Optimising Convolutional Neural Networks for Large-Scale Neuroimaging Studies



Nicola Katharine Dinsdale

Worcester College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2021

Acknowledgements

In the shelter of each other the people live

Although I would not recommend doing a large proportion of a dPhil during a pandemic, I have been fortunate to have been supported by many amazing people throughout my time studying in Oxford, and whilst here, I will thank many people; many more will be missed due to my thesis-addled brain and to how long the list would have to be.

Clearly first, thanks must be given to my supervisors, Prof Mark Jenkinson and Dr Ana Namburete for their constant support and guidance. You both invested so much time and energy into supporting me, and hours of talking through ideas, proof reading and cups of tea later, there is a thesis which I hope at least vaguely represents how amazing the support was that I was given. Fundamentally, I hope I have learned a little about MRI, a little about deep learning and a lot about how to be a supportive and caring advocate.

I was fortunate enough to be part of two great research groups and have been constantly grateful for the conversation and friendships, both in real life and then online, and especially to Rick, Amy and Sophie for countless cups of tea and Friday Fizz, commiserations and celebrations. This thesis has already been described as 'a bit long', so to list everyone who has supported me is not possible: most of the fMRIB analysis group have helped me at some point over the last three years, especially Vaanu and Ludo, and I would have undoubtedly floundered somewhere near the beginning of this process without all of the support. The Nambureteam have been the best group to work an idea through with and I am very grateful for the ideas and comments you constantly came up with, even when listening to the same work being presented for the fifteenth time. The ladies at GLOW have shown me what it means to be amazing physicists and engineers, and I've been so fortunate to get to know you all. ONBI, and all the lovely people that entailed, helped me start to learn how to research.

Eight years of Oxford have led to me being fortunate enough to meet some of the best people – and some of the dearest friends – who have supported me through the

dPhil and the pandemic. My undergrad friends from Brasenose coped with hours of me moaning about my dPhil and being in the wrong city, and spent hours doing crafts in real life and over Zoom. Rachel gave me many weekends of terrible TV and great baked goods. Sarah and Tobias have offered a constant resting place of calm and good cake (and tea). Lucy took me on rainy swims with copious amounts of tea. Stephen gave me and my bike a lift home after many great dinners with he and Renee. My housemates gave me three years of a home and a call of 'Hi Friend!' to come back to: Veerle, Charlie, Charles and Katie, you even managed to make some of Lockdown 3 fun!

My family have been the support everyone should have. My adopted Horrocks and Rhodes families have always been the terrific. Mum (special thank you for proof reading everything many, many times...), Dad, Paul and Tom: thanks forever for backing my decisions, being my biggest cheerleaders, and for the guaranteed place to wait out the storm (or the pandemic). And for lots and lots of tea.

I have been supported throughout my dPhil by funding from the Engineering and Physical Sciences Research Council (EPSRC) and Medical Research Council (MRC) [grant number EP/L016052/1]. The Wellcome Centre for Integrative Neuroimaging is supported by core funding from the Wellcome Trust [grant number 203139/Z/16/Z].

Abstract

Ageing has a pronounced effect on the human brain, leading to cognitive decline and an increased risk of neurodegenerative diseases. Thus, the ageing population presents a significant challenge for healthcare. The use of MRI and the availability of computational methods for analysing the MRI data is increasingly contributing to the understanding of healthy and diseased structural brain maturation and ageing. Increasingly, large cross sectional and longitudinal neuroimaging studies are becoming available, presenting opportunities for the application of deep learning to neuroimage analysis. There are however, many domain specific problems for applying deep learning to neuroimaging problems, which currently limit their wider applicability.

This thesis explores three distinct problems. First, a model is developed to explore brain ageing. Both normal ageing and neurodegenerative disease cause morphological changes to the brain, and deep learning models are well suited to capturing these patterns. A 3D CNN architecture is developed to predict chronological age, using T1-weighted MRI from the UK Biobank dataset. The proposed method shows competitive performance on age prediction, but, most importantly, the CNN prediction errors correlated significantly with many clinical measurements from the UK Biobank in the female and male groups. In addition, having used images from only one imaging modality in this experiment, the relationships between $\Delta_{BrainAge}$ and the image-derived phenotypes (IDPs) from all of the other imaging modalities in the UK Biobank are explored, showing correlations consistent with known patterns of ageing. The effect of the pre-processing is also explored. Specifically, it is shown that the use of non-linearly registered images to train the CNNs can lead to the network being driven by artefacts of the registration process, and therefore to miss subtle indicators of ageing, which would limit the clinical relevance of the model.

Increasingly large MRI neuroimaging datasets are becoming available but many of these are highly "multi-site, multi-scanner", which leads to an increase in variance due to nonbiological factors when these are combined. This increase in variance is due to factors such as differences in acquisition protocols and hardware, and can mask

the signals of interest, and this is known as the *harmonisation* problem. A deep learning based scheme, developed from domain adaptation techniques, is used to create harmonised outputs. An iterative update approach is used to create scanner-invariant features, whilst simultaneously maintaining performance on the main task of interest, thus reducing the influence of the acquisition scanner on the network predictions. The framework is demonstrated for regression, classification, and segmentation tasks with two different network architectures. It is shown that not only can the framework harmonise multi-site datasets, but it can also adapt to many data scenarios, including biased datasets and limited training labels. Finally, it is shown that the framework can be extended for the removal of other known confounds in addition to scanner or data source. The overall framework is therefore flexible and should be applicable to a wide range of neuroimaging studies.

Finally, the parameterisation of neural networks is considered: the vast number of parameters means that networks need large numbers of training examples and labels. Especially for medical image segmentation, large labelled datasets are rarely available. A method to train and prune UNet architectures simultaneously is developed, alongside an adaptive targeted dropout scheme that makes the network robust to the *pruning* – that is, the removal of filters from the model. It is shown that the pruned models outperform the standard UNet models, especially when working in very low data regimes, across medical imaging tasks. The framework is then applied to multisite MRI data: the standard UNet models trained on the data from one site suffer significant performance degradation when applied to the data from the other sites, but this is reduced when using the pruned models, due to a reduction in model overfitting. The generalisability is systematically explored, and it is shown that the pruned models have increased robustness, compared to the standard UNet models.

Contents

List of Figures **vii**

List of Tables **xii**

List of Abbreviations **xiii**

1 Introduction **1**
- 1.1 Thesis Contributions . . . 2
- 1.2 Thesis Structure . . . 4
- 1.3 Statement of Originality . . . 5
- 1.4 List of Publications . . . 6

2 Magnetic Resonance Image Analysis **9**
- 2.1 MRI Acquisition Basics . . . 10
- 2.2 Structural MRI . . . 12
  - 2.2.1 Partial Volume Effects . . . 14
  - 2.2.2 Intensity Normalisation . . . 14
  - 2.2.3 Image Artefacts . . . 15
  - 2.2.4 Brain Extraction or Skull Stripping . . . 16
  - 2.2.5 Registration . . . 18
    - 2.2.5.1 MNI Template . . . 18
    - 2.2.5.2 Linear Registration . . . 20
    - 2.2.5.3 Non-linear Registration . . . 21
  - 2.2.6 Bias Field Correction . . . 22
  - 2.2.7 Segmentation . . . 23
    - 2.2.7.1 Tissue Segmentation . . . 24
    - 2.2.7.2 Subcortical Segmentation . . . 25
    - 2.2.7.3 Other Measures Derived from Structural Images . . . 26
- 2.3 Diffusion MRI (dMRI) . . . 27

2.3.1 Diffusion Tensor Imaging (DTI) . . . 29
2.3.2 Tract-based Spatial Statistics (TBSS) . . . 31
2.3.3 Probabilistic Tractography . . . 31
2.4 Functional MRI (fMRI) . . . 32
2.4.1 Resting State fMRI . . . 34
2.4.2 Task fMRI . . . 34

3 Convolutional Neural Networks 36
3.1 Supervised Learning . . . 37
3.1.1 Back Propagation . . . 39
3.2 Convolutional Neural Networks . . . 41
3.2.1 Convolutional Layers . . . 42
3.2.2 Non-linearities or Activation Functions . . . 43
3.2.3 Pooling Layers . . . 45
3.2.4 Normalisation Layers . . . 46
3.2.5 Optimisers and Learning Rate . . . 47
3.2.6 Loss Functions . . . 48
3.3 Medical Imaging Tasks . . . 50
3.3.1 Classification . . . 50
3.3.2 Segmentation . . . 52
3.3.3 Image Generation or Reconstruction . . . 53
3.4 Challenges for the Application of Deep Learning to Medical Imaging . . 55
3.4.1 Data Availability . . . 56
3.4.1.1 Model Pre-training and Transfer Learning . . . 57
3.4.1.2 2D or Patch Based Approaches . . . 58
3.4.1.3 Overfitting . . . 59
3.4.1.4 Data Augmentation . . . 60
3.4.1.5 Differences between Datasets or Data Domain Shift . . 61
3.4.1.6 Data Composition and Algorithmic Biases . . . 65
3.4.2 Evaluation and Ground Truth . . . 66
3.4.3 Interpretability . . . 69
3.4.3.1 Visualisation . . . 71
3.4.3.2 Uncertainty . . . 73
3.4.4 Logistical Challenges . . . 75
3.4.4.1 Computational Power . . . 75
3.4.4.2 Data Sharing and Data Privacy . . . 76

4 Learning Patterns of the Ageing Brain in MRI using Deep Convolutional Networks 78
4.1 Introduction . . . 79
4.2 Methods . . . 86
4.2.1 Data and Pre-processing . . . 86
4.2.2 CNN Architecture . . . 87
4.2.3 Experimental Setup . . . 88
4.2.4 Correlation with Subject Biobank Measurements and IDPs . . . 90
4.2.4.1 Obtaining $\Delta_{BrainAge}$ . . . 90
4.2.4.2 Relationship with Biobank Measurements . . . 91
4.2.4.3 Relationships with IDPs . . . 91
4.2.4.4 Comparison with IDP-Based Age Prediction . . . 92
4.3 Results . . . 92
4.3.1 Age Prediction . . . 92
4.3.2 Relationships with UK Biobank Measurements . . . 100
4.3.2.1 UK Biobank Variables . . . 100
4.3.2.2 Image Derived Phenotypes . . . 100
4.3.3 Relationships with IDP-based Age Prediction . . . 103
4.4 Discussion . . . 106

5 Deep Learning-Based Unlearning of Dataset Bias for MRI Harmonisation and Confound Removal 113
5.1 Introduction . . . 114
5.2 Methods . . . 120
5.2.1 Data Inputs and Outputs . . . 120
5.2.2 Network Architecture . . . 121
5.2.3 Training Procedure . . . 122
5.2.4 Age Prediction Task . . . 126
5.2.4.1 Datasets . . . 127
5.2.4.2 Basic Fully Supervised Learning . . . 128
5.2.4.3 Biased Domains . . . 129
5.2.4.4 Extension to ABIDE Data . . . 130
5.2.4.5 Removal of Other Categorical Confounds . . . 131
5.2.4.6 Removal of Continuous Confounds . . . 133
5.2.5 Segmentation Task . . . 133
5.2.5.1 Basic Fully Supervised Learning . . . 135

5.2.5.2 Semi-supervised Learning . . . 135
5.3 Results . . . 136
5.3.1 Age Prediction Task . . . 136
5.3.1.1 Basic Fully Supervised Learning . . . 136
5.3.1.2 Biased Datasets . . . 144
5.3.1.3 Extension to ABIDE Data . . . 147
5.3.1.4 Removal of Other Categorical Confounds . . . 148
5.3.1.5 Removal of Continuous Confounds . . . 149
5.3.2 Segmentation Task . . . 151
5.3.2.1 Supervised Learning . . . 151
5.3.2.2 Semi-supervised Segmentation . . . 153
5.4 Discussion . . . 154

6 Simultaneous Model Pruning and Training for Low Data Regimes in Medical Image Segmentation 159
6.1 Introduction . . . 159
6.2 Methods . . . 166
6.2.1 Training Procedure . . . 167
6.2.2 Adaptive Channelwise Targeted Dropout . . . 169
6.2.3 Model Training . . . 171
6.2.4 Network Architecture . . . 171
6.2.4.1 Shallower Network . . . 173
6.2.5 Datasets and Implementation Details . . . 173
6.2.5.1 Datasets . . . 175
6.3 Results . . . 177
6.3.1 HarP Data - Evaluation of the Algorithm . . . 177
6.3.1.1 Choice of Metric . . . 178
6.3.1.2 Varying Numbers of Filters f . . . 181
6.3.1.3 Method Comparison . . . 183
6.3.1.4 Filter Magnitudes . . . 185
6.3.1.5 Base Dropout Value b_{drop} . . . 188
6.3.1.6 Number of Recovery Epochs . . . 189
6.3.1.7 Network Architecture Depth . . . 190
6.3.2 Application to Other Datasets . . . 193
6.3.3 Application to the IXI Subcortical Segmentation - Ablation Study 196
6.3.3.1 Amount of Training Data . . . 199

6.3.4 Training in Low Data Regimes . . . 201
6.3.4.1 Comparison to PruneFinetune . . . 203
6.3.4.2 Number of Recovery Epochs . . . 204
6.3.4.3 Training in Low Data Regimes – Application to Other Datasets . . . 205
6.4 Conclusion . . . 205

7 Model Pruning for Generalisability Across Imaging Sites 209
7.1 Introduction . . . 209
7.2 Methods . . . 211
7.2.1 Implementation Details . . . 211
7.2.2 Datasets . . . 212
7.2.3 Generalisability Across Sites: ABIDE - Tissue Segmentation . . . 214
7.2.3.1 Number of Recovery Epochs . . . 220
7.2.3.2 Exploring Potential Co-Adaptation . . . 221
7.2.3.3 Method Comparison . . . 222
7.2.3.4 Comparison to Dropout . . . 224
7.2.3.5 Network Architecture Depth . . . 228
7.2.4 Generalisability Across Sites: IXI - Brain Extraction . . . 230
7.2.5 Generalisability Across Sites: IXI - Subcortical Segmentation . . . 232
7.2.6 Robustness of Pruned Model . . . 235
7.3 Conclusion . . . 243

8 Conclusion 246
8.1 Contributions . . . 246
8.2 Limitations . . . 249
8.3 Future Directions . . . 252

A Dataset Details 254
A.1 UK Biobank . . . 254
A.2 OASIS . . . 256
A.3 Whitehall II . . . 256
A.4 ABIDE . . . 257
A.5 IXI . . . 258
A.6 HarP . . . 258
A.7 Medical Decathlon . . . 259
A.8 INTERGROWTH-21st Study . . . 259

B Appendix: Additional Background **260**
B.1 Classification Networks . . . 260
B.2 Segmentation Networks . . . 265
B.3 Image Generation or Reconstruction . . . 266
B.4 Brain Age Prediction Studies . . . 268
B.5 Harmonisation Studies . . . 271

C Additional Results **272**
C.1 Brain Age Prediction . . . 272
C.1.1 Deconfounding for Age . . . 272
C.1.2 Attention Network Design . . . 275
C.1.3 Training on Healthy Subjects . . . 276
C.1.4 Significant IDPS and Variables . . . 285
C.2 Harmonisation . . . 294
C.2.1 Comparison to Adding Site as Regressor . . . 294
C.2.2 DANN Results . . . 298
C.2.3 Visualisation of Confusion Loss . . . 298
C.2.4 p-Values . . . 299
C.3 Pruning . . . 301
C.3.1 Evaluation Metric - HarP Data . . . 301
C.3.2 ABIDE Dataset Distribution . . . 301
C.3.3 Choice of Evaluation Metric - ABIDE Data . . . 302
C.3.4 Pruning Percentage for PruneFinetune . . . 302

Bibliography **306**

List of Figures

2.1 T1-weighted MR Image . . . 12
2.2 T2-weighted MR Image . . . 14
2.3 Skull Stripping . . . 16
2.4 Registration . . . 19
2.5 Bias Field Estimation . . . 23
2.6 Tissue Segmentation . . . 24
2.7 Deep Grey Matter Structure Segmentation . . . 25
2.8 Diffusion Tensor Imaging Example . . . 27
2.9 Mean Diffusivity and Fractional Anisotropy . . . 30
2.10 FunctionaI MRI Scan . . . 32

3.1 Example Network Architecture . . . 41
3.2 Convolution . . . 42
3.3 Activation Functions . . . 44
3.4 Learning Rate . . . 47
3.5 VGG-16 . . . 51
3.6 UNet . . . 52
3.7 Autoencoder . . . 54
3.8 GAN . . . 55
3.9 Example Extracted Features . . . 59
3.10 Example Augmentations . . . 60
3.11 Domain Shift . . . 62
3.12 Age Prediction of Random Noise . . . 69
3.13 Saliency Schematic . . . 71
3.14 Uncertainty . . . 74

4.1 UK Biobank Age Distributions . . . 85
4.2 Brain Age Prediction Input Data . . . 87

4.3 Brain Age Prediction Network Architectures . . . 89
4.4 Density Plot of Age Prediction Results . . . 94
4.5 Comparison of Same Aged Subjects with Very Different Δ . . . 96
4.6 Average Brains . . . 97
4.7 Attention and Saliency Maps . . . 98
4.8 CNN model – Non-imaging UK Biobank Variables Correlations . . . 101
4.9 CNN model – IDP Correlations . . . 102
4.10 CNN – Tract Correlations . . . 104
4.11 Density Plots Comparing CNN Model and IDP-based Model . . . 105
4.12 IDP-based Method – Correlations with IDPs . . . 107
4.13 IDP-based Method – Correlations with Non-image Derived UK Biobank Variables . . . 108

5.1 General Unlearning Network Architecture . . . 121
5.2 Age Prediction Data Distributions . . . 126
5.3 Age Prediction Biased Dataset Distributions . . . 129
5.4 General Unlearning Network Architecture with Confound Removal . . . 131
5.5 Segmentation Unlearning Network Architecture . . . 134
5.6 Age Prediction t-SNE Plots . . . 138
5.7 Unlearning Training Graphs . . . 139
5.8 Effect of β . . . 141
5.9 Effect of Batch Size on Validation Loss . . . 142
5.10 Effect of Batch Size . . . 143
5.11 Biased Distributions Density Plots . . . 146
5.12 ABIDE Age Prediction Results . . . 147
5.13 Unlearning Categorical Confounds . . . 148
5.14 Unlearning Categorical Confounds t-SNE Plots . . . 150
5.15 Unlearning Continuous Confounds . . . 150
5.16 Segmentation Dice Scores by Tissue Type . . . 153
5.17 Semi-supervised Segmentation Dice Scores . . . 154

6.1 TrainPrune Algorithm Flow Chart . . . 167
6.2 Channelwise Convolutional Filter Pruning . . . 168
6.3 UNet Model Architecture . . . 172
6.4 Shallower UNet Model Architecture . . . 172
6.5 HarP Example Segmentation . . . 177
6.6 HarP Choice of Pruning Metric . . . 178

6.7 HarP Model Size Comparison . . . 180
6.8 HarP TrainPrune Filter Distribution . . . 181
6.9 HarP Method Comparison . . . 183
6.10 HarP Targeted Dropout Values . . . 184
6.11 HarP TrainPrune Filter Activation Values . . . 186
6.12 HarP Effect of Base Drop Value . . . 187
6.13 HarP Number of Recovery Epochs . . . 190
6.14 HarP Shallower UNet Model Size Comparison . . . 191
6.15 HarP UNet and Shallower UNet Comparison . . . 192
6.16 HarP Shallower UNet TrainPrune Filter Distribution . . . 192
6.17 Cardiac Segmentation . . . 193
6.18 Spleen Segmentation . . . 194
6.19 Prostate Segmentation . . . 194
6.20 Ultrasound Segmentation . . . 195
6.21 IXI Subcortical Segmentation TrainPrune Filter Distribution . . . 197
6.22 IXI Subcortical Segmentation Ablation Study Segmentations . . . 198
6.23 IXI Subcortical Segmentation Ablation Study Dice Scores . . . 199
6.24 IXI Subcortical Segmentation Dice Scores with Training Dataset Size . 200
6.25 HarP Segmentation in Low Data Regimes . . . 201
6.26 HarP Segmentation in Low Data Regimes for Varying Feature Numbers 202
6.27 Comparison of *PruneFinetune* and *TrainPruneTD* in Low Data Regime 203
6.28 Number of Recovery Epochs in Low Data Regime . . . 204
6.29 Heart Segmentation - Low Data Regime . . . 205
6.30 Spleen Segmentation - Low Data Regime . . . 206
6.31 Prostate Segmentation - Low Data Regime . . . 206
6.32 IXI Brain Extraction Segmentation - Low Data Regime . . . 207

7.1 ABIDE Example Segmentations . . . 215
7.2 ABIDE Tissue Segmentation Dice Results . . . 216
7.3 ABIDE TrainPrune Filter Distribution . . . 217
7.4 ABIDE Model Size Comparison . . . 218
7.5 ABIDE Training Epochs Comparison . . . 219
7.6 ABIDE Randomly Initialised Pruned Models . . . 221
7.7 ABIDE Method Comparison . . . 223
7.8 ABIDE Loss Graphs Comparison . . . 225
7.9 ABIDE Comparison to Dropout . . . 226
7.10 ABIDE Comparison to Channelwise Dropout . . . 227

7.11 ABIDE Shallower UNet Model Size Comparison . . . 228
7.12 ABIDE UNet and Shallower UNet Comparison . . . 229
7.13 IXI Brain Extraction Dice Results . . . 231
7.14 IXI Brain Extraction TrainPrune Filter Distribution . . . 232
7.15 IXI Subcortical Segmentation Ablation Study Dice Scores . . . 233
7.16 IXI Subcortical Segmentation Dice Scores with Training Dataset Size . 234
7.17 PCA Component 1 . . . 236
7.18 PCA Component 2 . . . 237
7.19 PCA Component 3 . . . 238
7.20 Augmentation at Test Time - Simulated Bias Field . . . 240
7.21 Augmentation at Test Time - Contrast Change . . . 241
7.22 Augmentation at Test Time - Gaussian Blurring . . . 241
7.23 Augmentation at Test Time - Anisotropic Sampling . . . 242
7.24 ABIDE Augmentation During Training . . . 243

B.1 LeNet . . . 260
B.2 AlexNet . . . 261
B.3 VGG-16 . . . 262
B.4 Inception-V1 . . . 263
B.5 ResNet-50 . . . 264
B.6 DenseNet-121 . . . 264
B.7 Fully Convolutional Network . . . 265
B.8 Autoencoder . . . 266
B.9 GAN . . . 267

C.1 Network with Attention Gate . . . 275
C.2 Healthy Subjects – Non-imaging UK Biobank Variables Correlations . . 277
C.3 Healthy Subjects – IDPs Correlations . . . 278
C.4 Healthy Subjects without Correlating Conditions – Non-imaging UK Biobank Variables Correlations . . . 279
C.5 Healthy Subjects Without Correlating Conditions – IDPs Correlations . 280
C.6 Healthy Subjects Train, All Test – Non-imaging UK Biobank Variables Correlations . . . 281
C.7 Healthy Subjects Train, All Test – IDPs Correlations . . . 282
C.8 Healthy Subjects – Density Plots . . . 284
C.9 Segmentation Results ABIDE . . . 295
C.10 GLM Results . . . 297

C.11 Confusion Loss Visualisation 299
C.12 HarP - Evaluation Metric Comparison 302
C.13 ABIDE Sites Data Distributions 303
C.14 ABIDE Example Slices 303
C.15 ABIDE TrainPrune Results with Different Metrics 304
C.16 ABIDE PruneFinetune with 1% Pruning 305
C.17 ABIDE PruneFinetune with 5% Pruning 305
C.18 ABIDE PruneFinetune with 10% Pruning 305
C.19 ABIDE PruneFinetune with 20% Pruning 305

List of Tables

4.1 Brain Age Prediction Train/Test Splits . . . 88
4.2 Ensemble Prediction Results . . . 92
4.3 Linear Registration Results . . . 93
4.4 Non-linear Registration Results . . . 93
4.5 Effect of Downsampling . . . 100

5.1 Age Prediction Results . . . 137
5.2 Comparison of Full Dataset vs Balanced Dataset . . . 144
5.3 Biased Dataset Results . . . 145
5.4 Unlearning Categorical Confounds when Correlated . . . 149
5.5 Segmentation Location of the Domain Classifier . . . 151
5.6 Segmentation Results . . . 152

6.1 Pruning Datasets . . . 175
6.2 IXI Subcortical Segmentation Dice Scores with Training Data Size . . . 200

7.1 Multisite Pruning Datasets . . . 213

B.1 Brain Age prediction Studies . . . 270
B.2 Harmonisation Studies . . . 271

C.1 Healthy Subject – Ensemble Network Predictions . . . 276
C.2 Healthy Subjects Without Correlating Conditions – Ensemble Network Predictions . . . 283
C.3 DANN Results . . . 298

List of Abbreviations

1D, 2D, 3D, 4D 1, 2, 3, 4 Dimensional

ABIDE . . . . Autism Brain Imaging Data Exchange

AD . . . . . Axial Diffusivity

ADNI . . . . . Alzheimer's Disease Neuroimaging Initiative

ANAT . . . . . Anatomical Processing Script

ANN . . . . . Artificial Neural Network

BEDPOSTX . Bayesian Estimation of Diffusion Parameters Obtained using Sampling Techniques for Crossing Fibres

BET . . . . . Brain Extraction Tool

BIANCA . . . Brain Intensity AbNormality Classification Algorithm

BOLD effect . Blood Oxygenated Level Dependent Effect

BraTs . . . . . Brain Tumour Segmentation

ComBat . . . . Combating Batch Effects when Combining Batches

CNN . . . . . Convolutional Neural Network

CSF . . . . . Cerebrospinal Fluid

CT . . . . . Computed Tomography

DANN . . . . Domain Adversarial Training of Neural Networks

DICOM . . . . Digital Imaging and Communications in Medicine

DL . . . . . Deep Learning

DNN . . . . . Deep Neural Network

DOF . . . . . Degrees of Freedom

dMRI . . . . . Diffusion Magnetic Resonance Imaging

DTI . . . . . Diffusion Tensor Imaging

EPI . . . . . Echo Planar Imaging

FA . . . . . Fractional Anisotropy

FAST . . . . . FMRIB's Automated Segmentation Tool

FCN . . . . . Fully Convolutional Network
FIRST . . . . . FMRIB's Integrated Registration and Segmentation Tool
FLAIR . . . . Fluid Attenuated Inversion Recovery
FSL . . . . . FMRIB Software Library
FLIRT . . . . . FMRIB's Linear Image Registration Tool
fMRI . . . . . Functional Magnetic Resonance Imaging
FNIRT . . . . FMRIB's Non-linear Image Registration Tool
GAM . . . . . Generalised Additive Model
GAN . . . . . Generative Adversarial Network
GLM . . . . . General Linear Model
GM . . . . . Grey Matter
GPR . . . . . Gaussian Process Regression
GPU . . . . . Graphical Processing Unit
HH . . . . . Hammersmith Hospital
HRF . . . . . Haemodynamic Response Function
ICVF . . . . . Intracellular Volume Fraction
IDP . . . . . Image Derived Parameter
IOP . . . . . Institute of Physics
ISOVF . . . . Isotropic Volume Fraction
IXI . . . . . Information eXtraction from Images
KL . . . . . Kullback-Leibler
MAE . . . . . Mean Absolute Error
MD . . . . . Mean Diffusivity
MICA . . . . . Multisite Image Harmonization by CDF Alignment
MICCAI . . . Medical Image Computing and Computer Assisted Interventions
MIUA . . . . . Medical Image Understanding and Analysis
ML . . . . . Machine Learning
MLP . . . . . Multi Layer Perceptron
MNI . . . . . Montreal Neurological Institute
MR . . . . . Magnetic Resonance
MRI . . . . . Magnetic Resonance Imaging
MS . . . . . Multiple Sclerosis

MSE . . . . . Mean Squared Error

NIFTI . . . . . Neuroimaging Informatics Technology Initiative

NN . . . . . Neural Network

OASIS . . . . . The Open Access Series of Imaging Studies

PCA . . . . . Principal Component Analysis

PD . . . . . Proton Density

PET . . . . . Positron Emission Tomography

PROBTRACKX Probabilistic Tractography for Crossing Fibres

RD . . . . . Radial Diffusivity

ReLU . . . . . Rectified Linear Unit

RF . . . . . Radio Frequency

rfMRI . . . . . Resting State Functional Magnetic Resonance Imaging

ROI . . . . . Region of Interest

RVR . . . . . Relevance Vector Regression

SIENA . . . . Structural Brain Change Analysis

SIENAX . . . Cross Sectional Structural Brain Change Analysis

SGD . . . . . Stochastic Gradient Descent

SNR . . . . . Signal to Noise Ratio

SVR . . . . . Support Vector Regression

TBSS . . . . . Tract Based Spatial Statistics

TE . . . . . Echo Time

tfMRI . . . . . Task Functional Magnetic Resonance Imaging

TR . . . . . Repetition Time

TrainPruneTD TrainPrune + Targeted Dropout

t-SNE . . . . . t Distributed Stochastic Neighbour Embedding

VBM . . . . . Voxel-Based Morphometry

VGG . . . . . Visual Geometry Group

WM . . . . . White Matter

WMH . . . . . White Matter Hyperintensities

1

Introduction

The ageing of the global population represents a significant challenge for all sectors of society, particularly healthcare. Ageing has a pronounced effect on the human brain, leading to cognitive decline and an increased risk of neurodegenerative diseases and dementia. With the use of non-invasive methods for *in vivo* imaging, especially Magnetic Resonance Imaging (MRI), and the availability of computational methods for analysing MRI data, cross-sectional and longitudinal neuroimaging studies on brain structure and function are increasingly contributing to the understanding of healthy and diseased structural brain maturation and ageing.

Across neuroimaging, the majority of datasets have been limited to small-scale collections, typically focusing on a specific research question or clinical population of interest. Recently, however, large scale collections have begun to be collated, many

of which are openly available to researchers, meaning that, if the acquisition protocol, demographic and non-imaging information of the data meets the requirements of the study, novel research can be completed without having to acquire new scans. The sharing of these large-scale datasets has had many benefits: not only do they enable the exploration of new research questions, but they also enable reproducibility and quicker method prototyping. UK Biobank [1] is the largest of these studies, a large scale prospective study with a brain imaging component aiming to collect brain imaging data from 100,000 volunteers, incorporating 6 MRI modalities, to study structure, function, and connectivity. UK Biobank also contains large quantities of lifestyle, genetic, and health measures. These allow researchers to create models of population ageing and model how genetic and environmental factors interplay with this process.

