for example, its shape, or atrophy. The fourth and final limitation posited was the strong dependence of the LRP method on the classifier network, shared with many heatmap methods; since LRP is a reflection of the network's reasoning for a given decision, it is obvious that better classifiers will produce better heatmaps, since they have learned more salient features.

Authors	**Problem**	**Data**	**Methods**	**Findings**
Prince et al. [10]	Intensity variability between scans inhibits quantitative comparison in ML implementations.	Various MRI volume types		ML dataset harmonisation eases the burden of qualitative comparisons.
Landman et al. [10]	Variability in cortical structure poses difficulty in brain atlasing	Various MRI brain volume types		Mapping pre-existing atlas segmentations to individual scans allows individual region labelling.
Erus et al. [10]	Small dataset sizes may contribute to lack of reproducibility in brain image analysis. Pooling datasets again poses the issue of dataset harmonisation.	Various MRI brain volume types		ENIGMA and iSTAGING projects pooled large MRI datasets with high degrees of success in dataset harmonisation.
Erus et al. [133]	Neurophysiological development to be modelled from childhood to early adolescence.	Diffusion Tensor Imaging (DTI) and T1-weighted imaging of brain volumes	Support vector regression used to develop the Brain Development Index (BDI)	Correlation coefficient of 0.89 and MAE of 1.22y for the SVM. Deviations from trajectory correlated to cognitive performance changes.

Habes et al. [137]	Quantifying patterns of brain ageing and of AD progression with respect to external risk factors.	T1-weighted brain volumes	Multivariate Pattern Analysis on the SHIP Dataset to create the SPARE-BA and SPARE-AD indices.	Smoking, use of anti-hypertensive or anti-diabetic drugs, and male waist circumference showed association with increased BA and risk for AD.
Eavani et al. [139]	Heterogeneity in brain ageing pathologies.	Resting-state fMRI and T1-weighted MRI brain volumes	Multivariate Pattern Analysis used to develop BA models. Classifiers and mixture models used to isolate ageing phenotypes.	Five distinct advanced brain ageing phenotypes were isolated.
Yang et al. [12]	Use of saliency maps to determine areas of interest for AD diagnosis by CNNs.	T1-weighted MRI volumes	3D CNN trained to classify MRI volumes as AD or not, and sensitivity backprop. methods used to create saliency maps.	Saliency maps pick up on known areas of AD pathology, although all slightly different from one-another by method. A mixture of methods is decided to be most efficacious
Rieke et al. [13]	Use of saliency maps to determine areas of interest for AD diagnosis by CNNs.	T1-weighted brain volumes	3D CNN trained to classify MRI volumes as AD or not, and sensitivity backprop. methods used to create saliency maps.	Saliency maps pick up on known areas of AD pathology, although all slightly different from one-another by method. A mixture of methods is decided to be most efficacious.
Bohle et al. [14]	Use of LRP (and Guided Backprop.) to determine areas of saliency for AD diagnosis in CNNs.	T1-weighted brain volumes	3D CNN trained to classify MRI volumes as AD or not, and LRP used to create saliency maps. Guided backprop. used as a comparison saliency mapping strategy.	LRP methods outperformed GP on localisation of known areas of AD pathology. Authors note the difficulty in quantitative assessment of saliency maps due to a lack of ground truth comparison.
Nigri et al. [15]	How to create saliency maps which avoid the issues raised by Bohle et al., specifically the ground truth issue. The proposed Occlusion method is implemented and tested.	T1-weighted MRI volumes	Various 2D and 3D CNNs trained on the same dataset to classify volumes as AD or not. The newly proposed Occlusion method was then applied to explain decisions, and the resulting saliency maps compared to existing techniques.	Occlusion saliency maps were deemed satisfactory compared to existing methods. Large computational overhead was identified as a major drawback, but the establishment of a proxy ground truth may serve as a better quantitative validation built into the framework.

Table 2.1: Summary of key brain imaging developments

Nigri et al. [15] used a novel method of heatmap production called the Swap Test to explain multiple CNN models trained to diagnose Alzheimer's through MRI scans. The authors aimed to perform heatmap implementations that avoided the limitations of LRP, DeepLIFT and others, as pointed out for example in [14], which they believe could be due to the nature of the data. The method, very similar to Occlusion Testing, compares a given image I to another image chosen randomly from within either the healthy control group (if I is classified as AD) or the AD group (if I is classified as healthy). Patches of the comparison image are then cropped into I in the same positions and the resulting image passed through the network to determine again the probability of AD. This is done for regions of a fixed size for every cubic region in the image, and thusly a heatmap is produced. For a more robust heatmap, this process can be applied multiple times with different comparison images. This serves as a proxy ground truth comparison, since the areas of interest are shown as compared to the opposing case, but are comparisons to a finite number of different subjects. Compared to known areas of saliency in AD, the resulting heatmaps were deemed satisfactory. It is clear though that since a forward pass is needed for every iteration of the heatmap procedure, it is extremely computationally intensive, especially in comparison to methods such as LRP and DeepLIFT. The findings of this section are summarised in Table 2.1.

2.7 Brain Ageing

We focus in particular on the effects of ageing on the brain. There are many contributing factors to brain ageing [49], including current and past physical activity, the presence of neurodegenerative disease or epilepsy, presence of type 2 diabetes or HIV, traumatic brain injury, and of course chronological age. There are also mitigating factors to brain ageing such as physical exercise, level of education, and diet.

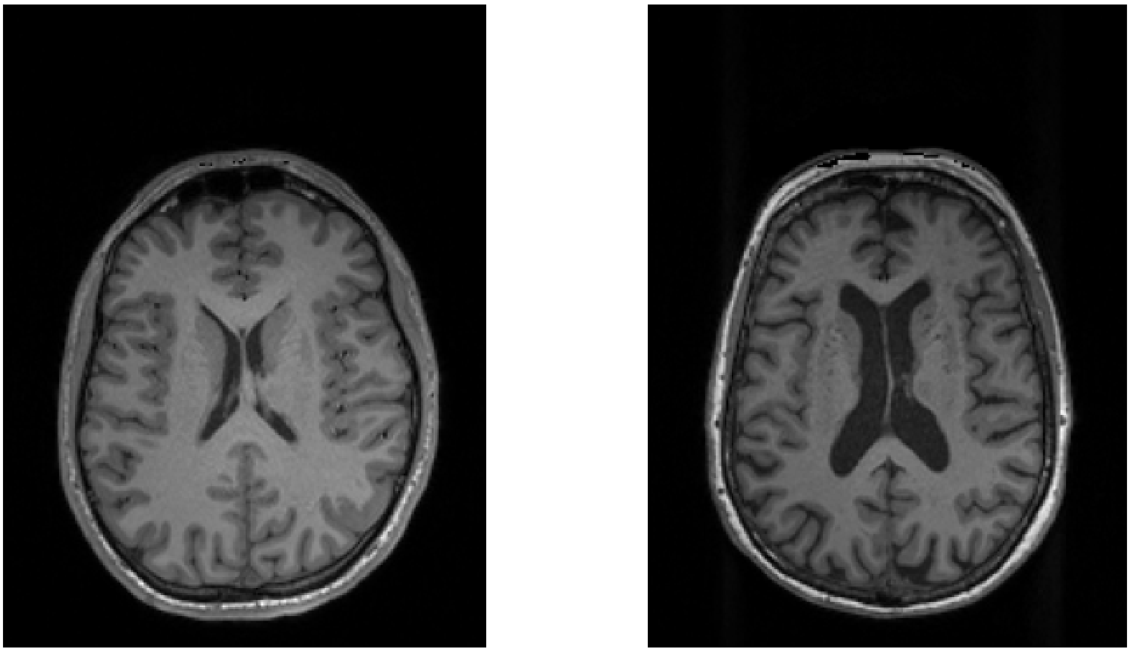
Analysing the brain through T1-weighted MRI scans limits our assessment to large-scale structural changes, and since factors such as neural plasticity and vascular health are not immediately visible at these scales, we cannot expect a network to focus on them.

There are many physiological changes occurring in the ageing brain. The first and most notable is overall shrinkage of the brain due to neuron loss and neuron shrinkage [140, 11, 141, 142] among possible other cumulative causes, especially in the frontal cortex [140, 141]. Another change we are interested by is the decrease in grey matter density [140, 141, 142], which occurs from early adolescence onward, but is accelerated in old age [141, 142]. We are also interested in the dilation of the cerebral ventricles with age [11, 142] and the filling of these spaces with cerebrospinal fluid. An example comparison between youthful and elderly brain structure is shown in Figures 2.11a and 2.11b below. Another large-scale age-related structural change is the decrease of white matter concentrations, which begins after about age 40 in most adults [140, 11, 141, 142]. Significant numbers of lesions occur in these areas as well with old age [140]. This is more difficult to quantify in T1-weighted images. Gunbey et al. [143] also examined the degradation of the limbic system with age. They found that the hippocampus, parahippocampus and fornix are affected significantly by ageing.

In assessing BA using CNNs, Cole et al. [49] examined only T1-weighted images, and looked at isolated white matter maps, isolated grey matter maps, combination white matter/grey matter maps and minimally processed, 'raw' T1 images. The authors found greatest success (assessed by network accuracy) with a CNN using the grey matter and white matter/grey matter combination maps, although there was relatively similar success with the other map types as well. Using a Gaussian Process Regression (GPR) approach as a contextualisation method, the authors again found greatest success with the grey matter and white matter/grey matter combination maps, with significantly less success in the raw maps. In both cases the lessened relative success of the raw maps was likely due to the complexity of the raw images over the other types. One may have more success on the raw images with a more complex architecture than used by the authors to capture the complexities of the raw structure of the brain. A trade-off may be present between simplification and the loss of information as produced by pre-processing (which yields, for example, the isolated grey matter maps). This is an important consideration to the model structure and any explanations thereof.

The Brain Age Delta (DBA) is the difference between predicted and chronological ages of individuals in the set, $\delta_1 = Y_{\text{pred}} - Y$. Smith et al. [144] discuss the reliability of DBA as a bio-marker for accelerated brain ageing. This measure of DBA, however, is biased by a model's tendency for regression towards the mean; the quantity δ_1 is not independent of age. Despite the presence of bias between δ and chronological age Y, a simple linear correction can yield a more reliable marker:

$$\delta_2 = M_Y X X^+ Y \tag{2.29}$$



(a) 18 year old male subject

(b) 87 year old female subject

Figure 2.11: Comparison of transverse brain volume sections of a young individual (left) and an elderly individual (right). These sections clearly show the individuals' lateral ventricles in the middle of the images (dark), filled with CSF. On the outside of the brain matter (dark grey), we see in both images high concentrations of grey matter in the cerebral cortex. From the images it is clear that the elderly individual's lateral ventricles are significantly more dilated, and that the density of grey matter in her cerebral cortex is significantly lower. In fact, there is clear shrinkage of the outermost layer of the cortex as compared to the younger individual.

where M_Y is a term which orthogonalises δ_2 with respect to chronological age, X is the input and $X^+ = \left(X^\intercal X\right)^{-1} X^\intercal$ is the pseudo-inverse of X. This is derived from the formulation of δ_2 as:

$$\delta_2 = \delta_1 - YY^+\delta_1 \tag{2.30}$$

where we treat Y as an $N \times 1$ matrix to form the pseudo-inverse. Since δ_2 removes the dependence of DBA on chronological age, and DBA dependence on chronological age is a result of regression towards the mean [144], the quantity $|\delta_1 - \delta_2| = YY^+\delta_1$ is a measure of the tendency of a model to regress towards the mean. It would be useful to analyse this as a metric for model reliability. No literature could be found to this end, but we would expect that lower values of the correction $|\delta_1 - \delta_2|$ would imply that a model does not badly regress toward the mean. Large positive or negative DBA has been shown to predict accelerated or slowed brain ageing respectively [144, 145, 146].

2.7.1 ML Applications of Brain Ageing

ML techniques to quantify BA have been used ubiquitously in the history of computer vision. In 2010 Dosenbach et al. [147] used fcMRI images to predict brain ages of 7- to 30-year-old participants using support-vector machine pattern analyses. Franke et al. [148] used T1-weighted images to predict the ages of healthy participants with a relevance vector machine, using the mean *Brain Age Gap Estimate* (BrainAGE). The BrainAGE framework has shown a high degree of efficacy in determining the brain ages of children and adolescents in particular [149].

Meier et al. [150] used a support-vector machine to examine the reorganisation of functional brain networks associated with brain ageing, by analysing fMRI images. The authors claim that this analysis method removes confounding factors such as strategy or motivation for performing the relevant tasks during the scanning process. The study aimed to investigate which functional connections best characterised brain ageing. The dataset used was extremely small ($n = 52$). 84% accuracy was achieved in BA classification.

Lin et al. [50] used a basic Neural Net architecture to classify T1-weighted and DTI MRI images according to age, reaching a correlation of $r = 0.8$ and MAE of 4.29y with chronological age. The models looked at grey matter concentrations and white matter connectivity.

These are insightful and useful early applications of BA regression, but in none of these has saliency mapping been used. Furthermore, since these are earlier attempts at the task, many of them do not use CNNs for the analysis. Since saliency mapping has not been applied to BA regression in CNNs before 2020, it is useful to look at past cases of saliency mapping with CNNs applied to diagnosis of diseases of ageing.

Mwangi et al. [51] used relevance vector regression on Diffusion Tensor Imaging (DTI) volumes to determine BA in a cohort of participants ranging in age from 4 to 85 years. Sensitivity maps were then extracted to localise areas of saliency for ageing in the regression models. The authors reported accurate age prediction on all markers, and that the sensitivity maps tended to highlight known areas of saliency to brain ageing.

Eitel et al. [75] used LRP to form a transparent CNN for the diagnosis of Multiple Sclerosis (MS). The typical presentation of MS in MRI images is the presence of white matter lesions in T2-weighted images. The cohort consisted only of 76 MS patients and 71 healthy controls. To compensate for the small number of test subjects, the authors employed the use of transfer learning with a set of 921 MRI scans initially to separate AD patients from normal controls, later fine-tuning to discriminate MS patients. Individual heatmaps were formed after training, as well as average heatmaps for MS individuals and healthy controls, both for grey matter and white matter regions. Using the 30 regions for each group with highest absolute relevance means, the areas of highest and lowest relevance on average were captured for the MS and healthy control groups. The authors showed that the transfer learning technique significantly increased the clarity of the LRP explanations. The explanations primarily focused on lesions (particularly in white matter and particularly in areas strongly associated with MS) as expected. The explanations also highlighted grey matter regions such as the Thalamus, an area well-known to be heavily affected early in the onset of MS. Lesions were removed from some of the test samples and the resulting images were tested and revealed significantly greater relevance scores in the corpus callosum, an area known to be affected by axonal loss and diffuse atrophy. The removal of lesions also slightly increased the relevance assigned to the fornix – lower anisotropy of which is exhibited in MS patients than the healthy controls. The authors conclude that LRP is useful not only for explaining a single network's decisions, but also for assessing the depth of learned features from a DNN.

Grigorescu et al. [16] used LRP in assessing a 3D CNN used to classify T2-weighted MRI scans of infants according to whether or not they had been born pre-term. They used values of α ranging from $\alpha = 1$ to $\alpha = 3$ and presented example heatmaps. The dataset was of 157 different scans, which is relatively small. The network obtained a true positive score of 100% and a true negative score of 86%. The resulting heatmaps highlighted in particular areas of cerebrospinal fluid, which agrees to some extent with previous work having shown that pre-term infants have a greater volume of cerebrospinal fluid, and less cortical folding.

Authors	**Task**	**Data**	**Model, Saliency Mapping**	**Results**
Cole et al. [49]	BA Regression	T1-weighted brain volumes; white matter isolated, grey-matter isolated, raw and combined grey- and white-matter maps. Data from 14 public sources validated on Brain-Age Healthy Control (BAHC) dataset.	CNN, None	Grey matter and combination grey matter/white matter maps yielded greatest test accuracy. Raw map performance was significantly lower, likely due to much more complex structure.
Meier et al. [150]	BA Regression	fMRI volumes from International Consortium for Brain Mapping (ICBM)	SVM, None	84% accuracy achieved in BA prediction, but with a very small dataset.

Lin et al. [50]	BA Regression	DTI and T1-weighted brain volumes, mapped onto region scores. Data gathered by authors (N=112)	Fully-connected NN, None	Correlation with chronological age of r=0.8, and MAE of 4.29 years.
Mwangi et al. [51]	BA Regression	DTI brain volumes from International Neuroimaging Data Sharing Initiative (INDI).	RVR, Sensitivity Mapping	High accuracy achieved in BA prediction across prediction variables. Sensitivity maps highlighted relevant anatomical regions on aggregate.
Eitel et al. [75]	MS Diagnosis	T2-weighted brain volumes from Alzheimer's Disease Neuroimaging Initiative (ADNI).	CNN, LRP	The saliency maps produced clearly showed areas of known importance to MS pathology. The removal of lesions highlighted by the saliency maps led to the highlighting of other known areas of saliency.
Grigorescu et al. [16]	Neonate term/pre-term classification	T2-weighted brain volumes from developing Human Connectome Project (dHCP).	3D CNN, LRP	100% sensitivity and 86% specificity reached in classification. The saliency maps accurately highlighted areas of greater CSF volume and less cortical folding in pre-term scans.
Gupta et al. [151]	AD/MCI diagnosis	Functional connectivity feature maps from fMRI scans from ADNI dataset	DNN and SVM, DeepLIFT	SOTA performance levels were reached by the DNN. DeepLIFT picked up on known areas of saliency for AD and MCI. Removal of areas reported as minimally salient increased the network's diagnostic accuracy.

Table 2.2: Summary of DL applications to brain imaging

Gupta et al. [151] used DeepLIFT to determine areas of saliency in the brain when diagnosing AD and Mild Cognitive Impairment (MCI). The diagnoses were carried out by a 5-layer DNN using functional connectivity features, and relevance was propagated to identify brain connections associated with the neurodegenerative diseases. The authors showed that limiting the input data to empirically relevant subsets improved the accuracy of the network. These subsets were isolated by recursively removing the 10% least relevant areas as measured by DeepLIFT. The authors compared the accuracy of predictions between AD/MCI samples, AD/-control samples, and MCI/control samples, as opposed to the simple accuracy of each in overall classification. The authors also achieved state-of-the-art classification accuracy with their network as compared to other uses of the same dataset, as a multi-diagnostic tool. The heatmaps revealed particular relevance of two regions of the uncus (an extremity of the parahippocampal gyrus). It is understood that early atrophy in this area is associated strongly with cognitive impairment [152]. Several other regions in the medial temporal lobe were also found to have high relevance, which have been previously reported. These findings have been summarised in Table 2.2.

Levakov et al. [17] used a CNN ensemble model to regress BA. Their model achieved 3.2y MAE and a Pearson Correlation Coefficient of 0.98, with a dataset of 10176 participants, 526 of whom were held out for testing. They used the SmoothGRAD saliency mapping technique to create population-level explanations on the input space. This was done by aggregating within each sub-ensemble the saliency maps of 100 subjects. These aggregated saliency maps were compared and contrasted between each of the trained models of the ensemble to assess the diversity in explanation among the CNNs. They showed that although by the Dice metric there was not a significant similarity in aggregated heatmaps between CNNs, there was a moderately low distance by the Modified Hausdorff Distance, showing that while there is overlap in the saliency attributions, there is significant enough distance between them that the overlap is only moderate. This may explain why the ensemble model performs significantly better than any of the individual CNNs. The authors also examined which brain regions corresponded with the highest attribution of saliency. To do this, they took the median value for each voxel across the CNNs in the ensemble to create a single population saliency map representative of the whole model. They then thresholded the top-1% of saliency voxels in the saliency map and examined clusters of such points with > 100 voxels. The regions in which the clusters lay were determined using an atlas and were ranked according to cluster size. The authors found that by far the greatest attribution of saliency went to the cisterns and the ventricles.

Hofmann et al. [18] used two multi-ensemble CNN models for BA regression. The first model made use of different imaging modalities in its sub-ensembles (T1, FLAIR and SWI), and the second made use of different brain regions in the sub-ensembles (cerebellum, subcortical and cortical). The LIFE-Adult dataset was used, which has 2637 subjects, aged $18 - 82$. The first model achieved an MAE of 3.72y on the held-out test set of 631 participants, while the second achieved an MAE of 3.38y. The authors noted that the SmoothGRAD technique used by Levakov et al. is not directional, while LRP can differentiate between areas of input that are supportive of the output, and areas which contradict the output. They used LRP$_{\text{CMP}}$ with $\alpha = 1$ to create individual saliency maps. Their first experiment was to verify the utility of the saliency mapping in the age regression task using a toy model with a ground truth. Ageing was simulated by adding accumulated lesions and atrophies to a 2D 'brain' (a torus) in proportion to an assigned age, with some random variation. The accumulated atrophies and lesions served as the ground truth for the explanations of age. They found that intact regions were assigned low relevance scores, whereas regions containing atrophies and lesions were assigned high relevance. This indicates that increased local ageing (lesions and atrophies) is characterised by increased relevance clustering. For the real BA regression task, the authors created one saliency map per individual in each of the sub-ensembles for both models, then averaged these across sub-ensembles to create two saliency maps per individual (one per model). The authors also found that the greatest amount of relevance was assigned in and around the ventricles. Significant relevance was also attributed to the grey-matter-dense outer cortex of the brain. The authors also conducted permutation-based one-sample t-tests to determine statistical significance to BA and found that in each of the MRI modality sub-ensembles, nearly the whole brain contained meaningful information. This does not, however address the concern of Geirhos et al. [85] and of Sixt et al. [82], that LRP tends simply to recreate the input of the model. The authors also contrasted the saliency maps of individuals in a younger and an older cohort to determine the difference in attribution of relevance. They also examined the change in relevance maps as a function of DBA in an older cohort. They found that all clusters indicating significant association corresponded spatially with increased relevance. In the T1-weighted images, large DBA was related more strongly with higher relevance in the frontal poles, the brainstem, the outer cerebellar borders, the cortical spinal tract, the putamen, caudate, amygdala, pre- and post-central gyri, and the cingulate gyri. The strong association between increased relevance and large DBA indicates further that large assignment of relevance is associated in the BA regression task with accelerated ageing in the given area.

As of the time of writing, there have been no publications of DeepLIFT being used for the BA regression task. Table 2.3 summarises the findings of [17] and [18].

Author	**Ref.**	**Model**	**Test MAE (y)**	**Dataset Size**	**Saliency Mapping**	**Regions of Greatest Saliency**
Levakov et al.	[17]	CNN Ensemble	3.2	10176	SmoothGRAD	Ventricles, Cisterns
Hofmann et al.	[18]	Modality multi-ensemble, Structural multi-ensemble	3.72, 3.38	2637	LRP	Ventricles, Cortical grey matter

Table 2.3: Summary of key BA saliency mapping studies

2.7.2 Architectures for Brain Imaging with CNN

One of the most important aspects of our task is to create as accurate a regression model as possible. For a task with data as complex and multivariate as ours, it is highly unlikely that the dataset can be learned – especially with a large enough dataset – in order to increase the accuracy of the model artificially and impede its generalisability.

It must necessarily be true that a change in network accuracy can only occur from a change in the areas upon which the network focuses most. This can happen by changing or shifting areas of interest in the input, or by changing the intensity with which it focuses on areas. In either case, this will affect the outputs of our decomposition methods, either by region location or relative region intensity. It is therefore imperative that we use an architecture optimally suited to the task of BA regression.

Cole et al. [49] in the early days of BA regression used a simple convolutional block architecture, which provided reasonable accuracy on the data they used, as previously discussed.

Kossaifi et al. [153] introduced Tensor Regression Networks, with the use of Tensor Contraction Layers (TCLs) and Tensor Regression Layers (TRLs). Using TCLs to replace pooling and TRLs to replace flattening and fully-connected layers, the authors were able to achieve state-of-the-art MAE of the time of publication on the BA regression task, or slightly above, while simultaneously reducing drastically the number of network parameters. The baseline comparison network was a 3D-ResNet which achieved 2.96 years MAE on the UK Biobank dataset [154]. In all cases using TRLs of varying core sizes on the model instead of fully-connected layers, the MAE achieved was lower than baseline, with the lowest achieved at 2.69 years. These TRL models also significantly outperformed the baseline on BMI prediction and gender prediction from the scans.

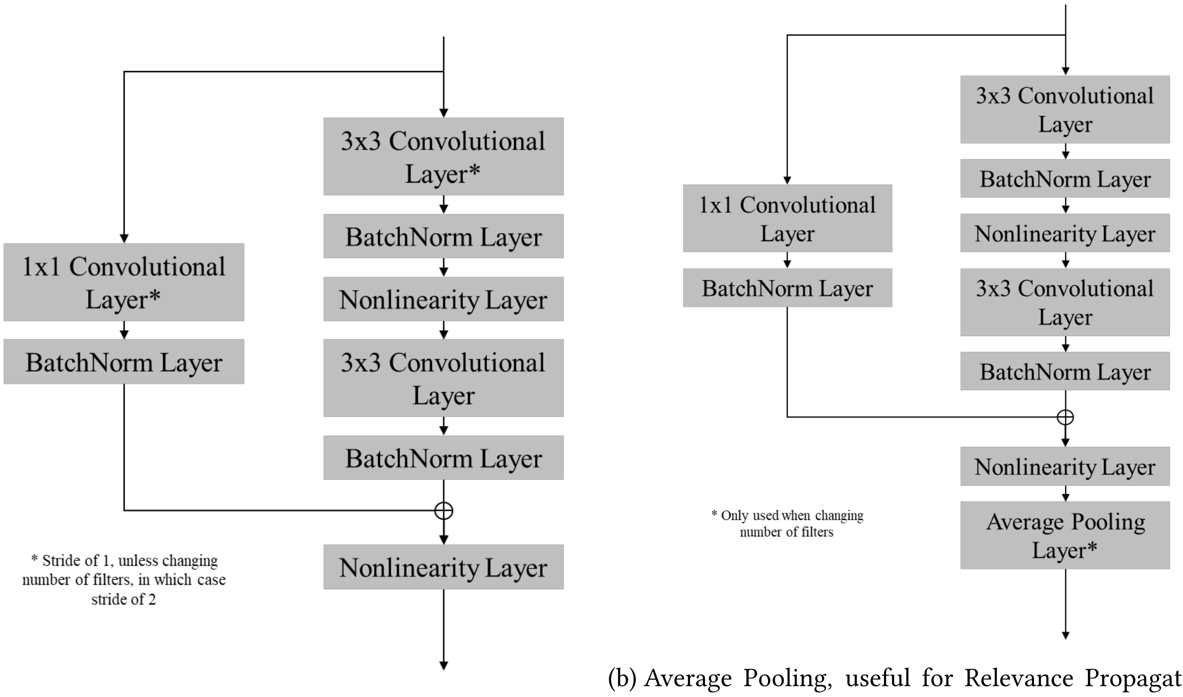
It would be interesting to test our relevance decomposition methods on the tensor regression methods, but this has not been done before and the methods of best practice are not clear for saliency mapping on the new architecture. For the main experiment, we will be using a standard 3D-ResNet, as used as the baseline by Kossaifi et al. This is a well-studied and reliable model for many computer vision tasks, and achieves close to State-of-the-Art (SOTA) performance on the BA regression task, as shown by the authors.

It was discovered after the experiments had been performed that both Levakov et al. [17] and Hofmann et al. [18] had used CNN ensemble models to great success in their BA regression tasks. Although they have achieved highly accurate models, SOTA performance is still held by models like that of Kossaifi et al. and the ResNet.

ResNet

The explanation and form of ResNet architectures that we provide closely follow those of Géron [155]. A ResNet typically cycles through modules of convolutional blocks. The blocks consist of a 3×3 convolutional layer followed by a BatchNorm layer and then by a nonlinear activation layer. The BatchNorm layer prevents gradients from becoming too large or too small during training, while the nonlinearities prevent sequences of blocks from becoming a single linear function. The modules, called residual modules, are formed from a pair of such blocks in a main branch, coupled with a skip connection. The skip connection contains another block, with only a 1×1 convolution, which takes the same input as the first layer of the other two blocks in the main branch. The output of the skip connection is then added to that of the final layer in the main branch. This is illustrated in Figure 2.12a.

The purpose of the skip connection in a residue block is to provide a copy of the input to the end of the computation of the block. This has been shown to increase learning speeds, especially early in training [155]. Down-sampling in a ResNet is traditionally performed when the number of filters is changed, by using a convolutional stride > 1 in the skip connection and the first convolution of the main branch. This saves on computational overhead as compared to down-sampling by pooling layers, since if pooling layers are implemented, the filter size must be increased before and with a unit convolutional stride to ensure minimal loss of information. As pointed out by Hui and Binder [80], using a convolutional stride > 1 creates a grid pattern if concentrated relevance in saliency maps. This is illustrated in Figure 2.13, showing results of two preliminary experiments. One model used convolutional strides > 1 for down-sampling, while the other used



(a) > 1 Convolutional Stride (Original)

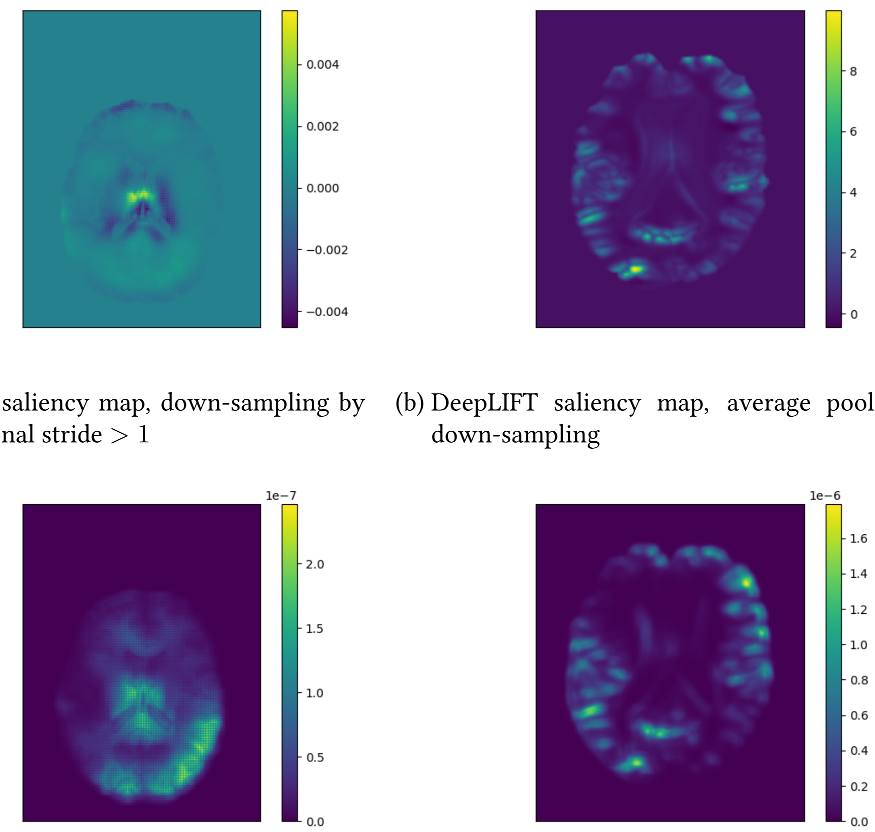
(b) Average Pooling, useful for Relevance Propagation to avoid grid patterns

Figure 2.12: Configurations of the Residual Module, depending on the method of down-sampling

average pooling. A residual module that uses average pooling to down-sample is shown in Figure 2.12b. The issue that this grid pattern presents is that regions of the brain are not smoothly assigned relevance, and so there is undue relevance assigned to parts of regions which otherwise would not be considered as relevant, and relevance removed from parts of regions which otherwise would be considered more relevant.

The input layer of a ResNet is typically followed by a reshaping layer, which is used to add an extra dimension to the data shape to initialise the number of convolutional filters. This is simply set to size 1. After this comes the first convolutional layer of the network, and the only one which is not part of a residual block. In the 3D case, this usually has a convolutional window of size $5 \times 5 \times 5$ or $7 \times 7 \times 7$, and 32 filters. After the first convolutional layer comes a BatchNorm layer, followed by a nonlinearity layer. After this comes an average pooling layer, to down-sample the data. Typically this is followed by a sequence of residual modules. Whenever the number of convolutional filters is increased, down-sampling occurs in the module. This is largely to save computational cost, but also helps to narrow down features of relevance to the model with greater network depth. A typical sequence of filter sizes for convolutional blocks will increase only in powers of 2 (doubling the number of filters occasionally). The number of residual blocks, and of course the sequence of filter sizes, are hyper-parameters of the model. The last of a ResNet's residual modules is typically followed by a Global Average Pooling layer which averages over the spatial dimensions of its input to give a single activation per convolutional filter from the residue module. This gives individual feature scores at a low dimensionality. This is followed by a fully-connected layer, and a final output layer. In the case of a regression network, the output layer will be a single neuron with a linear activation. Figure 3.5 in Section 3.2.2 shows a diagram of the ResNet we developed for our experiment.

An example of a light-weight ResNet implemented for the BA regression task is that of Jónsson et al. [156]. Their model is considered light-weight since it only has 5 residual blocks, while many have up to 25 [155]. The model achieved a test MAE of 4y on the UK Biobank T1 volume dataset [154]. The authors explain that the model has only 5 residual blocks, making it a ResNet-10, since it has 10 main-branch convolutional layers. They do not give the exact sequence of filter sizes for the main branch convolutions in the residual blocks, but they reveal that the final convolution has filter size 128. The number of filters generally increases in powers of 2 with depth in a ResNet.



(a) DeepLIFT saliency map, down-sampling by convolutional stride > 1

(b) DeepLIFT saliency map, average pooling down-sampling

(c) $\mathrm{LRP}_{\mathrm{CMP}}, \alpha = 1$ saliency map, down-sampling by convolutional stride > 1

(d) $\mathrm{LRP}_{\mathrm{CMP}}, \alpha = 1$ saliency map, average pooling down-sampling

Figure 2.13: Differences in saliency map smoothness as a result of the down-sampling technique. In both LRP and DeepLIFT, down-sampling by convolutional stride $>$ 1 causes a grid pattern overlay in the saliency maps, whereas no grid pattern appears in saliency maps produced with pooling down-sampling.

2.8 MRI Databases

For the task of creating an accurate baseline model for healthy brain ageing at a macro level, we need a dataset with the following characteristics:

1. A total number of scans as large as possible – it is well-understood that a model performs better over a specific domain with greater training data volume from that domain.

2. Even distribution of sex – it is well-established that the process of ageing is different between sexes [157]. It is also well-documented the females experience a greater risk of developing age-related cognitive impairments such as Alzheimer's [158]. To maximise the applicability of our model to patients of both sexes, it must understand the ageing process across both.

3. Maximally uniform representation across ages – in order for our model to perform well in understanding the age of a given brain, it must have exposure to as many individuals as possible in each age group within our range of ages. Furthermore, the age range must be as wide as possible, accommodating young adults up to elderly individuals.

It is a difficult task to accumulate large pools of MRI data for healthy individuals. Not only is this a storage-intensive datatype, but the issue of creating MRI data in the first place is extremely time-consuming and costly. Furthermore, people are far less likely to be subjected to MRI imaging if they are healthy than if they are not, and so the proportion of healthy individuals among brain MRI scans is low. We have also already discussed the difficulty in homogenising MRI intensities and contrasts. Many large MRI datasets are prohibitively difficult to access for researchers as well [49, 159, 160]. Table 2.4 shows some MRI datasets that have been used for BA regression in the literature, and compares their sizes, distributions of sex and age ranges.

Dataset	**Number of T1-Weighted Volumes**	**Sex Distribution (% Male)**	**Age Range**	**Pathology Present**	**Accessibility**
Alzheimer's Disease Neuroimaging Initiative [161] (ADNI 1), (ADNI 2)	189, 324	52.9, 44.8	59-89, 56-95	AD	By application
Cambridge Center for Ageing Neuroscience (Cam-CAN) [21, 162]	656	49.4	18-89	None	By application
Brain-Age Healthy Control (BAHC) [49]	2001	50.8	18-90	None	Proprietary
LIFE-Adult [163]	2016	(Unknown)	18-82	Various	By application
LIFESPAN [159]	10477	46.2	3-96	None	Proprietary
UK Biobank [160]	14503	46	42-82	None	By application

Table 2.4: Size, sex distribution and age range information of six popular MRI databases, as well as the presence of pathology, and the accessibility of these datasets.

From the table, we see that there are not many datasets containing large numbers of T1-weighted volumes that are both easily accessible and have age distributions which span the whole of adult life. From those that we have examined, we determine that the most suitable dataset for our purposes is the Cam-CAN dataset.

2.9 Summary and Conclusions

In this chapter we reviewed ML and subsequently its application to brain imaging, with specific focus on CNNs and BA regression. We also examined multiple saliency mapping techniques as methods of model explanation, and their uses and limitations, with specific focus on LRP and DeepLIFT. Our main focus was the application of saliency mapping techniques to the BA regression task.

From the literature we gather that brain ageing is a highly nonlinear process that takes many different trajectories. Particular areas of interest in T1-weighted imaging for brain ageing are the ventricles and cisterns, and cortical grey matter. Some subcortical areas such as the thalamus are also of high relevance to BA. ML has been used with great success for the regression of BA. The most commonly used tool for BA regression in ML is the CNN. ResNet architectures have shown particular utility, and have achieved SOTA performance in the task. We have established that for the task of brain age regression there are few suitable datasets which are both publicly available and large. We have determined that the most suitable dataset to our task is the Cam-CAN dataset.

LRP and DeepLIFT have both been used with great success in multiple classification tasks, including brain MRI analyses. Although there have not been many applications of either method to regression, LRP in particular has recently been used successfully for analysis of the BA regression task. It is clear that saliency mapping can reliably highlight key areas of brain ageing. Although there has been some criticism of XAI in general, and in particular of LRP, improvements have been made to mitigate some of these concerns. It is considered best practice, for example, to use LRP_{CMP} so as to minimise concerns of explanation faithfulness. Both DeepLIFT and LRP are computationally inexpensive as compared to other methods of saliency mapping, such as Occlusion [92], Integrated Gradients [89] and VDSM [103] as they are backpropagation techniques. They are also among the most commonly used backpropagation saliency mapping techniques [103, 82, 108], especially in medical imaging [14, 75, 16, 151].

No previous work has been found that compares the relevance maps produced for BA regression by different methods, and none have utilised DeepLIFT on BA regression. While the LRP method employed by Hofmann et al. [18] has directionality, it is unable to assign negative relevance without demeaning the data. Negative relevance values can be assigned by some LRP methods with different parameterisations, and by DeepLIFT. While Hofmann et al. compared a younger and older cohort in their study, no previous work has focused on the change in relevance attributions to brain regions as a function of subject age. Furthermore, the concern raised by [85] and [82] about modified backpropagation algorithms like LRP (that they attempt simply to recreate the input) has not been addressed in this context.

Saliency maps can be used to show the trajectories of regional saliency within the brain towards BA. We aim to do this with LRP and DeepLIFT methods, due to their computational efficiency and reliability shown from the literature. We aim also to compare the results from these methods. The link between DBA and brain region relevance can be further explored to this end as well.

Chapter 3

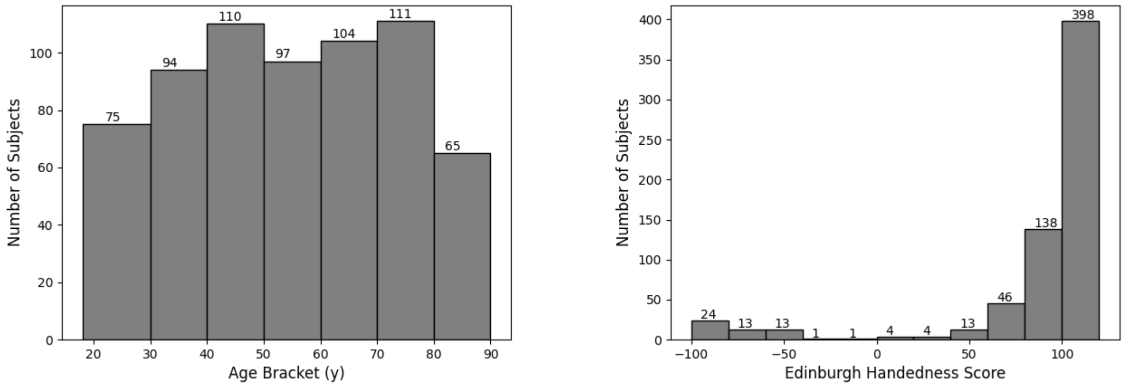
Experimental Design

In this chapter we discuss the layout of the experiment to be performed and the methods by which we will evaluate the results.

We created a BA regression model and applied our chosen saliency mapping methods to determine areas of saliency to BA. Using the data from the saliency maps, we evaluated their utility in comparison to known areas of BA relevance. We also analysed the differences between the maps produced through different methods, examined the link between region-specific saliency and accelerated brain ageing (positive DBA), and assessed region-specific trajectories of BA saliency over ages.

3.1 Dataset

The Cambridge Center for Ageing Neuroscience (Cam-CAN) has been assembling brain MRI volumes and accompanying data since before 2014 [21]. Currently held in their cc700 dataset are $N = 656$ T1- and T2-weighted pairs of MRI brain scans of healthy individuals (no symptoms of physical or mental disease, and no symptoms of cognitive decline). There are 75 subjects in the age range 18-29y, 94 in the range 30-39y, 110 in the range 40-49y, 97 in the range 50-59y, 104 in the range 60-69y, 111 in the range 70-79y, and 65 in the range 80-89y. There is relatively good age uniformity in the cohort, except for at the extremes[1]. The substantial size of this dataset (as compared to, say the ADNI datasets), its relative uniformity in sex and age distribution, and its homogeneity in scan size, scale and resolution make it an extremely useful tool for the purpose of creating a healthy brain ageing model.



(a) Age distribution of the subjects in the Cam-CAN cc700 dataset.

(b) Handedness distribution of the subjects in the Cam-CAN cc700 dataset.

Figure 3.1: Demographic information for the Cam-CAN cc700 MRI dataset

[1] This is to be expected – very young individuals are unlikely to receive MRI scans, and there are few individuals who are both healthy and very elderly

We chose to use the Cam-CAN dataset due to its relatively large size for a healthy MRI dataset, as well as its large span of ages and its ease of accessibility, as discussed in Section 2.8. We show in Table 2.4 the demographic data and accessibility of several popular MRI datasets in comparison to Cam-CAN. While the Cam-CAN dataset is not the largest available dataset, it is not so large as to cause issues with local storage. The dataset is also freely available upon request. Although the database is smaller than any that have been used before in the BA regression task within the literature, it is not of great concern that this will hinder the performance and generalisability of our model to any great degree.

Each subject in the dataset is assigned a patient ID. The following data is available on the subjects from the dataset:

- Age
- Sex
- Handedness (ranked on the 'Edinburgh scale' of -100 to 100 with -100 being strongly left-handed and 100 being strongly right-handed)
- Which MRI Repetition Time is used (30ms or 50ms). This is the amount of time between successive pulses while acquiring the scans (see Section 2.1).

Following the actions from within much of the literature [49, 50, 137, 139, 17], we chose to use only the T1-weighted images. We also chose to use all volumes of the dataset, regardless of handedness. Limiting the data only to right-handed individuals would decrease the size of the dataset further. The ages of subjects in the dataset range from 18 to 89, with a mean age of 55. The age and handedness distributions of the dataset is shown in Figures 3.1a and 3.1b. $N_m = 324$ of the subjects (49.39%) were male. 380 of the individuals (57.93%) had 30ms repetition times, 237 had 50ms repetition times (36.13%), and the rest did not have their repetition times recorded. Other than this data, the subjects were completely anonymised. The facial structures of the subjects were also masked in the raw MRI volumes. Like all previous studies on BA regression mentioned in the literature review, we chose not to consider the factors of sex, handedness and repetition time. Our model only had the pre-processed T1 MRI volume as input. This data is summarised in Table 3.1.

Number of Participants	**Age Range (Mean)**	**Sex Dist. (% Male)**	**Avg. Handedness Score**	**RT (%30ms, %50ms)**
656	18-89 (55)	49.39	76.10	57.93, 36.13

Table 3.1: Demographics data for the Cam-CAN dataset

3.2 Experimental Pipeline

The experimental pipeline is outlined in Figure 3.2. We started by pre-processing the data, using techniques discussed in Section 2.4.1. We then used the pre-processed data to train our ResNet model. The development of our model is discussed in Section 3.2.2 and the training and testing of the model is discussed in Section 3.2.3. Once we trained the model, we evaluated its performance on the test set, and executed all of our saliency mapping methods on the test set. For ease of storage and computation, we only performed the saliency mapping on the test set subjects. Storage was limited in the course of experimentation, and due to the large size of the data (volumes of shape $(233 \times 189 \times 197)$) there was a heavy computational burden. We used the subjects' pre-processed brain volumes as input to the regression model to make a prediction of their age. Using the activations at each layer of the model at inference time, we then applied the various methods to create saliency maps for each individual. This is described in Section 3.2.4. Once we have all the saliency maps for each individual in the test set, we commence with the evaluation of the data toward answering each of our research questions. This is detailed in Section 3.2.5.

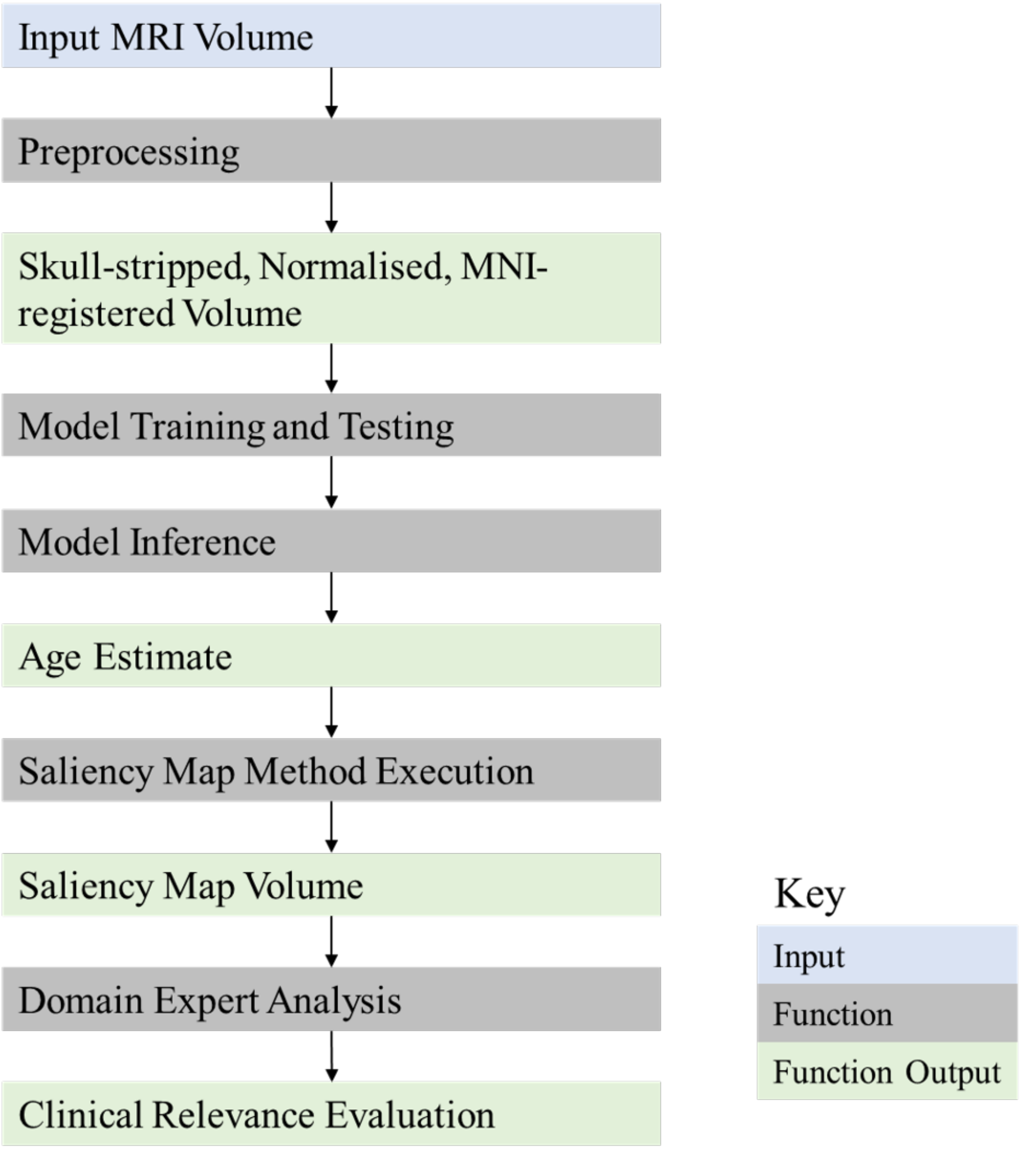

Input MRI Volume
Preprocessing
Skull-stripped, Normalised, MNI-registered Volume
Model Training and Testing
Model Inference
Age Estimate
Saliency Map Method Execution
Saliency Map Volume
Domain Expert Analysis
Clinical Relevance Evaluation
Key
Input
Function
Function Output

Figure 3.2: Experimental Pipeline for the creation and analysis of our BA regression model and the saliency maps

3.2.1 Pre-processing

The pre-processing pipeline for the experiment is outlined in Figure 3.3. It is common practice in the BA regression literature to perform skull-stripping [49, 50, 51, 17]. At the first stage of pre-processing, we use the Brain Extraction Tool (BET) provided by FSL [2] to skull-strip the volumes in the dataset, leaving only brain tissue and cerebrospinal fluid (CSF) in the volumes on a ‘black‘ background (voxel values of zero). A fractional intensity of 0.5 was used in this step; this is the value recommended by the developers. We do this to remove non-brain tissues such as bone and eyes from the volumes, such that only brain tissue is considered in the BA regression.