Alongside the growth of neuroimaging datasets, deep learning and convolutional neural networks have developed rapidly, and have been applied across medical imaging applications, including in neuroimaging. Neural networks automatically and adaptively learn spatial hierarchies of features; medical images encode rich textural and intensity information which can be learned by the network to complete the task of interest. They have been applied to many tasks and have found success across organised medical imaging challenges [2].

1.1 Thesis Contributions

There remain barriers, however, to deep learning based methods having broader impact, outside of the research domain and specific datasets. For instance, models are hard to interpret and the choice of pre-processing stages applied to the data can affect the model predictions. Therefore, chapter 4 explores the development of a predictive model for ageing using data from the UK Biobank. The effect of the registration

used in the pre-processing stage is explored, and it is demonstrated how non-linear registration (which is often naively applied as a data pre-processing step) can negatively impact model performance, by causing the model to be driven by artefacts from the registration process and to miss subtle but meaningful changes. The relationships between the resulting error in the brain age prediction and both the image-derived values and the non-imaging variables are obtained, and the results compared to an elastic net method trained on the image derived phenotypes (IDPs).

Chapter 5 then addresses the harmonisation challenge – that is, that data collected on different scanners have different characteristics due to the differences in scanner hardware and acquisition rather than biological variability, even when attempts have been made to harmonise across scanners. It is proposed that domain adaptation methods could be used to harmonise across scanners, and a domain adaptation framework is adapted for the harmonisation of the outputs of both age prediction and segmentation networks. The method is demonstrated across a range of data scenarios, and it is shown that the method enables the removal of scanner information while maintaining the performance on the main task.

Chapter 5 also explores the use of the harmonisation method when few labels are available for one of the scanners, and it is demonstrated that the unlearning domain adaptation approach leads to a large improvement in performance when few labels are available for a given site. This is a key challenge when applying deep learning across medical imaging tasks: acquiring labels is expensive, relies on expert knowledge and is time intensive, so there is often a lack of available labels for training. This is especially problematic for segmentation tasks, where dense labels covering all voxels are usually required. One of the causes of the need for a large number of data points is the vast number of parameters in neural networks, which cause the networks to overfit. Therefore, in chapter 6 a method to simultaneously train and

prune the UNet architectures is developed, meaning that the network decreases in size as the model converges. An adaptive targeted dropout algorithm is also developed, to encourage the model to be robust to the pruning. It is demonstrated that these models outperform the equivalent standard UNet in low data regimes and across medical imaging segmentation tasks, allowing acceptable segmentations to be produced when very low numbers of labels are available.

Chapter 5 also demonstrates how the proposed harmonisation method increases the performance on sites not seen during training, as the features which are produced generalise better. Model generalisability is another challenge for the application of deep learning based methods, as models trained on one site tend to suffer large decreases in performance when applied to other sites. This is true even when attempts have been made to match acquisitions across scanners, such as in multisite datasets. In chapter 7 the pruning method proposed in chapter 6 is applied to multisite MRI data, and it is shown that this leads to UNet models which are not only smaller, but also generalise better to unseen sites, across segmentation tasks.

1.2 Thesis Structure

This thesis is structured as follows:

- Chapter 2 discusses the MR background, focusing on the image analysis stages and structural images, but also discussing the aspects of functional and diffusion imaging used within the experiments in this thesis. Within structural imaging, the analysis pipeline commonly used and the tools which are used to generate segmentations from these processed images are explored, as these are used in later chapters, and so, understanding the model assumptions is vital.

- Chapter 3 then discusses the deep learning background required for this thesis, focusing on convolutional neural networks, as these are the most prevalent across image analysis tasks. This chapter also contextualises the challenges which arise when deep learning is applied to neuroimage analysis.

- Chapters 4 - 7 then present the contributions previously introduced. Chapters 4 - 6 take the same structure, beginning with motivation and a discussion of the literature specific to each contribution; subsequently, the methods are presented, alongside the datasets and implementation details. The results and discussion are then presented, followed by a conclusion which will focus on the strengths and limitations of the work. For chapter 7, the method follows that introduced in chapter 6.

- Chapter 8 is then the conclusion of the thesis, which will explore the strengths and weaknesses of the contributions, and will briefly explore possible future directions.

1.3 Statement of Originality

I declare that I am the sole author of this thesis document, and I produced all of the tables and figures included herein. Critique and comments were provided by Prof. Mark Jenkinson and Dr Ana Namburete. Unless otherwise specified, the pre-processing, code development and analysis was my own, individual work. Details of all datasets used can be found in Appendix A.

- Chapter 3: the challenges section is derived from a review paper of which I was the primary author. The original draft of the data privacy section was written by Emma Bluemke, but I have redrafted it based on the original.

- Chapter 4 utilises the UK Biobank data, which was processed according to the UK Biobank pipeline by Fidel Alfaro Almagro, apart from the application of the warp field to produce the linear registered images, which I completed. The work was a joint project with Emma Bluemke, who explored the initial methodology and initial model design. I reimplemented and trained all of the models in this thesis, and designed and implemented the visualisation experiments, and the exploration of the effect of health on the model. The scripts for the correlations were produced by Steve Smith, Zobair Arya and Diego Vidaurre, and the IDP based model was implemented by Steve Smith and Diego Vidaurre. I completed the analysis of the results. Advice for screening for health was provided by Ludovica Griffanti and I implemented the screening. I produced the final manuscript and am thankful to the reviewers for their constructive comments and suggestions throughout the review process.

- Chapter 5 again uses data from the UK Biobank. I screened the data from the OASIS site to remove subjects with relevant health conditions, based on the advice of Ludovica Griffanti. The screening of the Whitehall data was completed by Luca Melazzini. Initial advice for implementation of the method was given by Mohsan Alvi. I completed all other processing, method development and analysis. Thanks is again given to the reviewers for their constructive comments to help develop to the work.

1.4 List of Publications

Journal Articles

- Vaanathi Sundaresan, Giovanna Zamboni, **Nicola K Dinsdale**, Peter M Roth-

well, Ludovica Griffanti, Mark Jenkinson: *Comparison of domain adaptation techniques for white matter hyperintensity segmentation in brain MR images*, in Medical Image Analysis, 102215, August 2021 (In Press) doi:10.1016/j.media.2021.102215 (Contributions in Chapter 5.)

- **Nicola K Dinsdale**, Mark Jenkinson, Ana IL Namburete: *Deep Learning-Based Unlearning of Dataset Bias for MRI Harmonisation and Confound Removal* in Neuroimage, Volume 228, March 2021. doi:10.1016/j.neuroimage.2020.117689 (Contributions in Chapter 5.)

- **Nicola K Dinsdale***, Emma Bluemke*, Stephen Smith, Zobair Arya, Diego Vidaurre, Mark Jenkinson, Ana IL Namburete: *Learning patterns of the ageing brain in MRI using deep convolutional networks* in Neuroimage, Volume 224 January 2021. doi:10.1016/j.neuroimage.2020.117401 (Contributions in Chapter 4.)

Peer Reviewed Conference Proceedings

- **Nicola K Dinsdale**, Mark Jenkinson, Ana IL Namburete: *Unlearning Scanner Bias for MRI Harmonisation* in Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Lecture Notes in Computer Science, Volume 12262. doi:10.1007/978-3-030-59713-936 (Contributions in Chapter 5.)

- Jorge Corral Acero, Vaanathi Sundaresan, **Nicola K Dinsdale**, Vicente Grau, Mark Jenkinson: *A 2-Step Deep Learning Method with Domain Adaptation for Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Magnetic Resonance Segmentation* in Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges. STACOM 2020. Lecture Notes in Computer

Science, Volume 12592. doi:10.1007/978-3-030-68107-420 (Contributions in Chapter 5.)

- **Nicola K Dinsdale**, Mark Jenkinson, Ana IL Namburete: *Unlearning Scanner Bias for MRI Harmonisation in Medical Image Segmentation* in Medical Image Understanding and Analysis. MIUA 2020. Communications in Computer and Information Science, Volume 1248. doi:10.1007/978-3-030-52791-42 (Contributions in Chapter 5.)

2

Magnetic Resonance Image Analysis

Medical imaging is a vital part of medical practice, whether for diagnosing an illness or for monitoring the progression of the disease. Whilst medical knowledge and discernment form the basis of any diagnosis, medical imaging assists in the decision making process, providing insight into what is happening inside the body, which is otherwise difficult or impossible to observe. When tracking the progression of a disease, medical imaging is a key tool for monitoring the effectiveness of treatment and allowing the adjustment of protocols as necessary. Ultimately, the detailed information generated by medical imaging can provide patients with better and more comprehensive care.

Several modalities of medical imaging are common in clinical practice: x-ray, computed tomography (CT), positron emission tomography (PET), ultrasound, and magnetic resonance imaging (MRI) all form the standard repertoire of medical imaging.

MRI is a powerful and versatile modality, with the advantage over x-ray and CT of being non-ionising, and over ultrasound of being able to penetrate the calcified skull, allowing safe and non-invasive imaging of the brain. The images captured by MRI are able to provide information about gross anatomy, connectivity, pathologies and neuronal activity, due to the flexibility of the image acquisition. The received signals from the different tissues in the brain can be manipulated to create different image contrasts, providing varied information about the brain.

This thesis will focus on image analysis and so anything more than the most introductory aspects of the physics of image acquisition is beyond its scope. Instead, this section on MRI will explore the different acquisitions used in *in vivo* human neuroimaging and the processing pipelines that are commonly used to enable us to complete image analysis. The focus will be on structural MR images, the analysis pipeline used to allow us to analyse the images, and specifically on the tools in the FMRIB Software Library (FSL) toolkit [3]. The FSL toolkit (FSL 6.0.1) is used throughout this work, both to process the images and produce segmentation labels. Other packages could have been used, for instance The SPM[1] software package or FreeSurfer[2]. The choice of package used to process the data should not affect the results for chapters 5-7, where comparison is being made between methods using identical data. For chapter 4, the choice of processing would be likely to affect the results, especially consideration of the image derived phenotypes, due to underlying model assumptions. [269] details the choices of processing used for the UK Biobank data, and discusses decisions made in the pipeline.

[1] `https://www.fil.ion.ucl.ac.uk/spm/software/`
[2] `https://surfer.nmr.mgh.harvard.edu/fswiki`

2.1 MRI Acquisition Basics

MRI systems are formed of multiple parts that generate and receive the magnetic fields and signals. This includes the main magnetic coils, gradient coils, and radiofrequency (RF) coils. The main magnet creates the B_0 field, a strong, constant magnetic field – typically $1.5T$, $3T$ or $7T$ for modern human MRI scanners. The gradient coils act to slightly alter the B0 field at different positions in the scanner, providing spatial localisation. This spatial encoding means that the MR signal can later be transformed into an image. The RF coils then have two main purposes: the first is to transmit RF energy to the tissues of interest, and the second is to receive the induced signal back from the tissue being imaged. The RF fields are referred to as B_1 fields. When the RF transmitter coils are switched on, the B_1 transmit field produced combines with the B_0 field to manipulate the magnetisation of the tissues in the body, which in turn generates the MR signals that are detected by the receive coils.

The primary origin of the received MR signal in human brain imaging is from hydrogen nuclei in the water within tissues. Hydrogen nuclei are single protons and thus possess intrinsic magnetisation that interacts with magnetic fields induced by the scanner, and they in turn generate small, changing magnetic fields that are detected by the RF coils. Normally, the magnetisation of the protons is randomly orientated, meaning that there is no net magnetisation. Under the B_0 field, the protons align slightly, creating a net magnetisation and meaning that in the steady state the protons are aligned, on average, with the main magnetic field. To induce a detectable signal, the RF transmit coils create an excitation pulse that tips the average magnetisation away from B_0. Once excited, the magnetisation precesses (i.e. spins around the main field axis) at a resonance frequency that is proportional to the applied magnetic field, which is known as the Larmor frequency. As it precesses, it also relaxes back into alignment

with the B_0 field. The speed of relaxation is described by empirical time constants $T1$, $T2$ and $T2^*$, which are key in creating the image contrast. $T1$ describes the speed of recovery in the longitudinal (z) axis component of the magnetisation, where the B_0 field is conventionally aligned with the z axis, and $T2$ and $T2^*$ describe the speed of the magnetisation decay in the transverse (x-y) plane. $T2^*$ is always larger than $T2$ as it reflects both $T2$ decay as well as that from magnetic field inhomogeneities, which can often be refocused and hence recovered.

By varying the MR sequence, e.g. by changing the timing with which the RF pulses are applied and signal measurements are made, different types of images are created. *Repetition time* (TR) is the amount of time between successive RF excitation pulses applied to the same tissue, and *echo time* (TE) is the time between the delivery of the RF excitation pulse and the measurement of the signal. Broadly speaking, T1-weighted images are produced by using short TE and TR times, and therefore the contrast and brightness are dominated by the T1 properties of tissues. Conversely, T2-weighted images are produced by using longer TE and TR times, leading to the image being dominated by the T2 properties of the tissue. Using these properties different aspects of the tissues being studies can be captured, and thus different questions about the human brain explored.

2.2 Structural MRI

Structural MR images are normally used to show the gross anatomy of the brain. The most common type of structural scan is the T1 weighted image, an example of which is shown in Fig. 2.1 where the grey matter, white matter and CSF can be clearly seen surrounded by the scalp and other tissues in the head, since the skull is dark, due to a low concentration of hydrogenated molecules. In Fig. 2.1 the acquired

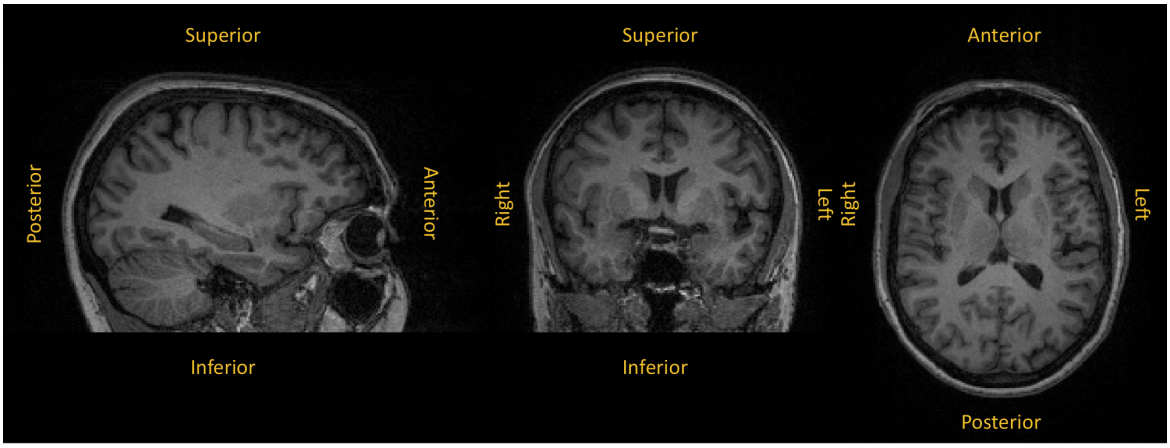


Figure 2.1: An example T1-weighted MRI scan, taken from the OASIS dataset [4]. The same subject will be used throughout the structural figures.

image has been taken off the scanner and converted from a DICOM to a NIFTI image format. The image has then been cropped so that the brain is central in the image and the neck, which we do not wish to analyse, has been removed. The image is in the standard coordinate frame such that the sagittal plane is the y-z plane, the coronal plane is the x-z plane and the transverse plane is the x-y plane. These pre-processing steps are common across most analyses and are central to ensuring data consistency when input into neural networks.

The structural image gives information about the individual anatomy allowing research into, for instance, changes in brain anatomy due to neurodegeneration; they are also vital for any group analysis, even for functional and diffusion studies, as they are a key part of registration. The image shown is a $1mm$ isotropic scan, meaning that the voxel size is $1 \times 1 \times 1mm$. This is common in the research setting but less common in a clinical setting where the acquired images often have very small in-plane voxel sizes but with thick through-plane slices. These anisotropic voxels are normally acquired in the clinical setting because they reduce both the acquisition time and the time taken for a radiologist to visually inspect the images, for instance, for pathological

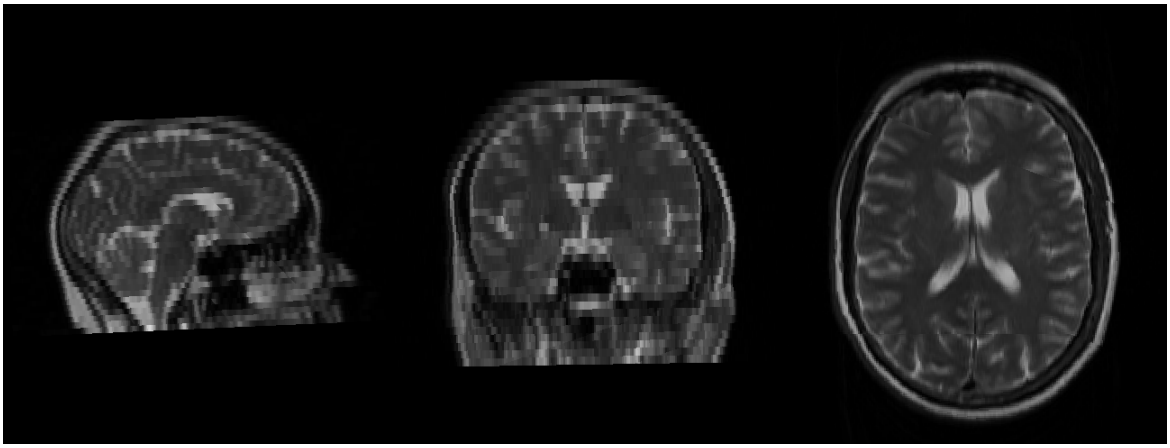


Figure 2.2: An example T2-weighted MRI scan showing the same subject as Fig. 2.1. It can be seen that the resolution is high in the axial plane (third column) but lower in the other two planes, due to the anisotropic voxels.

lesions and anatomical deformations. Figure 2.2 shows an anisotropic scan.

Different acquisition protocols lead to different types of structural scans being acquired, which help to show different aspects of the anatomy. Commonly acquired structural scans, in addition to the most common T1-weighted scans, are T2-weighted (Fig. 2.2), proton density (PD) and FLAIR, which are similar to T2-weighted but with additional fluid suppression. When working in research with only healthy participants, it is typical to acquire only the T1-weighted images as they are normally the best quality, enabling the clearest delineation of the healthy tissues. However, if we are interested in, for instance, research into Multiple Sclerosis, where lesions develop in the white matter, then these will be very faint and hard to distinguish on T1-weighted scans but are much clearer on FLAIR images.

2.2.1 Partial Volume Effects

Regardless of the spatial resolution of the acquired image, the underlying structures being imaged in the brain will be smaller than the voxel size achieved. This means

that many voxels contain signals from a range of tissues, and that the signal received is a weighted sum of the signal from all of the structures present. This effect is referred to as the partial volume effect, and it affects how precisely tissue boundaries can be located. For instance, this could affect the estimated volume of segmented structures in the image.

2.2.2 Intensity Normalisation

Voxel intensity values in T1 images do not hold intrinsic values; the value is expressed in arbitrary units, thus different scanners or acquisition protocols result in large intensity differences. This intensity variation can greatly undermine the performance of any subsequent MRI processing and thus there is a need for intensity normalisation. The normalisation process needs to produce values which have a common interpretation across locations in the same tissue type and preserve the rank of intensities [5]. Further, the normalised intensity values should not be influenced by biological abnormality, but need not to result in loss of information associated with pathology. Commonly, some modified version of histogram matching is used to match intensity values across datasets [5].

2.2.3 Image Artefacts

A number of artefacts can occur in MR images, many of which can be dealt with in the analysis stages, but which need to be identified through inspection of the data. Some more serious artefacts will result in the image needing to be discarded. Common artefacts include:

- *Motion artefacts*: caused by the subject moving in the scanner; it leads to blurring and ringing in the image.

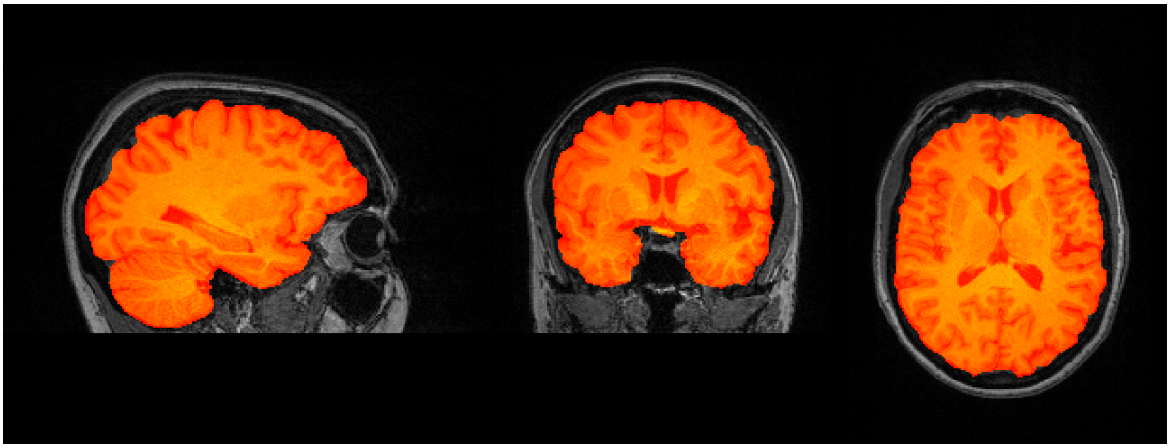


Figure 2.3: Example brain extraction applied to a T1-weighted image; the orange represents the brain mask. Brain extraction was completed using FSL BET [6].

- *Bias field inhomogeneity*: inhomogeneity in the B_1 field due to imperfections in the coils lead to low frequency variation in the intensity of the same tissue across the image, the correction of which is discussed in 2.2.6.

- *Ghosting*: caused by irregularities in k-space sampling, including motion-induced effects, leading to shifted, fainter versions of the image (the 'ghosts') to be superimposed on the original image.

- *Gradient non-linearity distortion*: imperfections in the linearity of the gradient coils leads to distortions in the final image, especially further away from the isocentre of the coils, such as in the neck and spine.

- *Wrap around distortion*: caused by the object being larger than the chosen field of view of the image, which is most commonly seen with the nose wrapped around to the back of the head, or even within the brain in extreme cases.

2.2.4 Brain Extraction or Skull Stripping

Once the data is cropped and in standard orientation, the first step of the main analysis pipeline is to complete brain extraction or skull stripping so that only brain tissue remains. Consequently, any subsequent analysis focuses only on what we are interested in, and prevents the irrelevant regions from having any impact on the downstream analysis, especially the registration stages.

As the brain extraction occurs early in the pipeline, and the result impacts all future steps, consideration should be given as to what is going to be most important for the downstream analysis. For many applications, removing a small amount of brain tissue or leaving behind small amounts of non-brain tissue will be sufficient. However, if the analysis requires very accurate results, such as voxel-based morphometry (VBM) that includes quantification of the volumes of different brain tissues, much more accurate and consistent brain extraction will be required. Similarly, the accuracy of the brain extraction will be more important for non-linear inter-subject registration far more than linear within-subject registration.

When brain extraction is performed throughout this thesis it is performed either using FSL BET [6] (within the processing pipelines) or the FNIRT-based extraction as part of FSL Anat[3] (for production of ground truth labels). An example of brain extraction using BET can be seen in Fig. 2.3. BET uses a deformable surface model that evolves to fit the brain's surface through the application of a set of locally adaptive model forces. The model initialisation is a triangular tessellation of a sphere's surface inside the brain, which is then allowed to deform slowly. The forces aim to keep the surface well-spaced and smooth, while attempting to move the sphere towards the brain's edge. If the model does not reach a suitably clean solution, then the process

[3] `https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/fsl_anat`

is re-run with a higher smoothness constraint. The FNIRT-based method used in FSL Anat – see section 2.2.5 for details of the registration process – transforms a standard-space mask to the input image, using non-linear registration. Through the brain extraction process, a brain mask is produced which is simply a binary mask representing valid brain voxels, which can also be used in downstream analysis.

The limitations of the brain extraction process must be acknowledged when interpreting any analysis results. The most clear is that the boundary at which something is included or not is somewhat arbitrary, so the amount of CSF, and brainstem especially, will depend on the choice of cut-off. Finally, it should be noted that brain extraction does not perform the same task as cortical modelling such as that performed by FreeSurfer [7]. Skull stripping or brain extraction aims only to remove the main non-brain structures and does not try to represent the fine detail of the cortical folds. This is sufficient for many of our analysis purposes, but would not be appropriate if, for instance, the aim were to characterise the cortical surface.

2.2.5 Registration

The primary purpose of the registration stage of the processing pipeline is to align the anatomy between images, whether between two images from different subjects, the same subject in different sessions, or an image of a subject and a template image. In group analyses, registration is typically used to align the structural image to an image template. Thus, before discussing registration in depth, templates must first be discussed.

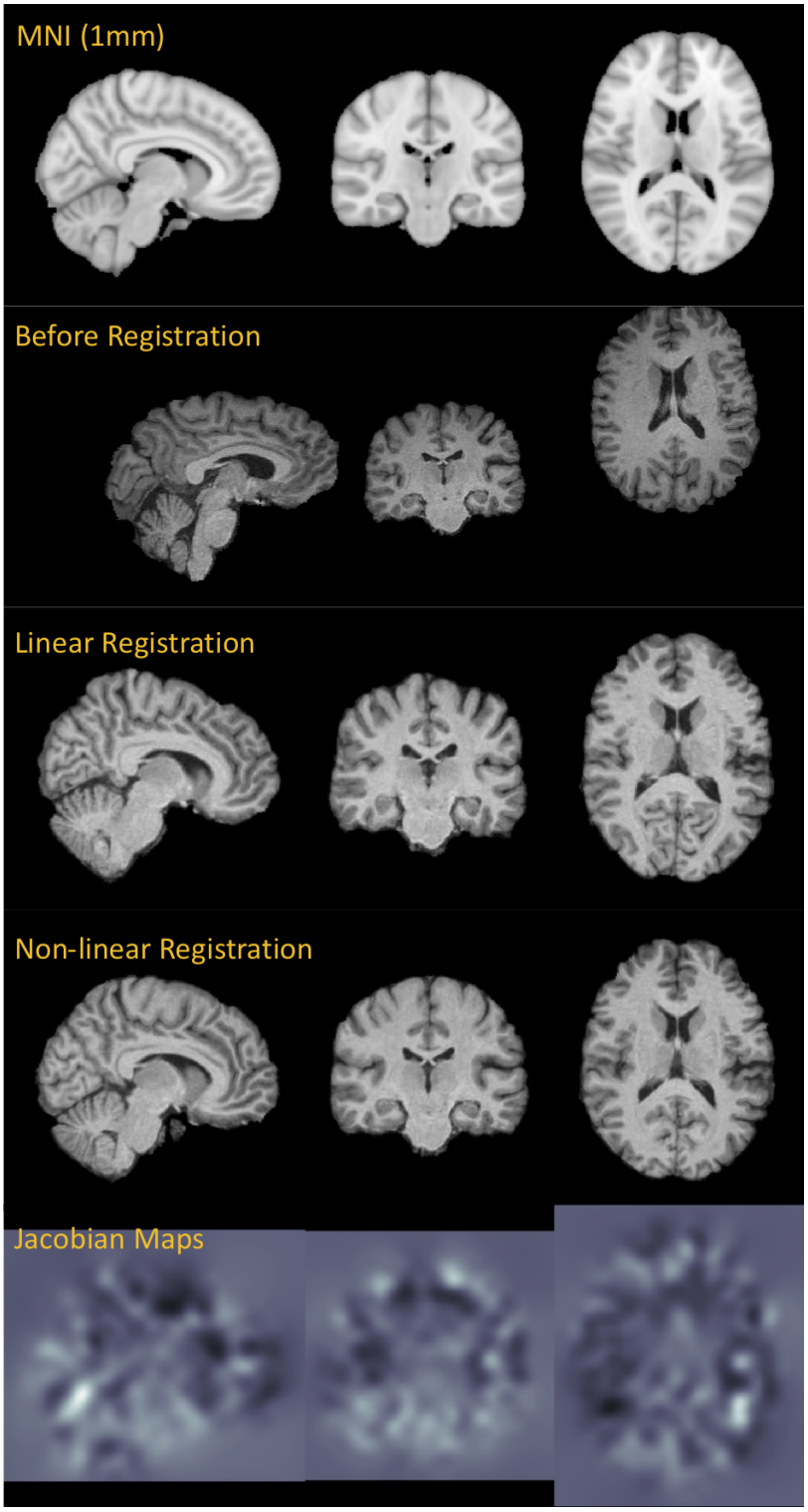


Figure 2.4: Demonstration of types of registration. a) MNI 1mm template, b) Image before registration, c) Linear registration with FSL FLIRT [8], d) Non-linear registration with FSL FNIRT [9], e) Jacobian maps produced by FNIRT.

2.2.5.1 MNI Template

The MNI template[4] is the most commonly used human brain MRI template. It is an average of 152 healthy adults registered to a common space using non-linear registration, and many versions exist spanning different resolutions (e.g. $2mm$, $1mm$) and different modalities. The version used in FSL, and so throughout this work, is an asymmetric version of the MNI ICBM 152 non-linear 6th Generation atlas [10]. Alternative, study specific templates can also be produced by creating an average image of the group of subjects in the study.

The MNI template is described as being in *standard space* whereas the acquired image will be in image space, where the term space is effectively used to describe the coordinate frame of the brain. The concepts of different spaces can be seen in Fig. 2.4 by comparing the first two rows showing the MNI template and the subject in image space before registration. Clearly, the two images are not aligned and the anatomy does not appear in the same region. Therefore, one of the purposes of registering an image to the standard template is to move each subject into the standard space or coordinate frame, thus facilitating meaningful voxelwise comparisons.

2.2.5.2 Linear Registration

Registration is about aligning the anatomy between different images, and so it is important to understand how to describe the transformation of an image in space. Spatially transforming an image can either require small transformations such as shifts and rotations (for example, when registering different sessions of the same subject) or more substantial transformations, such as changing the cortical folds. Consequently, the type of registration algorithm used relies on some judgement as to what is most

[4] `www.lead-dbs.org/about-the-mni-spaces/`

appropriate for the task. If the anatomy is expected to be the same between scans, then linear transformations are more appropriate than non-linear transformations that would enable, for instance, changes in tissue volumes that are not biologically motivated.