Using as reference a Standard MNI Volume [8], we then registered (align) all the skull-stripped volumes in the dataset to MNI space such that they were aligned with one-another, the reference MNI volume, and the corresponding region atlas [6, 7]. As discussed in Section 2.4.1, we do this so as to have all the volumes aligned in MNI space with the same orientation, such that the atlas can be used to determine areas of saliency by simple spatial correspondence. The registration was completed using FSL’s FLIRT tool [3, 4]. Initially, there was a 30% error rate (198 of 656 volumes) using the FLIRT tool, but after repeatedly rotating and re-registering the volumes, this was brought down to an error rate of $< 0.5\%$ (3 volumes). Errors were found by checking manually the alignment of each of the volumes. The three remaining volumes were manually

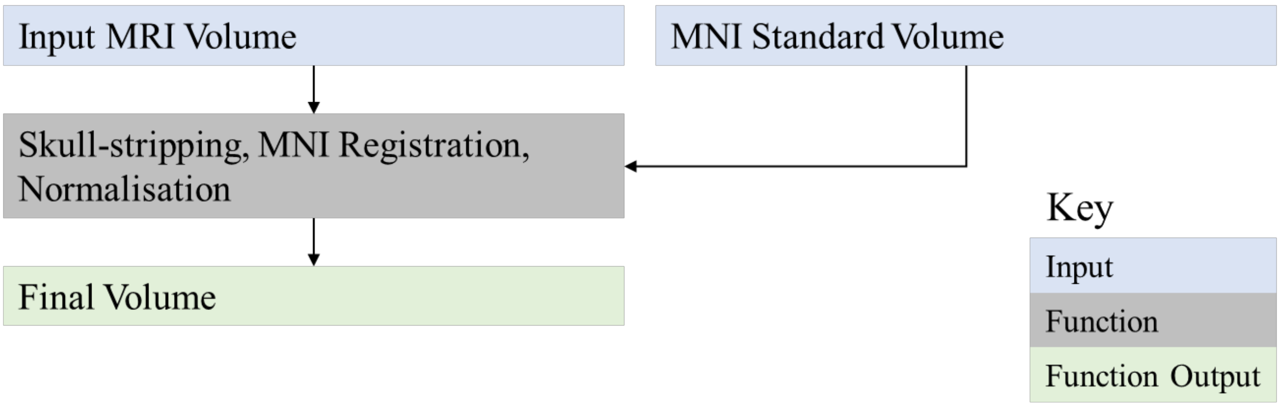
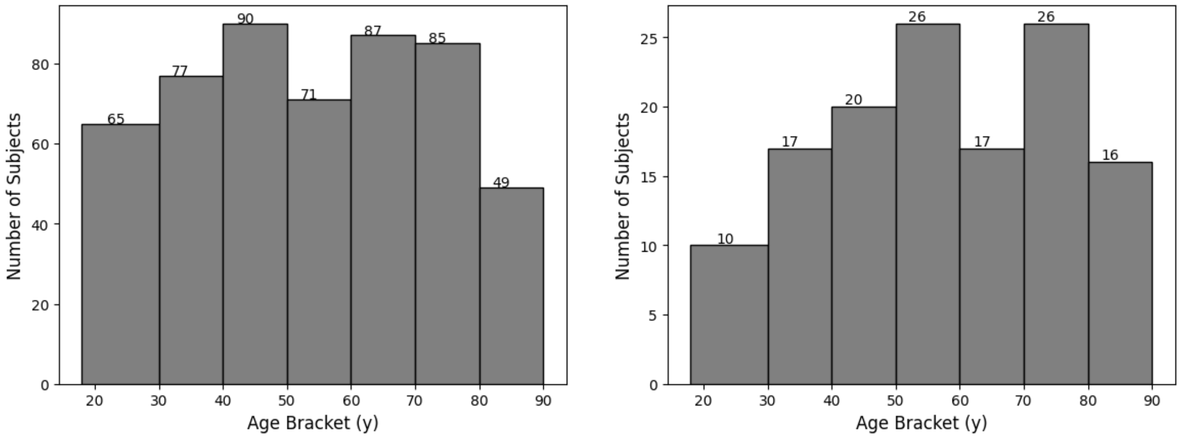


Figure 3.3: Pre-processing Pipeline

rotated and adjusted to be approximately aligned to the MNI volume. Finally, we use Global Contrast Normalisation (GCN) to normalise the volumes to have voxel activation contrasts of 1. As discussed in the Literature Review in Section 2.4.1, this is a normalisation step recommended by Goodfellow and Bengio [9], and is one of the most commonly implemented normalisations. We normalise the data this way to ensure that the voxel values lie in a similar distribution for all individuals. At the end of pre-processing, the volumes have shape $(233 \times 189 \times 197)$.

To create the training and test sets, we randomised the order of the patient IDs and split them into a training set comprised of 80% of the dataset (524 individuals), and a test set comprised of the other 20% (132 individuals). The corresponding ages and pre-processed MRI volumes were split into the appropriate training and test sets according to the ID sets. The true ages were the targets for the regression model. During training, 80 training volumes (15% of the training data) were set aside for validation. These were randomly chosen at the start of training as part of the training pipeline.

The age distributions of the training and test sets are shown in Figures 3.4a and 3.4b. 262 of the training set individuals (50%) were male, while 62 of the test set individuals (47%) were male. Minor disparities in the age and sex distributions between the training and test set are due to the fact that the split was completely random and there was not stratification for either variable.



(a) Age distribution of the subjects in the training set

(b) Age distribution of the subjects in the test set

Figure 3.4: Age distributions for training and test split

3.2.2 Model Development

We implemented a ResNet architecture, as shown in Figure 3.5. We did this for two reasons, the first being that this is an architecture that reaches SOTA BA regression performances on large datasets, as shown by Kossaifi et al. [153]. The second reason is that the ResNet architecture is comprised of layers that easily allow for relevance decomposition (as opposed to, say, the tensor regression layer, on which saliency mapping has not been performed before). For a detailed explanation of the ResNet structure, see Section 2.7.2.

The first convolutional layer – the only one not part of a residual module – was chosen to have a $7 \times 7 \times 7$ convolution window with 32 filters, as is often the configuration [155]. For all the nonlinearities in the network, we used a Softplus layer, as recommended by Dombrowski et al. [108] to avoid instabilities in the explanations. We also only used average pooling for down-sampling, so as to avoid the grid-like overlay on saliency maps.

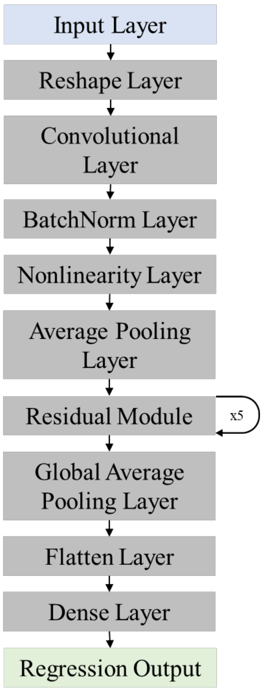


Figure 3.5: Architecture of our ResNet model

Our ResNet model used 5 consecutive residual modules, making it a ResNet-10 (the number of main branch convolution layers being 10). The main branches of our residual modules had convolutional layers with windows of size $3 \times 3 \times 3$ (the standard main branch window size), and filters of sizes $[32, 32, 64, 64, 128]$. This configuration of residual blocks was inspired by Jónsson et al. [156], who trained a ResNet to achieve a test MAE of 4y on the UK Biobank dataset [154]. Their model is discussed in Section 2.7.2, and used five residual blocks, the final of which had 128 filters. The skip connections of our model had convolutional windows of size $1 \times 1 \times 1$ (the standard skip connection window size). To avoid the overlaid grid pattern of relevance brought about by convolutional stride lengths > 1, we used average pooling to down-sample the data. The configuration of our residual modules is shown in Figure 2.12b of Section 2.7.2. We down-sampled resolution when the number of convolutional filters is increased.

As discussed in Section 2.7.2, the residual modules are followed by a Global Average Pooling layer that in our case produces an activation of 128 neurons (the number of filters at the final residual module). This is followed by a flattening layer to reduce the number of dimensions to 1 (since the global average pooling averages over the spatial layers but does not get rid of them, having output shape $(1, 1, 1, 128)$ per volume). This then connects to a fully-connected layer of 128 neurons with a Softplus activation. A final fully-connected single output neuron with linear activation delivers the prediction.

3.2.3 Training and Testing the Model

The model was created, trained and tested using Tensorflow Keras, on the Lambda Cloud Computing platform[2]. GPUs were necessary for computational time, as well as for the version of Tensorflow that needed to

[2] https://lambdalabs.com/

be run. The initial CPU version that was being run caused issues with the BatchNorm parameters at inference time, giving model predictions that were off by a large constant.

Hyper-parameter tuning was implemented to determine the optimal loss function, optimiser and starting learning rate. A grid search was performed across the Mean Squared Error (MSE) and Mean Absolute Error (MAE) losses, the Adam [164] and RMSProp optimisers, and starting learning rates of $\{5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}\}$. This gave us a total of $2 = 12$ hyper-parameter configurations. Each configuration was trained three times over 8 epochs each to determine the best performance. Configuration performance was measured by the average MAE over the three training runs. Early stopping was implemented to end training when the validation did not increase for 20 epochs, and the model was only saved on the best validation scores. A custom decaying learning rate was employed during training as well, whereby the learning rate would be halved after every 20 epochs. Decaying learning rates have been shown to increase model performances substantially [9]. Jónsson et al. [156] employ a decaying learning rate in the training of their model. The metrics we used for evaluation were MSE, Mean Absolute Percentage Error (MAPE) and MAE.

Memory use was continually an issue in training, testing and the production of the saliency maps. The size of the test set was 9.16GB uncompressed, and that of the training set was 36.37GB uncompressed. To this end we used two NVIDIA RTX A6000 GPUs with 48GB VRAM each for training, testing and saliency mapping. To maximise the batch size for training, we used a Tensorflow *ImageDataGenerator* object to feed volumes into the model from a directory as opposed to loading the entire dataset into memory. This allowed us to increase the batch size to 4. The highest batch size that we have found in the literature was 8, used by Peng et al. [146], who used volumes of size $160 \times 192 \times 160$ – about 60% of the total volume of our pre-processed scans. Jónsson et al. [156] also used a batch size of 4 in their model training.

3.2.4 Saliency mapping

Using the trained model, the saliency maps were created for DeepLIFT using two different reference volumes, and for LRP_{CMP} using three values of α. The first DeepLIFT reference volume was the MRI volume background, of all zeros. This was a suggestion offered by the authors of the DeepLIFT paper [86]. The second reference volume that we used was a composite image made of all the test set volumes. This was created simply by adding the volumes all together, and dividing through by the total number of individuals in the test set (that is, 132). Respectively we call these $\text{DeepLIFT}_{\text{bg}}$ and $\text{DeepLIFT}_{\text{comp}}$. For LRP_{CMP}, we used the three values of $\alpha = 1, 2, 3$, like Grigorescu et al. [16]. Algorithms 1 and 2 show how the saliency mapping methods were implemented on the trained model.

Algorithm 1 Pseudo-code for implementing the LRP methods on the trained model. The LRP function takes as input the current state of the saliency map, the parameters of the given layer, the activations at that position, and the relevant value of the parameter α. Text in italics, marked by an empty triangle, is a comment on a specific part of pseudo-code.

```
for $\alpha$ in $\{1, 2, 3\}$ do
    for subject in test_set do
        subject_activations $\leftarrow F$(subject) ▷ $F$ returns list of activations at each model layer, specific to input
        for layer in model do                                    ▷ Layers in order from output to input
            map $\leftarrow \text{LRP}_{\text{CMP}}$ (map, layer.parameters, subject_activations[layers.position], $\alpha$)
        end for
        save map
    end for
end for
```

We can visualise the 3D saliency maps as we would the 3D MRI volumes. One method of doing so is through the display of sections along different axes of the brain volume. This is commonly done in clinical and research settings efficiently using the FSLEyes application [1], which we utilised for our visualisations. It is easy to see relative assignments of relevance between and within large areas, and perhaps which areas were overall the most or least salient. However, to quantify the relevance assigned to each region of the brain, we must use numerical measures. To this end we threshold the saliency maps to determine the voxels of greatest

Algorithm 2 Pseudo-code for implementing the DeepLIFT methods on the trained model. For each of the DeepLIFT methods, we have to compute the difference-from-reference values. The DeepLIFT function takes as input the current state of the saliency map, the parameters of the given layer, and the difference-from-reference values at that position. Again, text in italics, marked by an empty triangle, is a comment on a specific part of pseudo-code.

```
for method in $\left\{\text{DeepLIFT}_{\text{bg}}, \text{DeepLIFT}_{\text{comp}}\right\}$ do
    if method=$\text{DeepLIFT}_{\text{bg}}$ then
        reference ← background_volume
    else
        reference ← composite_volume
    end if
    reference_activations ← $F$(reference)        ▷ $F$ returns list of activations at each model layer, specific to input
    for subject in test_set do
        subject_activations ← $F$(subject)
        subject_differences ← subject_activations − reference_activations        ▷ Difference-from-reference
        for layer in model do        ▷ Layers in order from output to input
            map ← DeepLIFT(map, layer.parameters, subject_differences[layers.position])
        end for
        save map
    end for
end for
```

saliency. For each individual the top 1% of activations in their saliency map is thresholded, for each method. These correspond spatially to the voxels in the individual's pre-processed MRI volume which were the most relevant (top-1%) to the regression output. We refer to such activations as Top-1% Relevance (T1R), and these were our primary measurement of relevance distribution. This measure of regional relevance attribution is similar to the method employed by Levakov et al. [17]. Using the region atlas, we determine the number of T1R activations which lie in each region.

While Hofmann et al. [18] showed statistical significance toward BA in their LRP saliency maps, they did not address the concerns of Geirhos et al. [85] and Sixt et al. [82]. The concern raised was that modified backpropagation techniques like LRP tend to recreate the input as opposed to focusing solely on regions of saliency. To determine statistically significant difference from the input, we create difference maps between the input volumes and their corresponding saliency maps (each normalised to have unit variance). These were the inputs to a permutation-based one-sample t-test to determine statistical significance. FSL's *randomise* tool [5] was used for this. 5000 permutations were performed for each method (the number recommended by the creators for publication), using threshold-free cluster enhancement (TFCE). The output of the test is a 3D volume of $(1 - p)$-values. We hope to show that the majority of the saliency mapping volume is significantly different from the input.

3.2.5 Evaluation Technique

To evaluate the saliency maps we turned to the analysis of a domain expert. In this analysis, we aimed to show whether or not the regions that were assigned most relevance were those expected of the ageing process. Dr Jonathan Ipser, our domain expert, is a Senior Research Officer in Neuroimaging at the University of Cape Town. His research directly concerning neuroscience dates back as far as his 2005-2010 PhD thesis, entitled 'The relationship between impulsivity, affect and a history of psychological adversity: A cognitive-affective neuroscience approach'. Dr Ipser's expertise in brain ageing has been acquired through familiarity with the relevant peer-reviewed literature, as well as his own studies. Both of these have guided his focus toward regions of interest particular to brain ageing in general, as well as in psychiatric populations. The most salient structural features of interest to Dr Ipser indicative of ageing are the cerebral ventricles (known to dilate with age [11, 142]), the density of grey matter in the frontal cortex (known to decrease with age [140, 141, 142]), and hippocampal structures (known to play a role in healthy ageing [143]).

Beyond the domain expert analysis, we wish to determine some other characteristics of the data. We performed the following analyses:

1. Determined on average over the test set, which regions were most heavily assigned relevance for each method, and which were least assigned relevance.
   - This was assessed by our domain expert. T1R was compared across regions for different methods. The similarities and differences between methods was analysed, as well as agreement with areas known to be relevant to BA.
2. Examined the relationship between high DBA and region-specific saliency.
   - To determine DBA, we used the definitions of Smith et al. [144]. The uncorrected score $\delta_1 = Y_{\text{pred}} - Y$ is age-dependent, where Y is the column vector of chronological ages in the test set and Y_{pred} is the predicted ages. The age-orthogonalised score $\delta_2 = (I - YY^+)\,\delta_1$ is independent of age, where $Y^+ = (Y^\intercal Y)^{-1}\, Y^\intercal$ is the pseudo-inverse of Y (treating Y like an $N \times 1$ matrix). Using the age-orthogonal measure of DBA, δ_2, the distribution of DBA becomes approximately normal (see Section 4.1.1). It is of interest to us to examine the correction difference, $|\delta_1 - \delta_2| = |YY^+\delta_1|$, as this quantifies the extent to which the model regresses towards the mean (regression towards the mean is the cause of chronological age-dependence of DBA; see Section 2.7).
   - Using δ_2 as a measure of DBA instead of δ_1 not only removes the dependence on chronological age (see Section 2.7), but also allows us to threshold 'high' DBA using the standard deviation of the DBA scores. This has not been done before in the literature, and so we make the simple choice of threshold DBA value $\delta^* = \sigma$ for which $\delta_2 \geq \delta^*$ determines 'high' DBA. σ represents the standard deviation of the value δ_2 over the test set. For each method we examined the distributions of T1R for three groups of individuals:
     - ◇ Individuals with high DBA $\delta_2 \geq \delta^*$ in an older age-group (> 50y)
     - ◇ Individuals with high DBA $\delta_2 \geq \delta^*$ in a younger age-group (< 50y)
     - ◇ Individuals with small-to-moderate DBA $|\delta_2| \leq \delta^*$

     We chose to split the high-DBA individuals into an older and younger cohort to examine the effect of age on the distribution of T1R in high-DBA individuals. We chose 50y as the old/young threshold since this is roughly the mean of our dataset ages.
3. Created region-specific trajectories of BA saliency.
   - We split the test set into seven age bins of equal age range, and examined for individual structures the trajectories of T1R for each method over the ages. For each region, the trajectory was calculated as the average of the proportions of T1R across all individuals lying within each age bracket. We chose seven bins because this was the largest number that afforded a non-trivial number of individuals per age bracket (with more brackets, some contained fewer than 5 individuals, which was deemed insufficient).
   - We are most interested in those highly-salient regions whose T1R changes most over the course of ageing, and those whose T1R changes least. These regions are of interest because they are the most illustrative of changes or the lack of changes in T1R over the course of ageing. The regions with greatest change were quantified by the highest Standard Deviations (SD) in T1R across the age brackets. The regions with lowest change were quantified by the lowest Coefficients of Variation[3] (CoV) in T1R across the age brackets. We use these separate metrics for the two quantities, since those regions with low SD tend to have very low total relevance, and so do the regions with high CoV. The latter is due to the fact that small perturbations in T1R for regions with very low average T1R lead to large CoVs. We wish to focus on regions that are highly salient overall, but either have a very high degree of change or a very low degree of change over the age brackets.

[3] SD normalised by the mean T1R across age brackets

3.3 Overview of Experiment

We train a ResNet model to regress BA on T1-weighted MRI volumes. We then apply LRP and DeepLIFT saliency mapping techniques to determine relevant areas to brain ageing. We are most interested in the highest contributions towards brain ageing. We therefore examine the distribution of top-1% relevance (T1R) in our analyses:

1. We compare the distributions of T1R between methods to determine the similarities and differences in explanations of BA. We also examine how each method's explanations compares to the expectations of our domain expert and the current literature.

2. To determine saliency trends for accelerated brain ageing, we examine the effect of large DBA on regional T1R distribution. We are interested in how this changes with age as well, and so we examine high-DBA subjects in an older and a younger group, in comparison to individuals with low to normal DBA.

3. We are interested in how regional relevance assignment changes over the course of ageing, especially in highly salient regions. We create population trajectories of T1R for each region to determine how saliency distributions change with age.

Chapter 4

Experimental Results

In this chapter we will discuss the results of our experiment and the associated data. To begin, we will examine the training and test performance of the regression model. We will then examine the results of applying the saliency mapping techniques to the model. In analysing the saliency mapping results, we focus on answering each of our research questions.

4.1 Brain Age Regression

We performed hyper-parameter tuning with a grid search over $2 \times 2 \times 3 = 12$ configurations, detailed in Section 3.2.3. We found that the best hyper-parameter configuration was with the Adam optimiser [164], the MSE loss function, and a starting learning rate of 5×10^{-3}. Over the three tuning runs of 8 epochs each with this hyper-parameter configuration, an average test MAE of 19.84y was achieved. The training and validation performance are shown in Figure 4.1.

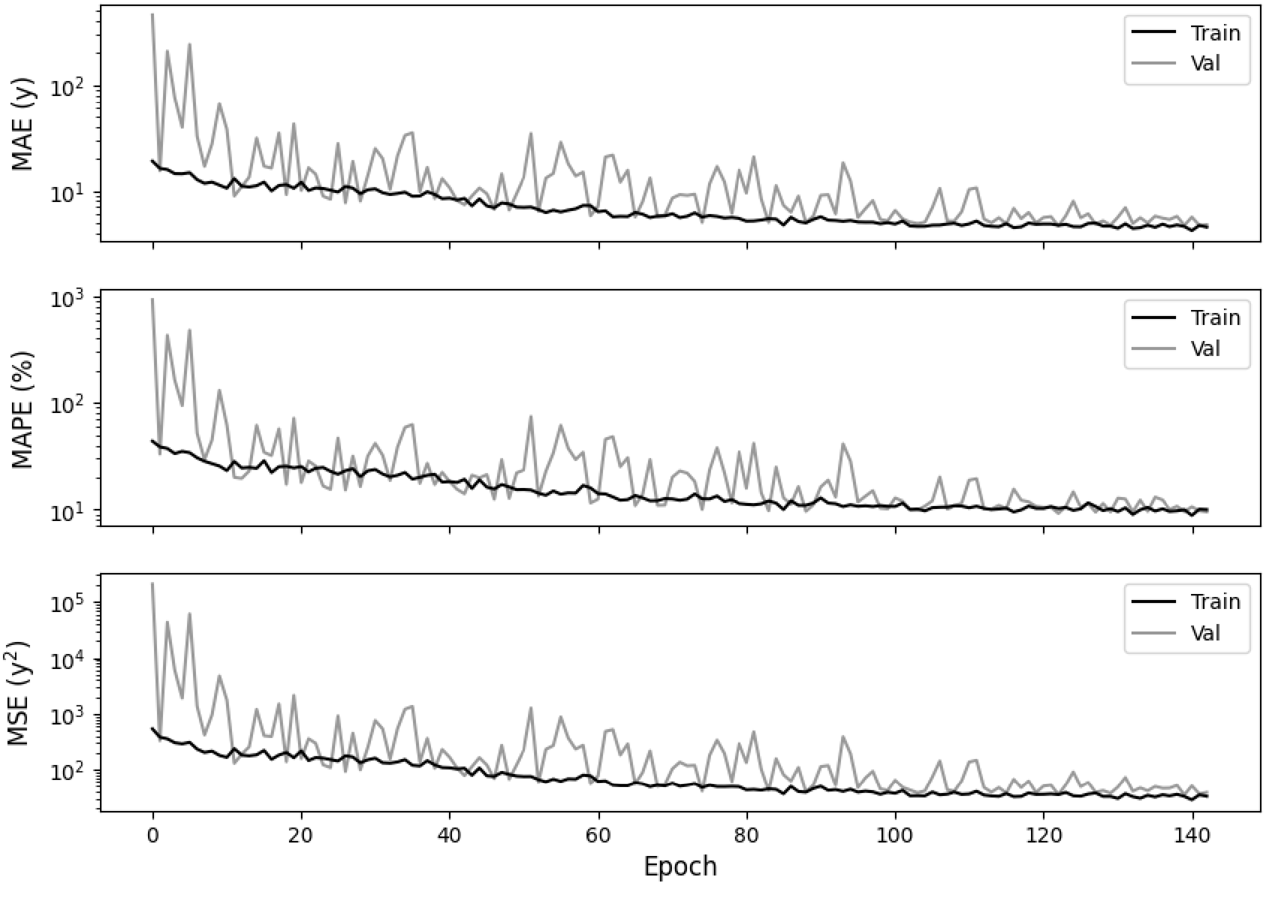


Figure 4.1: Metric performances of the model during training

Training ended at the 143rd epoch due to the early stopping callback, by which point the learning rate had decreased to 4×10^{-5}. The instance of the model with lowest loss was saved, which had 31.99y^2 MSE, 4.56y MAE and 9.72% MAPE on the training set. On the test set, the model achieved an MSE of 72.55y^2, an MAE of 6.55y, and an MAPE of 13.53%. The coefficient of regression corresponding to this performance was $r = 0.89$. The regression plot for the test set performance is given in Figure 4.2. Our metrics of model performance were the MAE, MSE and MAPE. As is most common in the literature, we report with greatest interest the MAE.

While this performance is far from SOTA, with Kossaifi et al. [153] achieving a test MAE of 2.69y (see Section 2.7.2), the BA regression task has not been performed on a dataset this small before, and a relatively good coefficient of regression has been achieved (see Lin et al. [50]). There are clearly several extreme misclassifications though. These individuals have very high absolute values of DBA. DBA is not a metric that interests us for the model performance, but rather it is an indication of what the model treats as deviations from a normal ageing trajectory.

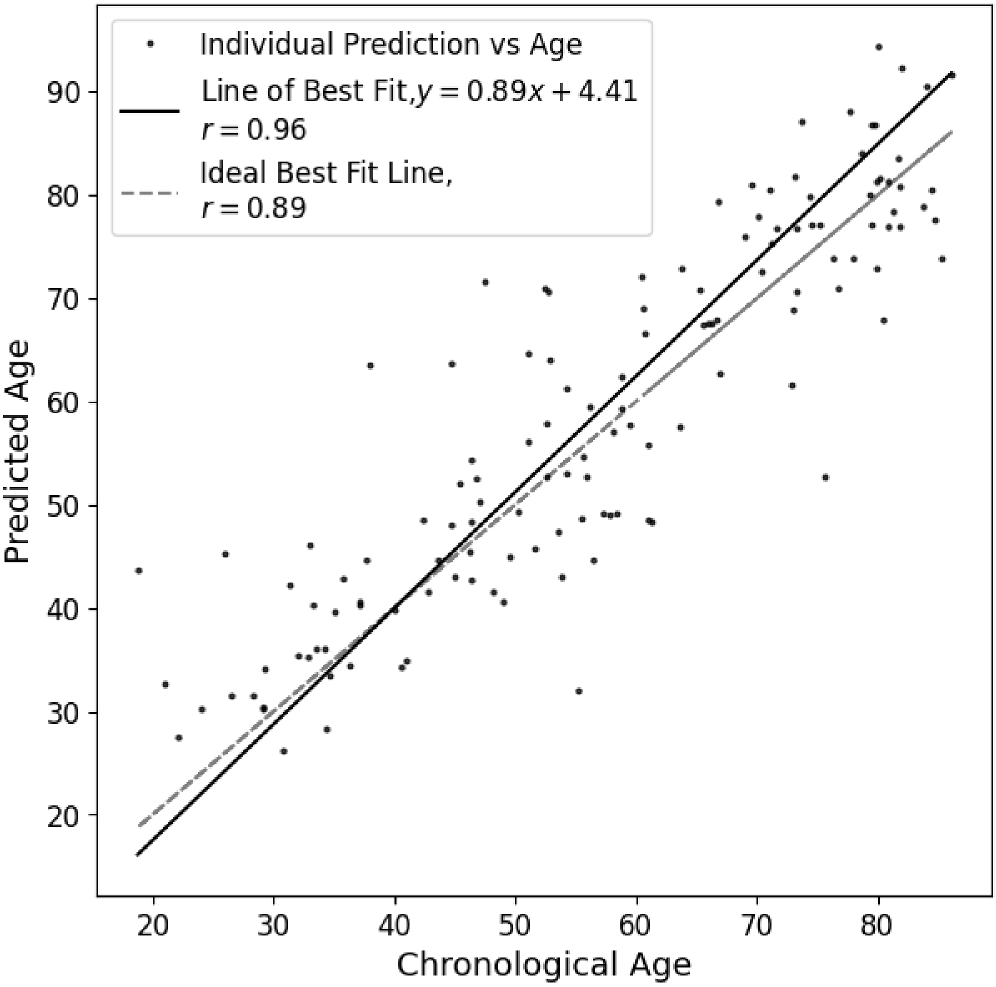


Figure 4.2: Regression plot of the trained model on the test set. Dotted line shows the ideal best fit line, where the predicted age equals the chronological age. The Pearson correlation coefficient for the test set predictions on this line is $r = 0.89$. Solid black line shows the actual line of best fit, with gradient 0.89 and offset 4.41y. The Pearson correlation coefficient for the test set predictions on this line is $r = 0.96$.

4.1.1 Test set DBA

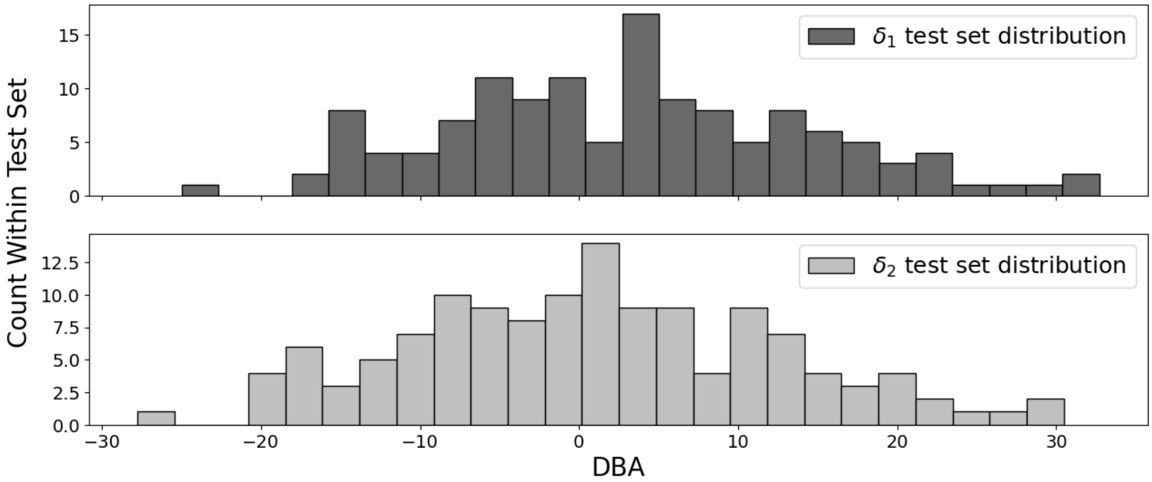


Figure 4.3: Distributions of the calculated values δ_1 and δ_2, as defined in Equation 2.29 in Section 2.7 and again in Section 3.2.5, within the test set. The age-orthogonality correction for the quantity δ_2 shifts the distribution to the left to become approximately normal.

The distribution of the DBA values in the test set is shown in Figure 4.3 for the quantities δ_1 and δ_2 as defined by Smith et al. [144], in Equation 2.29 in Section 2.7 and again in Section 3.2.5. The age-orthogonalisation correction of δ_2 shifts the distribution of DBA to become approximately normal. The mean of the δ_1 values is 3.30y with a range of $(-24.98, 32.70)$; the mean of the δ_2 values is 0.98y with a range of $(-27.79, 30.47)$. This allows us to define thresholds for 'extreme' DBA symmetrically, in terms of the standard deviation.

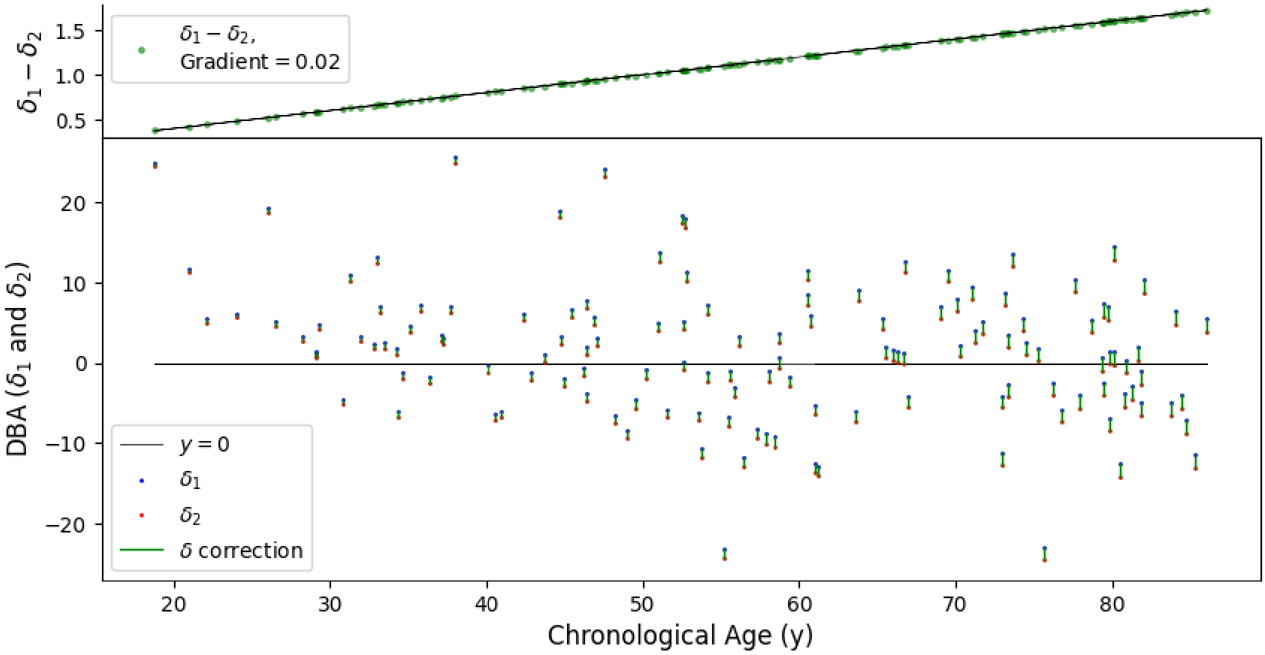


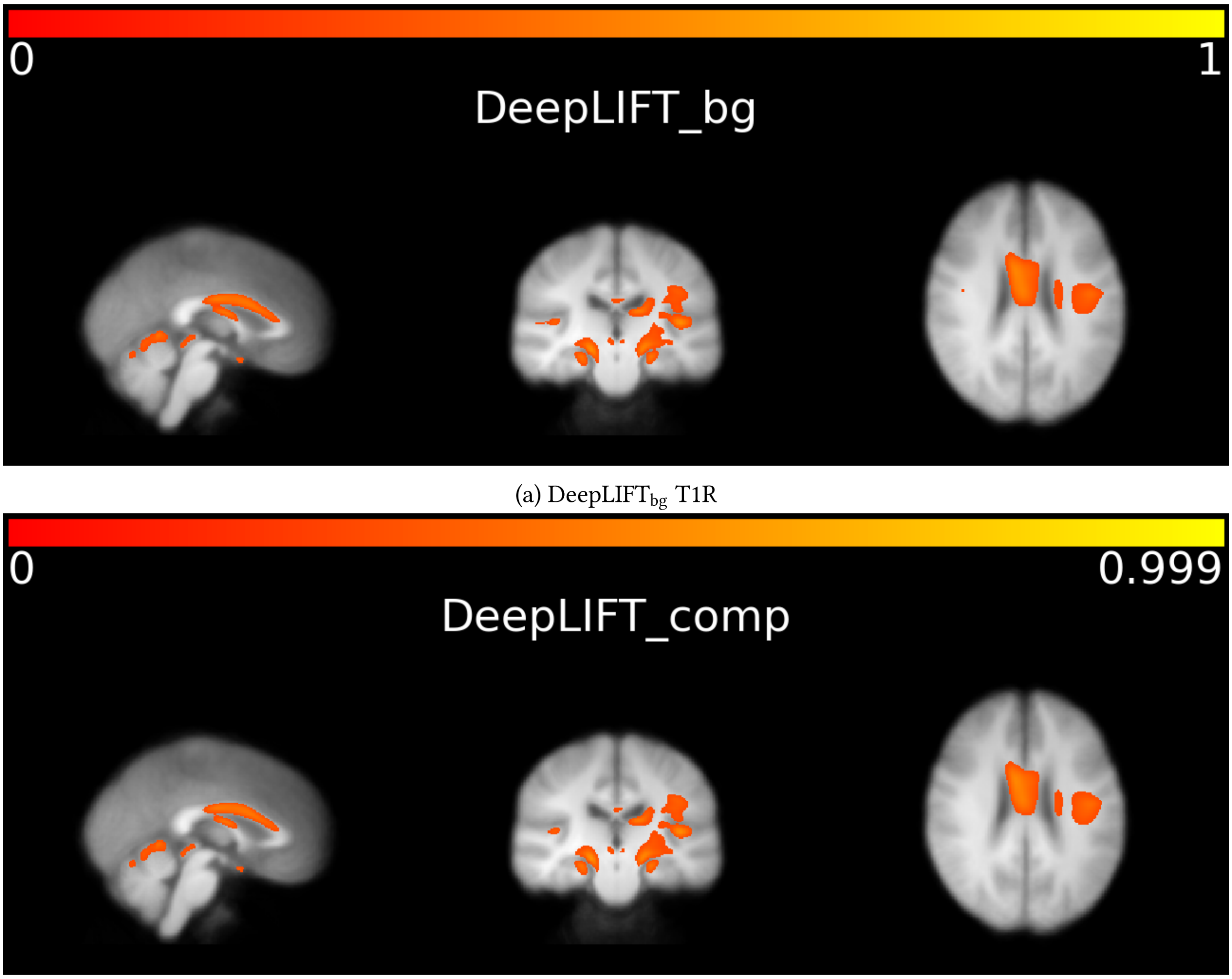
Figure 4.4: *Bottom*: DBA values δ_1 and δ_2 for the individuals in the test set, with the difference (green) increasing linearly in age. *Top*: The difference $\delta_1 - \delta_2 = YY^+\delta_1$ (see Section 2.7) as a function of chronological age. The linear increase illustrates the expected regression towards the mean. The small gradient of 0.02 shows that our model does not badly regress towards the mean.

The correction of individual DBA values is shown in Figure 4.4. The lower plot is a scatter plot of the DBA value pairs (δ_1 and δ_2 for each individual), ordered by chronological age. The correction ($\delta_1 - \delta_2$)

goes up linearly in age by Equation 2.30, and this difference is shown in green in the lower plot and in the upper plot. The linear increase in this difference is expected, and illustrates the extent to which the model regresses towards the mean (see Section 2.7). The regression of our model towards the mean is very moderate, considering the fact that the corrections to the DBA values are mostly small [144] (with a small correction gradient of 0.02) and that while our MAE of 6.55y is not small, we have a strong coefficient of correlation, $r = 0.89$.

4.2 Saliency Mapping

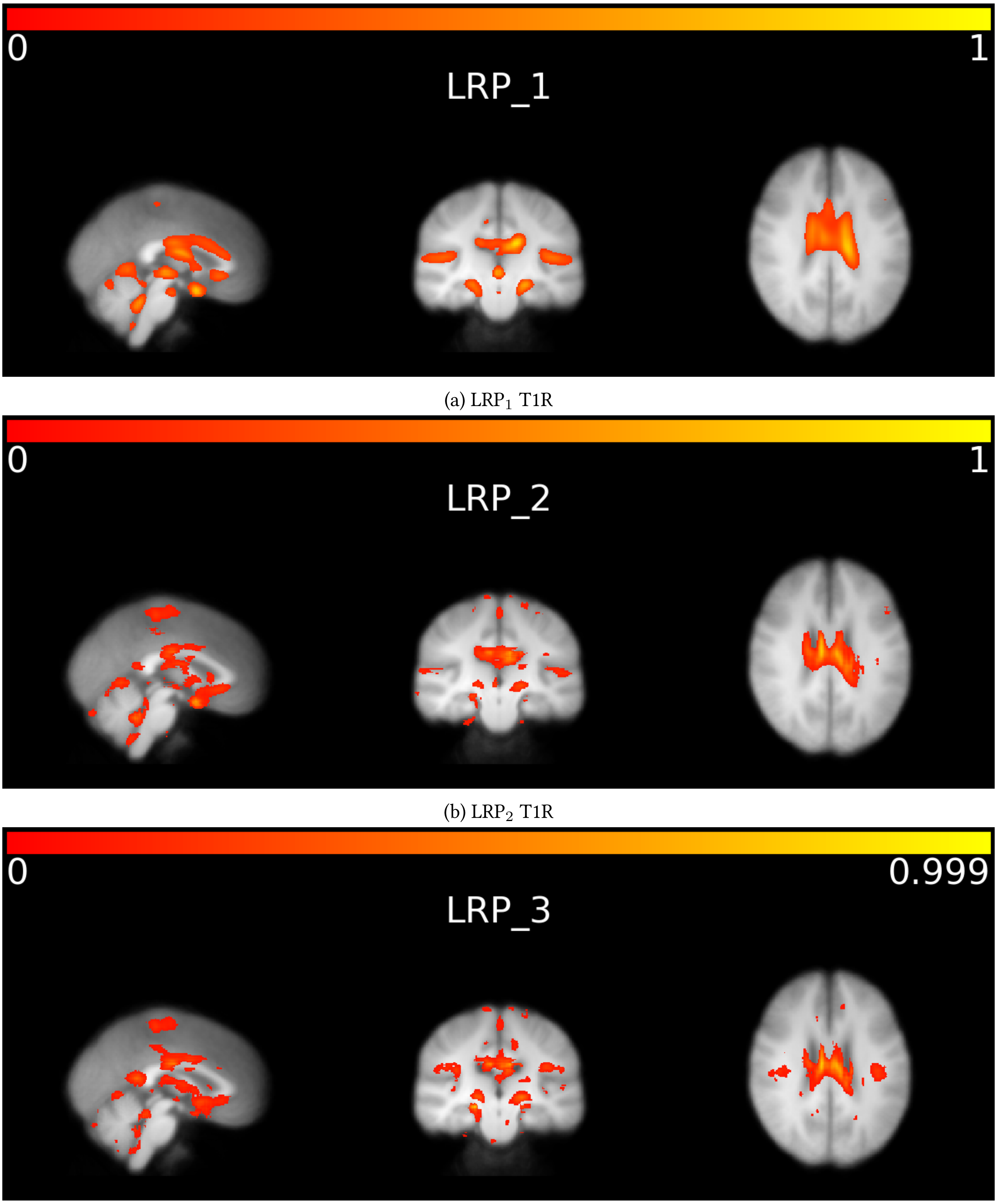
Saliency mapping was completed for the five methods on all individuals in the test set. Figures 4.5 and 4.6 show sections (slices) of the average T1R over the test set for each method, overlaid onto the composite MRI volume. The saliency maps and the composite volume correspond spatially, such that highlighted areas show population average T1R within the composite volume. This was done using FSLMaths and FSLEyes [1] tools. In Figure 4.7, we compare the T1R thresholded aggregated saliency maps for DeepLIFT$_{\text{comp}}$ and LRP1.



(a) DeepLIFT$_{\text{bg}}$ T1R

(b) DeepLIFT$_{\text{comp}}$ T1R

Figure 4.5: Coronal (left), transverse (middle), and mid-sagittal (right) sections of top-1% of relevance voxels from aggregated saliency maps for both DeepLIFT methods (yellow and red), overlaid on the composite test set volume (grey). The lateral ventricles (dark grey spaces in the center of each section of the composite volume) are surrounded by a high proportion of T1R.



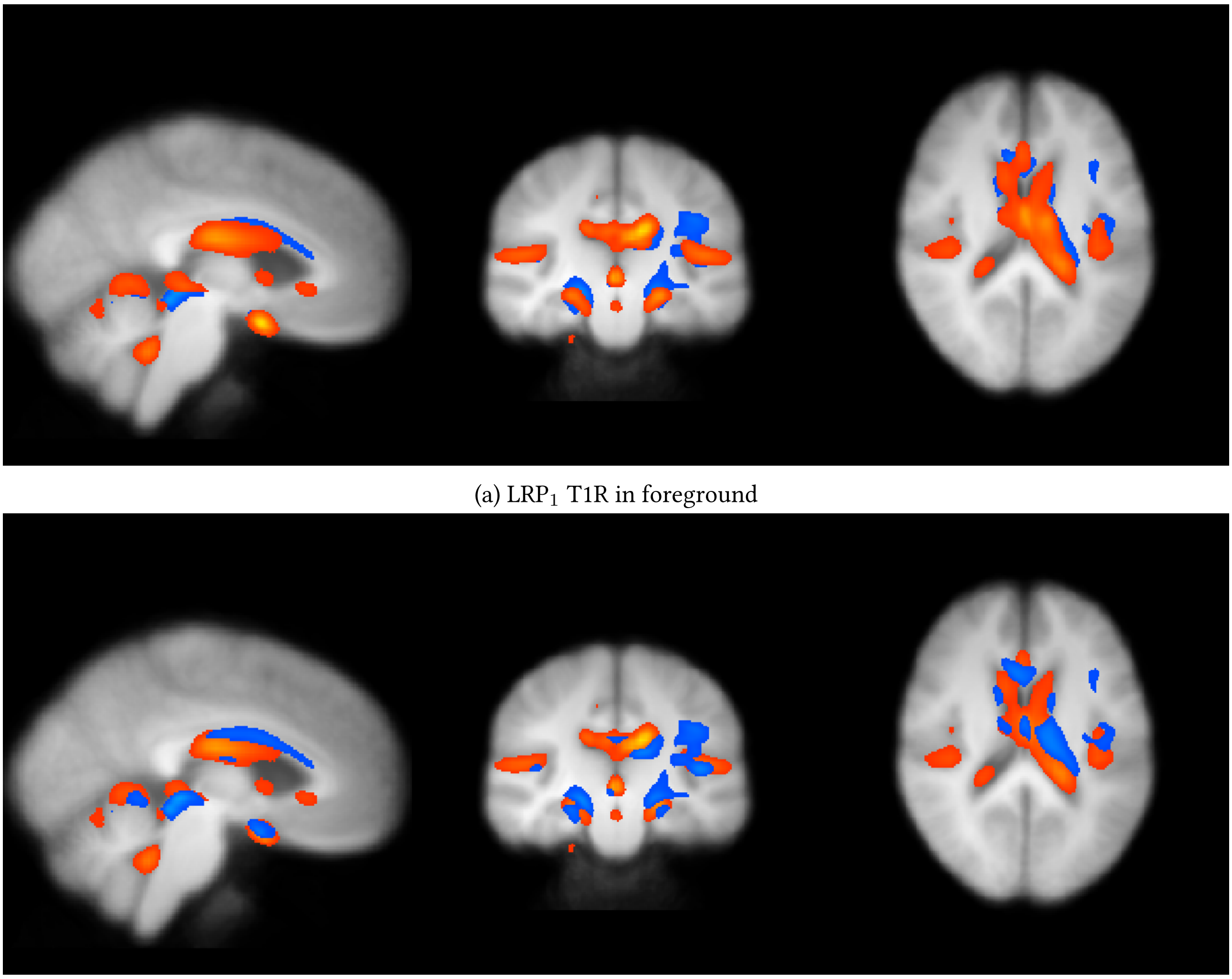
(a) LRP_1 T1R

(b) LRP_2 T1R

(c) LRP_3 T1R

Figure 4.6: Coronal (left), transverse (middle), and mid-sagittal (right) sections of top-1% of relevance voxels from aggregated saliency maps for each LRP method (yellow and red), overlaid on the composite test set volume (grey). The lateral ventricles (dark grey spaces in the center of each section of the composite volume) contain a high proportion of T1R.

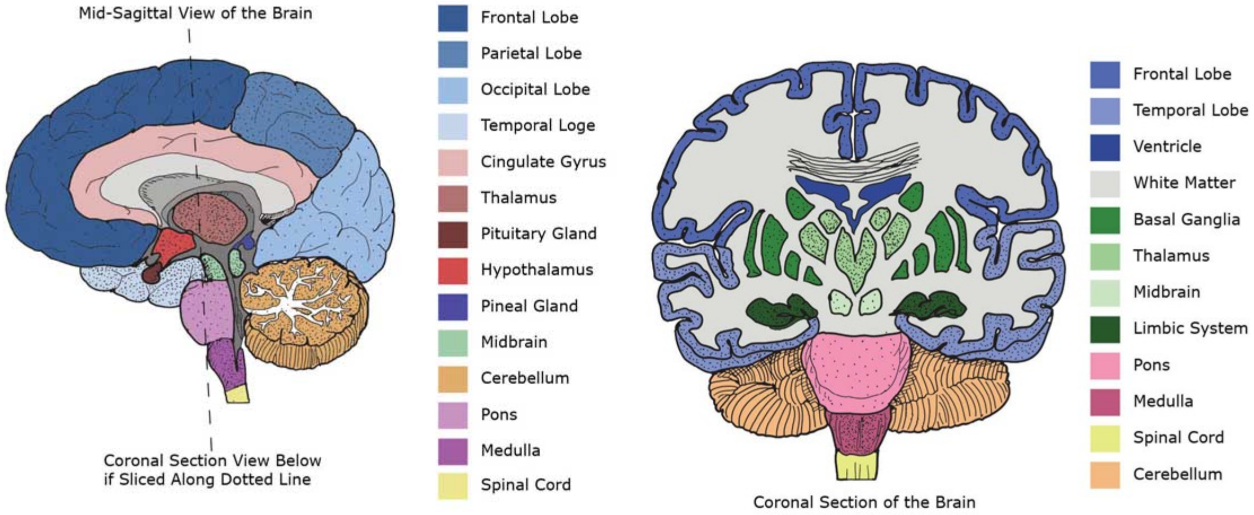
The permutation-based t-test yielded statistical significance in almost the entire brain volume ($p < 0.001$). Figure 4.8 shows two sections of a brain region atlas, illustrating the positions of several brain regions relevant to our discussion.



(a) LRP$_1$ T1R in foreground

(b) DeepLIFT$_{\text{comp}}$ T1R in foreground

Figure 4.7: Comparison between top-1% of saliency voxels for the aggregated maps of DeepLIFT$_{\text{comp}}$ (Blue) and LRP1 (Red-Yellow)



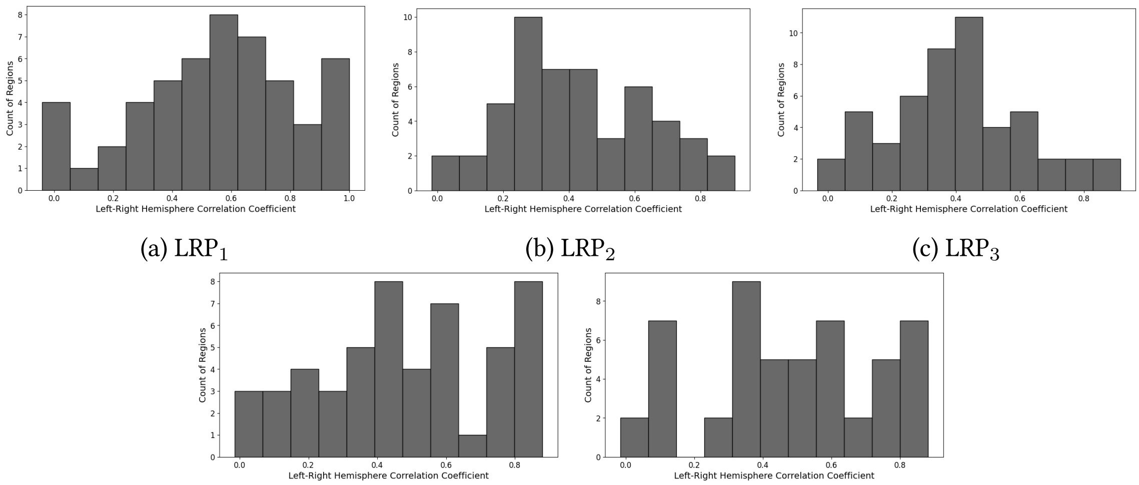
(a) Mid-sagittal section of brain region atlas.

(b) Coronal section of brain region atlas.