Linear registration methods can broadly be broken down into two types: rigid body and affine transformations. Rigid body transformations have 6 degrees of freedom (DOF): 3 rotations, one about each axis, and 3 translations, one along each axis. These transformations describe how the brain can move in space without changing in anatomy or size, and so are most appropriate when the anatomy will have not changed – so the same subject between sessions (over a short interval) or between 3D volumes acquired in a 4D session (4th dimension being time) for example in a functional scan. Affine transformations, in comparison, have 12 degrees of freedom: 3 rotations, 3 translations, 3 scaling and 3 shears (or skews). Fig. 2.4 shows the linear registration of a subject to the standard template using affine registration. It can clearly be seen that the registration was affine and not rigid body by considering the coronal slice that is clearly larger than it was before registration, indicating that it has been scaled to the MNI template.

2.2.5.3 Non-linear Registration

When there are larger changes in anatomy – such as different subjects with different folding patterns or the same subject in a longitudinal study with sufficient time between scans that there are notable changes (e.g., increase in ventricle size due to ageing) – there is often a need for non-linear registration. The key conceptual difference between linear and non-linear registration is that linear registration applies to the whole image equally and is not able to cause local transformations. Therefore, non-linear registration is required to align images where local differences exist, typically using thousands to millions of degrees of freedom. Fig. 2.4 shows the same subject registered non-linearly to

the template, using linear registration as the initialisation, completed using FLIRT [8].

Non-linear registration has to balance two aims: the image needs to be warped so as to match the reference image or template, but also the transformed brain needs to still be biologically plausible. Therefore, constraints are added to the deformation. First, the transformation must be diffeomorphic – that is, that it must maintain topology and not, for instance, allow a structure being removed or fractured as part of the registration process. The second major constraint is that the field is regularised – that is, that the field is encouraged to be smooth, reducing the sudden local transformations that can be applied. The main purpose of regularisation is to balance matching the anatomy with ignoring artefacts and noise in the image, which ideally should not drive the registration. The effect of the two constraints can be seen by considering the Jacobian map in Fig. 2.4 that illustrates the determinant of the Jacobian matrix that is calculated for each voxel, which is based on the local gradient of the transformation and expresses the change in volume for each local element. If we consider the region around the ventricles – clearest in the coronal slice – it can be seen that the Jacobian map contains little information, indicating that little transformation occurred in this region, as would be expected. Meanwhile, if we consider the cortex, regions of bright intensity and low intensity can be seen. Bright indicates expansion of the voxels, whereas dark indicates contraction of the voxels; however, there are no negative values that would indicate folding and thus mean that the transformation was not diffeomorphic.

Non-linear transformations are used to register individual subject T1-weighted images to the MNI template. As the non-linear registration has so many degrees of freedom, the optimisation is costly and takes a long time, with many possible solutions. Therefore, the registration is normally initialised using an affine registration, meaning that the non-linear registration has less alignment to complete, with the overall position, rotation and scale already corrected for.

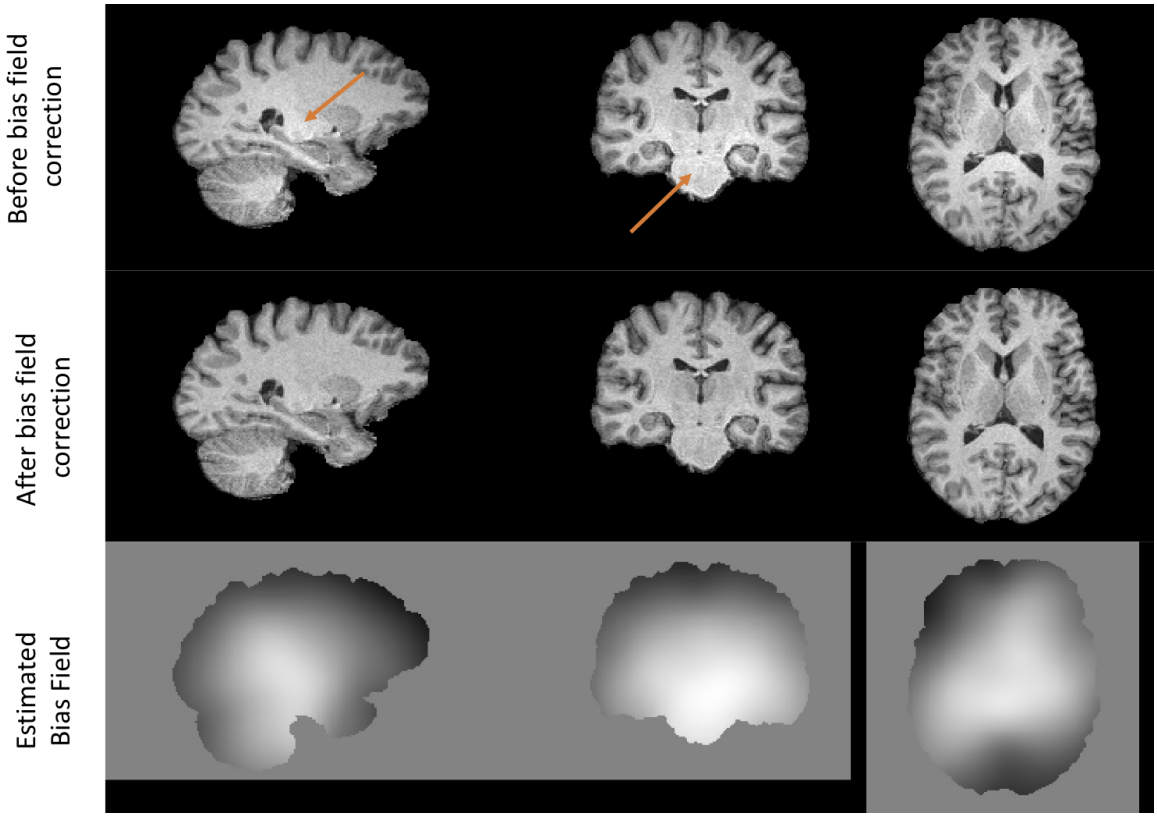


Figure 2.5: A T1-weighted image a) before bias field estimation, b) after bias field removal, c) estimated bias field. Arrows indicate regions of relatively high voxel intensity due to the bias field, which is not present after bias field estimation and removal. Bias field correction was completed with FSL FAST [11].

2.2.6 Bias Field Correction

If the sensitivity of the receive RF coil to detect the B1 field is not uniform across the whole brain, images are created that have lighter areas and darker areas, independent of the tissues present. These changes in intensity vary slowly across the image: that is, they have a low spatial frequency. This effect is often referred to as the bias field and an example can be seen in Fig. 2.5. Here the arrows indicate areas that are artefactually bright, as shown by the estimated bias field.

This bias field causes problems for downstream analysis, especially segmentation tasks, which are largely based on intensity information. Bias field can also cause

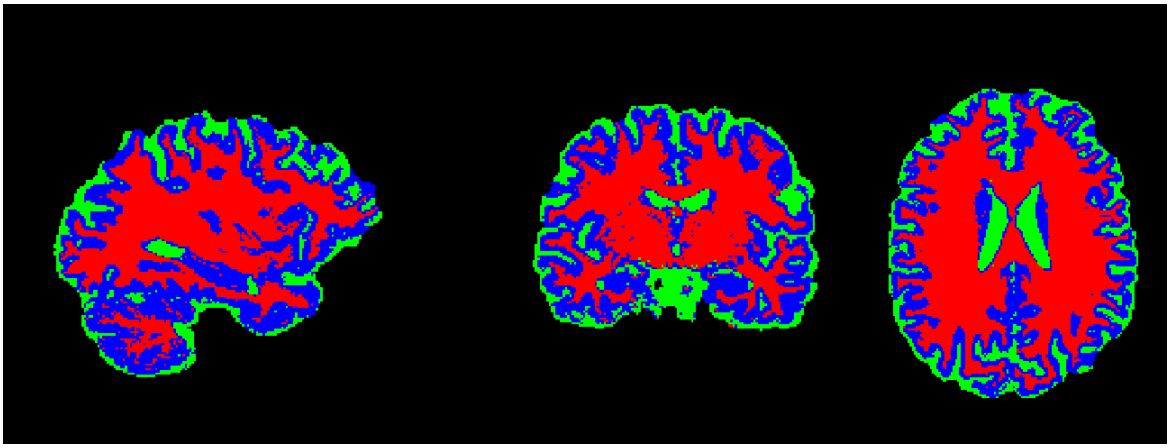


Figure 2.6: Tissue segmentations – white matter (red), grey matter (blue), CSF (green), generated using FSL FAST [11].

difficulties for brain extraction. Within FSL, bias field correction and tissue segmentation (white matter, grey matter, CSF) are completed together in an iterative process within the FAST tool [11]; the principles behind the methods will be explained together in section 2.2.7.

2.2.7 Segmentation

There are two main types of segmentation that need to be considered: tissue-type segmentation and structural segmentation. Segmentation aims to allocate a label to the voxel based on the anatomy at that location, and it is important to understand the assumptions behind the underlying models, to assess the validity of the produced segmentation labels.

2.2.7.1 Tissue Segmentation

FAST [11] performs tissue segmentation – that is, the classification of each voxel in the brain into white matter, grey matter or CSF. This is completed using a Gaussian mixture model with three components in the histogram of the voxel intensities and so is

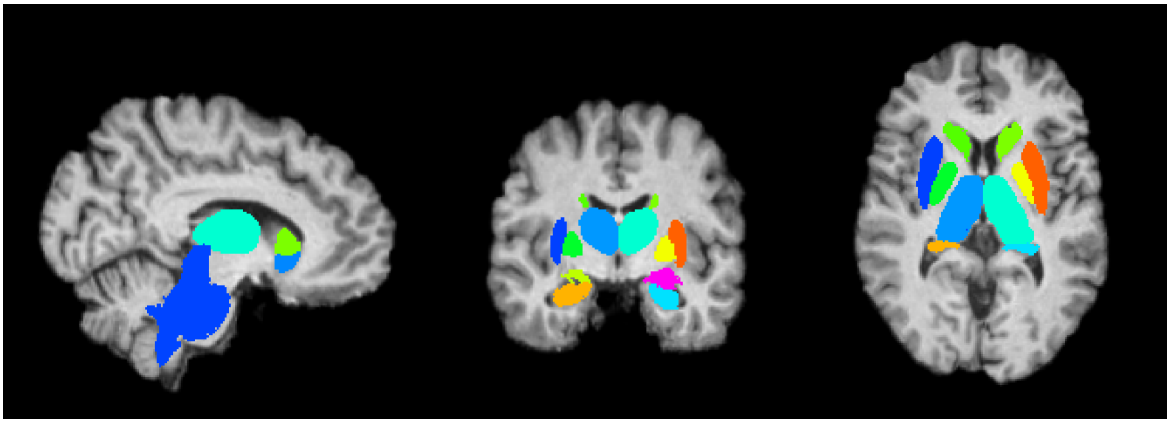


Figure 2.7: An example segmentation of deep grey matter structures produced using FSL FIRST [12]

sensitive to any bias field present in the image, which may cause the same tissue located in different regions of the brain to have very different intensity values. Therefore the bias field is estimated alongside the segmentation and the two are iteratively updated until convergence is reached. Alongside the Gaussian mixture model, the segmentation is also controlled by a Markov Random Field that incorporates spatial neighbourhood information into the segmentation, the contribution of which to the final segmentation is controlled using a weighting factor. By enforcing spatial similarity, i.e. that neighbouring voxels are more likely to have the same tissue, the segmentation result is less likely to be driven by noise. An example output segmentation from FAST is shown in Fig. 2.6, which was created alongside the bias field estimation shown in Fig. 2.5.

2.2.7.2 Subcortical Segmentation

FIRST [12] performs the segmentation of deep grey matter structures that cannot be achieved through only using voxel intensities. Therefore, anatomical information must be incorporated about where the structures are within the brain and what shapes

they have. Thus, the model for FIRST has training data with ground truth manual segmentation labels from T1-weighted MR images, which varied across an age range of 4-87, and a range of healthy subjects and patients with schizophrenia and Alzheimer's disease, helping the model to be general and to cope with varying demographics. All the training data was aligned to the MNI standard space using a two-stage linear registration process, concentrating on the alignment of the subcortical regions, possibly at the expense of the cortex. The registration to the MNI template normalises for brain size. The model uses the training data to learn a 'mesh' for each subcortical structure which describes the typical shape and location in standard space, along with the likely modes of variation. When applied to an unseen subject, these meshes are first warped to the subject space, and then deformed according to the model priors and intensity differences across boundaries for that particular subject. Boundary correction is also used to ensure that the neighbouring structures do not overlap, such that every voxel is only allocated a single label. An example segmentation, with all of the structures FIRST can segment visible, is shown in Fig. 2.7.

2.2.7.3 Other Measures Derived from Structural Images

SIENA and SIENAX: SIENA [13] is a tool for the analysis of brain change, particularly the estimation of atrophy. SIENA estimates a percentage volume change between two input images of the same subject, acquired at different points in time. The images are brain extracted, registered with the scaling determined by an estimate of the skull size, and the brain changes between time points are estimated.

SIENAX [14] allows for cross-sectional analysis of the total brain tissue volume. A single image is taken, the skull size estimated, and then the image is brain extracted and transformed to match the standard space image using the skull size to set the scaling. Tissue segmentation is then used to estimate the volume of the brain tissue,

and multiplied by the estimated scaling factor to reduce head-size-related variability between subjects.

Voxel Based Morphometry (VBM): VBM [15] is a method for voxelwise comparison of the local concentration of grey matter between two or more groups of subjects. The grey matter segmentations are spatially normalised and modulated by the Jacobians to compensate for volume changes made by the non-linear registration. The grey matter maps are then spatially smoothed and statistical tests performed between the groups to identify voxels where the GM concentration differs significantly between the groups.

Brain Intensity AbNormality Classification Algorithm (BIANCA): BIANCA [16] is a supervised method for white matter hyperintensity detection (WMH), using a k-nearest neighbour algorithm. The image voxels are classified based on their intensity and spatial features, and the output image represents the probability of each voxel being a WMH.

2.3 Diffusion MRI (dMRI)

Diffusion MRI (dMRI) sequences are typically used to examine the structural connectivity of the brain and to allow exploration of tissue microstructure. Importantly, for the analysis that will be considered in this thesis, diffusion MRI has also been proven to be effective at quantifying microstructural changes to tissue, including those relating to tissue damage. For instance, it can be useful for quantifying changes due to stroke. It provides us with two types of information: local estimates of diffusion properties which can reflect the integrity of microstructural compartments, and long

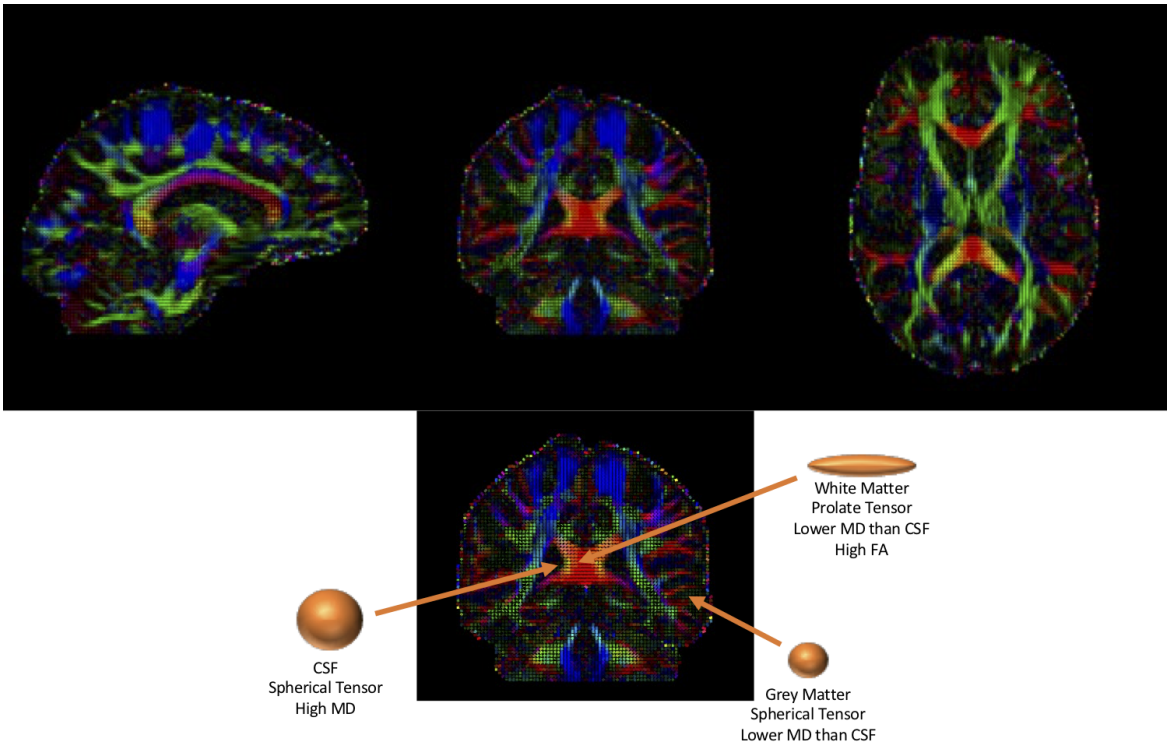


Figure 2.8: Diffusion Tensor Image (DTI) showing the primary direction of diffusion modulated by the fractional anisotropy (FA) map. The lower figure indicates the shape of the diffusion tensor in the different tissues.

range connectivity estimates based on tract tracing, reflecting the likely structural connectivity between brain regions.

Diffusion MRI makes inference on the movement of water and thus is very informative in white matter bundles, where the movement of water is restricted by the axons and is impeded by the membranes; thus, in the white matter, the flow of water becomes more anisotropic than in the other tissues, with diffusion primarily occurring along the axial direction. Therefore, dMRI quantifies the physical properties of the diffusing water molecules, such as how far water molecules move in a given time. To achieve this, the timings of the excitation and sampling are arranged, alongside additional gradient pulses, to be sensitive to the movement of water.

Having acquired the diffusion data, as with other modalities, the diffusion data

must be pre-processed. Diffusion data is normally acquired using a fast Echo-Planar Imaging (EPI) sequence and involves strong, rapidly changing gradients to produce the diffusion images. Therefore, the diffusion images suffer from substantial artefacts.

Diffusion images first have to be pre-processed to remove these artefacts. Fast imaging is essential for dMRI and so EPI acquisition sequences must be used. These sequences are particularly sensitive to susceptibility related inhomogeneities in the B_0 field such as those caused by the air in the sinuses. The distortion is at its worst in the inferior frontal and inferior temporal areas of the brain as they are nearest the sinuses. Another major artefact in diffusion images is eddy current distortion. To acquire diffusion MRI, strong diffusion encoding gradients must be used, which are turned on and off rapidly. This, therefore, induces eddy currents, which in turn generate unwanted magnetic fields which create distortions in the acquired image. Without correction, many anatomical locations across the brain would appear in the wrong location in the image.

2.3.1 Diffusion Tensor Imaging (DTI)

Diffusion tensor imaging (DTI) [17] uses a simple model for the diffusion of water in the brain, assuming that the diffusion in different directions can be modelled using a spatial Gaussian probability, usually visualised by an ellipsoid. The orientation of the long axis of the ellipsoid provides an estimate of the direction of the fibre (where there is maximum diffusion) and the relative width gives information about how constrained or free the diffusion is. The assumptions behind the model are reasonable when a single fibre bundle is present in a voxel. The model can be used to extract many physical values from the images. A single tensor, representing the direction and magnitude of the diffusion, is created for each voxel individually.

From this, different quantities can be extracted including mean diffusivity (MD) and fractional anisotropy (FA). Fig. 2.8 shows a DTI image along with the diffusion tensor in each voxel of the image, modulated by the FA, enabling visualisation of the major white matter bundles. Mean diffusivity describes the average mobility of the water molecules, whereas the fractional anisotropy describes the degree of anisotropy of the movement of the water present in that voxel, varying between 0-1, where 0 indicates a totally isotropic tissue – i.e. freely diffusing water – and 1 indicates totally anisotropic diffusion. Other values which are frequently considered are the axial diffusivity (AD), that is, the diffusion in the strongest direction, and the radial diffusivity (RD), the diffusivity perpendicular to the strongest direction. These values provide indirect information about the white matter microstructure: for instance, axonal loss leads to a reduction in diffusion restrictions and so reduced directionality, leading to a reduced FA and an increased MD.

An alternative model that can be fit to the diffusion data is the ball and stick model, which represents one or more axonal fibre bundles in addition to an isotropic free water component. The ball represents the proportion of the signal that is isotropic free diffusion, and the sticks represent the proportion of the anisotropic signal and the direction associated with each fibre. This allows the model to explicitly account for partial volume effects in the signal. The ball and stick model can therefore account for crossing fibres, and can be used to estimate the number of fibres present in each voxel.

2.3.2 Tract-based Spatial Statistics (TBSS)

TBSS aims to over-come artefacts due to misregistration by comparing diffusion metrics along a mean FA skeleton, representing the centre of the white matter tracts. The FA map of each subject is first registered to the mean FA image in standard space.

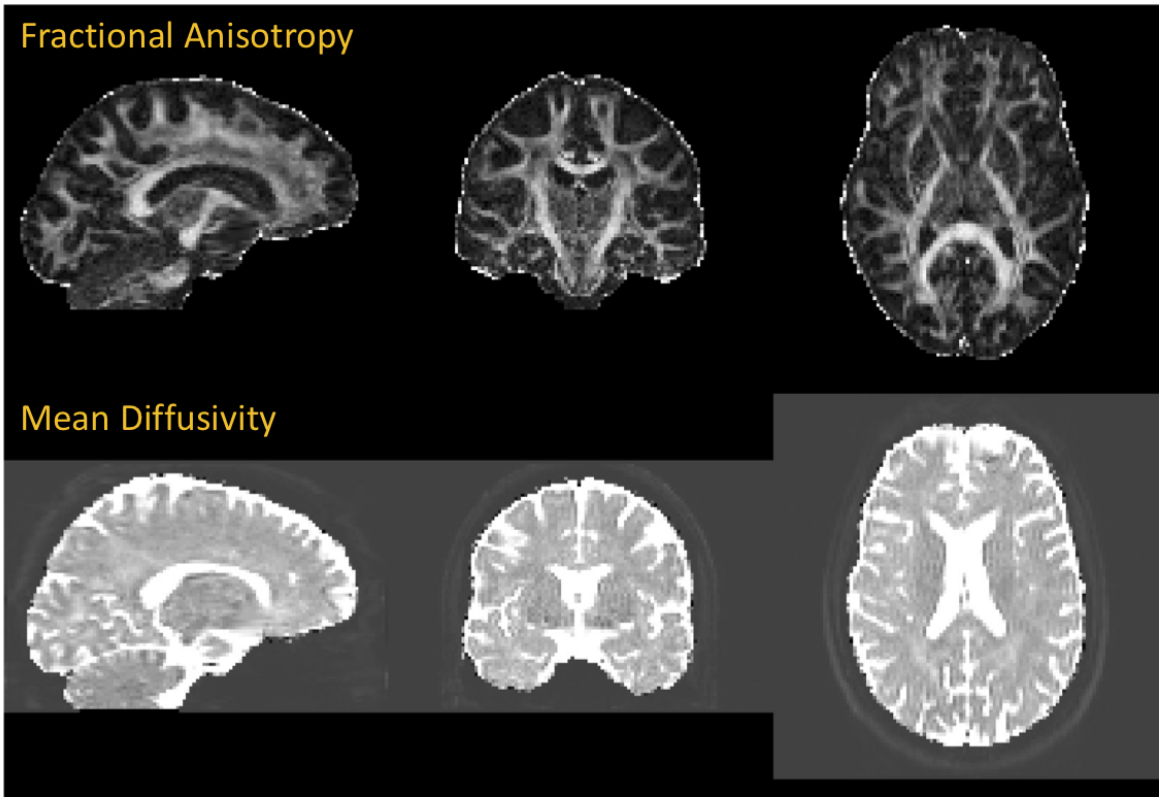


Figure 2.9: Example mean diffusivity and fractional anisotropy images. It can be seen that the fractional anisotropy is highest in the white matter tracts, whereas the mean diffusivity is highest in the ventricles, where water can relatively freely move within the CSF.

Using this transform, diffusion metrics such as FA and MD can be projected onto this skeleton. The projection aims to map values taken from the centre of each subjects tract onto the skeleton. Once each subject data has been mapped to the skeleton, voxelwise statistical tests can be performed across subjects.

2.3.3 Probabilistic Tractography

Tractography can be used to estimate structural connectivity in the brain, to show how one grey matter region is connected to others. Prior to tractography, the diffusion data is analysed (e.g. using DTI or the ball and stick model) to extract a set of fibre orientations per voxel. Tractography algorithms then trace a likely path between voxels following

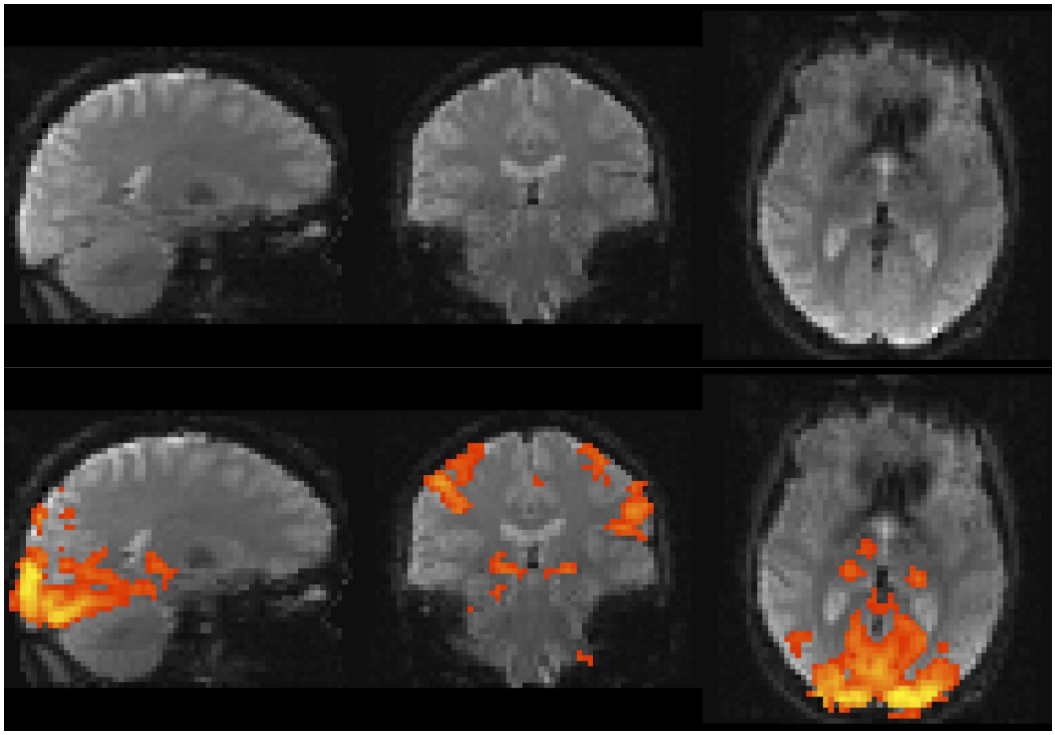


Figure 2.10: Example functional EPI image (top row) and thresholded activation from task fMRI example (bottom row), showing activation predominantly in the visual cortex.

these directions. In FSL PROBTRACKX uses the output from the BEDPOSTX [18] to perform probabilistic tractography. As BEDPOSTX estimates both multiple fibre populations per voxel, and the uncertainty associated with the fibre directions, PROBTRACKX can track though crossing fibre regions, and estimate the uncertainty of the connectivity pathways or streamlines. In the UK Biobank, AUTOPTX, a wrapper for PROBTRACKX, is used to perform anatomically constrained tractography with known start and stop masks to reconstruct 27 major tracts.

2.4 Functional MRI (fMRI)

Functional MRI (fMRI) focuses on dynamic changes in the brain in order to examine neuronal activity. Oxygenated and deoxygenated forms of haemoglobin interact with

magnetic fields differently. This creates inhomogeneities in the local magnetic field that depend on the concentration of deoxygenated haemoglobin, resulting in small measurable changes in the MRI signal. This is called the *blood oxygenated level dependent (BOLD) effect* and acts as an approximation to the neuronal activity. Because $T2^*$ reflects the magnetic field inhomogeneities as well as the pure $T2$ decay, it is typical to use a $T2^*$ sequence when imaging the bold effect.

Since fMRI is an indirect measure of neuronal activity, there is a delay between the neuronal activity (which is almost instantaneous with the stimulus) and the observed reduction in the MR signal. This is modelled using the haemodynamic response function. The response typically takes around 6 to peak after the stimulus, and 20 seconds to return to the baseline.

Images are acquired using Echo-Planar Imaging (EPI) sequences that allow images to be obtained quickly enough to detect the haemodynamic response. However, it also creates many artefacts in the images, such as those noted in the section on diffusion MRI, but with the exception of eddy current distortions, which are only notable in diffusion due to the extra gradients used in the diffusion acquisitions. Functional acquisitions suffer from additional loss in signal due to the B_0 inhomogeneities compared to diffusion MRI, since these inhomogeneities are the key part of the BOLD response, but they are also present near the sinuses where the air creates large inhomogeneities. Therefore, functional images are low resolution (2-3mm voxels) with low tissue contrast, together with noticeable geometric distortions and signal loss. Figure 2.10 shows an example functional EPI image, where the large voxels can clearly be seen compared to the T1 structural scan seen in Fig. 2.1.

Functional MRI is not quantitative, with the intensity values not being meaningful on their own. Instead, we analyse signal changes across time, either relative to a stimuli or other measured signals such as time series from other brain regions. Due

to the limited resolution and tissue contrast, it is often difficult to identify structures directly. It is also hard to register the image directly and so, whenever a functional study is completed, a structural image must also be acquired for an intermediate step in the registration process. This allows better alignment of the anatomical regions, and so, better statistical power when group analysis is completed. A B_0 field map should also be acquired allowing for correction of the geometric distortions, which again helps the localisation of the signal.