Figure 4.8: Labelled sections of brain atlases, showing some brain regions that we refer to. Images courtesy of article by Dr Matthew E. Bain (28-02-2011) in HealthPages.org

.

In our analyses of the saliency of individual regions[1], we average the proportion of T1R between hemispheres. This is common practice, and makes for less cluttered analyses. To justify this, we can analyse the correlation coefficients between the hemispheric region pairs, as in Figure 4.9. This shows the histograms for each method of correlation coefficients for T1R in each brain region. Every region has some positive correlation between left and right hemispheres, and those with the lowest correlation coefficients tend to be assigned T1R in very low proportion.



(a) LRP$_1$ (b) LRP$_2$ (c) LRP$_3$

(d) DeepLIFT$_{\text{bg}}$ (e) DeepLIFT$_{\text{comp}}$

Figure 4.9: Histograms of correlation coefficients of T1R between left- and right-hemispheric structures. Many regions have high correlation coefficients between hemispheres, and those that do not tend to be assigned very little relevance in total on average.

4.2.1 Relevance Assignment Between Methods

In Figures 4.6 and 4.5 we see many similarities and differences between the methodological assignment of T1R. There is a striking similarity between the T1R distributions of the two DeepLIFT methods. The LRP methods are very similar to one-another, although the clustering of T1R is increasingly noisy with greater values of the parameter α. This is to be expected, as we understand that larger values of α correspond to greater contrast in the resulting saliency maps. A key difference is that of assignment of T1R to the ventricles. The LRP methods tend to assign a high degree of relevance to the ventricles themselves, as shown in the overlay figures. The DeepLIFT methods on the other hand, tend to assign relevance not to the ventricles so much, but rather to the regions immediately surrounding the ventricles.

To assess quantitatively the differences in T1R distribution between methods, we examine for each method the proportion of each region containing T1R. This is shown in Figure 4.10. We see that largely the same downward trend is present across the methods, in descending order of T1R for LRP$_1$. The LRP methods assign T1R to a particularly large proportion of the ventricles; for example, over 10% of the lateral ventricle contains voxels in the top-1% of total brain age saliency by each of the LRP methods. While the DeepLIFT methods do so as well (with particular exception of the fourth ventricle, which is assigned almost no T1R for either DeepLIFT method), it is not nearly to the same extent. Large proportions of T1R are assigned by all methods to the surrounding limbic regions such as the caudate nucleus, hippocampus, thalamus, diencephalon, and parahippocampal gyrus. By the analysis of our domain expert, these are areas known to be involved in brain ageing. DeepLIFT tends to attribute greater proportions of T1R to these regions however, especially the parahippocampal gyrus. The parahippocampal gyrus is located near the fourth ventricle, and DeepLIFT tends to assign almost no T1R here; the disparity in these two regions between the LRP and DeepLIFT methods may be explained by DeepLIFT's proclivity to highlight the areas surrounding the ventricle as opposed to the ventricle itself.

[1] Apart from individual region-wise saliency trajectories

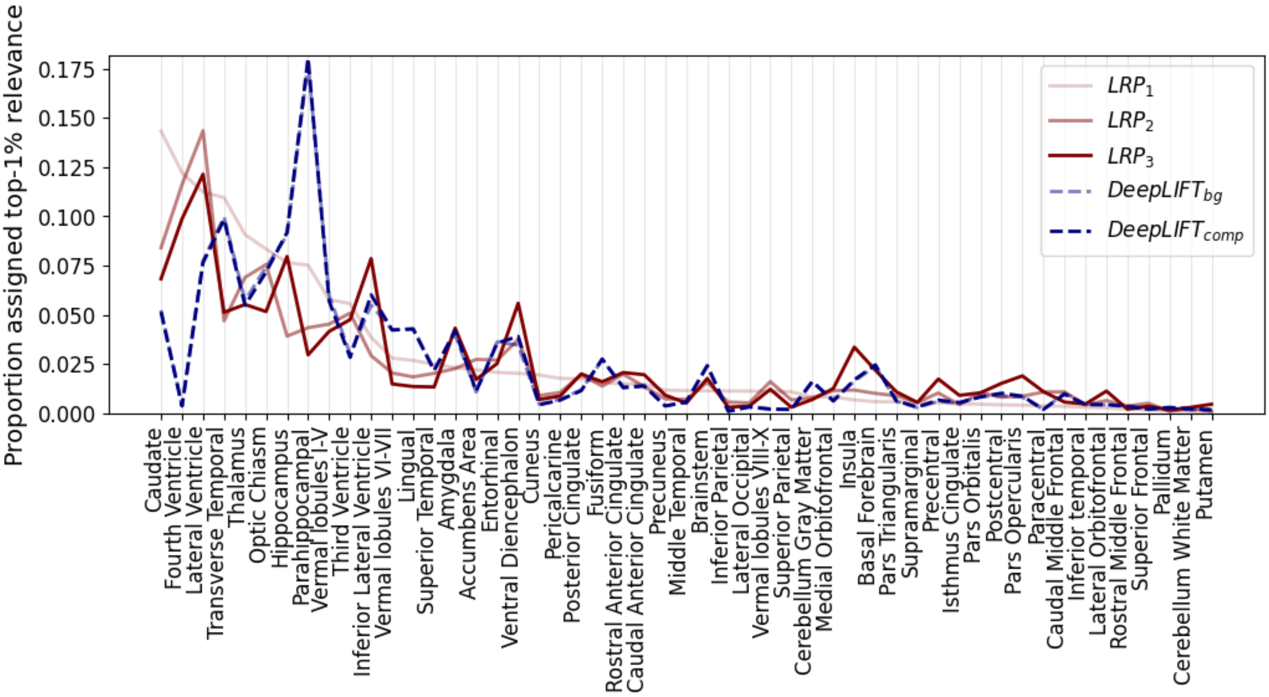


Figure 4.10: Proportion of each region assigned Top-1$\%$ Relevance per method. Regions are ordered in descending proportion of T1R as determined by LRP$_1$.

We refer the reader again to Figure 4.8 for reference to the positions of some regions within the brain. One region that is universally assigned T1R in large proportion is the transverse temporal gyrus (in the temporal lobe). This is to be expected according to our domain expert, as this structure comprises part of the auditory cortex, and is known to play a role in healthy ageing [152]. The optic chiasm is assigned T1R in large proportion too. Grey-matter-dense structures such as outer cortical regions and the vermal lobules of the cerebellum tend to be assigned T1R in large proportion. It is interesting to note however that the cerebellum white matter is universally assigned T1R in very low proportion. Also, as partly illustrated in Figures 4.6 and 4.5, there is generally very little T1R assigned to white matter regions in the saliency maps for the T1-weighted volumes.

Between the LRP methods there is a high degree of similarity in assignment of T1R to brain regions. There is an even higher degree of similarity between the DeepLIFT methods; Figure 4.10 shows that the curves of the two DeepLIFT methods almost perfectly overlap. There is a high degree of consistency between DeepLIFT and LRP in regional BA relevance assignment, apart from the major disparity in assignment between the ventricles and the parahippocampal gyrus.

4.2.2 Relevance Assignment for Large DBA

Since thresholding DBA for BA analysis has not been performed before in the literature as far as we could find, we chose for our experiment the simple DBA threshold value $\delta^* = \sigma$, where σ is the standard deviation of the quantity δ_2 in the test set. In the test set, we had $\sigma = \delta^* = 11.58$y. In Figures 4.11-4.15, we show the effect of large DBA on regional T1R distribution in older (> 50) and younger (< 50) individuals. Again we have averaged the relevance between the two hemispheres.

Several highly relevant structures show significant change in the proportion containing T1R between the older and younger individuals with high DBA, and as compared to the individuals with moderate DBA. One such region is the thalamus; all methods assign significant relevance to the thalamus in the younger high-DBA group, less relevance to the baseline group, and less still to the older high-DBA group. In the LRP methods, the same is true for the fourth ventricle.

The basal forebrain and (in the case of LRP) the transverse temporal gyrus tend to have the opposite trend, whereby more relevance is assigned to older high-DBA individuals, and less to younger high-DBA individuals.On the other hand, some structures tend not to change much in their assignment of relevance due to

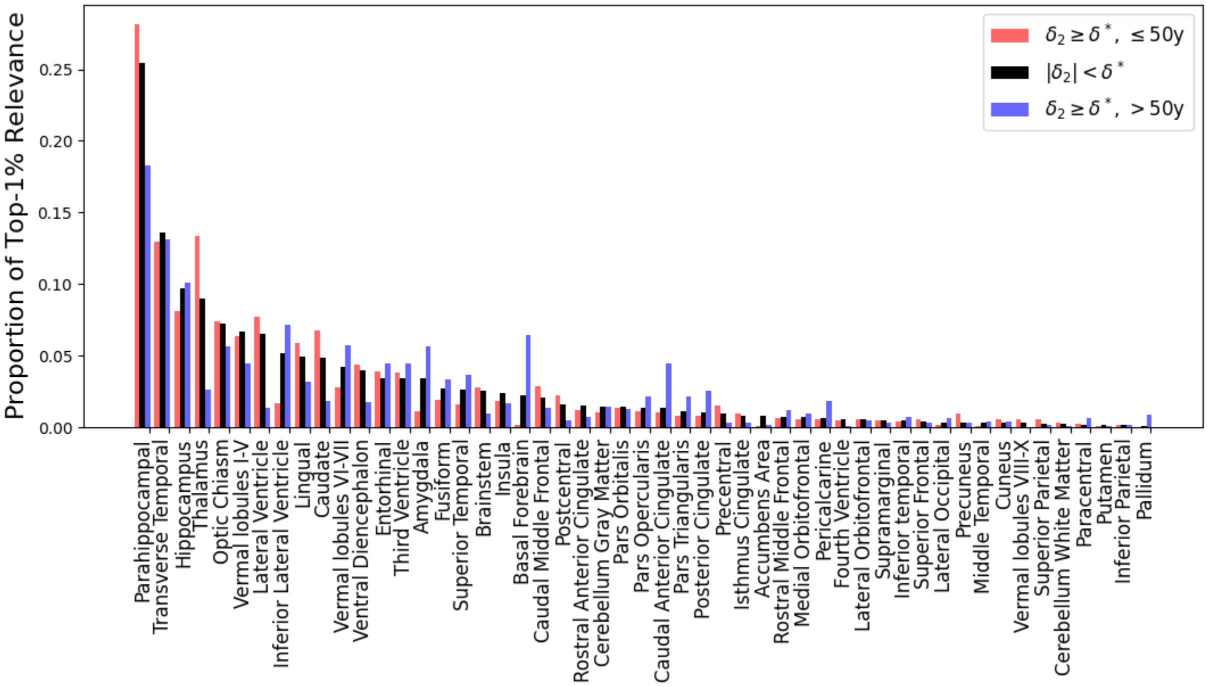


Figure 4.11: Distributions of Top-1% Relevance, via DeepLIFT$_{\text{bg}}$, in young ($\leq$ 50y) individuals with high DBA ($\delta_2 > \delta^*$), elderly individuals ($>$ 50y) with high DBA, and individuals with small-to-moderate DBA ($|\delta_2| < \delta^*$). Regions are ordered by descending proportion of T1R in the small-to-moderate DBA group.

high DBA or age. In the case of both DeepLIFT methods, the transverse temporal gyrus does not change in relevance to a great degree from one group to another.

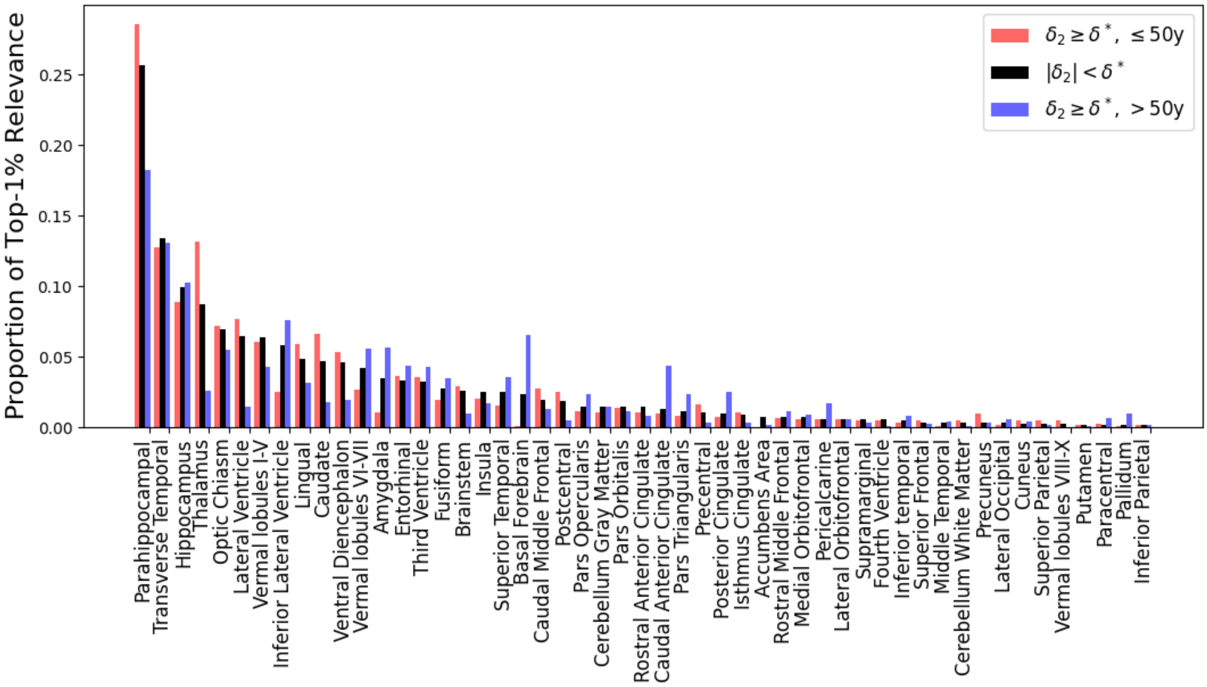


Figure 4.12: Distributions of Top-1% Relevance, via DeepLIFT$_{\text{comp}}$, in young ($\leq$ 50y) individuals with high DBA ($\delta_2 > \delta^*$), elderly individuals ($>$ 50y) with high DBA, and individuals with small-to-moderate DBA ($|\delta_2| < \delta^*$). Regions are ordered by descending proportion of T1R in the small-to-moderate DBA group.

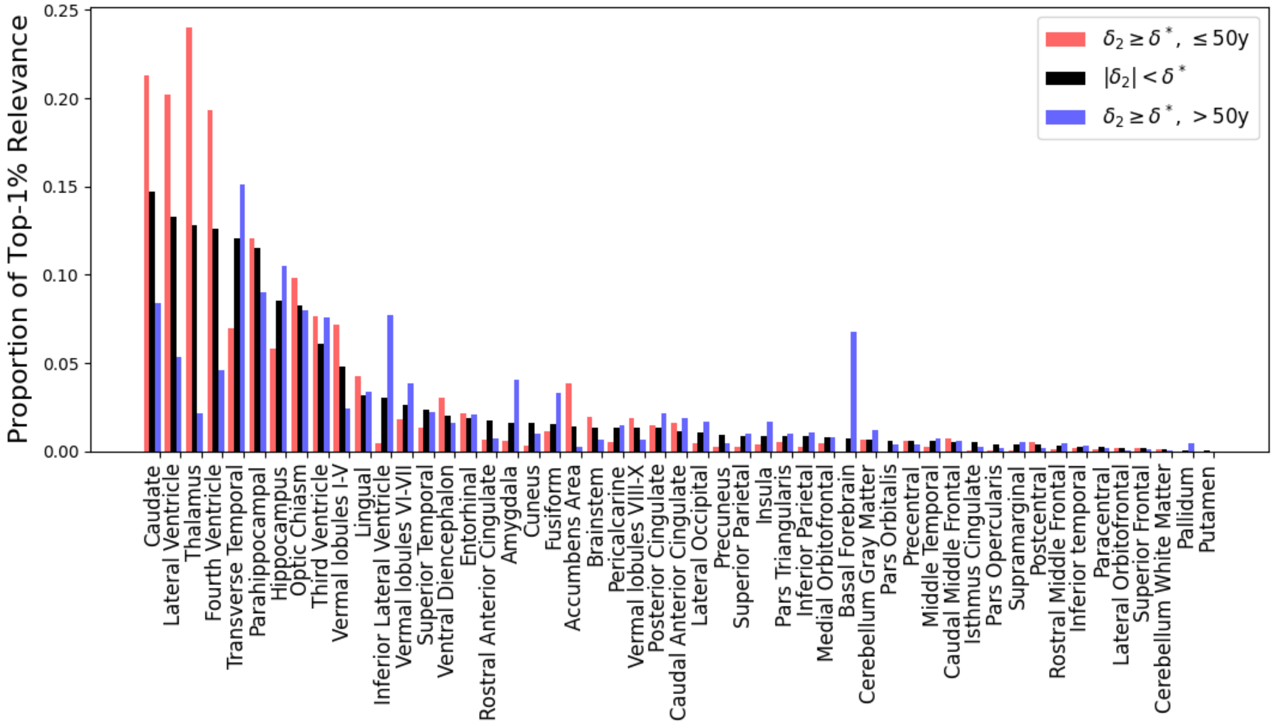


Figure 4.13: Distributions of Top-1% Relevance, via LRP$_1$, in young ($\leq$ 50y) individuals with high DBA ($\delta_2 > \delta^*$), elderly individuals ($>$ 50y) with high DBA, and individuals with small-to-moderate DBA ($|\delta_2| < \delta^*$). Regions are ordered by descending proportion of T1R in the small-to-moderate DBA group.

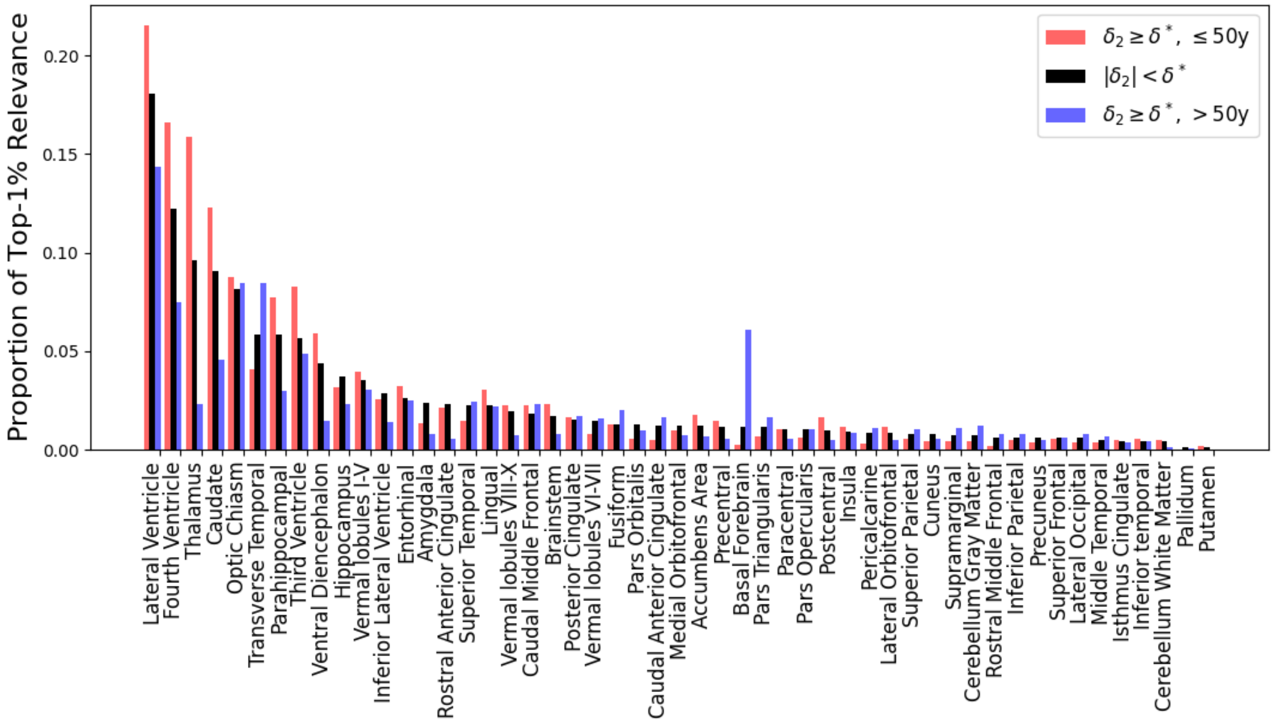


Figure 4.14: Distributions of Top-1% Relevance, via LRP$_2$, in young ($\leq$ 50y) individuals with high DBA ($\delta_2 > \delta^*$), elderly individuals ($>$ 50y) with high DBA, and individuals with small-to-moderate DBA ($|\delta_2| < \delta^*$). Regions are ordered by descending proportion of T1R in the small-to-moderate DBA group.

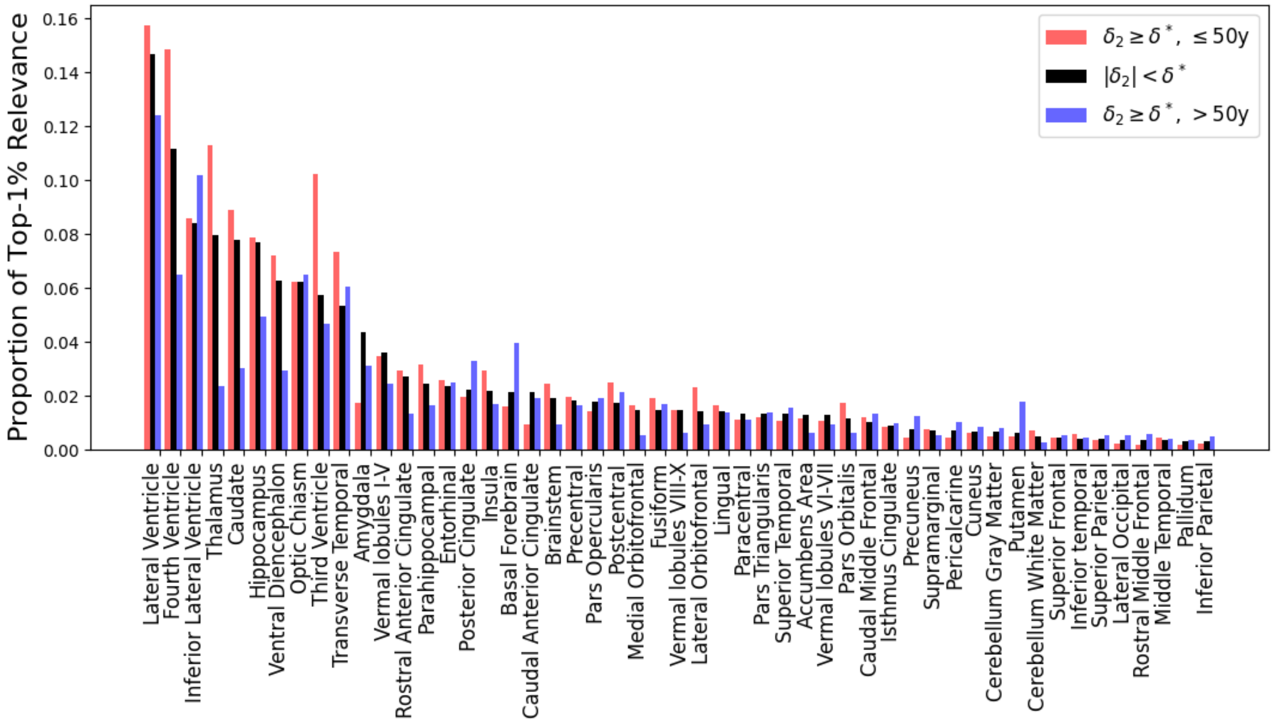


Figure 4.15: Distributions of Top-1% Relevance, via LRP$_3$, in young ($\leq$ 50y) individuals with high DBA ($\delta_2 > \delta^*$), elderly individuals ($>$ 50y) with high DBA, and individuals with small-to-moderate DBA ($|\delta_2| < \delta^*$). Regions are ordered by descending proportion of T1R in the small-to-moderate DBA group.