2.4.1 Resting State fMRI

In resting state fMRI the subject is not asked to do anything specific in the scanner. The resting state signals are then analysed to explore the functional relationships between different regions of the brain at 'rest', which are usually quantified by looking at the correlations between the time series from different locations in the brain.

Seed-based correlation can be used to calculate the correlation between a seed region of the brain and all other regions in the brain. Alternatively, ICA analysis can be used to analyse the resting state data by decomposing the data into components. Each component is described by a spatial component and a single common time course, so that for each component the spatial map represents the amount of that time course that is present in the data at different locations. Therefore, regions in a given spatial map with high voxel values share a large amount of the common signal. Finally, resting state fMRI can be analysed using network matrices, where the elements of the matrix represent the connectivity between a pair of functional regions and so allow the connections between different brain regions to be captured.

2.4.2 Task fMRI

In task fMRI, the subject is asked to complete a task during the scan, chosen to explore a particular response, for instance sensory (visual, auditory), cognitive (decision making), physiological (pain), or motor (button box, finger tapping). The task must be designed so as to minimise the subject's motion. In the UK Biobank, the task used is the Hariri faces/shapes emotion task [19].

When designing the study, it must have at least two conditions, such that a difference between the two states can be measured, as fMRI is not quantitative. Normally, comparison is made between a baseline condition (for instance, rest) and the stimulated state when then task is completed. There will still be activation at the baseline condition, but this allows comparison of the difference between the two conditions. There also needs to be sufficient repeats of the two conditions, such that sufficient statistical power is achieved. The images are noisy and so, averaging over multiple acquisitions leads to a less noisy final result.

The analysis pipeline for task MRI has several stages. First, the data needs to be processed to remove artefacts, such as those relating to the physics or hardware, and to the physiology, for instance head motion and breathing. The effects are largely removed through a mixture of pre-processing and confound regression during data analysis. After pre-processing, the next stage is that of developing an appropriate model of how we would expect the MR signal to change, given the study design. Here the HRF function is then convolved with the stimulus timings, which produces the predicted response to the known stimulus pattern. A General Linear Model (GLM) is then used to fit the predicted response to the measured MRI signal. The GLM determines which voxels have a significant relationship between the measured signal and the predicted response, from which we infer that these brain areas are involved

in the given task. An example thresholded activation map can be seen in Fig. 2.10 where there is significant activation, predominantly in the visual cortex.

3

Convolutional Neural Networks

Deep learning, via convolutional neural networks, has seen a huge growth in popularity across a wide range of applications, including medical imaging, due to their ability to automatically learn complex features at different levels of abstraction. Here, the necessary background to understand the motivation for the use of neural networks is discussed, followed by consideration of common applications of deep learning in medical imaging, and, finally, a discussion of some of the challenges which still exist for the full impact of these methods to be realised outside of the research domain. A systematic review of deep learning is too vast in scope for this thesis, but many good reviews can be found, such as [20].

The 'challenges' section of this chapter is adapted from the following review paper:

- **Nicola K Dinsdale**, Emma Bluemke, Vaanathi Sundaresan, Mark Jenkinson, Stephen Smith, Ana IL Namburete: *Challenges for machine learning in clinical translation of big data imaging studies*, Under Review - Available on ArXiv: arXiv:2107.05630v1

The code used in this chapter is available at: `github.com/nkdinsdale/challenges_review`.

3.1 Supervised Learning

The goal of *supervised learning* is to learn a mapping from an *input space*, $\mathcal{X}$, to an *output space*, $\mathcal{Y}$, given training data pairs of input and target data. For the input space $\mathcal{X}$, there exists a random variable $\boldsymbol{X}$ with *probability distribution* P_X, and similarly for the target space $\mathcal{Y}$ there exists a random variable $\boldsymbol{Y}$ with a probability distribution P_Y. For instance, considering Fig. 3.1, $\boldsymbol{X}$ could be the T1 MRI for a subject and $\boldsymbol{y}$ the corresponding diagnostic label. The availability of the labels, $\boldsymbol{y}$, is the key characteristic of supervised learning.

To learn the mapping, we have access to training data $\mathcal{D} = \{\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2 ... \boldsymbol{x}_N\}, \boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2 ... \boldsymbol{y}_N\}\}$, where the data pair samples are drawn from the joint distribution $P_{XY}(\boldsymbol{X}, \boldsymbol{Y})$. The joint probability distribution is defined as $P_{XY}(\boldsymbol{X}, \boldsymbol{Y})$ which by Bayes' rule can be written as: $P_{X|Y}(\boldsymbol{Y}|\boldsymbol{X}) = P_X(\boldsymbol{X})P_{Y|X}(\boldsymbol{Y}|\boldsymbol{X})$, which is the product of the *marginal* distribution $P_X(\boldsymbol{X})$ and the *conditional* distribution $P_{X|Y}$. The set of possible mappings from the input space to the target space $h : \mathcal{X} \Rightarrow \mathcal{Y}$ is called the *hypothesis space* $\mathcal{H}$ and the difference between the true label $\boldsymbol{y}$ and the prediction $h(\boldsymbol{x})$ is quantified using a *loss function*, $\mathcal{L}$, designed for the task of interest. The goal is to minimise $\mathcal{L}$.

The goal of supervised learning, therefore, is to learn a function to map from the input space to the target space, $h \in \mathcal{H}$, which takes the input data $\boldsymbol{x}$ and

predicts the most likely output label:

$$\hat{\boldsymbol{y}} = h(\boldsymbol{x}) = argmax_{\boldsymbol{y}^*} P(\boldsymbol{Y} = \boldsymbol{y}^* | \boldsymbol{X} = \boldsymbol{x}) \approx \boldsymbol{y} \tag{3.1}$$

and therefore the aim is to find a hypothesis (a mapping function) which minimises the *empirical risk*:

$$R_{emp}(h) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\boldsymbol{y}, h(\boldsymbol{x})) \tag{3.2}$$

where N is the number of samples available. This is an approximation to minimising the risk:

$$R(h) = \int_{\boldsymbol{x},\boldsymbol{y} \in (\boldsymbol{X},\boldsymbol{Y})} \mathcal{L}(\boldsymbol{y}, h(\boldsymbol{x})) dP(\boldsymbol{x}, \boldsymbol{y}). \tag{3.3}$$

when there is only access to a limited training dataset $\mathcal{D}$, drawn from the joint distribution, $P_{XY}(\boldsymbol{X}, \boldsymbol{Y})$. This is known as *empirical risk minimisation*.

The aim is that the model is *generalisable*, where minimising R_{emp} minimises R and thus the risk for the whole dataset, not just the points seen during training. Therefore, supervised learning aims to create a model, trained using the available data, which minimises the loss for the observed data, such that the model generalises to the unseen data – or, practically, performs well on the testing dataset.

Deep learning considers the case where the hypothesis space – or model used to learn the mapping between the input space $\mathcal{X}$ and the target space $\mathcal{Y}$ – takes the form of a *deep feed-forward neural network* (DNN), which is a series of fully-connected or convolutional layers stacked at different depths. Information is propagated through these layers, and they have learnable weights, which extract information from the data. These weights need to be optimised to complete the task of interest. We can

consider the neural network to be a function $f : \mathcal{X} \times \mathcal{W} \rightarrow \mathcal{Y}$ where $\mathcal{W}$ are the weights of the neural network to be learned through the training procedure. Each layer can be described as a function $f_l \ \forall l \in \{1, L\}$ where L is the number of layers in the network, and for each layer the learnable parameters are composed of the weights and a bias term $\{\boldsymbol{w}_l^{(k)}, b_l^{(k)}\}$. The mapping between layers is therefore given by:

$$f_l(\boldsymbol{z}_l; \mathcal{W}_l) = \sigma_l(\boldsymbol{z}_{(l-1)} * \boldsymbol{w}_l^{(k)} + b_l^{(k)}), \quad \mathcal{W}_l = \{\boldsymbol{w}_l, b_l\} \tag{3.4}$$

where $\boldsymbol{z}_l$ is the activation map for the lth layer or the output of applying f_l to the previous activation map; if this were the first layer $\boldsymbol{z}_0$ would be the input data $\boldsymbol{x}$. σ_l is the activation function, which is a non-linear function applied to the data, and can take many forms. Therefore, given that each layer can be considered to be a function of the previous layer activations and the weights, the function from $\mathcal{X} \rightarrow \mathcal{Y}$ can be considered to be a composite function of the functions for each layer:

$$f(\boldsymbol{x}; \mathcal{W}) = f_L(f_{L-1}(..f_1(\boldsymbol{x}; \mathcal{W}_1)..; \mathcal{W}_{L-1}); \mathcal{W}_L) \tag{3.5}$$

where the evaluation of this function is called *forward propagation*. This forward propagation stage updates the model predictions, allowing us to evaluate the current error in the model, where the error is found through evaluation of the loss function.

3.1.1 Back Propagation

To allow the network to improve the prediction, the weights must be updated, and they are updated as to reduce the loss using a gradient decent algorithm:

$$\mathcal{W} = \mathcal{W} - \alpha \nabla_{\mathcal{W}} R_{emp} \tag{3.6}$$

where R_{emp} is the empirical risk for the available training data, and α is termed the *learning rate* or step size, which controls the size of the change made in the weight search space. The aim of gradient descent is to iteratively find small perturbations which improve the network performance, reducing the value of the loss function evaluated on the training data $\mathcal{D}$ and therefore reducing the risk. Due to the network's multi-layer structure, and recalling Equation (3.5), the gradient can be calculated in a layer-wise fashion using the chain rule, such that the gradient of the weight $\mathcal{W}_i$ for the layer i is computed, using a process normally called *back propagation* [21], as:

$$\nabla_{\mathcal{W}_i} R_{emp} = \frac{\partial R_{emp}^T}{\partial f_l} \frac{\partial f_l^T}{\partial f_{l-1}} \dots \frac{\partial f_i}{\partial \mathcal{W}_i}. \tag{3.7}$$

The Jacobian of each layer $\frac{\partial f_l^T}{\partial f_{l-1}}$ can be computed by the chain rule also. If f_i is the output of the layer i and $\boldsymbol{z}_i = \boldsymbol{w}_i + b_i$ is the output activation map of the layer, then:

$$\frac{\partial f_i^T}{\partial f_{i-1}} = \frac{\partial \sigma(\boldsymbol{z}_i)^T}{\partial \boldsymbol{z}_i} \frac{\partial (\boldsymbol{w}_i f_{i-1} + b_i)^T}{\partial f_{i-1}} \tag{3.8}$$

where the bias values, b_i, are constants.

Evaluating this for all data points simultaneously would be computationally infeasible for large scale problems, so normally back propagation is evaluated for *mini-batches* with batch (sample) size m, where the size of m is generally constrained by the computational power available. The approximation is being made that the risk calculated for a mini-batch with batch size m will be close to that which would be calculated for the whole dataset. The learning rate is then set to be a small value ($\alpha < 1$), and is stochastic gradient descent, with the gradient being replaced with an estimate over a randomly selected sample of subjects.

Through this learning process, the weights of the network are optimised to complete the task of interest. The network f can therefore represent a wide range of functions,

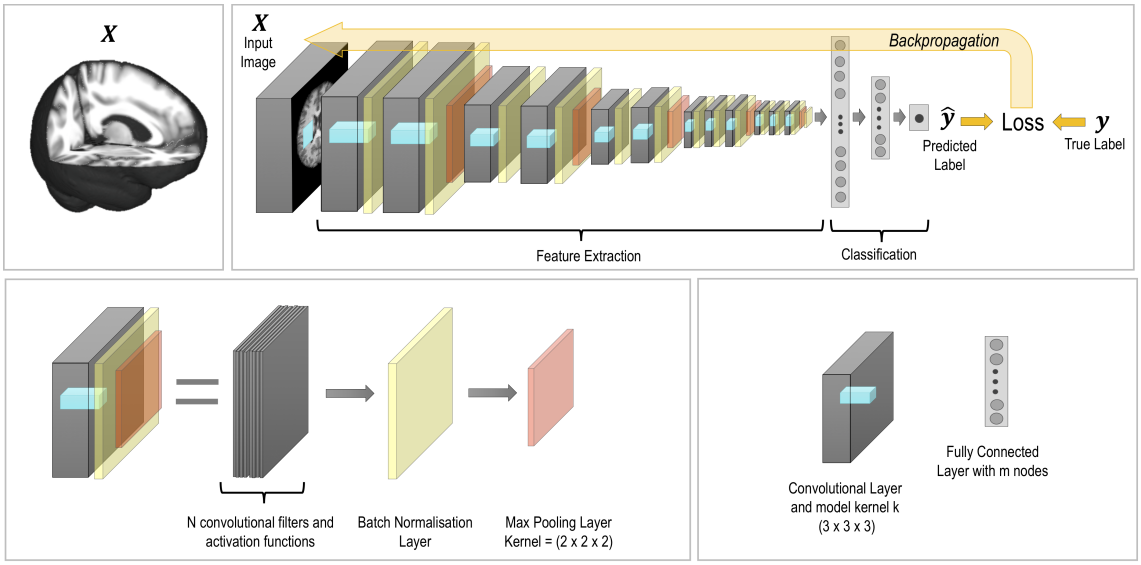


Figure 3.1: An example network architecture for a convolutional neural network (CNN) for a classification or regression task.

depending on the weights learned and the data used to train the network. The same network architecture can therefore be used to map between many different $\mathcal{X}$ and $\mathcal{Y}$ pairs, and is not specific to the data space. The choice of the architecture is therefore largely based on empirical evidence of design choices which have previously worked well, with the hyperparameter space of architecture design being too large to search exhaustively.

3.2 Convolutional Neural Networks

CNNs are normally built of several standard building blocks: convolutional layers, activation functions (mostly non-linearities), pooling layers, and, sometimes, fully connected layers. These components can all be seen in Fig. 3.1, where a regression or classification network, broadly based on VGG [22] is shown. The sequential nature of the layers can be clearly seen, where the information is propagated in the forward pass from the input $\boldsymbol{X}$ to the output prediction $\hat{\boldsymbol{y}}$, updating the network weights and, so, the model prediction. Then the loss, calculated between the true label and the predicted label, is

backpropagated, updating the gradients of the weights, ready for the next update stage.

3.2.1 Convolutional Layers

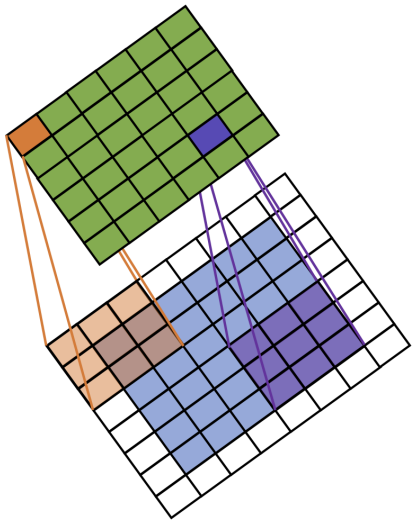


Figure 3.2: Demonstration of how the convolutional kernel moves over the image to create the response. The image is often zero padded to maintain the image dimensions between layers, as shown by the white region in the image. The convolved result is located at the middle of the kernel.

The first component of the CNN is the convolutional layer which, for an input tensor $\boldsymbol{x} \in \mathcal{R}^{N_x,N_y,N_c}$ where N_c is the number of channels of a 2D image, a convolutional layer f with a weight $\boldsymbol{w} \in \mathcal{R}^{k_x \times k_y \times N_c \times N_{c'}}$ (k_x is the x kernel dimension, k_y the y) and bias $b \in \mathcal{R}^{N_{c'}}$ is defined as

$$[f(\boldsymbol{x}; \boldsymbol{w}, b)]_{l,m,n} = \sum_{i=l-k_x/2}^{l+k_x/2} \sum_{j=m-k_y/2}^{m+k_y/2} \sum_{k=0}^{N_{c'}} \boldsymbol{w}_{ijkn} \boldsymbol{x}_{ijk} + b_n. \tag{3.9}$$

This shows that the weights are applied to the input image via convolution such that $\mathcal{W}^T \boldsymbol{x} = \boldsymbol{w} \circledast \boldsymbol{x}$. Figure 3.2 shows how the kernel moves over the image to create the response. The weights of the network are commonly referred to as the *convolutional kernel*, and the trainable parameters of the layer are the weights of the kernel and the

bias term. The size of the kernel (often 3×3 [22]) and the number of filters to be applied are hyperparameters that need to be set by the designer of the network. Clearly, it is trivial to extend this into $3D$ when working with $3D$ volumes in medical imaging, at the cost of an increase in parameters. In a convolutional layer, the filter weights are shared across the image spatially, reducing the number of parameters significantly, meaning that the layer has spatial invariance, and so that the same features lead to the same activation wherever they are located in the image. Convolutional layers, therefore, learn to extract local features and build complex relationships which are then aggregated before being passed down the network.

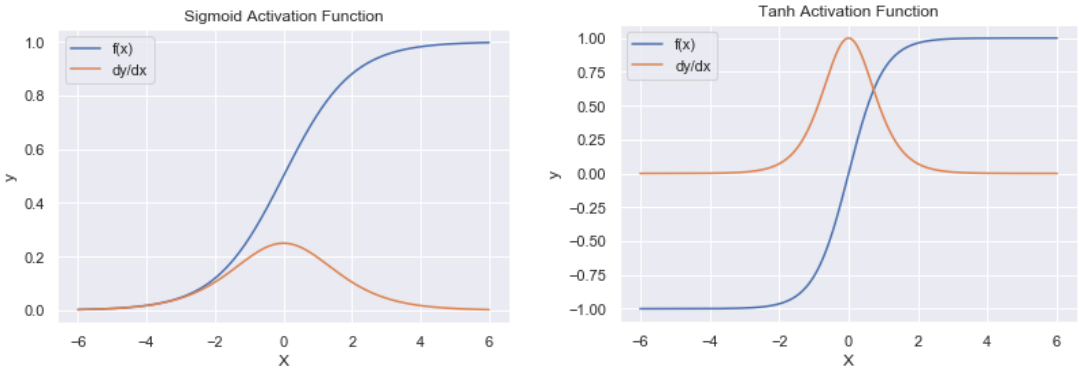
3.2.2 Non-linearities or Activation Functions

The next standard component is the non-linearity layer, or activation function. These functions rescale the data non-linearly, allowing the CNN to learn complex non-linear mappings between the input and the output, preventing the CNN from simply being a linear regression model. Non-linear activation functions are needed because if only linear activation functions were used then the neural network will just output a linear function of input.

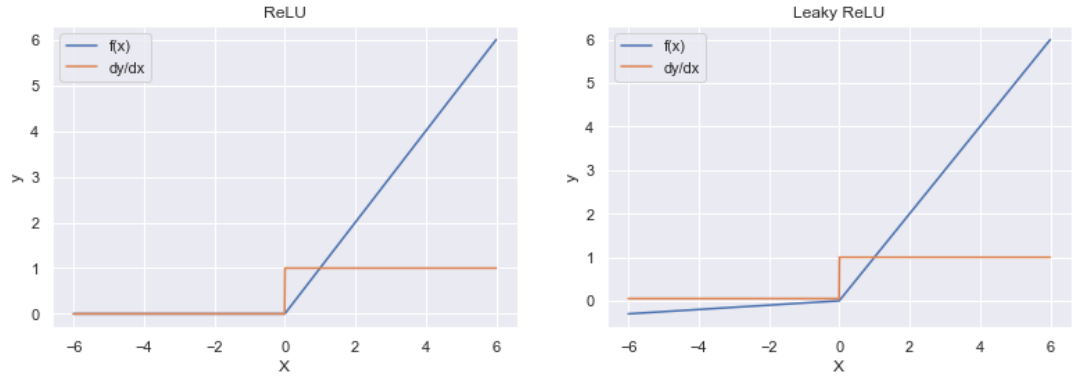
Commonly used activation functions include:

- **Sigmoid**: $f(\boldsymbol{X}) = \frac{1}{1+e^{-\boldsymbol{X}}}$

  The sigmoid loss function (Fig. 3.3a) is differentiable and has output values bounded between 0-1 meaning that it normalises the output values of each neuron; however, it suffers from vanishing gradients, especially in deep networks. This means that for very high or very low values of $\boldsymbol{X}$ there is almost no change to the prediction, causing very small gradients, leading to very slow convergence. The output values are also not zero-centred, and it is computationally expensive.



(a) Sigmoid **(b)** Tanh



(c) ReLU **(d)** Leaky ReLU

Figure 3.3: Example activation functions.

- **Tanh**: $f(\boldsymbol{X}) = tanh(\boldsymbol{X}) = \frac{e^{\boldsymbol{X}} - e^{-\boldsymbol{X}}}{e^{\boldsymbol{X}} + e^{-\boldsymbol{X}}}$

  The tanh activation function (Fig. 3.3b) is zero centred, making it easier to model inputs that have strongly negative, neutral, and strongly positive values, and works better in practice than the sigmoid loss function. However, it still suffers from the vanishing gradient problem and so still has slow convergence.

- **ReLU**: $f(\boldsymbol{X}) = max(0, \boldsymbol{X})$

  ReLU [23] (Fig. 3.3c) was introduced as a more computationally efficient activation function, which enables the network to converge more quickly. It is a piece-wise function, meaning it is non-linear and has a derivative function despite having a discontinuity, allowing back propagation. It does, however, have

problems when the inputs approach zero, the gradient becomes zero, and the ReLU 'dies', meaning it outputs the same value, zero, for any input. Once the gradient is zero, it is unable to recover from this state, which is usually entered because a large negative bias term has been learned. Despite this, ReLU is a very commonly used activation function.

- **Leaky ReLU**: $f(\boldsymbol{X}) = max(\epsilon\boldsymbol{X}, \boldsymbol{X})$

  Leaky ReLU [24] (Fig. 3.3d) deals with the dying problem of the ReLU function by replacing the zero with a very small value, for instance $\epsilon = 0.01$. This creates a small negative slope in the negative area, meaning that there is always some update. This does, however, mean that the results are not consistent for negative input values. A further variation is Parametric ReLU [25], where the small value is learned during network training.

- **Softmax**: $f(\boldsymbol{X})_i = \frac{e^{\boldsymbol{x}_i}}{\sum_{j=1}^{K} e^{\boldsymbol{x}_j}}$

  Softmax is normally used at the end of the network to output logits corresponding to each class which sum to one, allowing the prediction of the final class to be completed. It is a differentiable approximation to the argmax function.

3.2.3 Pooling Layers

Pooling layers are then added to locally aggregate the features within a certain pooling window. The most commonly used pooling layers are max pooling and average pooling. For a pooling window size of $p_x \times p_y$, a max pooling layer can be defined as:

$$[maxpool(\boldsymbol{X})]_{l,m,c} = \max_{\substack{i \in [l-p_x/2 : l+p_x/2] \\ j \in [m-p_y/2 : m+p_y/2]}} \boldsymbol{X}_{ijc} \tag{3.10}$$

where l and m are the size of the pooling window in the x and y dimensions, and c is the channel being considered.

Pooling layers provide basic invariance to translation and local perturbation. They are also used to reduce the spatial dimension of the input tensor. Commonly, the stride of the operation is matched by the pooling window, meaning that the output of the pooling layer is downscaled by a factor equal to the pooling window size. This causes the network to learn a representation which aggregates the local information and enables it to learn complex hierarchical representations.

3.2.4 Normalisation Layers

The final standard component is the normalisation layer. The most commonly used normalisation layer is the batch normalisation layer [26], which normalises the intermediate tensor using the mini-batch statistics, combined with two learnable parameters, γ and β. First, the mean and variance of the batch is computed and then they are used to rescale the data, whitening the features:

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i \tag{3.11}$$

$$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m} (\boldsymbol{x}_i - \mu_B)^2 \tag{3.12}$$

$$BN_{\gamma,\beta}(\boldsymbol{x}_i) = \gamma\left(\frac{\boldsymbol{x}_i - \mu_\beta}{\sqrt{\sigma_B^2 + e}}\right) + \beta \tag{3.13}$$

where m is the batch size. Batch normalisation can help to accelerate the learning by normalising the statistics of each layer every time, although the benefit is empirical and not well understood. Other normalisation schemes include group normalisation,

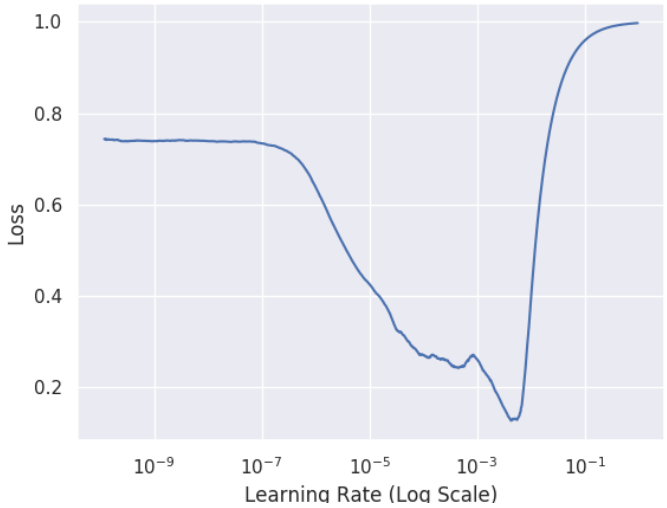


Figure 3.4: The learning rate was varied between 10×10^{-10} and 1 with the value being varied between batches. The learning rate was varied for each batch and the loss recorded, following the protocol set out in [29], for a segmentation task, with Dice as the loss function, hence the bounding at 1. When using learning rate finder, the recommended minimum learning rate is the value where the loss decreases the fastest, while the recommended minimum learning rate is 10 times the rate where the loss is the minimum, because the plot shows a smoothed version of the loss.

instance normalisation [27] and layer normalisation [28], which become important as the batch size becomes very small.

3.2.5 Optimisers and Learning Rate

To be able to gain from the expressiveness of CNNs, efficient training methodologies are needed. CNNs are highly non-convex functions, and therefore there are often many local optima, making their training potentially difficult. Most optimisers used in the training of deep learning networks are based on first-order stochastic optimisation techniques, aiming to find a good global minimum that will generalise well.

An important consideration is how to set the learning rate during training. If the learning rate is too slow, the training will be slow and will probably converge to a poor

local minimum; if the learning rate is too high, the network will diverge, and fail to find a good minimum. This is demonstrated in Fig. 3.4 where the loss across batches can be seen with varying learning rate, following the method set out in [29]. It can be seen that when the learning rate is very low, the loss does not change; as it increases, there is a range of learning rates for which the loss decreases and then, as it becomes too large, the loss increases rapidly. There is, therefore, a range of values for the learning rate for which the network will learn effectively, and many values where the network will not be able to be optimised effectively. While stochastic gradient descent (SGD) with a well-chosen annealing (reduction of the learning rate with training) schedule works well in practice to find a good global minimum, methods which use adaptive learning rates are gaining popularity, for instance Adam [30].

3.2.6 Loss Functions

When training networks and evaluating the model performance, a *loss*, or *cost function*, must be chosen which the network aims to minimise. This function determines the error in the network prediction by comparing the prediction ($\hat{\boldsymbol{y}}$); the true label ($\boldsymbol{y}$), and the choice of the loss function depends on the task being investigated. Although some works design bespoke, task-specific cost functions, the majority are based on standard functions, such as the following:

- **Categorical (or Binary) Cross-Entropy** is the most commonly used loss function for classification tasks. Given that each sample belongs to a single class with label y the output prediction of the network is a vector of softmax values, one per possible predicted class. The loss is then calculated as:

$$L_{crossentropy} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C} y_i(k) log(p_i(k)) \tag{3.14}$$

where k is the class in the set of C possible classes, $p(k)$ is the softmax value for the class, and N is the total number of subjects, such that we take the average loss over the batch of subjects considered. Weights can also be added to the loss function, especially if there is a high degree of imbalance between the classes.

- **Mean Squared Error (MSE)** is primarily used as the loss function for regression or reconstruction tasks. In the case of regression, such as for brain age prediction, the comparison is between a single true value and a single predicted value for each subject, whereas for reconstruction it is between each pair of voxels in the true and predicted images. It can be formulated as:

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{3.15}$$

  where y_i is the true label for subject i of N subjects, and $\hat{y}_i$ is the predicted label.

- **Dice Overlap** is a loss function primarily used for segmentation tasks, which evaluates the similarity of two sets, akin to accuracy but does not reward correctly identifying the background (or true negatives). It is given by:

$$L_{Dice} = 1 - \frac{1}{N}\sum_{i=1}^{N}\frac{2|\boldsymbol{y}_i \cap \hat{\boldsymbol{y}}_i|}{|\boldsymbol{y}_i| + |\hat{\boldsymbol{y}}_i|} \tag{3.16}$$

  where, again, $\boldsymbol{y}_i$ is the true label for subject i of N, $\hat{\boldsymbol{y}}_i$ is the predicted label, and in this case the label has the form of a dense segmentation label $\boldsymbol{y}_i \in \mathcal{R}^{W \times H \times D \times C}$ where C is the number of classes.

These metrics are normally chosen because of their well-understood and characterised behaviour [31], and, especially in the case of segmentation, their ability to produce visibly acceptable predictions. For clinical translatability, it may be important

to consider which characteristics of the predictions are the most important for the envisaged clinical application [32–34]. Metrics only tell part of the story and it is important that all vital information for clinical assessment is provided by the reported metrics. For example, when considering tasks such as tumour segmentation, the classification of the centre of the tumour is trivial and the utility for a clinician would be the identification of boundaries, where the classification is harder or ambiguous. Therefore, a metric such as the Dice overlap is not necessarily a good indicator of performance from a clinician's point of view, with the boundary having little impact on the metric if the region of interest is large, while being potentially critical for patient care. Developing networks and loss functions with the specific application in mind is vital and has been explored for some applications, such as focusing segmentations to the boundary [35–37] but needs to be considered for each application in turn.

3.3 Medical Imaging Tasks

Deep learning has been applied to many tasks in medical imaging, but the main three can be considered as classification (or regression), segmentation and generative tasks. Here, some architectures and approaches will be discussed for each of these tasks, but will not be an exhaustive discussion.

3.3.1 Classification

The development of deep neural networks has progressed rapidly as the availability of both processing power and large-scale, labelled datasets has increased, building on the previous work which lay the foundations of convolutional neural networks [21], back propagation [38] and the network architecture LeNet [39]. LeNet demonstrated that image features are distributed across the entire image, and convolutions with learnable

parameters are an effective way to extract similar features at multiple locations with relatively few parameters (compared to MLP).