4.2.3 Relevance Assignment Across Age brackets

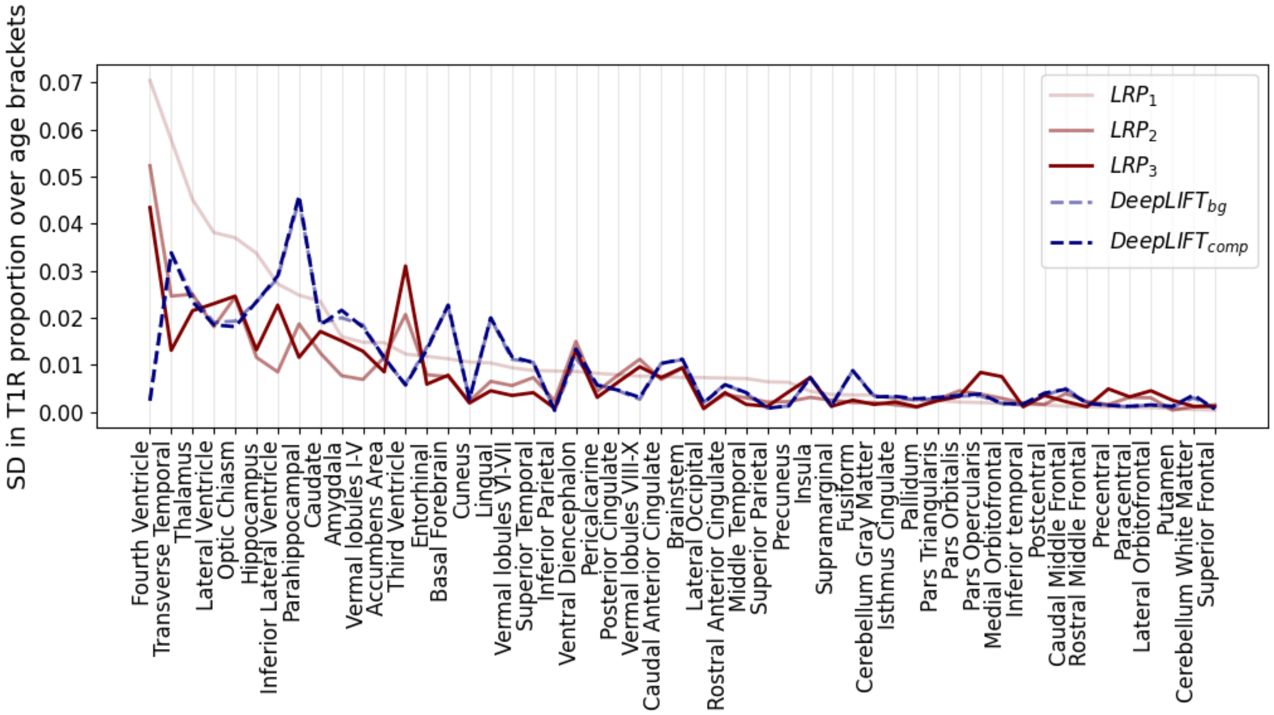


Figure 4.16: Standard Deviation in regional proportion of T1R over the age brackets for each method. Regions are ordered in descending SD of LRP$_1$.

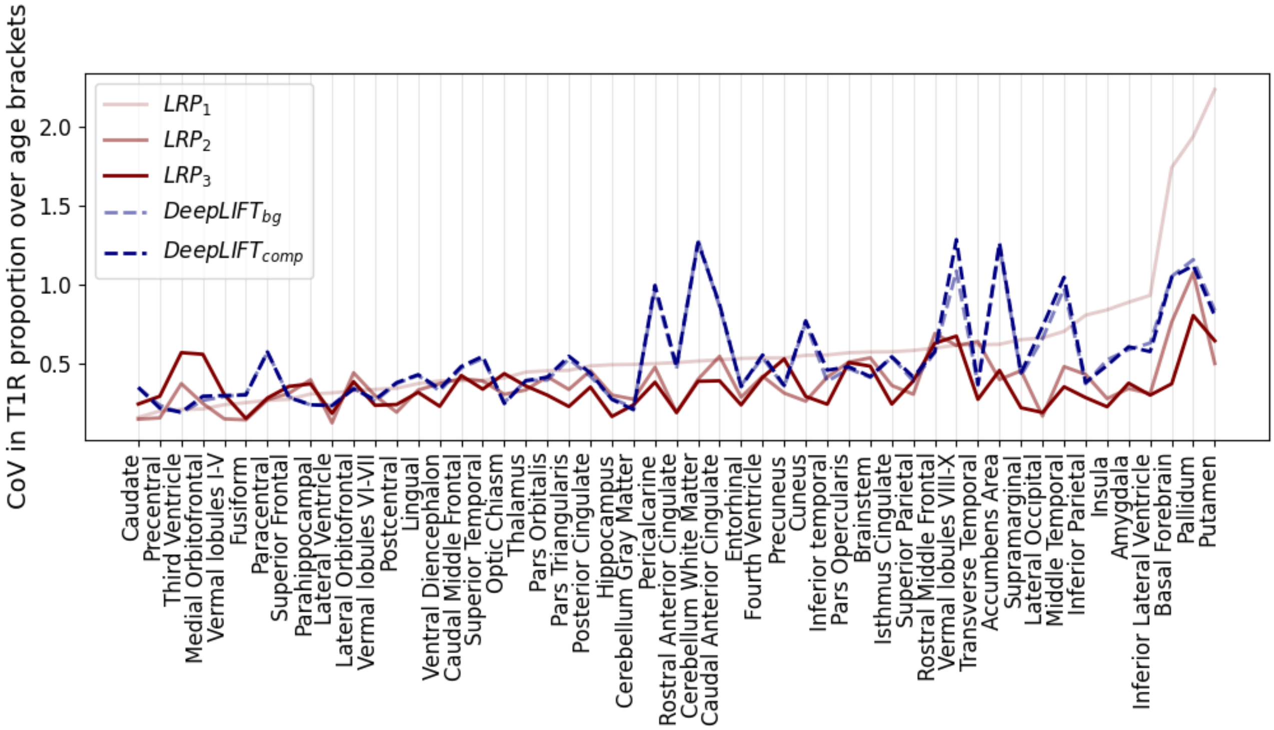
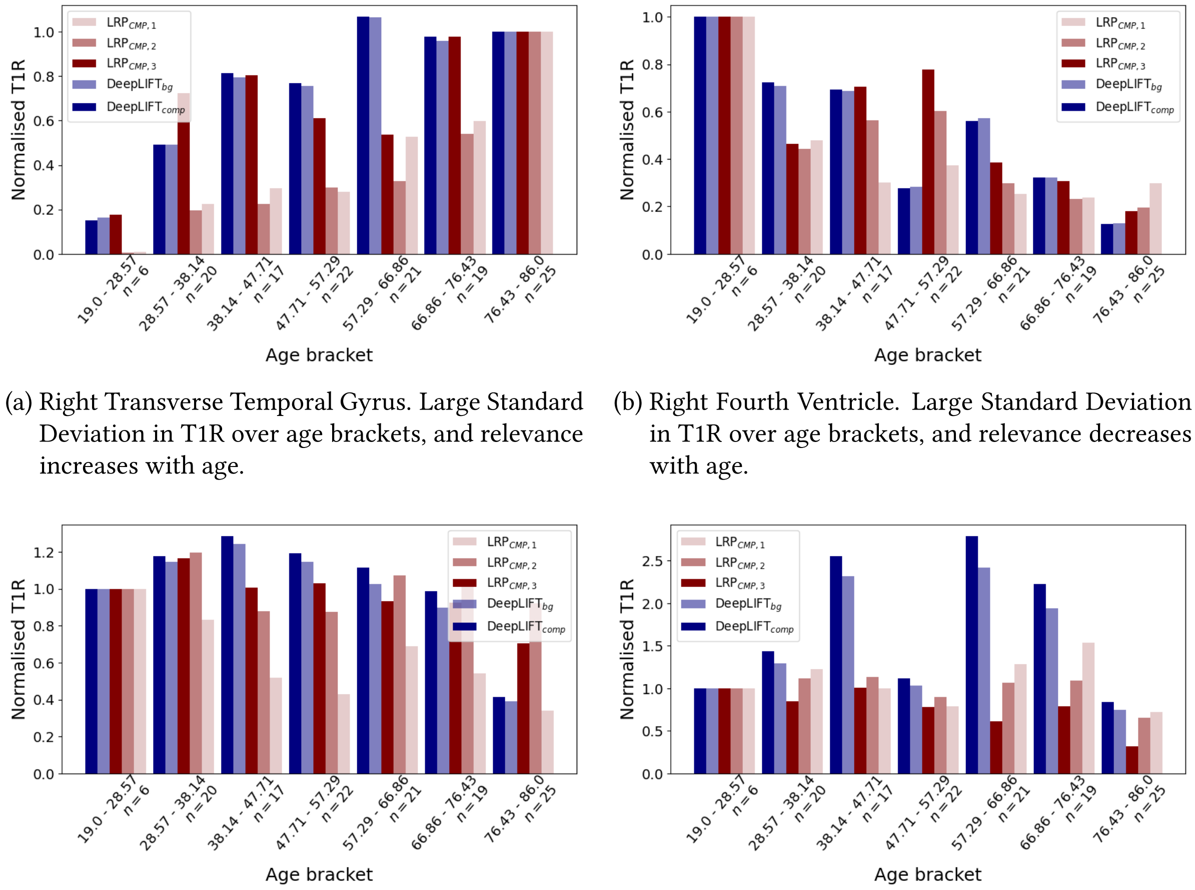


Figure 4.17: Coefficient of Variation in regional proportion of T1R over the age brackets for each method. Regions are ordered in ascending CoV of LRP$_1$.



(a) Right Transverse Temporal Gyrus. Large Standard Deviation in T1R over age brackets, and relevance increases with age.

(b) Right Fourth Ventricle. Large Standard Deviation in T1R over age brackets, and relevance decreases with age.

(c) Right Lateral Ventricle. Low CoV in T1R over age brackets.

(d) Right Caudate Nucleus. Low CoV in T1R over age brackets.

Figure 4.18: 'Proportion' of Top-1% Relevance per method over the age brackets. The proportion of T1R is normalised in each sub-figure such that either the youngest or oldest group have T1R assignment of 1. Brackets all have an age range of 9.57y. Four example regions are shown.

Trajectories of BA relevance are created for each region of the brain, for each method, by grouping individuals into seven bins of equal age range and determining within each bin the average proportion of the region is assigned T1R. Figures 4.16 and 4.17 show the Standard Deviations and Coefficients of Variation respectively of the proportion of T1R over these age brackets within each region of the brain. We argue the case for using two separate metrics for the greatest and least change in relevance in Section 3.2.5. We are interested in regions that have a high proportion of T1R but either change a lot over age brackets or change very little. Regions with small SDs tend to have very little relevance assigned overall, and the same is true for regions with high CoVs. An example of such a structure is the putamen. The CoV of the putamen is very high, while its SD is very low. This is due to its generally low proportion of T1R, and because of this low proportion, we are not interested in the small fluctuations in proportion of T1R.

Figure 4.18 shows the change with age in relevance assignment to four example regions in the test set. We normalise the proportion of assigned T1R such that the relevance of either the oldest or youngest bracket is unity for all methods in each region. Figure 4.18a shows the assignment of T1R to the right transverse temporal gyrus over the age brackets. For each of the methods we see a clear upward trend in relevance with age. Figure 4.18b shows the assignment across age brackets for the right fourth ventricle. We see in this case that the proportion of the region assigned T1R decreases with age. Figure 4.18c shows the age bracket distribution of T1R for the right lateral ventricle. Apart from in the final age bracket, this region maintains a uniform proportion of assigned T1R, especially by means of LRP_2 and LRP_3. In Figure 4.18d we see a similar trend for the right caudate nucleus. With the exception of some middle-aged to older age brackets by DeepLIFT, there is a very uniform distribution of assigned T1R across ages.

4.3 Summary of Results

We trained a BA regression model on a small dataset to a test MAE of 6.55y. While this is not SOTA, we note that the coefficient of regression achieved is high ($r = 0.89$), and no other work has used a dataset of this size.

The saliency mapping techniques were largely similar in their distribution of T1R within the brain volume, apart from a marked difference between LRP and DeepLIFT in allocation of T1R in and around the ventricles. While LRP tends to assign T1R to the ventricles in high proportion, DeepLIFT does so to the areas immediately surrounding the ventricles such as the parahippocampal gyrus.

We found that high-DBA individuals have T1R distributions that can vary greatly from those of medium-to-low-DBA individuals. Furthermore the direction and extent of this variation can be age-dependent. It was not the case that highly-relevant regions only increased in significance with age and high DBA. Indeed, many regions are assigned greater relevance in the younger high-DBA group than the older high-DBA group.

We found that regions with high SDs in age-bracket assignments of T1R tend either to increase in T1R proportion with age or decrease in T1R proportion with age. We found also that regions with low CoV in age-bracket relevance assignment tend to have uniform proportion of T1R across age brackets. It was not the case that highly-relevant regions only increased in relevance with age. In fact, there were many such regions which decreased in the proportion of assigned T1R with age.

Chapter 5

Discussion

The goal of this study was to train an accurate BA regression model from which we could create saliency maps to extract meaningful data about brain ageing. In this chapter we examine our results and discuss the extent to which each of our goals was met, and how the results answered our research questions.

5.1 BA Regression Model

Newer architectures were considered for the task of BA regression, such as the Tensor Regression Networks of Kossaifi et al. [153]. Since, however, they have not previously had saliency mapping techniques applied to them, the high-performing ResNet model was employed.

The hyper-parameter tuning phase indicated that for our model the best optimiser was Adam [164]. Some current BA regression literature makes use of this [18], while some others use RMSProp [145] or stochastic gradient descent methods [49, 17]. There does not seem to be a favoured optimiser in the literature. The loss function that was indicated as best was the MSE. Although the MAE is used commonly as the standard metric by which BA regression performance is measured, some authors choose to use MSE for the loss function in the task [145, 18]. The MSE and MAE seem to be the two most commonly used loss functions for the BA regression task. A major difference between the two is that the MSE penalises large errors to a much greater extent than the MAE, and small errors to a lesser extent. Peng et al. [146] use the Kullback-Leibler divergence loss to great success, which was not considered in the hyper-parameter optimisation. Performance was drastically improved by the use of a decaying learning rate as compared to a constant learning rate (which yielded an MAE of $>$ 9y on the test set), as well as increasing the batch size to 4 by many means of conserving RAM. Most methods used for conserving RAM are not discussed to a great extent in the literature. Instead, the methods we used come from advice given in online forums.

The hyper-parameter tuning phase was limited in the number of hyper-parameters that were considered. It may be the case that tuning other features of the model, such as the number and size of residual blocks would produce more accurate models. Furthermore, only one model was trained with the chosen hyper-parameters. It may easily be the case that many such models may perform better with similar training times but different random parameter initialisations. It would be advisable to train multiple models for the task in the future, and either select the best model, or create an ensemble model like Levakov et al. [17] and Hofmann et al. [18]. Our model's performance is compared to some other key works in BA regression in Table 5.1, taking dataset size into consideration.

The regression model was successfully trained to a correlation coefficient of $r = 0.89$ (strong correlation). While the MAE of 6.55y is not SOTA, no BA model in the literature has utilised a dataset of this size to this level of accuracy. This would indicate that while very large datasets may be necessary for SOTA performance, they may not be necessary simply to train a BA regression model of reasonable accuracy, such as our strong correlation of $r = 0.89$. As stated in Section 2.8 it can be prohibitively difficult to access large MRI datasets, and working with such datasets introduces considerable computational overhead. It may be preferable therefore for some researchers to use smaller datasets if SOTA accuracy is not necessary.

Our main reason for using such a small dataset was for ease of computational burden. Compared to other

Author	Ours	Cole et al. [49]	Jónsson et al. [156]	Kossaifi et al. [153]	Levakov et al. [17]	Hofmann et al. [18]	Peng et al. [146]
Dataset Size	656	2001	12378	19100	10176	2016	14503
Test MAE (y) on T1 Volumes	6.55	4.65	4.00	2.69	3.02	3.95	2.14

Table 5.1: The test MAE performances of several key BA regression studies in comparison with ours, considering the size of datasets used.

works using datasets of thousands or even ten of thousands of scans, this work can be seen as under-powered. Indeed, using a larger dataset would allow not only for better model performance, but statistically stronger saliency mapping results.

5.2 BA Regional Saliency

Regions which were expected to be of greatest saliency towards BA by our domain expert were the ventricles and surrounding regions, grey-matter-dense regions such as the outer cortex of the cerebrum and the cerebellum, and frontal and temporal structures of the brain. As expected, a significant proportion of T1R was assigned to the ventricles, especially by LRP. This is in accordance with not only the expectations of our domain expert and the relevant literature [11, 142], but also the findings of Levakov et al. [17] and Hofmann et al. [18].

One region that is universally assigned T1R in large proportion is the transverse temporal gyrus (in the temporal lobe). This agrees with the expectations of our domain expert, and is known to be affected by healthy ageing [152], and partly comprises the auditory cortex. Grey-matter-dense areas such as the vermal lobules of the cerebellum were also assigned high proportions of T1R. This is in accordance with many medical findings that grey matter density decreases with age, starting from late adolescence [140, 141, 142]. On the other hand, it was found that white matter regions generally were not assigned high proportions of T1R. This is despite the fact that white matter lesions are well-known markers of brain ageing [140, 11, 141, 142]. The lack of relevance assigned to white matter regions is likely due to the fact that we used T1-weighted volumes in our experiments. As discussed in Section 2.1, T1-weighted volumes are less telling of white matter lesions than T2-weighted volumes. The optic chiasm is assigned T1R in large proportion too. This was not expected by our domain expert; however, it agrees with the findings of Levakov et al. [17], due to the position of the optic chiasm within the interpeducular cistern. Many limbic structures such as the amygdala, caudate nucleus, thalamus and hippocampus were assigned T1R in large proportion. This is in agreement with the expectations of our domain expert and previous findings of brain age with regard to the limbic system by Gunbey et al. [143].

5.3 BA Saliency Mapping Methods

We implemented two DeepLIFT saliency mapping techniques and three LRP techniques for the BA regression task. For DeepLIFT we used a reference input of MRI background activations (voxel values of 0) and a composite MRI volume formed from all of the test set MRI volumes. For LRP, we used the composite method with $\alpha = 1, 2, 3$.

We analysed the regional distribution of top-1% of total brain volume relevance (T1R) to evaluate the utility of the saliency mapping methods. All methods strongly highlighted brain regions known to be key contributors to BA pathology. The ventricles and their surrounding regions were consistently assigned a very high proportion of T1R (except for the fourth ventricle by DeepLIFT), which agrees strongly with the findings of Levakov et al. [17] and Hofmann et al. [18], as well as the expectations set by the medical literature [11, 142]. Grey-matter-dense regions and limbic regions were also assigned a high proportion of T1R, especially by DeepLIFT. This too is in concordance with the medical literature [140, 141, 142, 143], but no previous BA regression analysis has produced such strong relevance in the limbic system. Since DeepLIFT has never been used for the BA regression task before in the literature, it is not too surprising that we find a different relevance distribution. Indeed, this may call for the implementation of other saliency mapping methods to this task in the future, to examine other possible differences in saliency distribution.

Although the trend of regional T1R distribution is largely similar among the methods used, there were some notable differences. The biggest difference was that the LRP methods assigned most relevance to the ventricles, while the DeepLIFT methods assigned most relevance to the regions immediately surrounding the ventricles. This trend is shown clearly in Figure 4.7, where the aggregated T1R maps from DeepLIFT$_{\text{comp}}$ and LRP$_1$ are compared. Initially this was thought to be due to the masking-like effect of the DeepLIFT methods, whereby contribution scores are computed as the direct product of the input difference-from-reference and the multipliers, $C_{\Delta x \Delta t} = m_{\Delta x \Delta t} \Delta x$. Since the values of the voxel activations of the ventricles tend to be close to 0 ('dark' CSF-filled volumes), this was thought possibly to mask out a significant amount of relevance assigned by the multipliers. Upon inspection of the multipliers, however, it was revealed that the same distribution emerges, and a masking effect did not take place.

While LRP assigns very high proportions of T1R to the fourth ventricle, for example, the DeepLIFT methods assign almost none here, and instead assign the highest proportion of T1R to the parahippocampal gyrus, which is very close. While LRP focuses on the dilation of the ventricles with age, it would appear that DeepLIFT focuses more on the atrophy of the surrounding regions with age. Interestingly, this would indicate that the model has developed a different concept of the CSF-filled ventricles as compared to other structures in the brain, since these are the only areas showing such great disparity in the different explanation methods. The explanations offered by DeepLIFT tend to highlight solid structures more favourably (non-CSF structures), while the LRP methods distribute relevance both to the solid brain matter and the CSF-filled ventricles. This does not appear to be a matter of choice of reference image for DeepLIFT, since the relevance distributions are almost identical for the different choices we have implemented.

While in the literature, LRP and DeepLIFT have been compared for classification and detection tasks [92, 82], we have not found any comparisons between the two in the case of regression. The two methods act similarly under classification and regression circumstances, since they both tend to highlight the relevant area of an input data point (an image for example) to the output. In the case of regression however, relevance must be distributed throughout the entire input region (at least, the parts of the input region that are subject to change from one data point to another) [165].

Given that the LRP and DeepLIFT methods both perform well by assessment of our domain expert, and that the major differences between the two are perspectival (ventricles versus surrounding regions), neither is necessarily more suitable toward the task of BA explanation than the other in this context. By way of recommendation, it may be best to employ both LRP and DeepLIFT methods towards BA explanation, as the differences in explanations can be complementary. While LRP highlights the ventricles for example, DeepLIFT highlights the exact areas of brain matter that recede with age. The DeepLIFT methods are highly consistent in their assignment of T1R. The distributions across regions are almost identical for the two reference inputs. Neither of the DeepLIFT methods is preferable over the other in this sense, and so a choice of reference input is at the user's discretion. In light of the DeepLIFT author's comments on what would constitute a reasonable choice of reference input [87], this would indicate that both reference inputs serve their function well. The LRP methods are also highly consistent, but not to the same degree. It seems that the disparities in T1R distributions between the LRP methods lies in the fact that for $\alpha \in \{2, 3\}$, LRP has very high contrast, and the saliency map volumes become somewhat noisy, even when thresholded for T1R. To this end, it may be best to recommend $\alpha = 1$ as the parameter of choice when utilising LRP$_{\text{CMP}}$. This will produce less noisy and generally more visually appealing saliency maps [69]. This is the most commonly used form of LRP in the current literature [73, 74, 71, 75, 16, 76, 18].

5.4 BA Relevance for Large DBA

With a threshold DBA value of $\delta^* = \sigma = 11.58$y, we showed that large DBA ($\delta_2 > \delta^*$) corresponds to significant changes in the proportion of T1R in many brain regions, and that this change can be age-dependent. The only examination in the literature that could be found of the relationship between DBA and BA relevance was that of Hofmann et al. [18]. Their findings showed that voxel clusters associated with large DBA corresponded spatially with increased relevance. However, only an older cohort was analysed ($\geq$ 50y).

We know that large relevance assignment to an area indicates that the area is predictive of accelerated BA. For individuals with large DBA, differences in proportional assignment of T1R to regions should indicate which regions are contributing most to the discrepancy. We showed that not only is the distribution of T1R different for those with large DBA and those with moderate DBA, but that the difference is often dependent on age-group. This was contrary to the findings of Hofmann et al.[18], since we found that in many regions lower T1R was assigned in older subjects with large DBA. Indeed this is the case for many regions via each method. Two such regions are the parahippocampal gyrus and the thalamus.

We found for example that both the thalamus and the fourth ventricle had a significantly greater proportion of assigned T1R in the younger high-DBA group ($\sim$ 11% and $\sim$ 15% respectively) than the low-to-moderate DBA group ($\sim$ 8% and $\sim$ 11% respectively), and a significantly lower proportion in the older high-DBA group ($\sim$ 2% and $\sim$ 6% respectively). Since increased relevance is predictive of higher DBA, this would suggest that in younger individuals, the thalamus and fourth ventricle are each more telling of accelerated brain age (and hence increased DBA) than in older individuals. We also found that the basal forebrain and (in the case of LRP) the transverse temporal gyrus tend to have more relevance assigned in older high-DBA individuals ($\sim$ 7% and $\sim$ 8% respectively), and less in younger high-DBA individuals ($<$ 1% and $\sim$ 4% respectively). This would suggest that these regions are more telling of accelerated BA in older individuals than in younger individuals. Severe neuronal loss has been found to occur in the basal forebrain in Alzheimer's Disease patients [166, 167]. Since not all elderly people experience Alzheimer's Disease, it may be the case that DBA is contributed to by the degradation of the basal forebrain only in some elderly individuals. This could partially explain why there is a marked increase in relevance in the region in elderly subjects with high DBA. The transverse temporal gyrus has been shown to be affected by healthy ageing [152], and this may be due in part to its partial comprising of the auditory cortex. Not all people experience dramatic hearing loss with age, but it is associated strongly with the onset of dementia and cognitive dysfunction [168]. This may be part of why there is an increase in relevance of the transverse temporal gyrus in older individuals with high DBA, as we would expect that degradation of the area would entail some attenuation in hearing, which we know to be associated with neurodegeneration. From these findings, it would appear that regional relevance can be more informative of DBA at some ages than at others.

In all of our analyses of saliency map data, we are faced with the issue of having only 132 subjects. It would be advisable in future to perform saliency mapping on the entire dataset, for the sake of statistical power. Alongside this, a larger dataset would allow for even greater statistical power in the analyses.

We have established that depending on the age of the individual, large DBA can correspond to lower-than-normal or higher-than-normal proportions of T1R in brain regions. In other cases, the proportion of assigned T1R does not change significantly, regardless of the age-group. These findings were novel, and gave precedent for the examination of region-wise trajectories of BA relevance over age.

5.5 BA Relevance Across Age Brackets

We examined the trajectories of the proportion of brain regions assigned T1R over age brackets. With special interest we examined which highly-relevant brain structures change the most and the least over age brackets with respect to the proportion assigned T1R. We found that while some highly-relevant regions changed significantly in assigned T1R over age brackets, other highly relevant regions did not change much. We then regarded some example regions which illustrate the tripartite pattern of relevance trajectories that appears across ages. Our primary source from the literature of what to expect from this analysis was again Hofmann et al. [18]. Although they did not examine the trajectories of relevance over age, they did compare regions of saliency between an older and younger cohort. The authors reported statistically significant increases in BA relevance in several regions in the older cohort, but did not report on any decreases in relevance. Our expectation was that brain regions that were assigned T1R in high proportion overall would increase that proportion in older individuals, and other, less salient regions would necessarily decrease in their proportion of assigned T1R.