The first major breakthrough for deep learning was for classification tasks, with AlexNet [40] outperforming the other methods on the ImageNet dataset, leading to its extensive use across a multitude of tasks. The successful deployment of convolutional layers in AlexNet [40] on the ImageNet dataset has led to their extensive use across a multitude of tasks.

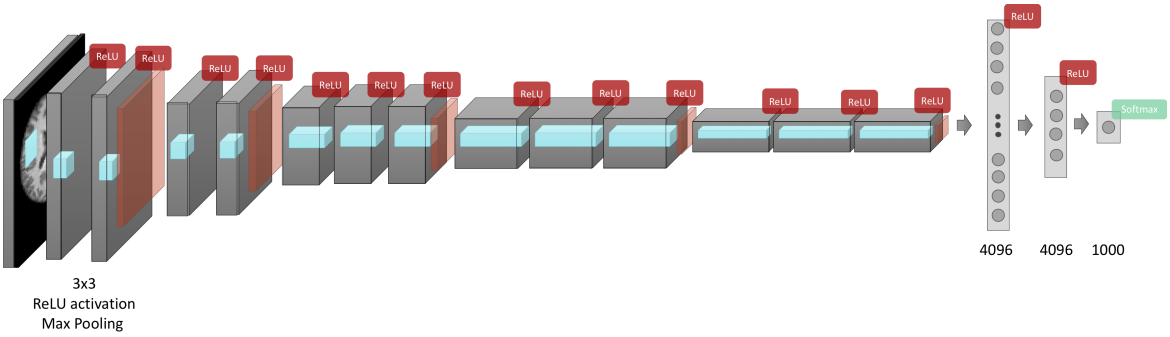


Figure 3.5: VGG-16 architecture

All classification or regression based networks used in this thesis are based on VGG-16. The VGG-16 [22] network – Fig. 3.5 – showed a large increase in model depth leading to a corresponding increase in model performance. The network combined 13 convolutional layers, 3 fully connected layers, and ReLU activation functions. It was also demonstrated in [22] that it is better in terms of performance to use two layers with kernels of 3×3 than one layer with a larger kernel size (5×5) with the same receptive field, largely due to the increase in the number of non-linear activation layers and, thus, the discriminative power of the network. This has led to many models being developed with 3×3 kernel sizes throughout.

For a more through summary of classification networks, see Appendix B.1.

3.3.2 Segmentation

Semantic segmentation considers the voxel wise labelling of an image, producing a dense prediction. For full resolution image segmentation, we require the feature size at the end of the network to be the same as the input image. The standard method to maintain the expressiveness of the network of increasing the number of feature maps (channels) as the network depth increases, is very computationally expensive if we are not able to downsample the features. Therefore, a popular approach for image segmentation models is to have an encoder-decoder structure, where the spatial resolution of the input is downsampled, creating lower resolution feature mappings which, ideally, are highly efficient at discriminating between classes, that can then be projected back into the image space by the decoder. These approaches, however, often lead to coarse segmentation maps, which often struggle with, for instance, tissue boundaries in medical images. This section will focus on the UNet architecture which is used across experiments, see Appendix B.2 for details of Fully Convolutional Network (FCN).

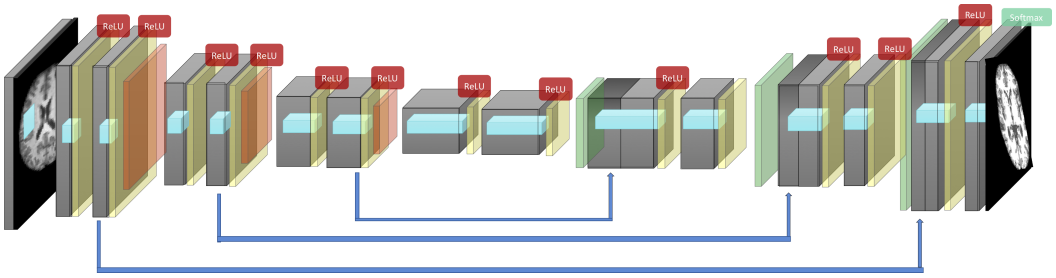


Figure 3.6: UNet architecture.

The use of the skip connections was developed in the UNet architecture [41], which has, along with its derivative networks, been the most successful network architecture designed specifically to work within the medical imaging domain – Fig. 3.6. First applied for cell segmentation, the UNet and its 3D equivalents [42, 43] have now been applied across segmentation tasks and imaging modalities for semantic segmentation.

The UNet has the encoder-decoder structure but also includes cross connections between representations at the same level of abstraction. The encoder captures context, and its symmetric expanding path decoder enables precise localisation. The cross connections (or skip layers) transfer the spatial information into the decoder and promote the flattening of the optimisation space, aiding the optimisation of the loss function and thus making it easier to train the network [44].

There have subsequently been many developments to the UNet architecture. In [45] the basic convolutional blocks are changed in favour of residual blocks. These residual blocks introduce short skip connections in addition to the existing long skip connections, leading to faster convergence of the training, and allowing deep UNets to be trained. These were then replaced with dense blocks in [46], again creating skip connections and multiscale features at each depth. The dense blocks carry low level features from previous layers directly alongside the higher-level features from more recent layers, allowing for efficient feature reuse.

The UNet has been used extensively across segmentation tasks; for a review see, for instance, [47]. Segmentation tasks are numerous across medical imaging and neuroimaging more specifically, for instance brain extraction [48], tissue segmentation [49] and lesion segmentation [50], and numerous deep learning-based approaches have been proposed, many based on the UNet architecture.

3.3.3 Image Generation or Reconstruction

Generative methods aim to find underlying patterns in the data to either compress data to reduced dimensions or create new data points from the same distribution, for either image generation or image reconstruction.

Autoencoders can be used for a range of tasks, such as classification via clustering in

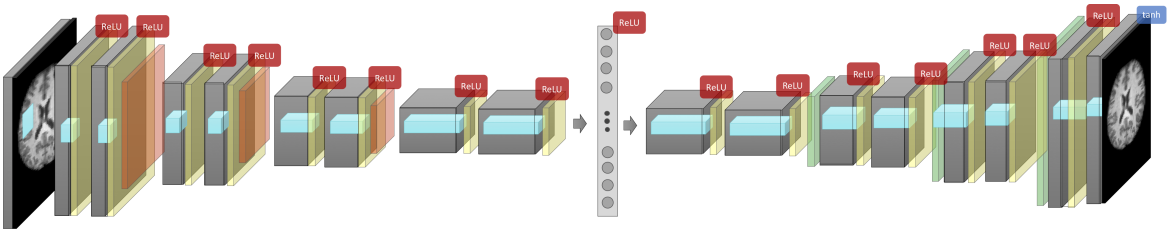


Figure 3.7: Example architecture for an autoencoder. A similar architecture could be used for a variational autoencoder, where the fully connected layer uses the reparameterisation trick to draw samples from the learned distribution [51].

the latent space [52], and anomaly detection, through observation of large reconstruction errors for subjects with unexpected pathologies [53], and unsupervised segmentation [54]. An *autoencoder* is an unsupervised learning technique for representation learning, where the model is trained to reconstruct the input by minimising some loss $L_{recon}(\boldsymbol{x}, \hat{\boldsymbol{x}})$. As can be seen from Fig. 3.7, the basic form of an autoencoder is an encoder branch which compresses the image information down to a fully connected latent space, and a decoder which projects this compressed information back up into the image space. The latent space is an information bottleneck, which forces the network to learn a compressed representation of the original input. *Variational autoencoders* [51] are then a development of autoencoders, which encode a probabilistic description of the data in the latent space, and thus hopefully a more meaningful latent space.

Generative adversarial networks (GANs) [55] aim to automatically discover and learn patterns in the input data in such a way that the model can be used to generate or output new examples that plausibly could have been drawn from the original dataset distribution. As can be seen in Fig. 3.8 they are formed of two models: the *generator* model is trained to generate new examples from a random vector initialisation, and a *discriminator* model is trained to classify the examples either as real data samples drawn from the data distribution, or fake ones produced by the generator. GANs have been used in medical imaging for a range of tasks, such as data augmentation [56],

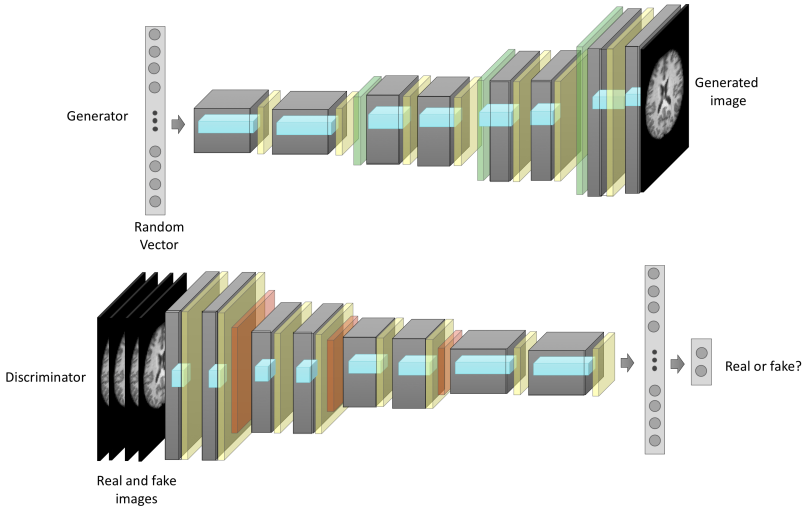


Figure 3.8: Example architecture for a Generative Adversarial Network (GAN) architecture.

generating MR images from CT [57] and segmentation [58].

For a longer summary of generative methods see Appendix B.3.

3.4 Challenges for the Application of Deep Learning to Medical Imaging

Due to the growth in size of the datasets, deep learning models are now an option for neuroimaging analysis. Applications of deep learning techniques to neuroimaging data have been explored in a research setting, with increasing numbers of novel methods proposed year on year. Powered by their ability to learn complex, non-linear relationships and patterns from data, deep learning methods have been applied to a wide range of applications. Across neuroimaging, the majority of datasets have been limited to small scale collections, focusing on a clinical population of interest. Recently, however, large scale 'big data' collections have begun to be collated, such as the imaging data from the UK Biobank study [1] and multisite studies such as ADNI [59].

Unfortunately, many challenges remain to enable the full impact of deep learning methods to be felt, especially in clinical settings. Recent developments in deep learning have begun to tackle the problems faced, but further developments are needed to overcome the clear challenges that remain.

3.4.1 Data Availability

For the application of deep learning techniques in medical imaging, the lack of available data for training the models is a hugely limiting factor. Despite the growth in the size of available datasets, the largest are still only of the order of tens of thousands of imaging subjects, with a thousand images being commonly regarded as a large dataset [59]. For many specific tasks, datasets exist only in the order of hundreds of subjects, due to many factors such as the monetary and time cost of acquiring data, the difficulties in sharing and/or pooling data across sites, and the fact that, for some conditions, there are insufficient patients to create a dataset of any great size [60]. For instance, the frequently explored Brain Tumour Segmentation (BraTS) dataset [61] only has data from 369 subjects[1] available for training, in stark contrast to the popular datasets from computer vision such as ImageNet [62] (1,281,167 training examples), CIFAR-10 [63] (50,000 training examples) and MNIST [39] (60,000 training examples) where many of the methods applied to medical imaging are developed. Simply by looking at dataset sizes, it is probable that the model training will be underpowered. Highly parameterised, deep neural networks are very dependent on the amount of available training data [64, 65], with performance generally improving as the number of data points is increased, and they are far more affected by the amount of available training data than classical machine learning techniques, due to the need to learn

[1] Training Data for 2020 challenge

the useful features as well as the decision boundary [65].

3.4.1.1 Model Pre-training and Transfer Learning

There has, therefore, been an increasing interest in developing techniques to facilitate the most effective use of the data available. A commonly used technique from computer vision is the use of large natural image datasets [66–68], with ImageNet [62] being the most popular [69] for many tasks, to *pre-train* the network. Pre-training involves training the weights on a related task with more available data, such that the optimisation starts from an informed place, rather than a random initialisation. It can seen why this might be useful by considering the information learned by the network at the different stages [70]. The early layers learn features such as edges and simple textures, largely resembling Gabor filters – see Fig. 3.9 – and are therefore very general and applicable across different images, regardless of the target tasks [71]; the deeper convolutional layers then tend to learn features such as more complex textures and object parts, and the final layers learn features which are far more task- and dataset-specific (e.g. fully connected layers learn discriminative features for a classification task). Therefore, a network pre-trained on the large dataset can be chosen, then the deep layers can be fine-tuned to the specific task. This requires less data, as not only is the optimisation process starting from an informed point in the *parameter space*, but also the earlier layers can often be frozen (kept at their value and not updated during training) greatly reducing the number of weights in the model that need to be optimised. This process is referred to as *transfer learning* and is a step frequently used to allow networks to be trained with lower amounts of training data. Transfer learning can be performed across data domain (dataset), task, or both, depending on the datasets available for pre-training.

The standard practice is to use the very large datasets of natural images for the

pre-training. However, natural images have very different characteristics from medical images and thus, the features learned are not necessarily the most appropriate for the tasks being considered [66]. For instance, natural images are often stored as RGB images, whereas MR images are encoded as greyscale. Similarly, in medical images the location of structures could be informative, which is rarely true in natural images, where most images are cropped so that the region of interest is central. Creating pre-trained networks for medical images has therefore been a focus, with Model Genesis [72] creating a flexible architecture trained to complete multiple tasks, to create features which aim to generalise across medical imaging tasks. Similarly, some works pre-train on large datasets such as the UK Biobank for tasks such as age prediction or sex prediction, where obtaining labels is relatively trivial [73] or on datasets for the same task with a dataset where more labels are available [74–76]. Here again, the aim is to learn features from the prediction task which are useful for the task of interest: that is, features that generalise across tasks. Other approaches are based on self-supervised learning, where some kind of supervisory signal is automatically generated to pretrain the model weights on a related task, for instance [77, 78].

3.4.1.2 2D or Patch Based Approaches

Other approaches to solving the shortage of available data focus on breaking the input data down into patches e.g. [80, 81] or slices (where 3D data is available) with many studies treating MRI data as 2D inputs e.g. [82, 83]. This approach can vastly increase the amount of available data and can be especially effective for segmentation tasks where we have voxel-wise labels. However, by fragmenting the image, these approaches lead to the loss of global information. Although this can be compensated for to some degree using optical flow [84] or conditional random fields [85], patch-wise or slice-wise approaches cannot necessarily be applied to classification tasks, where

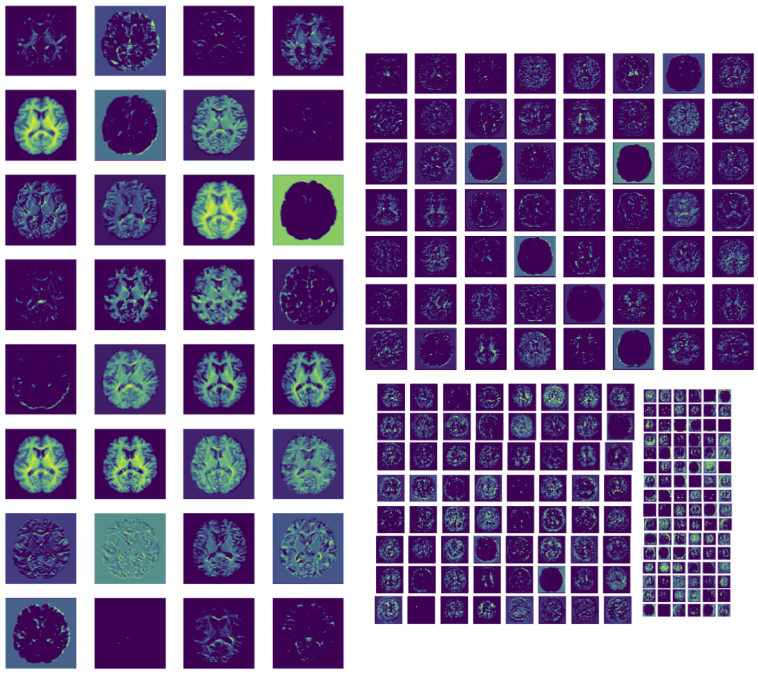


Figure 3.9: Example extracted features, $\boldsymbol{z}_l$, for brain age prediction (see [79]) at different network depths and spatial resolutions. It can be seen that several different features are extracted and some features are repeated, showing redundancy in filters.

a single label is provided for the whole image (which may not be valid for all slices) on a given patch or slice of the image [86]. When these approaches can be applied, say for a segmentation task, care must be taken when combining the patches at the output to reconstruct a single output image, so that artefacts do not appear at the boundaries. Fully 3D networks have been found in most cases to provide better results when they can be implemented [85].

3.4.1.3 Overfitting

Due to the highly parametrised nature of deep neural networks, they are prone to overfitting to the training data – that is, memorising the training data such that they perform badly on unseen data. Therefore, several different tools are employed to reduce the degree of overfitting. The first is the use of a validation dataset to

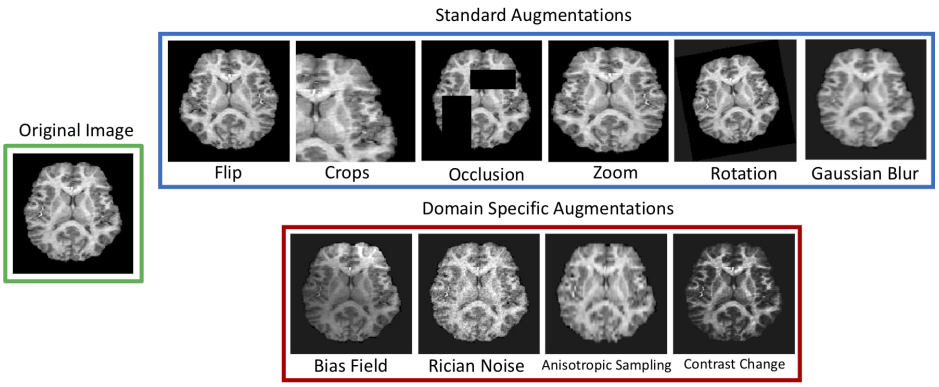


Figure 3.10: Example augmentations that can be applied to the training data.

check the performance of the model during training. This is used within the training loop to select the best performing model before testing on the held-out test set. The validation set can also be used for early stopping, that is, training can be stopped when the performance on the validation set decreases, preventing overfitting. Weight regularisation is also used to constrain the magnitude of the weights, with the intuition being that the highly non-linear function represented by the network will not be robust to small perturbations if the weights have extremely large coefficients.

Dropout [87] is another technique regularly applied to reduce overfitting. Dropout is commonly applied to the fully connected layers of classification networks, where the majority of the parameters are in the network, but can also be applied to convolutional layers. It randomly, with probability p, drops out nodes, preventing the layers from learning complex co-adapted relationships, forcing the learned features to be more robust and less specific. This, therefore, helps to prevent the network from overfitting to the training data.

3.4.1.4 Data Augmentation

Data augmentation is the final technique that will be discussed which aims to prevent overfitting. Ultimately, especially for medical imaging tasks, the amount of data

available for training is much less than is required for training the highly parameterised network. Therefore, data augmentation is normally applied, artificially increasing the size and diversity of the training dataset through applying transformations, creating slightly perturbed versions of the data, working under the assumption that more information can be extracted from the original dataset through these transformations.

These augmentations can be general or domain specific – Fig. 3.10. Basic transformations include intensity transformations (pixelwise noise, image brightness, contrast adjustment), geometric transformations (affine and elastic transformations), cropping, occlusion and application of filters, as are standardly applied in computer vision tasks [22, 40]. More extreme augmentations such as combinations can also be applied – for instance, mixup [88], where images are combined from two different labels to produce hybrid images.

Domain specific approaches aim to augment the data in ways expected using domain knowledge: for instance, [89] propose an adversarial intensity transformation model to model bias field inhomogeneities, and many approaches use deformable methods, such as registration to an atlas, to create biologically meaningful shape variation [90, 91]. Generative networks, such as conditional Generative Adversarial Networks (GANs) [92] are also used for augmentation, being trained to generate simulated data in a data driven manner [93, 94]. CycleGAN [95] and Pix2Pix GAN [96] are style transfer methods which can also be used for data augmentation, transferring structural information from one domain (for instance scanner) to another target domain. For a review of augmentation methods in medical imaging, see for instance [97].

3.4.1.5 Differences between Datasets or Data Domain Shift

Model overfitting leads to models whose performance degrades quickly once it has to extrapolate to even slightly out of distribution data, and even perturbations which are

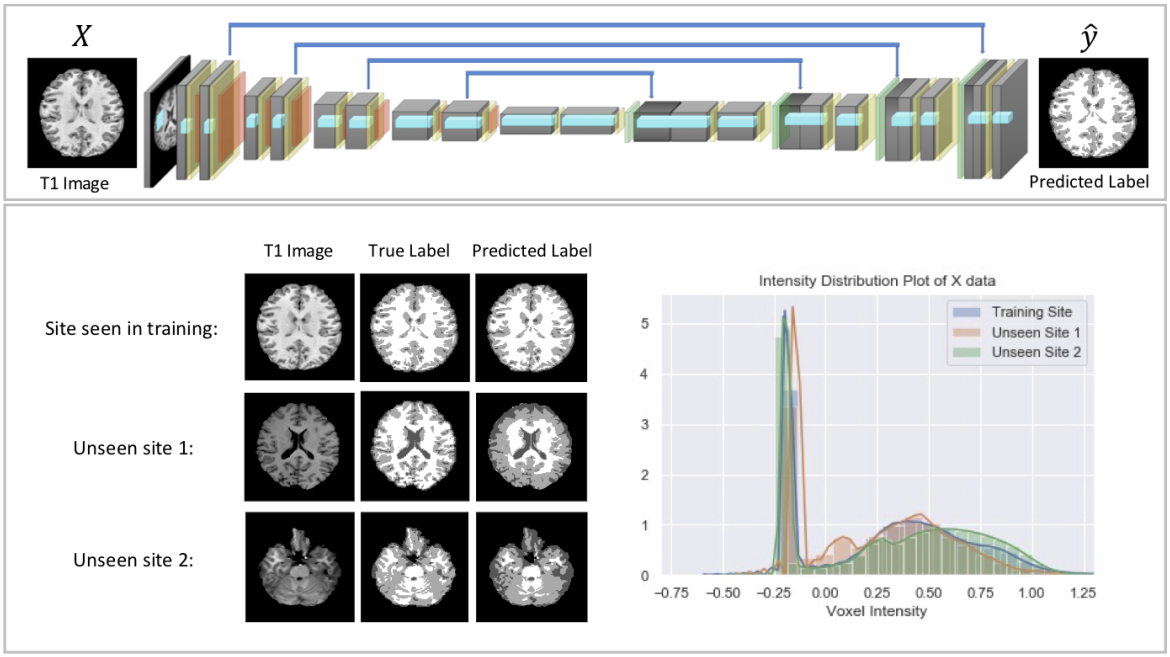


Figure 3.11: To demonstrate the effect of the difference between domain datasets or domain shift, tissue segmentation is completed on data from three sites collected as part of the ABIDE [98] multisite dataset. Although the data was collected as part of the same study, there are differences between the data collected at different sites due to being collected on different scanners. The architecture used was a 3D UNet [42] with T1 as the input image, and only images from one site were used during training. The predictions can be seen for example images for three sites, the one seen during training and two unseen. It is clear that the segmentation for the site seen during training is good but suffers significant degradation when applied to the unseen sites, despite their being collected for the same study and having similar (normalised) voxel intensities, demonstrating the potential difficulties caused by domain shift.

not noticeable by the human eye can cause the network performance to collapse [99]. Considerable research effort within computer vision has focused on generalisability from the dataset seen during training to the dataset only seen during testing, where both datasets are drawn from the same distribution. Ideally, generalisability is needed from the training set to all other reasonable datasets, including those not yet collected.

If, for instance, multisite datasets are considered, such as from the ABIDE study [98], where attempts have been made to harmonise acquisition protocols and to use identical phantoms across imaging sites, there is still an increase in non-biological variance when

data is pooled across the sites and scanners [100]. A demonstration of this variance leading to performance degradation for a segmentation task is shown in Fig. 3.11. Multiple studies have confirmed this variation, identifying causes (batch effects) from scanner and acquisition differences, including scanner manufacturer [101, 102], scanner upgrade [101], scanner drift [103], scanner strength [101], and gradient non-linearities [104]: this problem is commonly known as the data harmonisation problem.

The removal of this scanner-induced variance is therefore vital for many neuroimaging studies. The majority of deep learning approaches try either to produce harmonised images [105, 106], or to remove the scanner related information from the features used to produce the predictions [107, 108]. Both approaches aim for any results obtained for use downstream to be invariant to the acquisition scanner and protocol. These methods succeed in removing the scanner effects from the predictions, but hold no guarantees for scanners not seen during training, and, as the results are very hard to verify without *'travelling heads datasets'* (images from the same subjects acquired on the scanners to be harmonised) the results obtained from the generated harmonised images are hard to validate [107].

The domain shift experienced when multisite data is used is much less than might be expected when moving between research data and clinical data, or even just two datasets collected independently [109]. The domain shift here can come from two clear sources: the scanner and acquisition, and the demographics of the studies. Firstly, MRI scans collected for research are often at a higher resolution with smaller spacing. On the other hand, clinical scans are designed to be much more efficient – both in terms of acquisition time and in time required for visual inspection – and are often collected at a lower resolution, for instance at 1.5T, often with anisotropic voxels. Additionally, the majority of deep learning approaches also expect a specific acquisition protocol. Unfortunately, due to the lack of data already discussed, it is unlikely that

it would be possible to train directly on the clinical data.

Therefore, methods are being developed that consider this domain shift, which, similarly to the harmonisation problem, focus either on domain adaptation approaches to create shared feature representations for the different datasets, or in synthesising data to enable training to be in the clinical domain. Unlike the harmonisation paradigm, they do not wish to enable the combination of datasets, but simply to be able to harness the shared information from one domain to the other. Domain adaptation techniques normally consider the situation where there is a large dataset – say a research dataset such as the UK Biobank [1] – and a much smaller target dataset – say the clinical dataset that is the dataset of interest [110, 111]. Domain adaptation then asks the question: *can a shared embedding be found which is discriminative for the task of interest, while invariant to the domain of the data*? This can take the form of a fully supervised problem, where task labels are available for both datasets, semi-supervised [112], where only a small number of labels are available for the target dataset, or unsupervised [113, 114], where no labels are available for the target dataset. The vast majority of approaches perform domain adaptation across domains only; however, some also consider adaptation across related tasks [115] and have been applied for segmentation [114, 116] and classification problems [117]. These methods are clearly closely related to transfer learning from a large dataset, but a single feature representation is found for both source and target dataset, rather than creating a new one for the target dataset. These methods perform well on the target data, but further work is required to enable them to adapt reliably to higher numbers of datasets simultaneously. Domain adaptation methods, at the extreme, essentially have the end goal that the network would work regardless of the acquisition, which is an active area of research [118–120]. The other approach which has been explored is to use generative methods to convert the data from one domain to the other [121], such that the transformed data can be used

in the existing model. Any generated images must be carefully validated to ensure that they convey the same information as the original and that the outcomes are the same.

3.4.1.6 Data Composition and Algorithmic Biases

Finally, and potentially most concerning, is the consideration that the demographics of study data frequently do not fully represent the population as a whole, and so a domain shift is experienced when moving from the research domain to the clinical domain. Research data is usually acquired with a certain study question in mind and therefore subjects are selected so as to try to allow targeted exploration of that question. Therefore, research datasets rarely contain subjects with co-morbidities and also subjects with incidental findings would possibly be excluded from the study. An example being that patients with advanced Alzheimer's disease are unlikely to be recruited for an imaging study, due to the ethical implications [122], such as the inability to consent and the potential trauma of being scanned.

In addition, there exists a strong selection bias, both in relation to the people who volunteer for studies and those who see the study through to the end, with studies having demonstrated associations to age, education, ancestry, geographic location, and health status [122, 123]. Furthermore, people with family connections to a given condition are more likely to volunteer for a study as a healthy control, leading to certain genetic markers being more prevalent in a study dataset than in the population as a whole [124].

Therefore, any trained model will suffer from data bias: that is, the outcomes of the model will potentially be systematically less favourable to, or have systematically lower performance on, individuals within a particular group, where there is no relevant difference between groups that justifies such harms [125]. Erroneous or unsuitable outcomes will likely be produced for the groups who are less likely to be represented in the training data. Data bias is therefore a function of the creator and the creation

process, and fundamentally, of the data which drove the model training. When considering bias, there are two issues that must be considered: first, *does the model have different accuracy rates for different demographic groups*? Second, *does the model make vastly different decisions when applied to different populations*? As networks simply learn the patterns in the data, any bias, such as racial bias [126], in the data will be learnt and encoded into the models.

Inevitably, when considering complicated questions with extremely heterogeneous populations, the datasets used to train the deep learning methods will be incomplete and insufficient in terms of spanning all of the possible modes of variability [127]. For instance, pathologies will occur against a background of normal ageing, with differences being present between individuals due to both processes. Sufficiently encompassing all of this variation is infeasible, not only due to the number of subjects which would be required, but also to the difficulty in recruiting subjects from some specific groups. There is, therefore, often inadequate data from minority groups, and consequently any model found through the optimisation process will probably be inadequate for these groups, especially for conditions with higher rates in these groups [128]. When models are developed for clinical translation, therefore, the limitations of the models must be understood: wherever groups are under-represented, the appropriateness of the application of the model must be considered, and any limitations identified. Where these limitations mean that minorities will receive a lower standard of care, the models are inappropriate.

3.4.2 Evaluation and Ground Truth

The evaluation of model performance requires labels – that is, a *ground truth* annotation created by a domain expert. In medical imaging, the ground truth is regarded as

labels created by domain experts and these labels are key for training models, but do not necessarily form part of standard clinical practice. Firstly, the labels are required for evaluation of the model performance, and secondly, they are required to allow the training of supervised methods. This, therefore, exacerbates the problem of the shortage of data, as not only are large amounts of data needed, but also they need equal amounts of correct manual labels. These labels are expensive to obtain, requiring a large allocation of expert time to curate, and expert domain knowledge. Therefore, there is a need to develop methods which can work in data domains where low numbers of labelled data points are available.