For some regions, such as the right transverse temporal gyrus, the proportion assigned T1R increased with age. Greater attribution of relevance in older individuals suggests that while the region is generally highly informative of BA, it bears more information about BA and DBA in older individuals than in younger individuals [18]. This agrees with our findings about relevance distribution based on DBA. These regions, and those highly-relevant regions whose proportion of T1R decreased with age, had high SDs in assigned T1R over age brackets. For some other regions, such as the right fourth ventricle, the proportion assigned T1R decreased with age. This would indicate that the region is more informative about BA and DBA in younger individuals than in older individuals. We did not expect this finding, and indeed there are many regions which exhibit this characteristic.

One may make sense of decreasing relevance trajectories by considering the fact that these regions do not necessarily indicate that a young subject is older than expected, but simply that the region is more indicative of DBA in younger individuals than older individuals. For example, it is well-established that the ventricles dilate with age [11], and these structures are assigned a large proportion of T1R, especially by LRP. As we have seen however, the right fourth ventricle is far more relevant to BA in younger individuals than older individuals. This makes sense if we consider that an elderly person is expected to have large ventricles; on the other hand, if a younger person has severely dilated ventricles, this would be indicative of accelerated brain ageing. Indeed, Cannon et al. [169] showed that clinically high-risk individuals who later went on to develop psychosis showed greater ventricular dilation at young ages as compared to high risk individuals who did not go on to develop psychosis. With this in mind, we might expect many regions to have such decreasing relevance trajectories. Other structures remained roughly uniform in their proportion of assigned T1R across ages, such as the right lateral ventricle and the right caudate nucleus. This would suggest that the region is uniformly informative of BA and DBA between young and old individuals. These are the regions with low CoV.

It was found that each region followed one of these three trajectories over the age brackets. Given the fact that increased relevance is indicative of accelerated BA [18], these trajectories imply the following:

1. **Increased regional relevance with age** indicates that the region is more informative of BA and more salient towards large DBA in older individuals than in younger ones. The structure of the right transverse temporal gyrus, for example, is more indicative in older people than in young people of pathological brain ageing.

2. **Decreased regional relevance with age** indicates that the region is more informative of BA and more salient towards large DBA in younger individuals than in older ones. The structure of the right fourth ventricle, for example, is more indicative in young people than in older people of pathological brain ageing. It may be the case that since the ventricles dilate with age for all individuals they are not very telling of DBA in old individuals, whereas a young person with highly dilated ventricles is clearly ageing at an accelerated rate. This would agree with our findings that for most of the ventricles, much higher relevance is attributed in those younger individuals with high DBA than older ones.

3. **Uniform or roughly uniform regional relevance with age** indicates that the region is consistent in its relevance toward BA and DBA.

The creation of region-specific trajectories of saliency over ages is the primary contribution of this work. The establishment of these trajectories serves two purposes, as discussed in Section 1.6:

1. Determining the saliency of a given brain structure and the change thereof over time.

2. Allow for individual comparisons to a baseline relevance trajectory on a region-specific level. This can be done to assess BA in a clinical setting.

We have discussed our contributions in regard to the first point. We now discuss how our findings have served the second listed purpose and the utility to be found therein.

In a clinical application of the BA regression and saliency mapping tool, we can expect the following procedure to take place:

- An MRI scan is performed on a patient to acquire their brain volume.
- The brain volume is fed through the pre-processing/regression/saliency mapping pipeline to produce BA and DBA predictions and a saliency map corresponding to the pre-processed volume, in just over a minute (inference and DBA calculation is close to real-time; saliency mapping takes $\sim$ 1 minute on a modern GPU).
- Regional distribution of T1R is determined (again, in close to real-time).
- The proportions of T1R per brain region can be compared to the baseline trajectories with particular attention to the relevant age bracket. Significant deviation from the trajectories within the age bracket will warrant further investigation, especially in light of high individual DBA.

In this way we propose that this technology can be utilised in a clinical setting to determine regional contributions to pathological brain ageing.

Chapter 6

Conclusion

The primary aim of this work was to examine changes in BA saliency with age. We created a BA regression model and performed LRP and DeepLIFT saliency mapping methods on unseen data. We then analysed the similarities and differences between the saliency maps of the different methods, in light of the literature on BA regional importances and the knowledge of a domain expert. We then examined the difference between relevance distributions for individuals with high DBA and those with low to moderate DBA, and further examined the effect of age on the high-DBA relevance distributions. Our primary contribution was the examination of the trajectories of saliency across ages for individual regions. We set about this aim through five objectives.

Objective 1: Create an accurate BA regression model using DL techniques

We have created a DL regression model that successfully ($r = 0.8925$) predicts chronological age based on T1-weighted structural volumes. We were able to do this with a smaller dataset than has been used previously in the BA regression literature. The main contributions of this work come from the creation and analyses of saliency maps for the BA regression task.

Objective 2: Apply saliency mapping techniques to the BA regression model and compare the results to known characteristics of brain ageing

This work compares and evaluates five different saliency mapping methods – two DeepLIFT methods and three LRP methods. The analysis of saliency map utility was performed through the assessment of T1R distribution by a domain expert. It was found that although there were differences in some distributions of T1R between the methods, the areas of the brain deemed most salient were areas known to be affected by brain ageing. These were the ventricles (particularly by LRP), grey-matter-dense areas such as the vermal lobules, and surrounding limbic structures such as the thalamus, hippocampus and parahippocampal gyrus (especially by DeepLIFT).

Objective 3: Analyse the differences between saliency mapping techniques specific to the BA regression task, to determine the strengths and limitations of each

This objective addressed the first of our three research questions: *What are the differences and similarities between the explanations of BA from different saliency mapping methods?* Given the fact that all methods focused on known areas of brain ageing, but that the LRP methods differed from the DeepLIFT methods in their distribution of relevance relative to the ventricles, it is not of interest to recommend one group of methods over another. In fact, the use of both methods may be preferable as an explanation for BA. Our findings proved incorrect the hypothesis that was posed to this research question. Indeed there were meaningful differences between the saliency mapping methods in their distribution of T1R, most notably in and around the ventricles. This suggests that it may be of interest in the future to compare more methods of saliency mapping for the BA regression task.

Objective 4: Examine the link between region-specific saliency and accelerated brain ageing both in older and younger individuals

Our fourth objective addressed the second research question: *How does accelerated brain ageing affect the distribution of BA relevance?* The regional distribution of T1R by each method was used to examine the effects of large DBA on regional relevance. Individuals with large DBA were sub-categorised by age (older or younger) and the distributions of T1R were compared to that of individuals with moderate DBA. It was found that some regions increased or decreased dramatically in the proportion assigned T1R with high DBA, often depending on age. The thalamus, for example, was found to have greater relevance in young individuals with high DBA than those with low-to-moderate DBA by all methods, and less relevance in older individuals with high DBA. On the other hand, LRP_1 and LRP_2 assign to the transverse temporal gyrus greater proportions of T1R in older individuals with high DBA than in individuals with low-to-moderate DBA, and lower T1R proportions in younger individuals with high DBA. This served as precedent to examine the distribution of relevance in regions across age brackets. Our hypothesis was only partially correct in addressing this research question. We do indeed see that relevance increased in some regions known to be affected by brain ageing, but there was a strong age dependence in the change in relevance, and there were many regions which decreased in relevance for specific age groups with high DBA compared to baseline.

Objective 5: Create region-wise trajectories of BA saliency over ages from a population study

This addressed our third research question: *How does BA saliency change with age on a region-wise basis?* The distribution of T1R allowed us to examine region-wise trajectories of BA relevance over age brackets. By grouping subjects into seven age brackets, we were able to determine that a tripartite pattern of relevance trajectories emerges. Regions tend to increase in relevance with age, decrease in relevance with age or remain uniformly relevant with age, for all individuals, regardless of DBA. Our analyses suggest that this points toward differential informativeness toward DBA on the basis of brain regions, as a function of age. Some regions are more informative of DBA in younger individuals, such as the fourth ventricle; some are more informative in older individuals, such as the transverse temporal gyrus; and some are uniformly informative of BA and DBA across ages, such as the caudate nucleus. Our hypothesis to this end was again only partially correct. While some regions decreased in relevance with age and others increased in relevance with age, we found highly salient regions following each of these trends, as well as others which remained uniformly salient with age.

6.1 Future Work

In the development of this work, the execution of experiments and the collation and analysis of data, several avenues for future work have emerged. Due to constraints on time, not all the limitations of the work could be addressed, and not all the ideas could be explored. In this section we describe avenues for future research on the topics addressed in this work.

Dataset Size

The dataset used in this study was only of size $n = 656$. As we have discussed, this study can be seen as under-powered. The main reasons for using such a small dataset was for ease of computational and storage burden, and ease of access, since the dataset is freely available on application.

A larger dataset would allow for several improvements to the model and the analyses that were performed:

- Greater training data volume would allow for a more accurate BA regression model.
  - A greater number of data points from a distribution allows for more accurate modelling of the distribution (in this case, healthy brain volumes).
  - Exposure to a necessarily more diverse set of individuals allows the model to generalise better in its performance, and gives better representation of the population.

- Greater test data volume would allow for greater robustness of the saliency mapping analyses.
  - The distribution of DBA values would be more narrowly peaked with a better-performing model, but the proportion of individuals lying outside of a threshold DBA value $\delta^* = \sigma$ is expected to remain roughly the same. We would expect that a larger number of individuals with 'large' DBA (relative to the DBA distribution) would be made available for analysis.
  - While the test set size of 132 is not too limiting for the comparison of regional T1R distribution between methods broadly, the grouping of individuals into age brackets greatly diminishes the power of regional relevance attribution analysis. A larger dataset (with a similar range and distribution of ages) would allow for a more robust modelling of region-wise brain age relevance trajectories. With a big enough dataset, the age bracket size may be made smaller, such that a finer gradation of ages could be examined for these trajectories.
  - Such improvements could also be made by using the entire dataset for saliency mapping as opposed to the test set only. We discuss this more below.

Stratification of Age and Sex in the Training-Test Split

We see in Figure 3.4 of Section 2.4.1 that the distributions of age in the training and test sets exhibit some proportional disparity. The same is true of the proportional differences in sex between the two sets (50% male in the training set and 47% male in the test set). This is because we did not stratify for age or sex in the training-test split. In many brain imaging studies [49, 17, 18], the split between training and test sets stratifies for age and sex, bringing the distributions into closer alignment. This would be advisable in future, so as to expose the model to a closer representation of the underlying distributions both in training and testing.

Hyper-parameter Tuning

The hyperparameter tuning phase only focused on three elements of the model: the optimiser, the loss function and the starting learning rate. In future work, the model may benefit from testing multiple values for the following other hyper-parameters:

- The sequence of convolutional filters.
- The number of residual blocks.
- The number and sizes of fully-connected final layers of the network.
- The method by which the three-dimensional data is flattened (e.g. flatten layer vs global average pooling layer).

Number of Trained Models

Only one model was trained on the selected hyper-parameters. We do not know whether the model would on average perform better or worse with different random initial parameterisations. If more such models were trained, we would be able to choose the best performer, or take a combination of results in an ensemble model to create a better model than any single performer.

Ensemble Model

Levakov [17] and Hofmann [18] showed that not only does the BA regression task benefit from the ensemble method, but that saliency mapping works well on these models as well. As we discussed, it may be of great utility to use the powerful ResNet architecture within ensemble models to boost performance further.

Hofmann et al. also showed that using multiple imaging modalities within an ensemble model greatly enhances performance. Furthermore this allows for the production of modality-specific saliency maps. By way of example, saliency maps from T2-weighted images are able to highlight white matter lesions much better than those from T1 images. Including these different modalities offers insight into a greater number of aspects of BA.

The ensemble model also offers quantification of model uncertainty for regression, as well as for saliency mapping. The ability to measure uncertainty would greatly improve the validity of results not only for the brain age model, but for its explanations and the analyses of DBA and relevance trajectories. Combined with a greater dataset size, this would greatly increase the power and reliability of the model and the analyses.

Saliency Mapping on the Entire Dataset

In future work, it would be advisable to perform the saliency mapping on the entire dataset, not only the test set. Although we faced a burden of computational load and storage space, we could gain a lot of statistical power by utilising the data from far more saliency maps. Using our entire dataset for the saliency mapping techniques would multiply the number of assessed subjects by five. The statistical significance of our findings would be greatly strengthened by this. There would be no problem with inference on training data, since the saliency mapping is not concerned with whether or not the data has been seen before. This improvement would be supplemented by the use of a larger dataset too.

Other Saliency Mapping Methods

Considering the differences in explanations brought forth by DeepLIFT and LRP, it would be of great interest to examine the explanations provided by other methods of saliency mapping. Having laid groundwork for qualitative comparisons between methods, more computationally expensive but reliable methods such as Integrated Gradients [89] could be employed to examine possibly different patterns of explanation.

6.2 Closing Remarks

This work has shown new methods of examining BA saliency and its changes with age. We have provided clinically relevant tools for the analysis of brain age saliency and believe that they can show regional contributions to individual DBA. To our knowledge, this is the first work to compare saliency mapping techniques for the BA regression task, and the first to use DeepLIFT for BA regression reasoning. This is also the first work to examine the change in relevance distribution as a result of large DBA in both younger and older individuals, as well as to create region-wise trajectories of BA saliency across ages.

Bibliography

[1] SM Smith, M Jenkinson, MW Woolrich, CF Beckmann, TEJ Behrens, H Johansen-Berg, PR Bannister, M De Luca, I Drobnjak, DE Flitney, R Niazy, J Saunders, J Vickers, Y Zhang, N De Stefano, JM Brady, and PM Matthews. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23(S1):208–19, 2004.

[2] SM Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17(3):143–155, nov 2002.

[3] M Jenkinson and SM Smith. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2):143–156, 2001.

[4] M Jenkinson, PR Bannister, JM Brady, and SM Smith. Improved optimisation for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–841, 2002.

[5] AM Winkler, GR Ridgway, MA Webster, SM Smith, and TE Nichols. Permutation inference for the general linear model. *NeuroImage*, 92:381–397, 2014.

[6] DL Collins, AP Zijdenbos, WFC Baare, and AC Evans. ANIMAL+INSECT: Improved Cortical Structure Segmentation. *IPMI Lecture Notes in Computer Science*, 1613/1999:210–223, 1999.

[7] VS Fonov, AC Evans, CR Almli, RC McKinstry, and DL Collins. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47:s102, jul 2009.

[8] AL Manera, M Dadar, VS Fonov, and DL Collins. CerebrA, registration and manual label correction of Mindboggle-101 atlas for MNI-ICBM152 template. *Scientific Data*, (7(1)):1–9, 2020.

[9] Aaron Courville Ian Goodfellow, Yoshua Bengio. Deep Learning Book. *Deep Learning*, 2015.

[10] S. Kevin Zhou, Gabor Fichtinger, and Daniel Rueckert. *Handbook of medical image computing and computer assisted intervention.* 2019.

[11] Brian H. Anderton. Ageing of the brain. *Mechanisms of Ageing and Development*, 2002.

[12] Chengliang Yang, Anand Rangarajan, and Sanjay Ranka. Visual Explanations From Deep 3D Convolutional Neural Networks for Alzheimer's Disease Classification. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2018:1571–1580, 2018.

[13] Johannes Rieke, Fabian Eitel, Martin Weygandt, John Dylan Haynes, and Kerstin Ritter. Visualizing convolutional networks for MRI-based diagnosis of alzheimer's disease. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11038 LNCS, pages 24–31. Springer Verlag, 2018.

[14] Moritz Böhle, Fabian Eitel, Martin Weygandt, and Kerstin Ritter. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Frontiers in Aging Neuroscience*, 10(JUL), 2019.

[15] Eduardo Nigri, Nivio Ziviani, Fabio Cappabianco, Augusto Antunes, and Adriano Veloso. Explainable Deep CNNs for MRI-Based Diagnosis of Alzheimer's Disease. apr 2020.

[16] Irina Grigorescu, Lucilio Cordero-Grande, A David Edwards, Jo Hajnal, Marc Modat, and Maria Deprez. Interpretable Convolutional Neural Networks for Preterm Birth Classification. pages 1–4, 2019.

[17] Gidon Levakov, Gideon Rosenthal, Ilan Shelef, Tammy Riklin Raviv, and Galia Avidan. From a deep learning model back to the brain—Identifying regional predictors and their relation to aging. *Human Brain Mapping*, 41(12):3235–3252, aug 2020.

[18] Simon M Hofmann, Frauke Beyer, Sebastian Lapuschkin, Markus Loeffler, Klaus-Robert Müller, Arno Villringer, Wojciech Samek, and A Veronica Witte. Towards the Interpretability of Deep Learning Models for Human Neuroimaging. *bioRxiv*, page 2021.06.25.449906, jan 2021.

[19] David A. Sinclair and Matthew D. LaPlante. *Lifespan: Why We Age—and Why We Don't Have To - David A. Sinclair, Matthew D. LaPlante.* Atria Books, 2019.

[20] Daniel Taylor, Jonathan Shock, Deshendran Moodley, Jonathan Ipser, and Matthias S Treder. Brain Structural Saliency Over The Ages. *Proceedings of Machine Learning Research-Under Review*, pages 1–19, 2022.

[21] Meredith A. Shafto, Lorraine K. Tyler, Marie Dixon, Jason R. Taylor, James B. Rowe, Rhodri Cusack, Andrew J. Calder, William D. Marslen-Wilson, John Duncan, Tim Dalgleish, Richard N. Henson, Carol Brayne, Ed Bullmore, Karen Campbell, Teresa Cheung, Simon Davis, Linda Geerligs, Rogier Kievit, Anna McCarrey, Darren Price, David Samu, Matthias Treder, Kamen Tsvetanov, Nitin Williams, Lauren Bates, Tina Emery, Sharon Erzinçlioglu, Andrew Gadie, Sofia Gerbase, Stanimira Georgieva, Claire Hanley, Beth Parkin, David Troy, Jodie Allen, Gillian Amery, Liana Amunts, Anne Barcroft, Amanda Castle, Cheryl Dias, Jonathan Dowrick, Melissa Fair, Hayley Fisher, Anna Goulding, Adarsh Grewal, Geoff Hale, Andrew Hilton, Frances Johnson, Patricia Johnston, Thea Kavanagh-Williamson, Magdalena Kwasniewska, Alison McMinn, Kim Norman, Jessica Penrose, Fiona Roby, Diane Rowland, John Sargeant, Maggie Squire, Beth Stevens, Aldabra Stoddart, Cheryl Stone, Tracy Thompson, Ozlem Yazlik, Dan Barnes, Jaya Hillman, Joanne Mitchell, Laura Villis, and Fiona E. Matthews. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: A cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC Neurology*, 2014.

[22] Feng Hsiung Hsu. IBM's Deep Blue chess grandmaster chips. *IEEE Micro*, 1999.

[23] Michael Laris. Waymo self-driving taxi service launches in Arizona - The Washington Post, 2018.

[24] Stephanie Kanowitz. Speech recognition tech cuts paperwork for police, 2019.

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019.

[26] Harsh Agrawal and Shweta B Guja. Reinforcement learning in OpenAI five. Technical report, 2020.

[27] Github copilot · your ai pair programmer, 2021.

[28] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 1943.

[29] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 1958.

[30] C. M. BERNERS-LEE, A. G. Ivakhnenko, and V. G. Lapa. Cybernetics and Forecasting (Translation). *Nature*, 1968.

[31] HENRY J. KELLEY. Gradient Theory of Optimal Flight Paths. *ARS Journal*, 1960.

[32] P.J. Werbos. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. *PhD Thesis, Harvard U.*, 1974.

[33] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning Internal Representations Error Propagation. *Cognitive Science*, 1986.

[34] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 1986.

[35] Augustin-Louis Cauchy. Méthode générale pour la résolution des systèmes d'équations simultanées. *Comptes rendus hebdomadaires des séances de l'Académie des Sciences*, 1847.

[36] Ding Xuan Zhou. Universality of deep convolutional neural networks, 2020.

[37] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 1980.

[38] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 1968.

[39] Y. T. Zhou and R. Chellappa. Computation of optical flow using a neural network. 1988.

[40] Alexander Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J. Lang. Phoneme Recognition Using Time-Delay Neural Networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1989.

[41] Geoffrey E. Hinton, James L. McClelland, and Geoffrey J. Goodhill. LEARNING REPRESENTATIONS BY RECIRCULATION. 1987.

[42] J S Denker, W R Gardner, H P Graf, D Henderson, R E Howard, W Hubbard, L D Jackel, H S Baird, and I Guyon. Neural network recognizer for hand-written zip code digits. In *Advances in Neural Information Processing Systems (NIPS 1989)*, 1989.

[43] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1989.

[44] Ian J. Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. In *30th International Conference on Machine Learning, ICML 2013*, 2013.

[45] Eleanor M. Caves, Nicholas C. Brandley, and Sönke Johnsen. Visual Acuity and the Evolution of Signals, 2018.

[46] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015.

[47] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with Noisy Student improves ImageNet classification. nov 2019.

[48] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy: FixEfficientNet. mar 2020.

[49] 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, 2017.

[50] Lan Lin, Cong Jin, Zhenrong Fu, Baiwen Zhang, Guangyu Bin, and Shuicai Wu. Predicting healthy older adult's brain age based on structural connectivity networks using artificial neural networks. *Computer Methods and Programs in Biomedicine*, 125:8–17, 2016.

[51] Benson Mwangi, Khader M. Hasan, and Jair C. Soares. Prediction of individual subject's age across the human lifespan using diffusion tensor imaging: A machine learning approach. *NeuroImage*, 75:58–67, 2013.

[52] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, U C Berkeley, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:5000, 2014.

[53] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004.

[54] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, 2005.

[55] Mark Everingham, Luc Van Gool, Christopher K.I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 2010.

[56] Sanja Fidler, Roozbeh Mottaghi, Alan Yuille, and Raquel Urtasun. Bottom-up segmentation for top-down detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013.

[57] J. R.R. Uijlings, K. E.A. Van De Sande, T. Gevers, and A. W.M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 2013.

[58] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. 2009.

[59] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 2015.

[60] João Carreira, Rui Caseiro, Jorge Batista, and Cristian Sminchisescu. Semantic segmentation with second-order pooling. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012.

[61] Ross Girshick. Fast R-CNN.

[62] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 2015-Janua, pages 91–99, 2015.

[63] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN.

[64] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 2019.

[65] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*, 2016.

[66] Anthony W. Flores, Kristin Bechtel, and Christopher T. Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals. And it's biased against blacks". *Federal Probation*, 2016.

[67] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias. *ProPublica*, 2016.

[68] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences, 2017.

[69] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7):1–46, 2015.

[70] Jacek M. Zurada, Aleksander Malinowski, and Ian Cloete. Sensitivity analysis for minimization of input data dimension for feedforward neural network. *Proceedings - IEEE International Symposium on Circuits and Systems*, 6:447–450, 1994.

[71] Grégoire Montavon, Wojciech Samek, and Klaus Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing: A Review Journal*, 73:1–15, 2018.

[72] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus Robert Müller. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65(May 2016):211–222, 2017.

[73] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017.

[74] Weizheng Yan, Sergey Plis, Vince D. Calhoun, Shengfeng Liu, Rongtao Jiang, Tian Zi Jiang, and Jing Sui. Discriminating schizophrenia from normal controls using resting state functional network connectivity: A deep neural network and layer-wise relevance propagation method. *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, 2017-Septe:1–6, 2017.

[75] Fabian Eitel, Emily Soehler, Judith Bellmann-Strobl, Alexander U. Brandt, Klemens Ruprecht, René M. Giess, Joseph Kuchling, Susanna Asseyer, Martin Weygandt, John Dylan Haynes, Michael Scheel, Friedemann Paul, and Kerstin Ritter. Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. *NeuroImage: Clinical*, 24, 2019.

[76] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1–8, 2019.

[77] Wojciech Samek, Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, and Klaus-Robert Müller. Interpreting the Predictions of Complex ML Models by Layer-wise Relevance Propagation. (Nips), 2016.

[78] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus Robert Müller. Layer-Wise Relevance Propagation: An Overview. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2019.

[79] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9887 LNCS, pages 63–71. Springer Verlag, 2016.

[80] Lucas Y. W. Hui and Alexander Binder. *BatchNorm Decomposition for Deep Neural Network Interpretation*, volume 2. Springer, Cham, 2019.

[81] Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. Towards best practice in explaining neural network decisions with LRP. pages 1–5, 2019.

[82] Leon Sixt, Maximilian Granz, and Tim Landgraf. When Explanations Lie: Why Many Modified BP Attributions Fail. (December 2019), 2019.

[83] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, 2018-Decem(NeurIPS):9505–9515, 2018.

[84] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004.

[85] Robert Geirhos, Claudio Michaelis, Felix A. Wichmann, Patricia Rubisch, Matthias Bethge, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *7th International Conference on Learning Representations, ICLR 2019*, (c):1–22, 2019.

[86] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. *34th International Conference on Machine Learning, ICML 2017*, 7:4844–4866, 2017.

[87] Avanti Shrikumar, Peyton Greenside, Anna Y. Shcherbina, and Anshul Kundaje. Not Just a Black Box : Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 33rd International Conference on MachineLearning*, 2016.

[88] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Gradients of Counterfactuals. nov 2016.

[89] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *34th International Conference on Machine Learning, ICML 2017*, 7:5109–5118, 2017.

[90] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings*, pages 1–8, 2014.

[91] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings*, 2015.

[92] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, pages 1–16, 2018.

[93] Theerasarn Pianpanit, Sermkiat Lolak, Phattarapong Sawangjai, Apiwat Ditthapron, Sanparith Marukatat, Ekapol Chuangsuwanich, and Theerawit Wilaiprasitporn. Interpreting deep learning prediction of the Parkinson's disease diagnosis from SPECT imaging. aug 2019.

[94] 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. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[95] Scott M. Lundberg and Su In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 2017.