Few- and zero-shot learning methods work in very low data regimes and are beginning to be applied to medical imaging problems [129]. They are, however, very unlikely to generalise well to images from other sites and scanners, as the variation seen will not span the expected variation of the data, but they can help to begin to learn clusters within subjects where few labels are available. Unsupervised domain adaptation has been applied more widely, including for neuroimaging problems, to help to cope with a lack of labels, with information from one dataset being leveraged to help the same or a related task on another dataset being performed [114, 116].

Other methods to overcome the lack of available labels focus on working with approximations for labels, which are cheaper to acquire [130]. For instance, many methods propose pre-training the network using auxiliary labels generated using automatic tools (e.g., traditional image segmentation methods) and then fine-tuning the model on the small number of manual labels [35, 131], or registration of an atlas to propagate labels from the atlas to the subject space [132]. Other approaches are weakly supervised, utilising quick annotations such as image level labels [133], bounding box annotations [134], scribbles [135, 136] or point labels [137].

Other approaches to allow the use of deep learning when there is access to a limited

number of training labels include active learning and omnisupervised learning, where both try to make the most effective use of the limited number of labels available for a task. Active learning aims to minimise the quantity of labelled data required to train the network by prompting a human labeller only to produce additional manual labels where they might provide the greatest performance improvements. This minimises the total number of annotations that need to be provided, and results in better performance than randomly annotating the same number of samples [138, 139]. In omnisupervised learning [140] automatically generated labels are created to improve predictions, starting from a small labelled training set. By combining data diversity through applying data augmentation, and model diversity through the use of multiple different models, a consensus of labels is produced, which can be used to train the final model. For both approaches, the labels used are chosen using various different approaches such as uncertainty estimation [141, 142].

The difficulty in acquiring good quality manual labels, is of course, exacerbated by the variance caused when data is pooled, as discussed above. This increase in variance limits the impact that any produced labels can have, and so again, methods of pooling data across datasets without getting an increase in variance due to the scanner effects will be necessary. The labels themselves, however, will provide another source of variance [143]: when working with medical images, the labels are frequently complicated and ambiguous [144], often open to interpretation, or with subjects having multiple labels that could be attributed due to co-morbidities [145] – yet despite this they are assumed to be 100% accurate [146] – the 'gold' standard. Furthermore, this ambiguity also leads to inter-*rater* variability between the people creating the labels, which leads to a degree of uncertainty in the produced ground truth. The effect that this variability has on the predictions of the network needs to be understood and

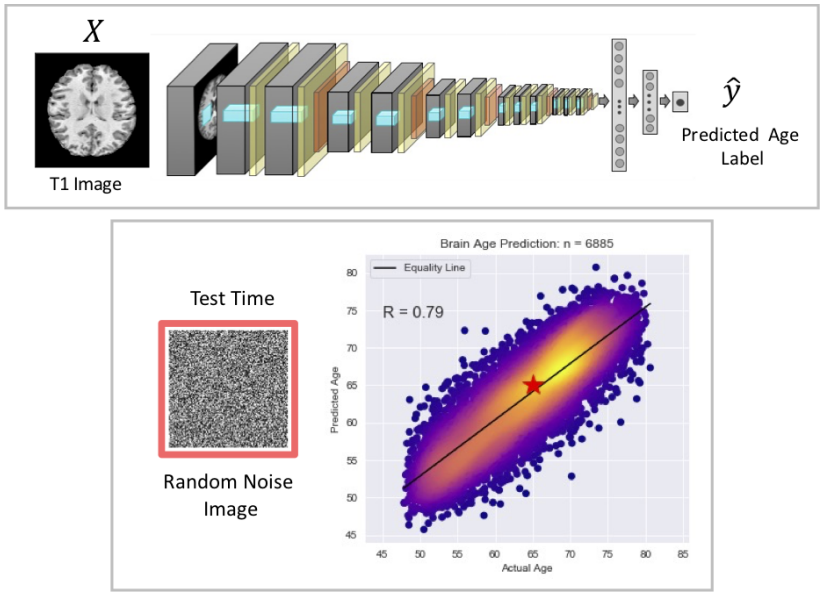


Figure 3.12: If a model trained to predict brain age [79] from T1 structural images is used, and the network is presented with an image of random noise, as it was unable to output an unknown class, the network predicted the random noise to have an age of 65, around the average age of the subjects of the dataset. While it is trivial to identify the random noise image, there are many situations where the model could be presented with an image outside of the distribution it was trained on, but would still output a result, which would be much harder for a user to identify.

mitigated against, with the challenge being that obtaining labels for multiple raters for datasets is deeply infeasible, and has been shown to limit the performance of algorithms [147]. The uncertainty in the labels is also amplified by the lack of available data for rare conditions, which are therefore less represented in datasets, resulting in raters having less experience in assessing them. This is especially problematic if trying to quantify longitudinal changes with different raters [148].

3.4.3 Interpretability

The performance degradation experienced with domain shift would be potentially less problematic were it not for another problem associated with deep learning methods: that is, models will output a prediction for any data fed in, but that prediction may

not necessarily be meaningful. Lacking a 'do not know' option, a neural network will output a prediction, even if that prediction is meaningless or the input is nonsensical. For instance, if a random noise image is fed into a network trained to predict brain age, the network will predict an age which falls within the distribution of expected ages for the random noise (see Fig. 3.12). While in this example, visually identifying the pure-noise image is trivial, were the network trained for a more complicated classification task, identifying erroneous results would be more difficult and require large amounts of clinical and domain knowledge.

Were deep learning methods to be used to determine patient care, they would need to be interpretable and interrogable, providing logical decision rules which lead to a given diagnosis or system of patient care being chosen. This is especially important because neural networks have no semantic understanding of the problem. If there exists spurious information in $\boldsymbol{X}$ which can aid in this mapping (or *confounders*), then this information will probably be used, misleading the predictive potential of the network. Consider, for instance, the case where all subjects with a given pathology were collected on the same scanner. A network could then achieve 100% recall (correctly identify all examples) for this pathology by fitting to the scanner signal, rather than learning any information about the pathology [149]. It would then, in all probability, identify a healthy control from the same scanner as having the same pathology. The effect of confounders would not be observed without further probing of the behaviour of the trained model, and the probing of networks is non-trivial. There is, therefore, a need for interpretable networks, allowing both understanding and scrutiny of any decision made. Approaches have broadly focused on two main areas: visualisation and uncertainty.

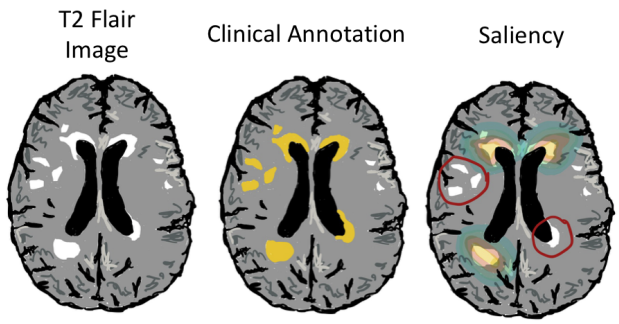


Figure 3.13: Schematic of the limitation of using saliency: when identifying the presence of white matter hyperintensities, the neural network might only need to focus on a few of them to make the prediction. This would lead to not all of the white matter hyperintensities being indicated in the saliency map and so the prediction not matching the clinician's expectation.

3.4.3.1 Visualisation

Visualisation methods generally attempt to show which aspects of the input image led to the given classification, often by creating a 'heat map' of importance of the input image. Many of these methods are post-hoc, taking a pre-trained model and trying to understand how the decision was made. Most commonly, these methods analyse the gradients or activations of the network for a given input image, such as saliency maps [150], GradCAM [151], integrated gradients [152], and layerwise relevance propagation [153]. They have been applied in a range of MRI analysis tasks to explain decision-making, such as in Alzheimer's disease classification [154], brain age prediction [79], and brain tumour detection and segmentation [155, 156]. Other methods are occlusion- (or perturbation-) based, where parts of the image are removed or altered in the input image, then heat maps are generated which evaluate the effect of this perturbation on the network's performance [157, 158]. Most of these methods, however, provide coarse and low-resolution attribution maps and can be extremely computationally expensive [159], especially when working with 3D medical images. Furthermore, care must be taken to ensure that any interpretation of the maps is

reasonable, as they have been found in some applications to be largely a function of the input images, rather than the models [160].

These posthoc methods do not require model training in addition to the original network; however, they have been shown in some instances to be unable to identify all of the salient regions of a class, including in medical imaging applications [159, 161]. It has been shown that classifiers base their results on certain salient regions, rather than the object as a whole, and therefore, a classifier may ignore a region if the information there is redundant – i.e. if it can be provided by a different region of the image, which is sufficient to minimise the loss function. Therefore, the regions of interest highlighted by these methods may not fully match the expectations of a clinician (see Fig. 3.13). Generally, although many methods have been developed to produce saliency or 'heatmaps' from CNNs, limited effort has been focused on their evaluation with actual end-users [162]. Furthermore, these methods, at their best, only highlight the important content of the image, rather than uncovering the underlying mechanisms of the model.

Attention gates aim to focus a CNN on the target region of the image (the salient regions) by suppressing irrelevant feature responses in feature maps. This focuses the attention of the network onto the information critical for the specific task, rather than learning non-useful information in the background or surrounding areas [163–165]. These methods provide the user with attention maps, which again highlight the regions of the input image driving the network predictions. However, these methods, similarly to saliency or gradient based methods, may not highlight all of the expected regions in the image. Attention gates have been applied to a range of medical image analysis tasks, both for classification [79, 166] and segmentation [166]. Other methods have been developed to allow the visualisation of the differences between classes directly, rather than analysing the model post-hoc [159, 161]. Other methods aim to increase their interpretability by breaking down the task into smaller, more understandable

tasks, such as first segmenting a region known to be a biomarker for a given condition, and then classifying based on this region [167].

The methods discussed so far enable the visualisation of the regions of the input image which drive the predictions, but they do not provide insight into how the underlying filters of the network create decision boundaries. In addition, in brain imaging, the class phenotypes are typically heterogeneous and any changes they cause probably occur simultaneously, with significant amounts of healthy and normal variation in shape and appearance, meaning that the interpretation of feature attribution maps to understand network predictions is often difficult. Given the millions of parameters in many deep learning networks, despite the ability to visualise individual filters and weights helping understanding of the hierarchical image composition, it is very difficult to interrogate how decision boundaries are formed.

3.4.3.2 Uncertainty

The use of uncertainties, then, is an approach which aims to address the problem that, regardless of the input image, neural networks will always output a prediction, even when they are very unsure of the prediction. An example is when the test data falls far away from the domain of data the network was trained on. Furthermore, the softmax values frequently output by a neural network are not true probabilities [170], and networks often output high, incorrect softmax values, especially when presented with noisy or ambiguous data, or when the data presented to them differs from the distribution of the training data. Therefore, the development of models aware of the uncertainty in their predictions is key for providing confidence and trust in systems – and this is not provided by traditional deep learning algorithms.

The majority of methods in the literature focus on epistemic uncertainty, using Monte Carlo dropout [170] to approximate the uncertainty. Dropout [171] is applied to

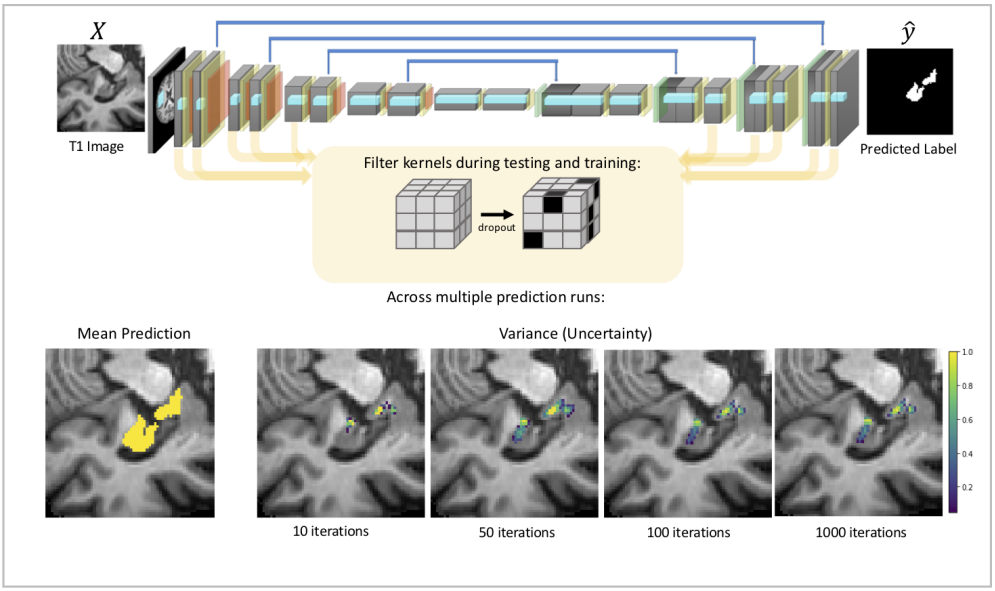


Figure 3.14: Methods which use Monte Carlo dropout have dropout applied at training and test time, where weights in the convolutional kernels are removed, which is approximated to represent the distribution of possible model architectures at test time. To demonstrate this, a standard 3D UNet was trained to complete hippocampal segmentation, with a dropout value of 0.5 applied on all convolutional layers. The HarP dataset [168] was used in this experiment, pre-processed as in [169]. For each subject, a mean prediction and an uncertainty map were obtained, indicating the regions where the predictions between models were the most varied and so approximated to be the least certain.

each of the convolutional layers and kept at test time, and thus the model architecture is random at test time. Although other methods exist, in the majority of approaches for medical imaging, the uncertainty is then quantified through the variance of the predictive distribution, resulting from multiple iterations of the prediction stage with dropout present at test time, as demonstrated in Fig. 3.14. This approach can be simply applied to existing convolutional neural networks and in medical imaging has primarily been used for segmentation tasks [172, 173], where the segmentation is predicted alongside an uncertainty map.

There is, however, need for further development of these methods to ensure that the uncertainties produced are meaningful in all of the circumstances in which they could be deployed. For instance, further study is required to ensure that the uncertainty is

meaningful for all possible dataset shifts and to provide a calibration for the uncertainty values so that they are comparable across methods [174, 175].

3.4.4 Logistical Challenges

3.4.4.1 Computational Power

The final category of challenges are far more logistical. Firstly, most deep learning models are developed using GPUs (graphical processing units), hardware that is rarely available in a clinical setting, and producing predictions on standard hardware often takes infeasible amounts of time or requires prohibitive amounts of computational power. Many of the most successful methods applied in imaging challenges involve large ensemble models such as [176, 177], leading to a very large amount of parameters and therefore calculations that must be stored and computed.

Therefore, for clinical translatability, methods need to be developed with consideration of the computational limitations that will be present on deployment, and which seek to create solutions which work within these constraints. Student-teacher networks [178, 179] and model distillation [180, 181] aim to create smaller networks which are able to mimic the performance of the original large model, thus reducing the number of parameters in the final network which is deployed. Other approaches use separable convolutions which drastically reduce the number of parameters in the network. Model pruning [182, 183] acknowledges that the parameters in neural networks are very sparse and, therefore, by removing those that are contributing the least to the final prediction, the size of the model architecture can be reduced, with limited impact on the final performance of the model.

3.4.4.2 Data Sharing and Data Privacy

Clearly from the above discussion of data and labels, if the aim is to develop CNNs which will work for patients in real clinical applications, it will be necessary to train models on medical data which is relevant, realistic and representative. Many current approaches focus on pooling anonymised data from across sites and patient groups, and, in order to be able to do this, concentrate on removing identifiable features such as name, birth data and faces from the images. However, neural networks are still capable of extracting identifiable features from these anonymised images such as age and sex [184], which, in combination with other features such as hospital location and illness, could be identifying [185].

Medical research is not the only field to face difficulties regarding the handling of sensitive and personal information, and many techniques have been developed for privacy-preserving data analysis. The most popular of these techniques are *federated learning* and *differential privacy*. Federated learning means training or testing your model on data that is stored on different devices, without having to centrally collect the data samples into one local aggregated dataset [186, 187]. Instead of moving the data to the model, copies of the global model are sent to where the data is located; the data remains on the individual servers. The model is then trained on the local data, and the model sent back to be aggregated with the main model. This preserves privacy in the sense that the data has not been moved from the device. It is therefore gaining popularity in various healthcare applications [188] – for instance [189, 190].

Differential privacy techniques work by injecting a controlled amount of statistical noise to obscure the data contributions from individuals in the dataset [191], and has been applied to medical imaging [192]. This is performed while ensuring that the model still gains insight into the overall population, and thus provides predictions that are

accurate enough to be useful. Differential privacy is a mathematical framework, and thus enables the degree of privacy loss to be calculated and evaluated based on the concept of a privacy 'budget'. Therefore, ultimately the use of differential privacy is a careful trade-off between privacy preservation and model utility [191, 193].

For a recent review discussing this topic in more detail see, for example, [194].

4

Learning Patterns of the Ageing Brain in MRI using Deep Convolutional Networks

The work presented in this chapter has been previously published in[1]

- **Nicola K Dinsdale***, Emma Bluemke*, Stephen Smith, Zobair Arya, Diego Vidaurre, Mark Jenkinson, Ana IL Namburete: *Learning patterns of the ageing brain in MRI using deep convolutional networks* in Neuroimage, Volume 224 January 2021.

The code used in this chapter is available at:

`github.com/nkdinsdale/LearningPatternsofAgeing`.

[1]* The authors contributed equally.

4.1 Introduction

The ageing of our global population is predicted to become one of the most significant social transformations of the twenty-first century [195]. This demographic shift has implications for all sectors of society, particularly healthcare. A range of neurodegenerative diseases, as well as cognitive decline, are challenges associated with this demographic change. In order to provide effective care and treatment, it is increasingly important to understand which individuals are most at risk of age-associated deterioration and how this deterioration will progress [196].

Even the process of healthy ageing results in changes to the structure and function of the brain. For instance, if we just consider the three tissue volumes, the grey matter volume increases from birth to the age of four and thereafter decreases consistently; white matter volume steadily increases until around the age of twenty and then remains constant; CSF remains constant until around twenty and then increases with ageing [197]. Neuroimaging modalities enable further characterisation of the ageing brain: in addition to the observed changes to the grey and white matter tissues and brain volume [198–201], T1-weighted MR images show evidence of age-related cortical thinning [202, 203], sulcal widening [199] and morphological changes to sub-cortical nuclei [204]. Similarly, T2-weighted MR images have shown an increased prevalence and magnitude of white matter hyperintensities with ageing [205, 206]. Diffusion-weighted MRI (dMRI) has demonstrated the effect of ageing on the microstructure of the white matter tracts, decreased fractional anisotropy and increased mean diffusivity, reflecting demyelination [207, 208] and decreasing structural connectivity [209], reflected also in findings from resting state fMRI [210, 211].

The process of ageing is biologically complex and varied across the population, but neuroimaging provides several methods of viewing the structural and chemical changes.

Although there are general trends, the ageing process is both temporally and spatially heterogeneous, and a large range of 'healthy' variation exists [198–200, 212, 213]. Distinguishing normal ageing from the early stages of neurological disease is challenging, as the early stages of several disease states can resemble advanced ageing [199, 214, 215]. The ability to differentiate early signs of these diseases from normal ageing would hold clinical value, as many current therapies are most effective with early intervention [216]. A better understanding of the healthy ageing process can help us to discriminate between normal age-related changes and neurodegenerative diseases.

Several different approaches have been used to gain an understanding of the healthy ageing process. Most take the approach of developing a model of a healthily ageing brain and measuring how far a new patient differs from the model's prediction of age: a greater divergence could indicate accelerated ageing, or possibly the onset of disease. This divergence has been named multiple things across the literature including BrainPAD [217], BrainAGE [218] and Brain Age Delta [219]. It is first necessary to produce a model which accurately predicts the age of healthy subjects [220], in order to have a meaningful divergence measurement. To date, the most accurate methods have prediction errors in the range of 2.5-5.0 years, depending on the subject group [221].

There is, however, little clinical relevance to predicting age from a medical image: if the subject can be scanned, it is probable that we can ascertain their age by other means. However, error in age prediction in some models has been related to other physiological measurements of ageing, mortality, and general physical health [195, 222]. The fact that $\Delta_{BrainAge} = Age_{Predicted} - Age_{True}$ relates to cognitive performance, ageing fitness, and mortality [218, 222, 223] strongly supports the idea that $\Delta_{BrainAge}$ can provide clinically important insights [222].

The methods and types of neuroimaging data used to obtain the models in the

literature vary widely; however, the vast majority use Magnetic Resonance Imaging (MRI) data. The majority of reported methods extract image derived phenotypes (IDPs) from the MR images, which are then used as the input features to regress brain age using classical machine learning techniques. The majority of methods utilise values derived from T1 MR images, such as the tissue maps – grey matter [224–230] or a combination of tissues [217, 231–235] – while others extract other values such as cortical thickness [221, 236–242], surface area [221, 237] and volumes of ROIs [236, 243, 244]. A comparatively small number of studies utilise features derived from other modalities, such as functional connectivity from fMRI [228, 245–248], measures of fractional anisotropy (FA) and mean diffusivity (MD) or structural connectivity from dMRI [249–251] and combinations of modalities [212, 252–255]. The best performing methods have been based on T1 weighted images [256].

Having extracted these IDPs, they are then used as input features to learn a model to predict brain age. Many different techniques have been used to learn a model of brain age including ridge regression [243, 244, 254], relevance vector regression (RVR) [224, 225, 227, 251, 257], support vector regression (SVR) [245, 247] and Gaussian process regression (GPR) [217, 232]. Although the underlying mechanisms of the models vary, they approach the task in the same way: the input to the model is information derived from the MR images and the model regresses the true age label. This can also be achieved using a CNN as the model; this has been the most explored method recently, producing the lowest weighted MAE results (MAE scaled by the age range of the test subjects as proposed for method comparison in [256]).

The use of a CNN was first explored in [217] – although previous works explored the use of ANNs [221, 249] – which demonstrated that CNNs could be used to accurately predict brain age from raw (minimally pre-processed) T1 MR images. Various additional studies have applied CNNs to different datasets [76, 220, 258–261]. Some approaches

apply the CNNs to IDPs [220, 260]; the trend is towards applying the network to the raw, non-feature extracted images themselves, thus removing the need to identify the features which will be the most informative for the task. This not only reduces the amount of pre-processing required, but also allows unknown relationships to be identified as the model learns, through the optimisation process, which features to use.

To be of use, the $\Delta_{BrainAge}$ values must be meaningful, and contain information corresponding to clinical outcomes. Age-related changes to the brain are subtle, non-linear, and spatially and temporally heterogeneous, both within a subject and across a population [217]. Deep learning models are particularly suited to addressing tasks with these properties, and, as demonstrated by the previous studies [217], they are able to produce models which are sensitive to the large variety in healthy brain appearance. This can be more effective than comparing population 'averages' per age, which are unlikely to be representative of any single individual. Deep learning approaches are able to extract features and create non-linear functions of these features to map from the input images to the true ages which are used as the training labels. However, to achieve this, due to the vast number of parameters which form the networks, they require large quantities of training data – far more than classical machine approaches. Within neuroimaging, and medical imaging more generally, the availability of training data, and further training data with labels for the task of interest, is a limiting factor: 'large' neuroimaging databases often contain samples in the magnitudes of 10^2s, not 10^6s as found in the commonly used computer vision datasets. In addition, although a CNN allows the whole image to be used, allowing the model to identify the salient regions rather than the user, this is computationally very expensive, requiring significant amounts of computational power and GPU time.

Projects such as the UK Biobank and the Human Connectome Project are now

making larger amounts of neuroimaging data accessible to researchers [1] than has ever previously been available. UK Biobank contains data from participants' lifestyle, clinical measurements, genetic, and imaging data. The longitudinal aspect of the UK Biobank makes it well-suited for examining age prediction and health outcomes, as the $\Delta_{BrainAge}$ correspondence to later clinical outcomes will be measurable. There are, however, limitations associated with the dataset, which are discussed in chapter 8.

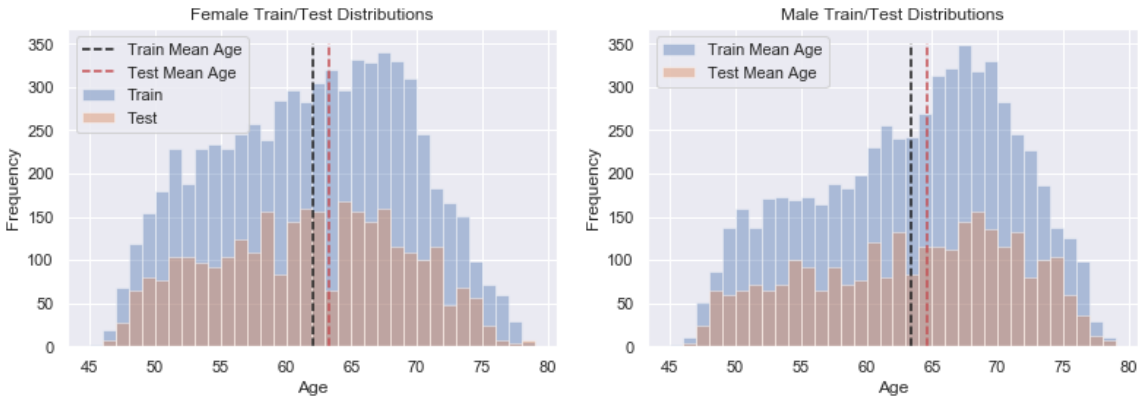
Therefore, the contributions of this chapter are as follows:

- A 3D CNN architecture was developed, designed to predict age, and then tested for $\Delta_{BrainAge}$ correlation with other measurements (image derived values and non-imaging factors) from the UK Biobank. A training dataset of $12,802$ T1-weighted MRI images was used and a further $6,885$ images for testing (limited by the subjects released when this work was completed). The choice was made to utilise a 3D architecture as this preserves the structural information from all of the dimensions of the image, and has been shown to outperform 2D CNNs where sufficient training data is available [262], as is the case here due to the size of the UK Biobank.

- Since sex-specific effects have been demonstrated in both the ageing process and in previous prediction models [263–265], the subjects are stratified by sex and separate models developed for the two groups. Sex is considered as stated by the participant: no information on gender is available. This allowed examination as to whether the resulting $\Delta_{BrainAge}$ corresponded to different UK Biobank measurements for each sex. This reduces the number of subjects available to train each model, and so risks reducing the accuracy of the final trained model, but increases the likelihood of the trained model being meaningful, as it is known

that the trends of ageing differ.

- Any produced model is only meaningful if the predicted Δs hold physiological meaning and our goal is not simply to produce the lowest error on the age prediction task but also to show independent biological associations with the deltas. To investigate the performance of the network for subjects with a high predicted Δ, a subject with a high $\Delta_{BrainAge}$ was compared to an age-matched subject with a low $\Delta_{BrainAge}$. This provided an indication of whether the Δs held physiological relevance, or if they were simply evidence of the model's failure to predict accurately. Since images from only one imaging modality in this experiment were used, the relationships between the $\Delta_{BrainAge}$ and IDPs from all of the imaging modalities in the UK Biobank were examined. The predictions were also compared to those derived from a second model that used only these multi-modality IDPs as training input [266], to give insight into which modalities could be most useful for future age prediction models. The use of a CNN means that the created model has freedom to investigate spatial signatures, rather than relying on spatial priors or target structures, in contrast to other methods which have previously explored brain age in the UK Biobank. In both [266] and [267] image-derived phenotypes were used to train a linear model to predict brain age, with the former using IDPs to create a single estimate of brain age, and in the latter, IDPs being used to explore whether multiple modalities can produce distinct modes that relate to different processes involved in ageing and disease. These methods allow exploration of different aspects of ageing using multimodal data but they require prior choice of the IDPs, whereas by only using the T1 images as the input, there is potential for the CNN to pick up on different changes to the IDP-based model, especially subtle, distributed changes. In [220]



(a) **(b)**

Figure 4.1: Training and test data histograms for male and female subjects. Maximum age = 80.6, minimum age = 44.6. Female mean (training) = 62.1, Female mean (testing) = 63.3, Male mean (training) = 63.4, Male mean (testing) = 64.6.

an ensemble of CNNs is used to predict brain age using the UK Biobank as the training data, but focuses on the creation of the best performing model possible rather than exploration of the biological relevance.

- These possible subtle changes are also a motivation for the exploration of the effect of registration on the model predictions. Registration is a standard pre-processing stage, but the registration process may suppress morphological changes, affecting the brain age prediction. Therefore, separate models are trained not only for each sex, but also for linear and non-linear registration in the pre-processing stage, and the effect on the produced models is explored. It is demonstrated that the most accurate models are not necessarily those which appear the most biologically meaningful and what drives the models is investigated.

4.2 Methods

4.2.1 Data and Pre-processing

A total of 12,802 T1-weighted structural MRI from the UK Biobank database [1] were used for training, 1424 for validation, and a further 6,885 for testing in this experiment, split by sex: the age distribution can be seen in Fig. 4.1. Due to existing findings about sex differences, such as post-menopausal accelerated brain ageing [267], stratification by sex was conducted in this study.

The images were acquired at 1mm isotropic resolution using a 3D MPRAGE Sequence [268] and processed using the UK Biobank pipeline [269]. For these experiments, both the linearly registered and non-linearly warped versions of the images (registered to the standard MNI template) were used to train the network separately, and the results compared. Both linearly and non-linearly registered images were considered because, whilst part of many standard pipelines including the UK Biobank Pipeline [269], non-linear registration to the MNI template suppresses morphological information. These morphological changes are likely to be used by neural networks as a biomarker for age prediction and so it is important to understand the effect of the registration process in the age predictions and $\Delta_{BrainAge}$ values calculated.

The original image dimensions were $182 \times 218 \times 182$ voxels; however, to reduce the memory needed for computation, the x and y dimensions were resized to 128×128 voxels, and only 20 slices out of 182 were used, as can be seen in Fig. 4.2. To include a sufficient amount of anatomical information and resolution whilst reducing redundancy, every fourth slice was used from a region of 80 slices. The region was chosen so as to include as much of the brain as possible while still including some information on the cerebellum, whose structure and function is also impacted by ageing [270].

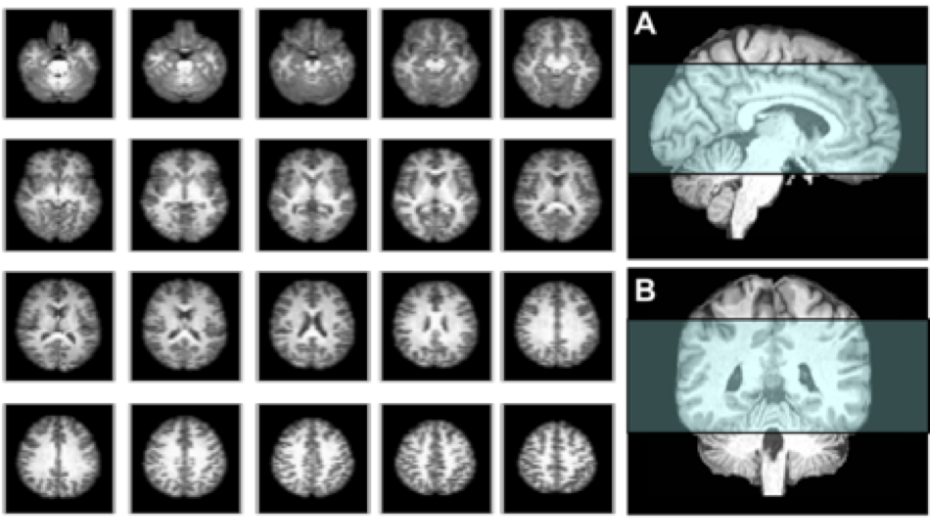


Figure 4.2: An example of structures contained in the axial view (128x128 pixels) of the 20 slices used. The region covered by the extracted slices is illustrated in the (A) sagittal and (B) coronal planes.

4.2.2 CNN Architecture

For this experiment, a 3D CNN was constructed. The sections of brain-extracted T1-weighted MRI scans were used as input to produce a single scalar regression output that represented the predicted age of the subject. Adapting the successful VGG-16 [22] architecture for age regression, the network used a mean squared error (MSE) loss function and consisted of 12 convolutional layers which convolved the input with a varied number of filters (f) of kernel size $3 \times 3 \times 3$. The number of filters increased as the max-pooling decreased the spatial dimensions: this allowed the collection of a rich representation of features. The outputs of each convolutional layer were processed by a rectified linear unit (ReLU) activation function and then they were batch normalised [26]. The network contained four max-pooling layers with kernel sizes of $2 \times 2 \times 2$ and a stride of 2. The output of the final convolution was then flattened and followed by two fully-connected layers of size $3f \times 1$ and $f \times 1$, respectively. The final output layer of size 1×1 contained a linear activation function

in order to predict age as the output scalar. To create the final network, three of these networks were combined to create an ensemble network, such that the final age prediction was the mean of the three network predictions. A complete schematic of both of these architectures is illustrated in Fig. 6.3.

4.2.3 Experimental Setup

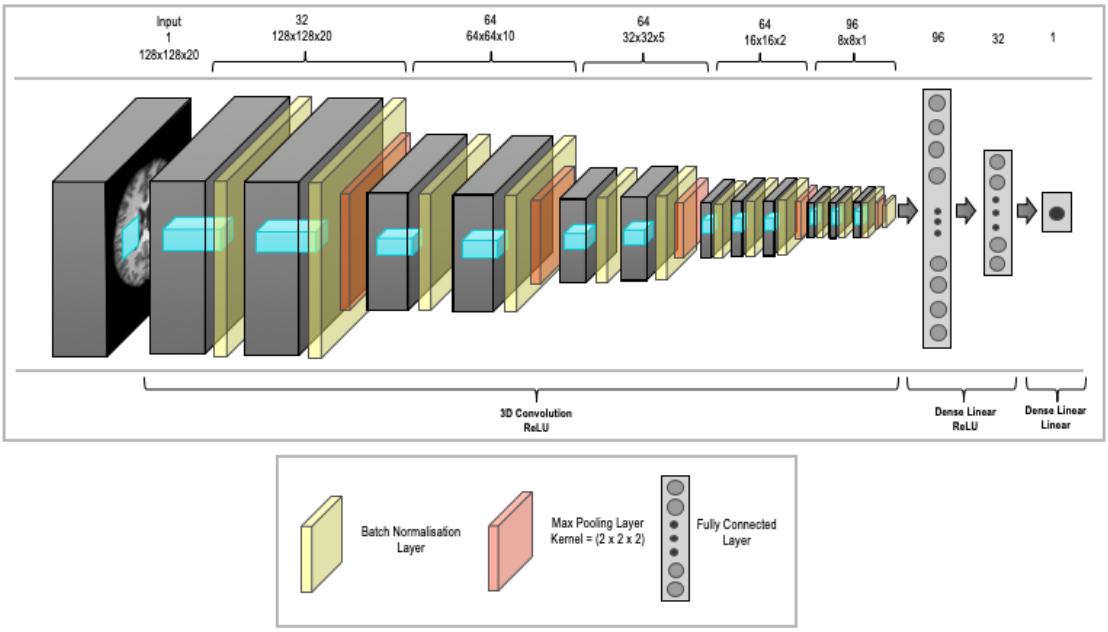
A set of three networks were trained, for men and women separately, and the results were combined to form an ensemble network by finding the mean result. Ensembling was used because it increases the stability of the network predictions by reducing the variance. This, therefore, helps to make sure that the predicted $\Delta_{BrainAge}$ values are being driven by biological differences and not by the random network initialisation or stochasticity in training.

The final ensemble networks were trained on the full dataset with a 90%/10% training/validation split and five-fold cross validation, and the final ensemble's performance was evaluated on a hold-out test dataset (subject numbers can be seen in Table 4.1). For each network and the final ensemble networks, the mean squared error (MSE), mean absolute error (MAE), and standard deviation of $\Delta_{BrainAge} = Age_{Predicted} - Age_{True}$ were calculated. Separate networks were trained and tested on the female and male subjects, as well as on the linearly registered and non-linearly warped image datasets, leading to four ensembles in total.

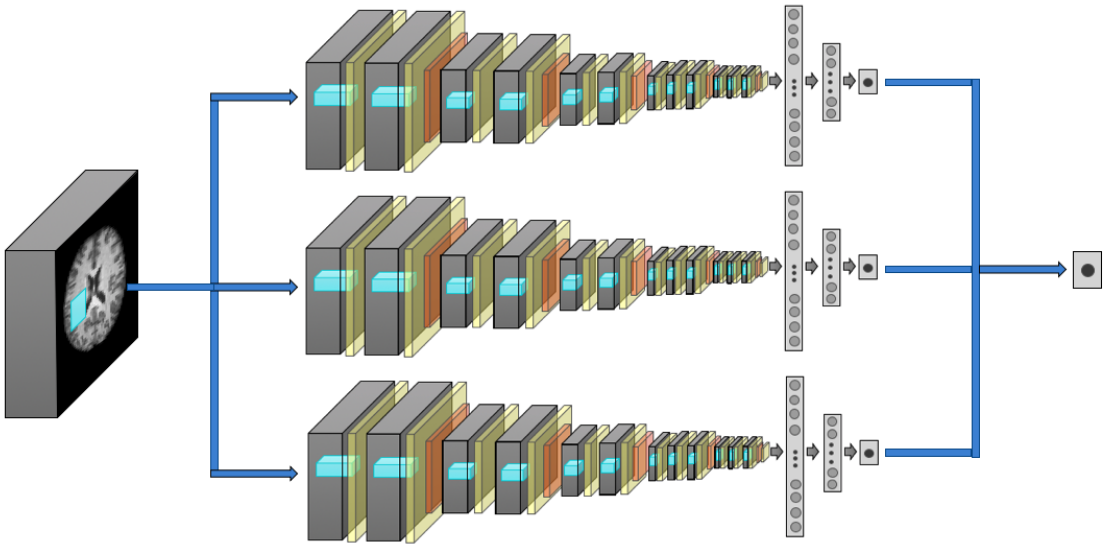
Sex	Train	Validation	Test
Female	6579	732	3654
Male	6223	692	3231

Table 4.1: The number of subjects used in each step, separated according to sex.

The training of the network was implemented using the Keras (v2.1.5) framework



(a) The architecture of the fundamental network used in this work. As illustrated, each convolution is followed by batch normalization and ReLU activation. The number of filters in each of the twelve layers increases as f, f, 2f, 2f, 2f, 2f, 3f, 3f, 3f, 3f, 3f, 3f.



(b) The ensemble network architecture. Each of the networks has the same structure but independent weights, learnt from randomly initialised weights. The aggregation is completed through averaging the outputs of the three individual networks.

Figure 4.3: The network architectures used in this work.

with a Tensorflow (v1.8.0) backend [271, 272] and performed on an NVIDIA Tesla P100 GPU. Optimization was achieved using the Keras RMSprop algorithm and mean squared error loss function over a maximum of 200 epochs. To minimize overfitting, the MSE on the validation set was monitored during training and the network was set to stop training if the validation MSE did not improve for 30 epochs. The initial learning rate was 10^{-2} and was set to decrease by a factor of 2 if the validation MSE did not decrease over 15 epochs, with a minimum learning rate of $10^{\breve{}6}$. In every epoch, the network was trained with a batch size of 16.

4.2.4 Correlation with Subject Biobank Measurements and IDPs

4.2.4.1 Obtaining $\Delta_{BrainAge}$

The final predicted age was the output of the ensemble network, found by averaging the outputs of the individual networks in the ensemble. Once this mean age prediction was obtained, the difference between the network's predicted age and true age was calculated with a simple subtraction: $\Delta_{BrainAge} = Age_{Predicted} - Age_{True}$. In this convention, a positive delta implies that the network predicted the subject's brain to be older than the subject was, and a negative delta would imply that the brain was predicted to be younger. Before performing any correlations, but after calculating MAE and similar performance metrics, consistent biases associated with linear and quadratic age were removed from the $\Delta_{BrainAge}$ acquired from the CNN – see Appendix C.1.1 for explanation of the deconfounding process.

4.2.4.2 Relationship with Biobank Measurements

To examine the relationship between the CNN-derived deltas and UK Biobank measurements, univariate statistics were carried out using Pearson correlation analyses between the deltas and 8787 non-image derived variables from the UK Biobank database. These variables fell under various categories of lifestyle factors and physiological measurements and have been categorized in Miller et al. [268]. Sixteen of these categories were chosen as being of interest in this research, which included six lifestyle factors (Early Lifestyle Factors, Lifestyle General, Exercise, Work, Alcohol, and Tobacco), eight physiological or medical measurements (Physical General, Bone Density, Size, Cardiac, Blood Assays, Eye Test, Physical Activity Measures, Cognitive Phenotypes), and all variables contained in the Medical History and Mental Health Self-Report categories. The variables examined had all confounding variables removed (such as sex, age, head size, head motion, scanner table position, and imaging centre) and were then Gaussianised (quantile normalisation) [268]. The resulting p-values were sorted by category and plotted together with the Bonferroni correction and false discovery rate (FDR) correction thresholds. The correlations were performed separately for each sex.

4.2.4.3 Relationships with IDPs

To examine the relationship between the CNN-derived deltas calculated from T1-weighted images and the image derived phenotypes (IDPs) collected from all of the imaging modalities, univariate statistics were again performed using Pearson correlation analyses between the deltas and 941 IDPs as used by Miller et al. [268], more details can be found in Appendix A. Confound variable removal, Gaussianisation, and p-value calculation were all carried out in the same way as in Section 4.2.4.2.

4.2.4.4 Comparison with IDP-Based Age Prediction

A separate model by Smith et al [266] for age prediction was used for comparison, where all of the IDPs, rather than images, were used to predict age. This was a linear regression model using all IDPs, a matrix of confounds, and a vector of ages as input variables. The model was regularised using an elastic net penalty, which is designed to promote sparsity in the model – i.e. to use a subset of the variables only – as well as leading the regression coefficients of correlated variables to have similar values [273]. The elastic net dominates other regularisation choices, given an adequate choice of hyperparameters, which were selected using cross-validation. The prediction was then performed using multiple regression. Univariate pre-feature selection was completed [274] and only the 12.5% most correlated features were selected using nested five-fold cross correlation. In this paper, comparisons were made with all subjects for whom these features were available: 3212 female subjects were retained and 2717 male. Before performing any correlation between the IDP model-based deltas, the variance associated with age and quadratic age was removed following the method proposed by Smith et al (2019) [266].

4.3 Results

4.3.1 Age Prediction

	Linear Registered Input		Non-linearly Registered Input	
	Female	Male	Female	Male
MAE	2.86±2.22	3.09±2.37	2.71±2.10	2.91±2.18
MSE	13.12±19.1	15.13±22.3	11.77±16.85	13.28±18.98
R	0.870	0.860	0.887	0.882

Table 4.2: Ensemble Network results (years) split by sex and image registration method.

	Female			Male		
	Network 1	Network 2	Network 3	Network 1	Network 2	Network 3
MAE	3.09±2.39	3.23±2.45	3.19±2.43	3.26±2.52	3.41±2.62	3.48±2.61
MSE	15.26±22.6	16.44±23.4	16.07±23.1	17.00±25.5	18.53±26.9	19.00±23.1
R	0.847	0.840	0.837	0.841	0.826	0.822

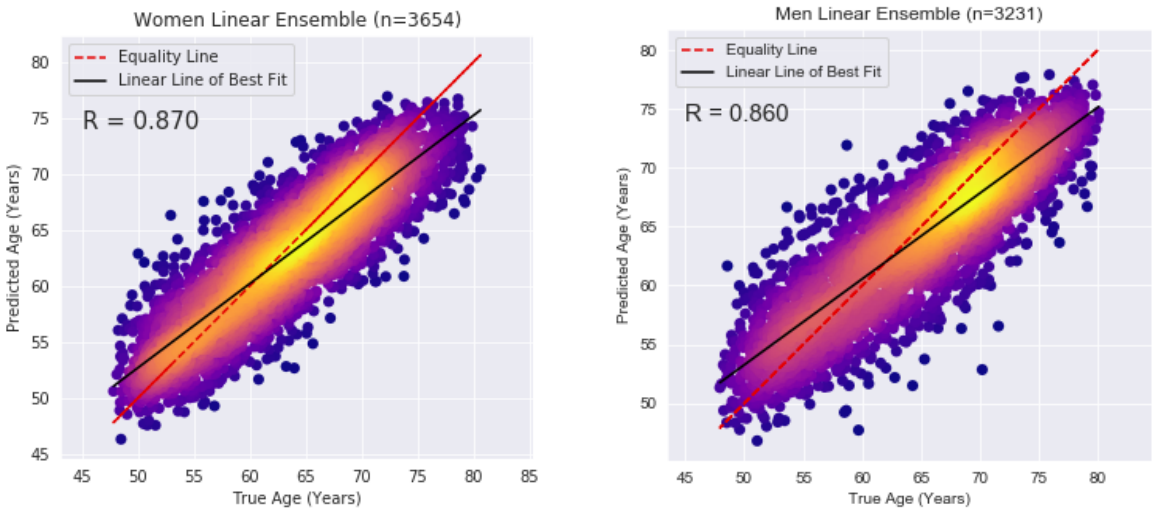
Table 4.3: Linear Registered Results (years) for each network in the ensemble.

	Female			Male		
	Network 1	Network 2	Network 3	Network 1	Network 2	Network 3
MAE	3.09±2.37	3.17±2.42	2.87±2.22	3.32±2.56	3.23±2.41	3.21±2.47
MSE	15.15±22.0	15.88±22.8	13.15±19.0	17.06±27.8	16.27±23.0	16.42±24.2
R	0.847	0.844	0.869	0.839	0.849	0.844

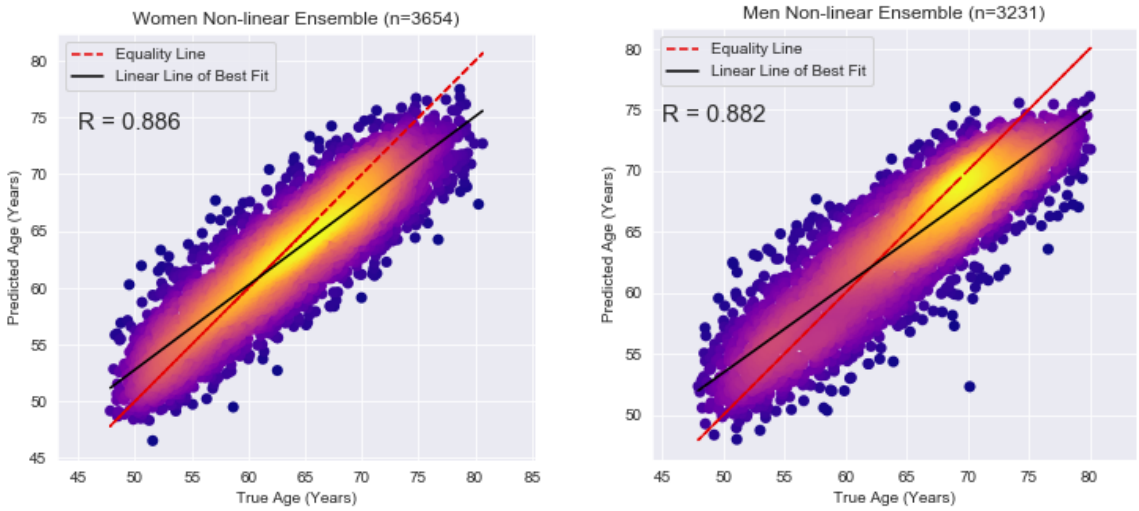
Table 4.4: Non-linear Registered Results (years) for each network in the ensemble.

The results of the CNN's age prediction on both the female and male groups with both linearly registered and non-linearly warped images can be found in Table 4.2. Although the aim of this work is not necessarily to create the model with the lowest MAE, rather, to create a meaningful model of brain ageing, for it to be meaningful it is important that the model performs well at the age prediction task [220]. MAE and MSE are sufficient for comparison between our experiments on the same dataset: for comparison to other methods tested on other datasets, weighted MAE has been proposed [256] which scales the MAE by the age range of the data. The age range of the test set in all of our experiments was 44.6 - 80.6 years. It can first be seen, as expected, that the ensemble model outperformed the three individual models, even though all three models had the same architecture and the same training data. It can also be seen that the model trained on the non-linearly registered data had the lowest overall MAE and MSE.

Figure 4.4 shows the density plot of the ensemble results, stratified by sex and registration type. For each experiment, the predicted age is plotted against true age with the linear line of best fit shown. It can be seen that for all four experiments, the



(a) **(b)**



(c) **(d)**

Figure 4.4: Density plot of the predicted ages versus the true subject age. a) predictions resulting from linearly registered images for the female subjects, b) predictions resulting from linearly registered images for the male subjects c) predictions resulting from non-linearly registered images for the female subjects, d) predictions resulting from non-linearly registered images for the male subjects.

model systematically over-estimates the age of the subjects below the average age of the dataset and under-estimates the age of the subjects who fall above the average age of the dataset. This clearly shows that the predicted $\Delta_{BrainAge}$ value varies as a function of the true age, clearly motivating the need to deconfound for age before the associations are explored.

Having obtained the models for the linearly and non-linearly registered data, the effect of the registration on the model predictions was investigated, and, thus, the potential validity of the underlying model. First, two subjects with the same chronological age were considered who had very different predicted ages and, thus, very different $\Delta_{BrainAge}$ values. In Figure 4.5, two subjects are compared, both of whom were 55 years old when they were scanned. Subject 1 was predicted to be 72.31 using the linearly registered data and 68.28 using the non-linearly registered data, while Subject 2 was predicted to be 55.44 using the linearly registered data and 58.43 using the non-linearly registered data. In Figure 4.5B and D, the same slice, linearly and non-linearly registered, can be seen. Figures 4.5A and C show slices from the linearly registered brain for comparison between the two subjects.

Subjects 1 and 2 in Fig. 4.5 have the same true age but very different levels of atrophy and predicted ages; for instance, the ventricles in subject 1 are clearly much more dilated than the those of subject 2, and more so than might be expected for someone of that age. If both subjects were predicted to be the true age, this would result in a lower MAE but would not highlight the large differences in anatomy between the two subjects. The age prediction with the non-linear registration is worse for subject 2, whose visible atrophy is closer to what might be expected and better for subject 1, potentially suggesting that the non-linear registration is suppressing differences to conform the brain to the template. Whilst certain details and pathologies are still

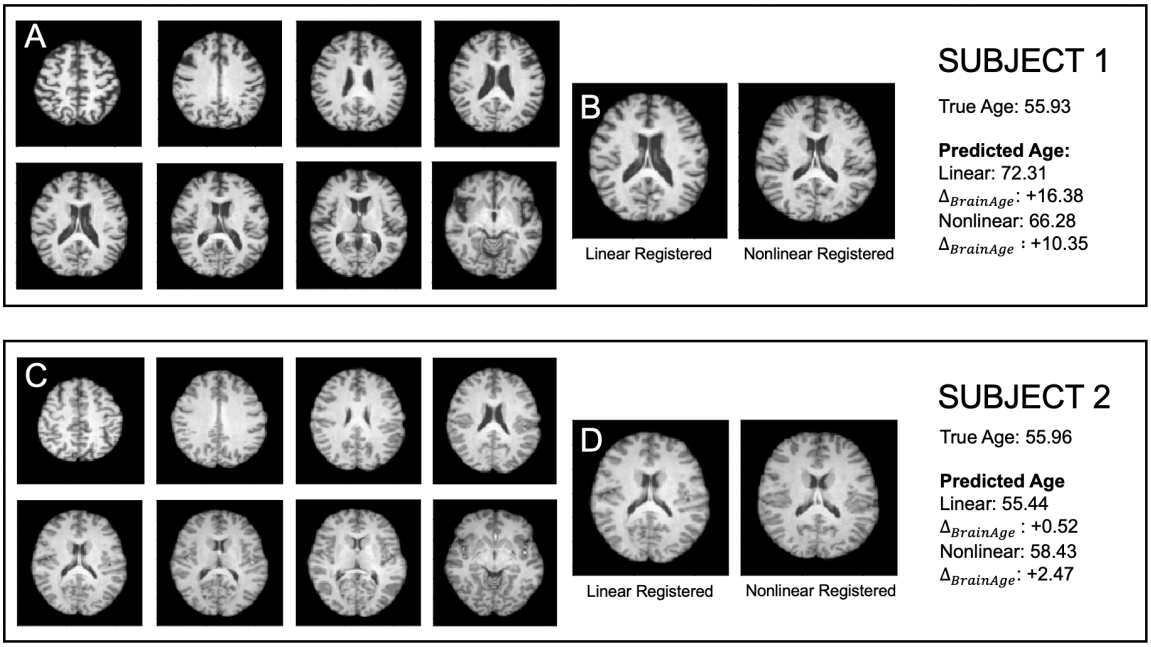


Figure 4.5: MRI Images from two subjects both with true age 55. A and C are example linear slices from the two subjects. B and D are the same slice for both subjects, both linearly and non-linearly registered.

preserved, it is evident in Figs. 4.5 C and D that the ventricles in the non-linearly registered slices do not appear as dramatically different as they do in the linearly registered slices. This is notable, as change in cortical volume is one of the most common effects of age reported, and many of the methods mentioned have cortical volume or thickness measurements as input for successful age prediction [221, 233, 275].

This example is striking due to the large structural differences. However, it is entirely possible that there are subjects with more subtle pathologies which lead to larger predicted Δ values. Similarly, there might be subjects with pathologies which are not visible in the T1 images, that the network fails to account for.

To examine which brain features are associated with subjects that have a higher predicted brain age, the non-linearly registered images of all subjects predicted to be 75-80 or 45-50 by the CNN in each sex group were averaged to form two mean

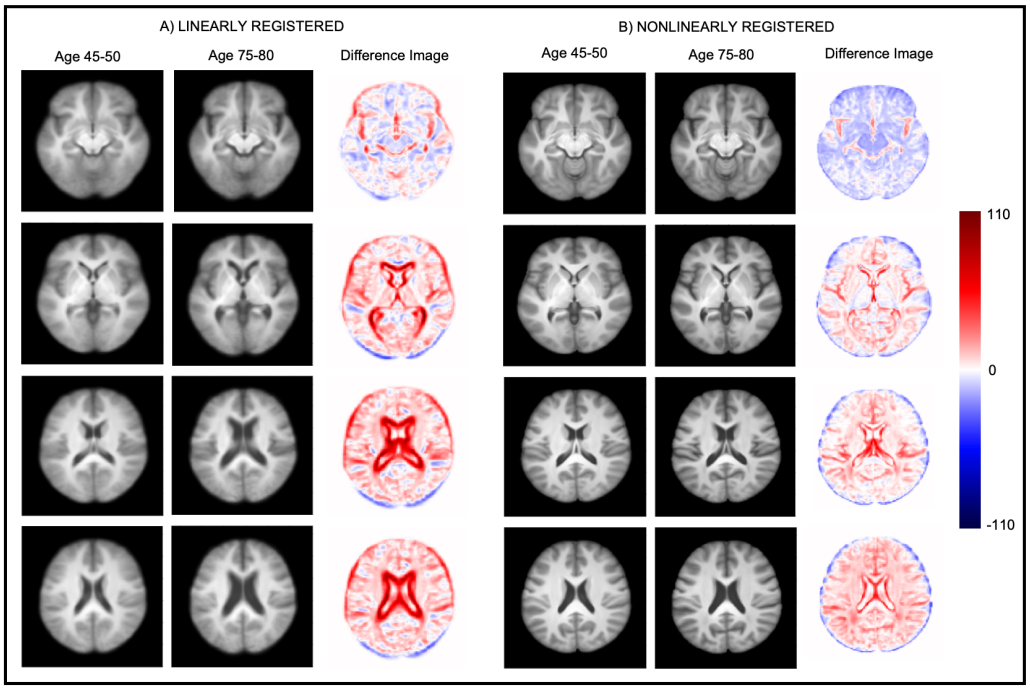


Figure 4.6: The mean of all the T1-weighted MRI of the subjects predicted by the CNN to be 45-50 and 75-80 in the female subject group: average of linearly registered images in (A) and non-linearly registered images in (B). The male subjects showed similar changes. The four slices shown are sampled from the inferior (top row) to superior (bottom row) regions of the brain.

composite images (Figure 4.6B). This process was repeated for the linearly registered images in Figure 4.6A. Although the largest difference can be seen around the ventricles for both types of registration, the differences between the 75-80 and 45-50 age groups on the averaged non-linearly registered images are subtle (Fig. 4.6B); however, the composite image of the linearly registered images of those same subjects (Figure 4.6A) shows very obvious differences between the age groups. As in Fig. 4.5, the dramatic differences are most evident in the linearly registered images and suppressed by the registration process when non-linear registration is used.

Despite the visual loss of information, the network trained with non-linearly registered images was able to predict age, and so volumetric information is probably still encoded in the images. Although the non-linear registration has significantly warped

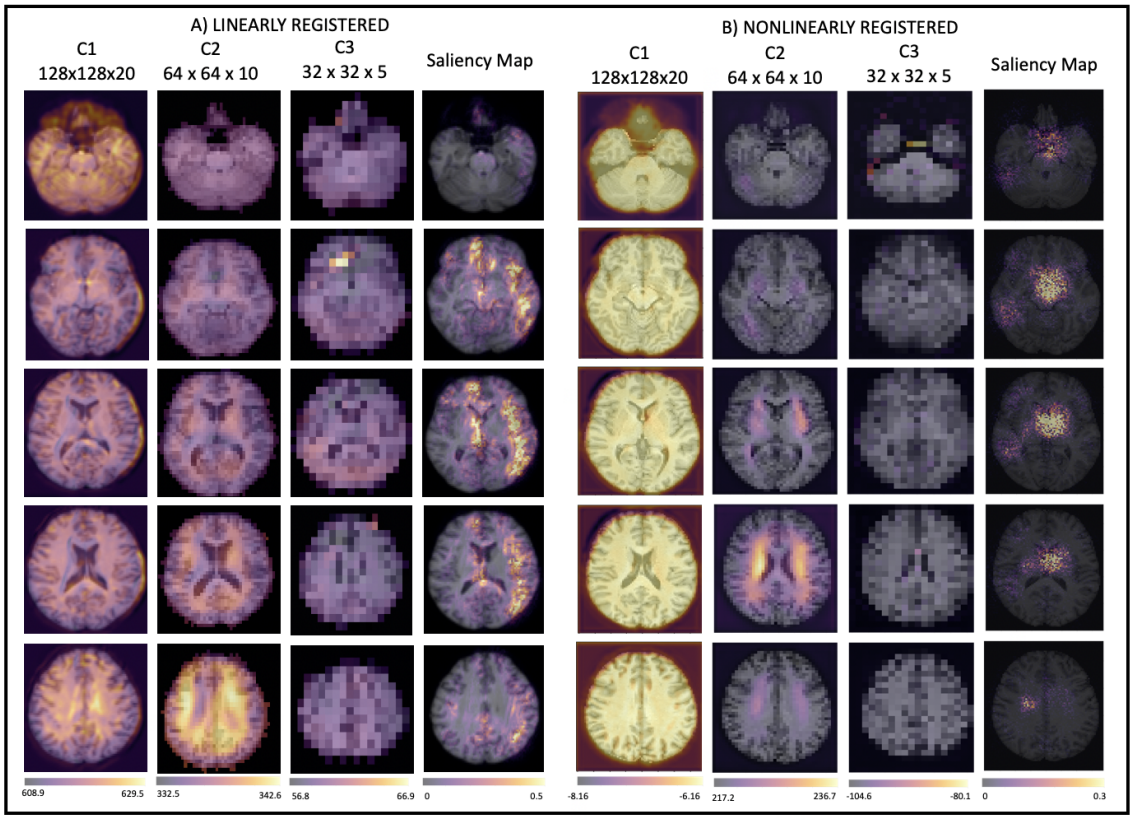


Figure 4.7: Attention and saliency maps for the two registration types, averaged across all test subjects (n=6885). The attention maps are from three resolutions as indicated by Fig. C.1. The saliency maps were produced from the network trained without attention gates.

the ventricles, it is still possible to see relatively clear differences in the caudate nucleus, even in the non-linearly registered composite (Fig. 4.6B). It is possible that this is true for other structures affected by ageing and therefore used by the CNN as features.