[96] Hongyoon Choi, Seunggyun Ha, Hyung Jun Im, Sun Ha Paek, and Dong Soo Lee. Refining diagnosis of Parkinson's disease with deep learning-based interpretation of dopamine transporter imaging. *NeuroImage: Clinical*, 16:586–594, 2017.

[97] Soumick Chatterjee, Fatima Saad, Chompunuch Sarasaen, Suhita Ghosh, Rupali Khatun, Petia Radeva, Georg Rose, Sebastian Stober, Oliver Speck, and Andreas Nürnberger. Exploration of Interpretability Techniques for Deep COVID-19 Classification using Chest X-ray Images. jun 2020.

[98] Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, and Marzyeh Ghassemi. COVID-19 Image Data Collection: Prospective Predictions Are the Future. jun 2020.

[99] Daniel S. Kermany, Kang Zhang, and Michael H. Goldbaum. Labeled optical coherence tomography (oct) and chest x-ray images for classification. 2018.

[100] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016.

[101] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-ResNet and the impact of residual connections on learning. In *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 2017.

[102] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017.

[103] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Interpreting Neural Network Classifications with Variational Dropout Saliency Maps. *Conference on Neural Information Processing Systems (NIPS)*, 2017.

[104] Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2017.

[105] Matthew D Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks arXiv:1311.2901v3 [cs.CV] 28 Nov 2013. *Computer Vision–ECCV 2014*, 2014.

[106] Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining explanations in AI. In *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 2019.

[107] G.E.P. Box. Robustness in the Strategy of Scientific Model Building. In *Robustness in Statistics*. 1979.

[108] Ann Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems*, 2019.

[109] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 2019.

[110] Stefano Teso. Toward faithful explanatory active learning with self-explainable neural nets. In *CEUR Workshop Proceedings*, 2019.

[111] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv*, page arXiv:1602.04938, 2016.

[112] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *Advances in Neural Information Processing Systems*, 2017.

[113] Mégane Millan and Catherine Achard. Explaining Regression Based Neural Network Model. apr 2020.

[114] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *Bernoulli*, (1341):1–13, 2009.

[115] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom In: An Introduction to Circuits. *Distill*, 2020.

[116] Geoffrey E. Hinton, Alex Krizhevsky, and Sida D. Wang. Transforming auto-encoders. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6791 LNCS, pages 44–51, 2011.

[117] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):3857–3867, 2017.

[118] Geoffrey Hinton, Sara Sabour, and Nicholas Frosst. Matrix capsules with EM routing, 2018.

[119] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*, 2018.

[120] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42(October):60–88, 2017.

[121] Ying Liu, Brent Logan, Ning Liu, Zhiyuan Xu, Jian Tang, and Yangzhi Wang. Deep Reinforcement Learning for Dynamic Treatment Regimes on Medical Registry Data. In *Proceedings - 2017 IEEE International Conference on Healthcare Informatics, ICHI 2017*, 2017.

[122] Viola Biberacher, Paul Schmidt, Anisha Keshavan, Christine C. Boucard, Ruthger Righart, Philipp Sämann, Christine Preibisch, Daniel Fröbel, Lilian Aly, Bernhard Hemmer, Claus Zimmer, Roland G. Henry, and Mark Mühlau. Intra- and interscanner variability of magnetic resonance imaging based volumetry in multiple sclerosis. *NeuroImage*, 2016.

[123] R. T. Shinohara, J. Oh, G. Nair, P. A. Calabresi, C. Davatzikos, J. Doshi, R. G. Henry, G. Kim, K. A. Linn, N. Papinutto, D. Pelletier, D. L. Pham, D. S. Reich, W. Rooney, S. Roy, W. Stern, S. Tummala, F. Yousuf, A. Zhu, N. L. Sicotte, and R. Bakshi. Volumetric analysis from a harmonized multisite brain MRI study of a single subject with multiple sclerosis. *American Journal of Neuroradiology*, 2017.

[124] Blake E. Dewey, Can Zhao, Aaron Carass, Jiwon Oh, Peter A. Calabresi, Peter C.M. van Zijl, and Jerry L. Prince. Deep harmonization of inconsistent mr data for consistent volume segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018.

[125] J. Dubois, M. Benders, C. Borradori-Tolsa, A. Cachia, F. Lazeyras, R. Ha-Vinh Leuchter, S. V. Sizonenko, S. K. Warfield, J. F. Mangin, and P. S. Hüppi. Primary cortical folding in the human newborn: An early marker of later functional development. *Brain*, 2008.

[126] Bruce Fischl, Niranjini Rajendran, Evelina Busa, Jean Augustinack, Oliver Hinds, B. T.Thomas Yeo, Hartmut Mohlberg, Katrin Amunts, and Karl Zilles. Cortical folding patterns and predicting cytoarchitecture. *Cerebral Cortex*, 2008.

[127] Stephanie Sandor and Richard Leahy. Surface-Based Labeling of Cortical Anatomy Using a Deformable Atlas. *IEEE Transactions on Medical Imaging*, 1997.

[128] Gabriele Lohmann. Extracting line representations of sulcal and gyral patterns in MR images of the human brain. *IEEE Transactions on Medical Imaging*, 1998.

[129] Maryam E. Rettmann, Xiao Han, Chenyang Xu, and Jerry L. Prince. Automated sulcal segmentation using watersheds on the cortical surface. *NeuroImage*, 2002.

[130] Paul M. Thompson, Jason L. Stein, Sarah E. Medland, Derrek P. Hibar, Alejandro Arias Vasquez, Miguel E. Renteria, Roberto Toro, Neda Jahanshad, Gunter Schumann, Barbara Franke, Margaret J. Wright, Nicholas G. Martin, Ingrid Agartz, Martin Alda, Saud Alhusaini, Laura Almasy, Jorge Almeida, Kathryn Alpert, Nancy C. Andreasen, Ole A. Andreassen, Liana G. Apostolova, Katja Appel, Nicola J. Armstrong, Benjamin Aribisala, Mark E. Bastin, Michael Bauer, Carrie E. Bearden, Ørjan Bergmann, Elisabeth B. Binder, John Blangero, Henry J. Bockholt, Erlend Bøen, Catherine Bois, Dorret I. Boomsma, Tom Booth, Ian J. Bowman, Janita Bralten, Rachel M. Brouwer, Han G. Brunner, David G. Brohawn, Randy L. Buckner, Jan Buitelaar, Kazima Bulayeva, Juan R. Bustillo, Vince D. Calhoun, Dara M. Cannon, Rita M. Cantor, Melanie A. Carless, Xavier Caseras, Gianpiero L. Cavalleri, M. Mallar Chakravarty, Kiki D. Chang, Christopher R.K. Ching, Andrea Christoforou, Sven Cichon, Vincent P. Clark, Patricia Conrod, Giovanni Coppola, Benedicto Crespo-Facorro, Joanne E. Curran, Michael Czisch, Ian J. Deary, Eco J.C. de Geus, Anouk den Braber, Giuseppe Delvecchio, Chantal Depondt, Lieuwe de Haan, Greig I. de Zubicaray, Danai Dima, Rali Dimitrova, Srdjan Djurovic, Hongwei Dong, Gary Donohoe, Ravindranath Duggirala, Thomas D. Dyer, Stefan Ehrlich, Carl Johan Ekman, Torbjørn Elvsåshagen, Louise Emsell, Susanne Erk, Thomas Espeseth, Jesen Fagerness, Scott Fears, Iryna Fedko, Guillén Fernández, Simon E. Fisher, Tatiana Foroud, Peter T. Fox, Clyde Francks, Sophia Frangou, Eva Maria Frey, Thomas Frodl, Vincent Frouin, Hugh Garavan, Sudheer Giddaluru, David C. Glahn, Beata Godlewska, Rita Z. Goldstein, Randy L. Gollub, Hans J. Grabe, Oliver Grimm, Oliver Gruber, Tulio Guadalupe, Raquel E. Gur, Ruben C. Gur, Harald H.H. Göring, Saskia Hagenaars, Tomas Hajek, Geoffrey B. Hall, Jeremy Hall, John Hardy, Catharina A. Hartman, Johanna Hass, Sean N. Hatton, Unn K. Haukvik, Katrin Hegenscheid, Andreas Heinz, Ian B. Hickie, Beng Choon Ho, David Hoehn, Pieter J. Hoekstra,

Marisa Hollinshead, Avram J. Holmes, Georg Homuth, Martine Hoogman, L. Elliot Hong, Norbert Hosten, Jouke Jan Hottenga, Hilleke E. Hulshoff Pol, Kristy S. Hwang, Clifford R. Jack, Mark Jenkinson, Caroline Johnston, Erik G. Jönsson, René S. Kahn, Dalia Kasperaviciute, Sinead Kelly, Sungeun Kim, Peter Kochunov, Laura Koenders, Bernd Krämer, John B.J. Kwok, Jim Lagopoulos, Gonzalo Laje, Mikael Landen, Bennett A. Landman, John Lauriello, Stephen M. Lawrie, Phil H. Lee, Stephanie Le Hellard, Herve Lemaître, Cassandra D. Leonardo, Chiang shan Li, Benny Liberg, David C. Liewald, Xinmin Liu, Lorna M. Lopez, Eva Loth, Anbarasu Lourdusamy, Michelle Luciano, Fabio Macciardi, Marise W.J. Machielsen, Glenda M. MacQueen, Ulrik F. Malt, René Mandl, Dara S. Manoach, Jean Luc Martinot, Mar Matarin, Karen A. Mather, Manuel Mattheisen, Morten Mattingsdal, Andreas Meyer-Lindenberg, Colm McDonald, Andrew M. McIntosh, Francis J. McMahon, Katie L. McMahon, Eva Meisenzahl, Ingrid Melle, Yuri Milaneschi, Sebastian Mohnke, Grant W. Montgomery, Derek W. Morris, Eric K. Moses, Bryon A. Mueller, Susana Muñoz Maniega, Thomas W. Mühleisen, Bertram Müller-Myhsok, Benson Mwangi, Matthias Nauck, Kwangsik Nho, Thomas E. Nichols, Lars Göran Nilsson, Allison C. Nugent, Lars Nyberg, Rene L. Olvera, Jaap Oosterlaan, Roel A. Ophoff, Massimo Pandolfo, Melina Papalampropoulou-Tsiridou, Martina Papmeyer, Tomas Paus, Zdenka Pausova, Godfrey D. Pearlson, Brenda W. Penninx, Charles P. Peterson, Andrea Pfennig, Mary Phillips, G. Bruce Pike, Jean Baptiste Poline, Steven G. Potkin, Benno Pütz, Adaikalavan Ramasamy, Jerod Rasmussen, Marcella Rietschel, Mark Rijpkema, Shannon L. Risacher, Joshua L. Roffman, Roberto Roiz-Santiañez, Nina Romanczuk-Seiferth, Emma J. Rose, Natalie A. Royle, Dan Rujescu, Mina Ryten, Perminder S. Sachdev, Alireza Salami, Theodore D. Satterthwaite, Jonathan Savitz, Andrew J. Saykin, Cathy Scanlon, Lianne Schmaal, Hugo G. Schnack, Andrew J. Schork, S. Charles Schulz, Remmelt Schür, Larry Seidman, Li Shen, Jody M. Shoemaker, Andrew Simmons, Sanjay M. Sisodiya, Colin Smith, Jordan W. Smoller, Jair C. Soares, Scott R. Sponheim, Emma Sprooten, John M. Starr, Vidar M. Steen, Stephen Strakowski, Lachlan Strike, Jessika Sussmann, Philipp G. Sämann, Alexander Teumer, Arthur W. Toga, Diana Tordesillas-Gutierrez, Daniah Trabzuni, Sarah Trost, Jessica Turner, Martijn Van den Heuvel, Nic J. van der Wee, Kristel van Eijk, Theo G.M. van Erp, Neeltje E.M. van Haren, Dennis van 't Ent, Marie Jose van Tol, Maria C. Valdés Hernández, Dick J. Veltman, Amelia Versace, Henry Völzke, Robert Walker, Henrik Walter, Lei Wang, Joanna M. Wardlaw, Michael E. Weale, Michael W. Weiner, Wei Wen, Lars T. Westlye, Heather C. Whalley, Christopher D. Whelan, Tonya White, Anderson M. Winkler, Katharina Wittfeld, Girma Woldehawariat, Christiane Wolf, David Zilles, Marcel P. Zwiers, Anbupalam Thalamuthu, Peter R. Schofield, Nelson B. Freimer, Natalia S. Lawrence, and Wayne Drevets. The ENIGMA Consortium: Large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging and Behavior*, 2014.

[131] Carrie E. Bearden and Paul M. Thompson. Emerging Global Initiatives in Neurogenetics: The Enhancing Neuroimaging Genetics through Meta-analysis (ENIGMA) Consortium, 2017.

[132] Christos Davatzikos. P4-101: BRAIN AGING HETEROGENEITY ELUCIDATED VIA MACHINE LEARNING: THE MULTI-SITE ISTAGING DIMENSIONAL NEUROIMAGING REFERENCE SYSTEM. *Alzheimer's & Dementia*, 2018.

[133] Guray Erus, Harsha Battapady, Theodore D. Satterthwaite, Hakon Hakonarson, Raquel E. Gur, Christos Davatzikos, and Ruben C. Gur. Imaging patterns of brain development and their relationship to cognition. *Cerebral Cortex*, 2015.

[134] Fabrice Crivello, Nathalie Tzourio-Mazoyer, Christophe Tzourio, and Bernard Mazoyer. Longitudinal assessment of global and regional rate of grey matter atrophy in 1,172 healthy older adults: Modulation by sex and age. *PLoS ONE*, 2014.

[135] Hiroshi Matsuda. Voxel-based morphometry of brain MRI in normal aging and Alzheimer's disease. *Aging and Disease*, 2013.

[136] Henry Völzke, Dietrich Alte, Carsten Oliver Schmidt, Dörte Radke, Roberto Lorbeer, Nele Friedrich, Nicole Aumann, Katharina Lau, Michael Piontek, Gabriele Born, Christoph Havemann, Till Ittermann, Sabine Schipf, Robin Haring, Sebastian E. Baumeister, Henri Wallaschofski, Matthias Nauck, Stephanie Frick, Andreas Arnold, Michael Jünger, Julia Mayerle, Matthias Kraft, Markus M. Lerch, Marcus Dörr, Thorsten Reffelmann, Klaus Empen, Stephan B. Felix, Anne Obst, Beate Koch, Sven Gläser, Ralf Ewert,

Ingo Fietze, Thomas Penzel, Martina Dören, Wolfgang Rathmann, Johannes Haerting, Mario Hannemann, Jürgen Röpcke, Ulf Schminke, Clemens Jürgens, Frank Tost, Rainer Rettig, Jan A. Kors, Saskia Ungerer, Katrin Hegenscheid, Jens Peter Kühn, Julia Kühn, Norbert Hosten, Ralf Puls, Jörg Henke, Oliver Gloger, Alexander Teumer, Georg Homuth, Uwe Völker, Christian Schwahn, Birte Holtfreter, Ines Polzer, Thomas Kohlmann, Hans J. Grabe, Dieter Rosskopf, Heyo K. Kroemer, Thomas Kocher, Reiner Biffar, Ulrich John, and Wolfgang Hoffmann. Cohort profile: The study of health in Pomerania. *International Journal of Epidemiology*, 2011.

[137] M. Habes, D. Janowitz, G. Erus, J. B. Toledo, S. M. Resnick, J. Doshi, S. Van Der Auwera, K. Wittfeld, K. Hegenscheid, N. Hosten, R. Biffar, G. Homuth, H. Völzke, H. J. Grabe, W. Hoffmann, and C. Davatzikos. Advanced brain aging: Relationship with epidemiologic and genetic risk factors, and overlap with Alzheimer disease atrophy patterns. *Translational Psychiatry*, 2016.

[138] M. Habes, J. B. Toledo, S. M. Resnick, J. Doshi, S. Van Der Auwera, G. Erus, D. Janowitz, K. Hegenscheid, G. Homuth, H. Volzke, W. Hoffmann, H. J. Grabe, and C. Davatzikos. Relationship between APOE genotype and structural MRI measures throughout adulthood in the study of health in pomerania population-based cohort. *American Journal of Neuroradiology*, 2016.

[139] Harini Eavani, Mohamad Habes, Theodore D. Satterthwaite, Yang An, Meng Kang Hsieh, Nicolas Honnorat, Guray Erus, Jimit Doshi, Luigi Ferrucci, Lori L. Beason-Held, Susan M. Resnick, and Christos Davatzikos. Heterogeneity of structural and functional imaging patterns of advanced brain aging revealed via machine learning methods. *Neurobiology of Aging*, 2018.

[140] Ruth Peters. Ageing and the brain, 2006.

[141] Elizabeth R. Sowell, Bradley S. Peterson, Paul M. Thompson, Suzanne E. Welcome, Amy L. Henkenius, and Arthur W. Toga. Mapping cortical change across the human life span. *Nature Neuroscience*, 2003.

[142] Naftali Raz and Karen M. Rodrigue. Differential aging of the brain: Patterns, cognitive correlates and modifiers, 2006.

[143] Hediye Pjnar Gunbey, Karabekir Ercan, Ayge Serap Fjndjkoglu, H Taner Bulut, Mustafa Karaoglanoglu, and Halil Arslan. The Limbic Degradation of Aging Brain: A Quantitative Analysis with Diffusion Tensor Imaging. 2014.

[144] Stephen M. Smith, D. Vidaurre, F. Alfaro-Almagro, Thomas E. Nichols, and Karla L. Miller. Estimation of brain age delta from brain imaging. *NeuroImage*, 200:528–539, oct 2019.

[145] Nicola K. Dinsdale, Emma Bluemke, Stephen M. Smith, Zobair Arya, Diego Vidaurre, Mark Jenkinson, and Ana I.L. Namburete. Learning patterns of the ageing brain in MRI using deep convolutional networks. *NeuroImage*, 224:117401, jan 2021.

[146] Han Peng, Weikang Gong, Christian F. Beckmann, Andrea Vedaldi, and Stephen M. Smith. Accurate brain age prediction with lightweight deep neural networks. *Medical Image Analysis*, 68:101871, feb 2021.

[147] 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, Deanna M. Barch, Steven E. Petersen, and Bradley L. Schlaggar. Prediction of individual brain maturity using fMRI. *Science*, 2010.

[148] Katja Franke, Gabriel Ziegler, Stefan Klöppel, and Christian Gaser. NeuroImage Estimating the age of healthy subjects from T 1 -weighted MRI scans using kernel methods : Exploring the in fl uence of various parameters. *NeuroImage*, 50(3):883–892, 2010.

[149] Katja Franke, Eileen Luders, Arne May, Marko Wilke, and Christian Gaser. Brain maturation: Predicting individual BrainAGE in children and adolescents using structural MRI. *NeuroImage*, 63(3):1305–1312, 2012.

[150] Timothy B. Meier, Alok S. Desphande, Svyatoslav Vergun, Veena A. Nair, Jie Song, Bharat B. Biswal, Mary E. Meyerand, Rasmus M. Birn, and Vivek Prabhakaran. Support vector machine classification and characterization of age-related reorganization of functional brain networks. *NeuroImage*, 60(1):601–613, 2012.

[151] Sukrit Gupta, Yi Hao Chan, and Jagath C. Rajapakse. Decoding Brain Functional Connectivity Implicated in AD and MCI. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019.

[152] Juergen Lutz, Felix Hemminger, Robert Stahl, Olaf Dietrich, Martin Hempel, Maximilian Reiser, and Lorenz Jäger. Evidence of Subcortical and Cortical Aging of the Acoustic Pathway: A Diffusion Tensor Imaging (DTI) Study. *Academic Radiology*, 14(6):692–700, jun 2007.

[153] Jean Kossaifi, Arinbjörn Kolbeinsson, Aran Khanna, Tommaso Furlanello, and Anima Anandkumar. Tensor Regression Networks. *Journal of Machine Learning Research*, 21:1–21, 2020.

[154] Karla L. Miller, Fidel Alfaro-Almagro, Neal K. Bangerter, David L. Thomas, Essa Yacoub, Junqian Xu, Andreas J. Bartsch, Saad Jbabdi, Stamatios N. Sotiropoulos, Jesper L.R. Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Thomas W. Okell, Peter Weale, Iulius Dragonu, Steve Garratt, Sarah Hudson, Rory Collins, Mark Jenkinson, Paul M. Matthews, and Stephen M. Smith. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, 2016.

[155] Aurélien Géron. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow (Early Release)*. 2019.

[156] 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), 2019.

[157] B. M. Barer. Men and women aging differently, 1994.

[158] Mary Ellen I. Koran, Madison Wagener, and Timothy J. Hohman. Sex differences in the association between AD biomarkers and cognitive decline. *Brain Imaging and Behavior*, 2016.

[159] Raymond Pomponio, Guray Erus, Mohamad Habes, Jimit Doshi, Dhivya Srinivasan, Elizabeth Mamourian, Vishnu Bashyam, Ilya M. Nasrallah, Theodore D. Satterthwaite, Yong Fan, Lenore J. Launer, Colin L. Masters, Paul Maruff, Chuanjun Zhuo, Henry Völzke, Sterling C. Johnson, Jurgen Fripp, Nikolaos Koutsouleris, Daniel H. Wolf, Raquel Gur, Ruben Gur, John Morris, Marilyn S. Albert, Hans J. Grabe, Susan M. Resnick, R. Nick Bryan, David A. Wolk, Russell T. Shinohara, Haochang Shou, and Christos Davatzikos. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, 2020.

[160] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, 12(3):e1001779, mar 2015.

[161] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L Whitwell, Chadwick Ward, Anders M Dale, Joel P Felmlee, Jeffrey L Gunter, Derek LG Hill, Ron Killiany, Norbert Schuff, Sabrina Fox-Bosetti, Chen Lin, Colin Studholme, Charles S DeCarli, Gunnar Krueger, Heidi A Ward, Gregory J Metzger, Katherine T Scott, Richard Mallozzi, Daniel Blezek, Joshua Levy, Josef P Debbins, Adam S Fleisher, Marilyn Albert, Robert Green, George Bartzokis, Gary Glover, John Mugler, Michael W Weiner, and CRJ Jr. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI Methods. *www. interscience.wiley.com). JOURNAL OF MAGNETIC RESONANCE IMAGING*, 27:685–691, 2008.

[162] 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, jan 2017.

[163] Markus Loeffler, Christoph Engel, Peter Ahnert, Dorothee Alfermann, Katrin Arelin, Ronny Baber, Frank Beutner, Hans Binder, Elmar Brähler, Ralph Burkhardt, Uta Ceglarek, Cornelia Enzenbach, Michael Fuchs, Heide Glaesmer, Friederike Girlich, Andreas Hagendorff, Madlen Häntzsch, Ulrich Hegerl, Sylvia Henger, Tilman Hensch, Andreas Hinz, Volker Holzendorf, Daniela Husser, Anette Kersting, Alexander Kiel, Toralf Kirsten, Jürgen Kratzsch, Knut Krohn, Tobias Luck, Susanne Melzer, Jeffrey Netto, Matthias Nüchter, Matthias Raschpichler, Franziska G. Rauscher, Steffi G. Riedel-Heller, Christian Sander, Markus Scholz, Peter Schönknecht, Matthias L. Schroeter, Jan Christoph Simon, Ronald Speer, Julia Stäker, Robert Stein, Yve Stöbel-Richter, Michael Stumvoll, Attila Tarnok, Andrej Teren, Daniel Teupser, Francisca S. Then, Anke Tönjes, Regina Treudler, Arno Villringer, Alexander Weissgerber, Peter Wiedemann, Silke Zachariae, Kerstin Wirkner, and Joachim Thiery. The LIFE-Adult-Study: Objectives and design of a population-based cohort study with 10,000 deeply phenotyped adults in Germany. *BMC Public Health*, 15(1):1–14, jul 2015.

[164] Diederik P. Kingma and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, dec 2014.

[165] Simon Letzgus, Patrick Wagner, Jonas Lederer, Wojciech Samek, Klaus-Robert Müller, and Gregoire Montavon. Toward Explainable AI for Regression Models. 2021.

[166] Stéphane Lehéricy, Étienne C. Hirsch, Pascale Cervera-Piérot, Louis B. Hersh, Serge Bakchine, François Piette, Charles Duyckaerts, Jean-Jacques -J Hauw, France Javoy-Agid, and Yves Agid. Heterogeneity and selectivity of the degeneration of cholinergic neurons in the basal forebrain of patients with Alzheimer's disease. *Journal of Comparative Neurology*, 330(1), 1993.

[167] K. M. Cullen and G. M. Halliday. Neurofibrillary degeneration and cell loss in the nucleus basalis in comparison to cortical Alzheimer pathology. *Neurobiology of Aging*, 19(4), 1998.

[168] Richard F. Uhlmann, Eric B. Larson, Thomas S. Rees, Thomas D. Koepsell, and Larry G. Duckert. Relationship of Hearing Impairment to Dementia and Cognitive Dysfunction in Older Adults. *JAMA*, 261(13):1916–1919, apr 1989.

[169] Tyrone D Cannon, Yoonho Chung, George He, Daqiang Sun, Aron Jacobson, Theo G M Van Erp, Sarah Mcewen, Jean Addington, Carrie E Bearden, Kristin Cadenhead, Barbara Cornblatt, Daniel H Mathalon, Thomas Mcglashan, Diana Perkins, Clark Jeffries, Larry J Seidman, Ming Tsuang, Elaine Walker, Scott W Woods, and Robert Heinssen. Archival Report Progressive Reduction in Cortical Thickness as Psychosis Develops: A Multisite Longitudinal Neuroimaging Study of Youth at Elevated Clinical Risk.