As a means for further investigation, attention gates [276] were placed before the first, second, and third max pooling layers, fed with the final layer output, training with both the non-linearly and linearly registered images. Attention gates enhance regions of the image which are useful for the task and suppress the other regions. Fig. 4.7 shows the output of the attention gates. For the non-linear case, it indicates that the network is being driven by the areas where the non-linear registration has had to

expand the brain tissue and contract the CSF in the ventricles to match the template (C2). The first gate at the highest resolution locates the brain, and the third gate found few regions of interest at that resolution. More interestingly, C2 was looking at the region around the ventricles, which most probably contains information about how much warping was required to register the ventricles to the MNI template. Conversely, it can be seen that when training with the linearly registered images, the network considers the brain more as a whole. This would indicate that the network trained on linearly registered images would be more likely to notice subtle changes located away from the ventricles, and it is possible that the network trained on non-linearly registered images might miss cortical changes, which would limit the utility of the network in detecting the early stages of neurological disease. The network used for this can be seen in Fig. C.1 in appendix C.1.2. For confirmation and comparison, saliency maps [150] were produced from the network trained without attention gates. In agreement with the attention networks, the saliency maps also indicate that with the non-linearly registered images, the network focuses only on the area around the ventricles, whereas with the linearly registered images, far more of the image influences the predictions.

Finally, before exploring the correlations with the model, the effect of downsampling the input image on the model predictions was explored. Model ensembles were trained with no downsampling and with downsampling factors of 2, 4 and 8; the results can be seen in Tab. 4.5. It can be seen that the choice to downsample by a factor of 4 did not have a negative impact on the performance of the ensemble, with no significant difference between the predictions with no downsampling, factor 2 and factor 4.

Downsampling Factor	MAE	MSE	R
Full	3.03 ± 2.22	14.09 ± 21.32	0.864
2	3.04 ± 2.19	14.04 ± 21.02	0.870
4	2.97 ± 2.19	13.64 ± 18.99	0.867
8	3.32 ± 2.40	16.74 ± 23.5	0.834

Table 4.5: Effect of downsampling on the brain age prediction results.

4.3.2 Relationships with UK Biobank Measurements

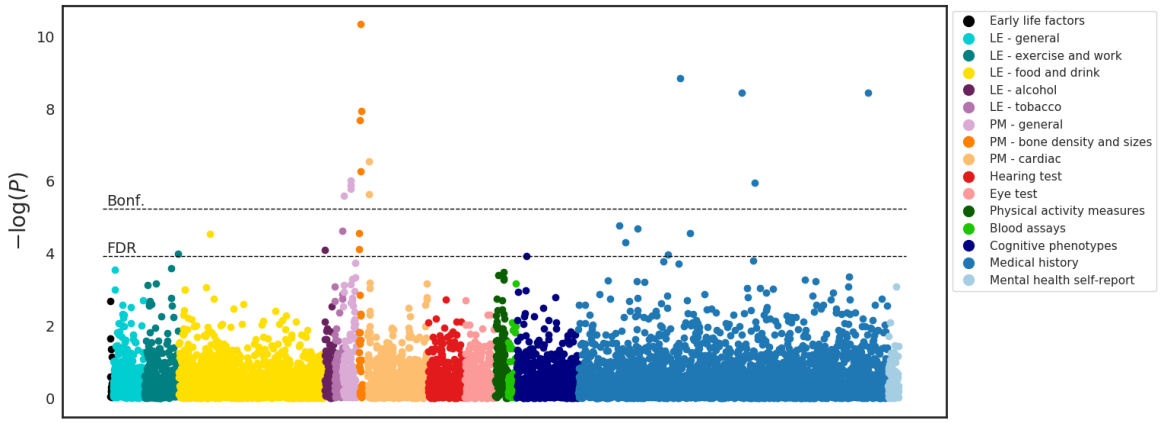
4.3.2.1 UK Biobank Variables

Having created the model of ageing, the relationship between the $\Delta_{BrainAge}$ values and the UK Biobank measurements was to be explored. Here, the results with the linearly registered data will be presented because, although the model was less accurate with the age predictions, the previous exploration suggests that the model is being driven in a way which is more biologically meaningful and better follows our understanding of ageing.

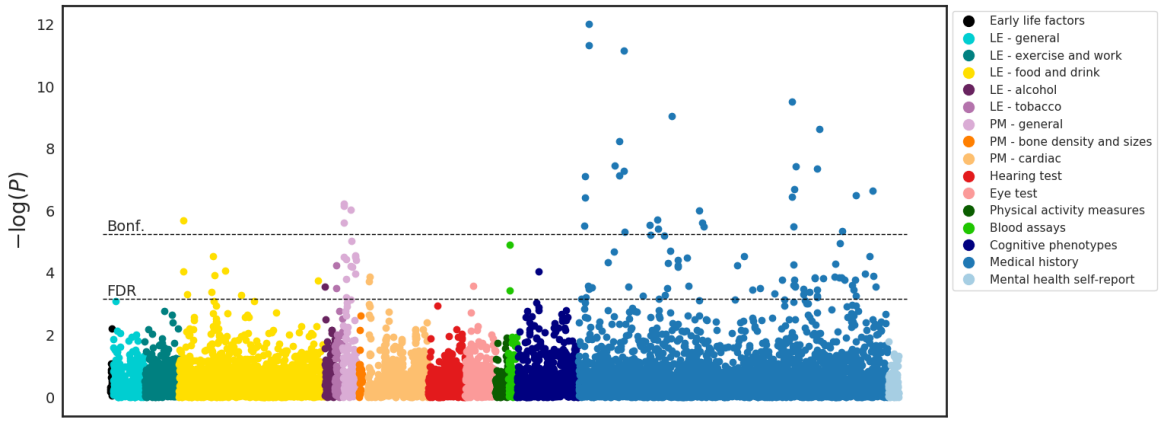
The relationship between the deltas and 8787 non-image derived variables from the UK Biobank are plotted in Figure 4.8. For the female group, using Δs from 3654 subjects, 14 variables passed the Bonferroni threshold including 4 variables from the Medical History, 2 from Physical Measurements (Cardiac), 4 from Physical Measurements (Bone Density) and 4 from Physical Measurements (General). Similarly, for the male group, using Δs from 3231 subjects, 33 passed the Bonferroni threshold including 28 variables from the Medical History, 1 from Early Life Style (Food and Drink) and 4 from Physical Measurements (General).

4.3.2.2 Image Derived Phenotypes

Using the linearly registered images as input, 596 and 622 correlations passed the Bonferroni threshold in the female and male groups respectively. It must, however, be noted that the threshold for the T1-related IDPs is not strictly valid, as the variables

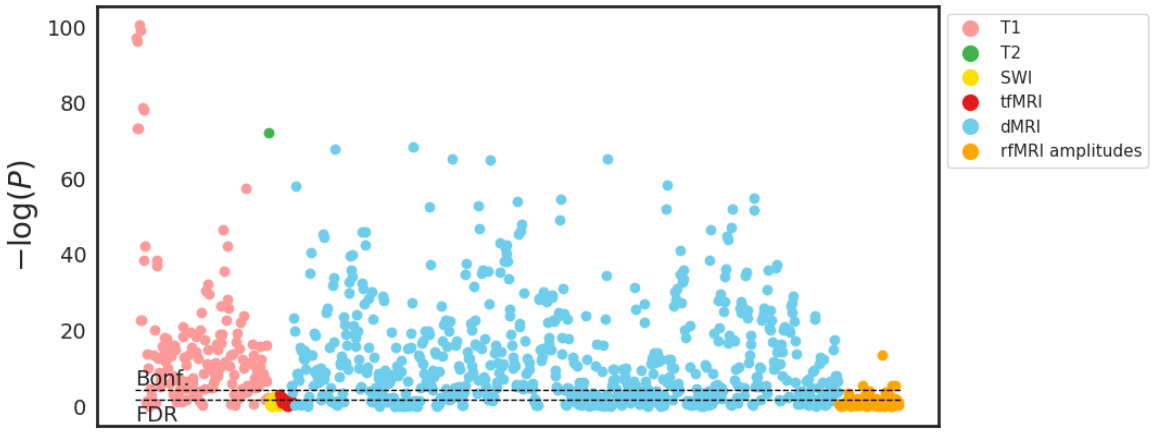


(a) Female Correlations (n=3654)

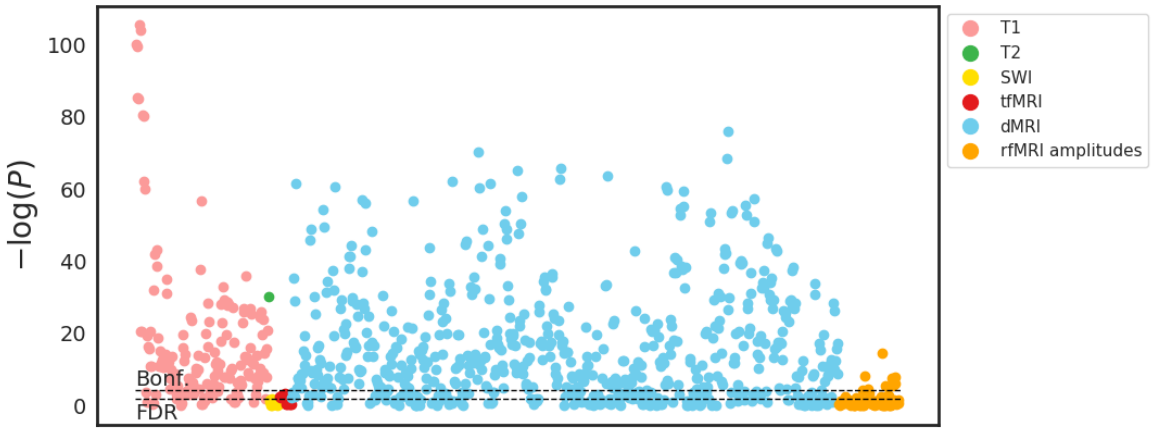


(b) Male Correlations (n=3231)

Figure 4.8: Manhattan plot relating each of the non-imaging UK Biobank variables with the delta measurements from the CNN (with linearly registered images as input) on female and male subjects. Each dot represents the statistical significance of the correlation between one UK Biobank variable and the deltas. A total of 8787 variables were tested, of which 14 and 33 passed the Bonferroni threshold for the female and male groups respectively.



(a) Female Correlations (n=3654)



(b) Male Correlations (n=3231)

Figure 4.9: Manhattan plot relating each of the IDPs with the delta measurements from the CNN (with linearly registered images as input) on female and male subjects. Each dot represents the statistical significance of the correlation between one IDP type and the deltas. A total of 941 were tested, of which 596 and 622 passed the Bonferroni threshold for the female and male groups respectively.

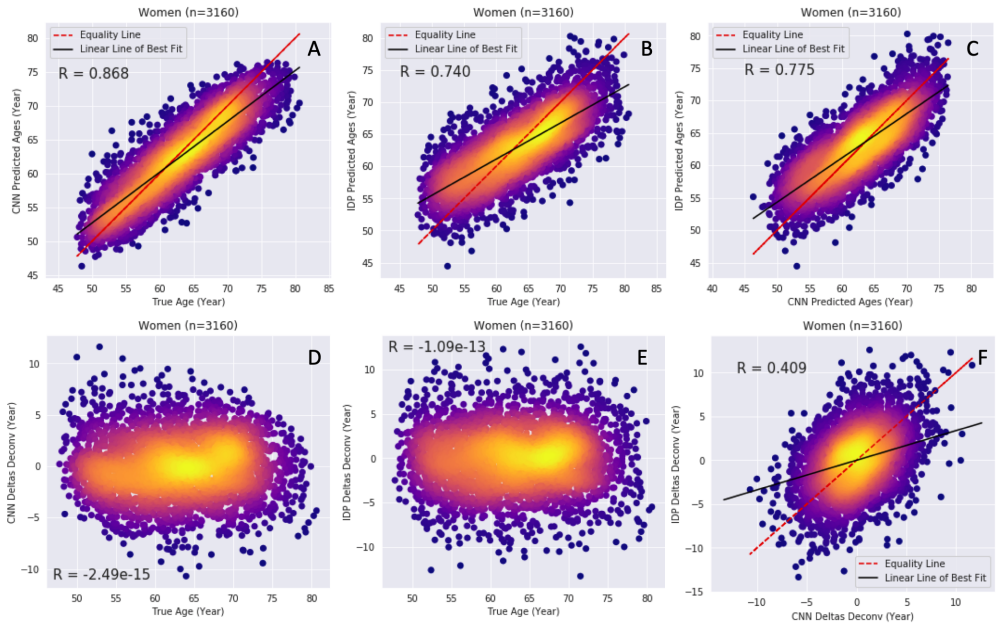
and CNN predictions were derived from the same source and thus are not independent. The relationship between the $\Delta_{BrainAge}$ and the 941 IDPs can be seen in Figure 4.9. To examine these more closely, the anatomical location of each Tract-Based Spatial Statistic (TBSS) IDP is illustrated in Figure 4.10. These TBSS IDPs are derived from white matter tracts and quantities calculated from diffusion MRI: fractional anisotropy (FA; the directional variation of water molecule diffusion), mean diffusivity (MD; the average magnitude of water molecule diffusion), intra-cellular volume fraction (ICVF; showing neurite density), isotropic volume fraction (ISOVF; showing extracellular water diffusion) and tract complexity/fanning (OD)[277].

To confirm the utility of the network, the network was also trained only on healthy subjects by two definitions: the first, the removal of conditions known to impact on the brain, and the second, the removal of these subjects plus those with conditions found to correlate with $\Delta_{BrainAge}$ – diabetes and hypertension – that had not already been removed. These were also found to correlate with $\Delta_{BrainAge}$ in [266]. Further details of conditions removed and the results can be seen in appendix C.1.3. The results were consistent across the experiments, showing that the results are not driven by subjects with known conditions.

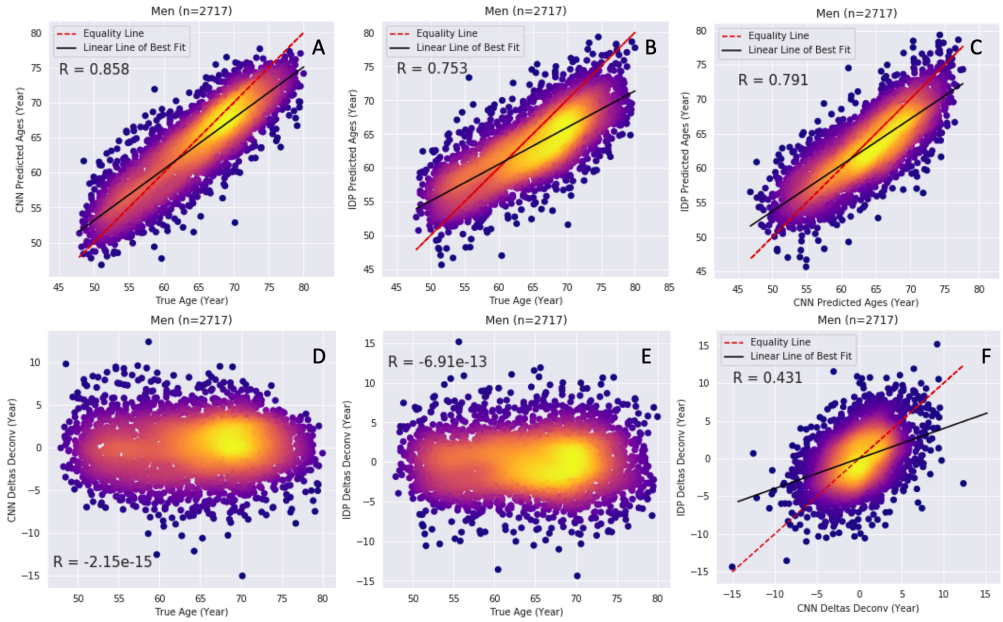
4.3.3 Relationships with IDP-based Age Prediction

Illustrated in Figure 4.11, the CNN-derived age predictions (using linearly registered images) and IDP-derived age predictions were strongly correlated (female r=0.775, male r=0.791). The deltas from each model were also moderately correlated after being deconfounded for age (female r=0.409, male r=0.431). The CNN-derived age predictions using non-linearly registered images (not shown) resulted in very similar correlations (female r=0.740, male r=0.799; deconfounded deltas: female r=0.399,

Figure 4.10: A chart of each IDP passing the Bonferroni threshold in Figure 4.9 and the corresponding correlation coefficients. The white matter tract associated with each listed IDP is displayed, based on the JHU white-matter tractography atlas [277] on top of the FMRIB58-FA standard-space FA template [3]. Acronyms: fractional anisotropy (FA), mean diffusivity (MD) intra-cellular volume fraction (ICVF); isotropic volume fraction (ISOVF), tract complexity/fanning (OD); L1, L2, and L3 are the three eigenvalues of the diffusion tensor.



(a) Female Correlations (n=3212)



(b) Male Correlations (n=2717)

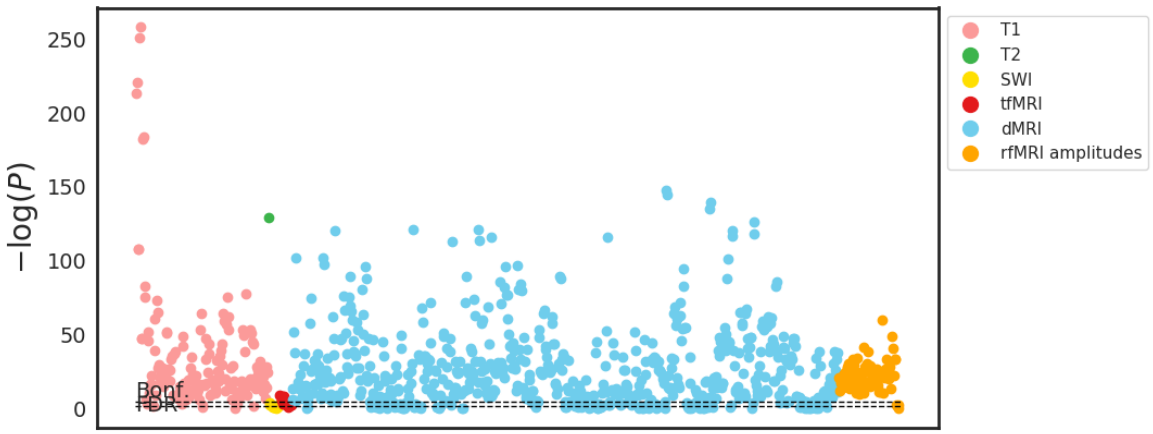
Figure 4.11: Density plots of the predicted ages from the CNN model (A) and IDP-based model (B) versus true age. (C) shows the relationship between the predicted ages of the CNN model and the IDP-based model. The deltas (deconfounded for age) from the CNN model and IDP-based model are plotted against true age in (D) and (E), respectively. (F) shows the relationship between the deltas from each model. The results are shown separately for each sex group. These results are based on the results from the CNN, trained on linearly registered images.

male r=0.446). Although high correlations were found, the results from the two methods were not the same, as discussed below. The correlations with the IDPs and non-imaging variables can be seen in Figures 4.12 and 4.13. The Elastic Net achieved an MAE of 3.87 ± 2.88 on the female group and 4.15 ± 2.99 on the male group compared to the CNN achieving 2.89 ± 2.17 on the female group and 2.71 ± 2.09 on the male group on the same subset of subjects.

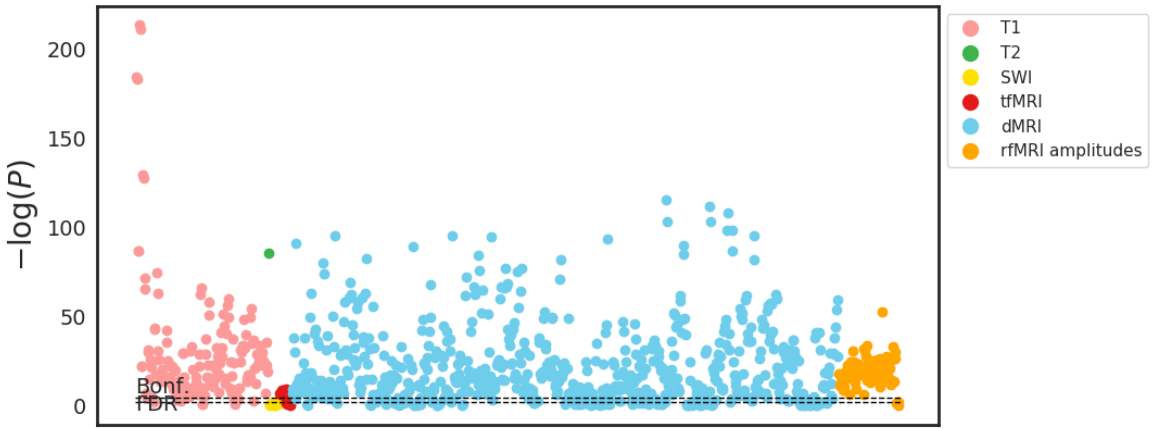
There were biological variables with significant correlations in common between the IDP-based method [266] and the CNN predictions, alongside the clear difference between the two groups. For women, notable common associations were bone density (density, ultrasound attention and mineral density), blood pressure, height, weight, and the presence of MS. For men, the strongest correlations in common were diabetes, hypertension, weight and height. The presence of bone density being a highly correlating factor for women, but not men, follows many reported results from the literature [278, 279]. This was also true when the model was only trained on healthy subjects and tested on all subjects. The common associations between the CNN method and the IDP-based method also remained constant.

4.4 Discussion

In this work, a CNN was used to predict brain age using $14,226$ subjects from the UK Biobank as training and a further 6885 subjects for testing. The network performed well on both sex groups, using either the linearly registered or non-linearly warped images as input. Using the linearly registered image dataset as input, the 3D ensemble network achieved a MAE – for an age range of 44.6 to 80.6 years – of 2.86 ± 2.22 and 3.09 ± 2.37 on the female and male groups respectively ($r = 0.870$ and 0.860). On the non-linearly warped data the network achieved a MAE of 2.71 ± 2.10 and 2.91 ± 2.18

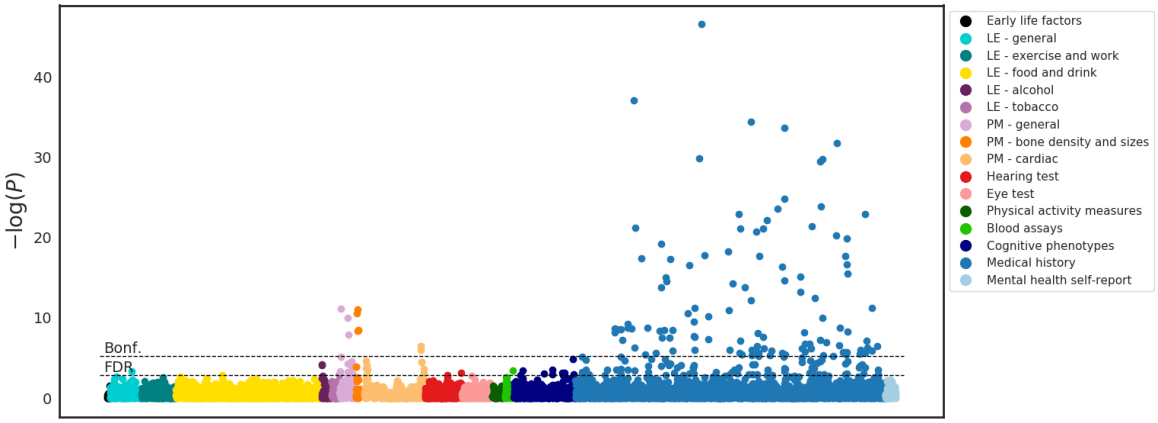


(a) Female Correlations (n=3212)

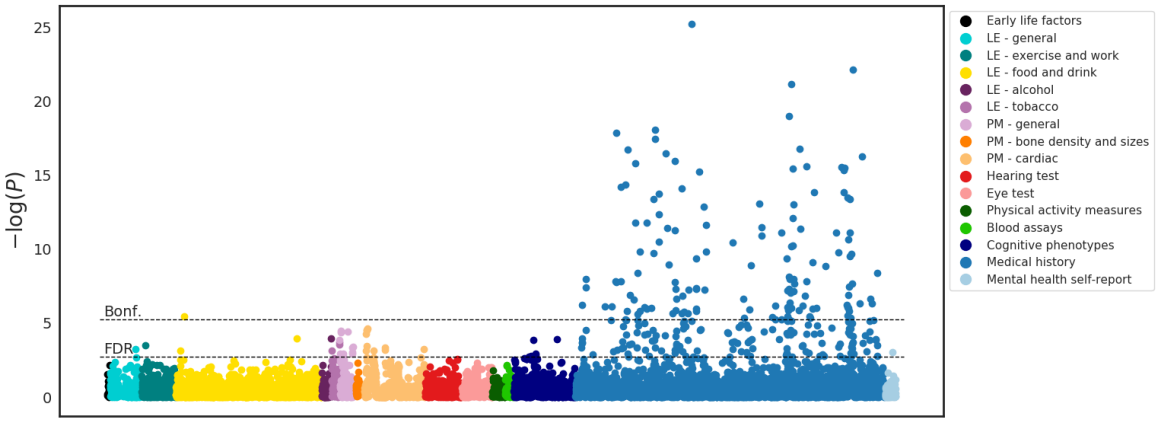


(b) Male Correlations (n=2717)

Figure 4.12: Manhattan plot relating each of the IDPs with the delta measurements from the CNN (with linearly registered images as input) on female and male subjects for the IDP-based method. Each dot represents the statistical significance of the correlation between one IDP type and the deltas. A total of 941 were tested, of which 757 and 751 passed the Bonferroni threshold for the female and male groups, respectively.



(a) Female Correlations (n=3212)



(b) Male Correlations (n=2717)

Figure 4.13: Manhattan plot relating each of the non-image derived variables with the delta measurements from the CNN (with linearly registered images as input) on female and male subjects for the IDP-based method. Each dot represents the statistical significance of the correlation between one variable type and the deltas. A total of 8787 were tested, of which 120 and 137 passed the Bonferroni threshold for the female and male groups, respectively.

on the female and male groups respectively ($r = 0.887$ and 0.882). Both of these results are competitive with the MAE range of the 2.5 - 5 years achieved for other methods reported in the literature. Although the MAE and MSE scores are lower than most in the current literature, it should be observed that achieving the lowest MAE on the UK Biobank dataset would not yield the most clinically relevant model since not all of the subjects are healthy. Subjects 1 and 2 in Fig. 4.5 have the same true age but very different levels of atrophy and predicted ages.

The clinical utility of the model could potentially be improved by including information from more imaging modalities. Combining modalities would potentially be better at capturing different aspects of the ageing brain such as brain atrophy, iron deposition, and alterations in white matter microstructure [212, 280] and so improve the model. It is important to note that the ageing process presents in multiple ways: mineralization, microstructure damage, anisotropy loss, and iron deposition [212], and that a T1-weighted MRI is unable to capture all of this information. Despite this, the CNN trained on only T1-weighted images outperformed the multi-modality IDP-based model. Furthermore, the network was trained using only the T1-weighted MRI and so independent correlations with the IDPs from the other modalities can be computed. Our model only showed mild correlation with the IDP-based model; however, this is not surprising as the IDP-based method is a linear model based on IDPs (i.e., a reduction of voxelwise image data to summary imaging-derived phenotypes) from multiple modalities, whereas our model does not have explicitly defined spatial priors or target structures and, rather, searches the whole image space during the training process. The degree of correlation does indicate a broad agreement between the two methods and the two methods together provide complementary analyses of the data. An interesting further line of investigation could be to explore whether these differences are purely random, driven by noise and non-biological factors, or whether interesting structure

exists in the differences, related to different modalities and degrees of non-linearity.

It is important to consider not only the modality of the input image but the pre-processing applied: in this work, using the non-linearly registered image as input performed as well as using the linearly registered images (p values from paired t-test between the results from the linearly registered images and non-linearly registered images: female $p = 0.710$, male $p = 0.810$). This seems surprising because the step of non-linear warping essentially suppresses differences by making the brain conform to the standard template. Whilst certain details and pathologies are still preserved, it is evident in Figs. 4.5 C and D that the ventricles in the non-linearly registered slices do not appear as dramatically different as they do in the linearly registered slices. Despite the visual loss of information, the network trained with non-linearly registered images was able to predict age, and so volumetric information is probably still encoded in the images. Although the non-linear registration has significantly warped the ventricles, it is still possible to see relatively clear differences in the caudate nucleus, even in the non-linearly registered composite (Fig. 4.6B). It is possible that this is true for other structures affected by ageing and therefore used by the CNN as features. Exploration with saliency maps and attention gates clearly show that the model is able to use the artefacts of the registration process, especially around the ventricles to make model predictions with the non-linear registration and does not consider information in other locations. This may indicate that the model trained with the non-linearly registered data will be missing some indicators of ageing, limiting the utility of the model. With the linearly registered data, the model clearly considered a far greater proportion of the image.

Interestingly, although our network was trained using only T1 images, the deltas showed significant correlations with 596 and 622 IDPs (in the female and male groups, respectively) from dMRI, T1, and T2 FLAIR images (Figs. 8 and 9). After

viewing the anatomical locations of the Tract-Based Spatial Statistics and Probabilistic Tractography IDPs that correlated with the deltas, it was evident that the correlations were distributed throughout the brain and were not particularly localized to a certain region. Overall, one very consistent finding was that intracellular volume fraction (ICVF) showed some of the strongest correlations with the deltas, and the association was always negative. ICVF is related to neuronal density, which means that a higher predicted age than true age (i.e. positive delta) correlated with decreased neuronal density. In addition, IDPs relating to T1 grey matter and white matter volumes correlated negatively with delta, while CSF and the volume of white matter hyperintensities (WMH) correlated positively. This would be expected, since CSF volume in the brain will either increase with age, disease, or loss of other brain tissue, and there is a well-known association between WMH and ageing [281]. The CNN-derived deltas and IDP-derived deltas correlated with mostly similar categories of the UK Biobank variables. In the female subject group, both methods showed correlations passing the Bonferroni threshold with UK Biobank medical measurements from the Physical General, Bone Density and Size, Medical History and Cardiac Physical Measurements categories, as defined by [268]. For the male group, both methods correlated with Medical History and Early Lifestyle (Food and Drink). Only the CNN-based model correlated with the Physical General Measurements.

Overall, the variables that correlate with deltas from the CNN are consistent with current knowledge about ageing, and encourage further exploration into the use of brain age as an indicator of healthy or unhealthy ageing. Furthermore, the associations were consistent between training on all subjects or just training on healthy subjects.

The presented method has used one of the greatest number of subjects for CNN-based age prediction, producing accurate age predictions, the errors from which have

been to shown to correlate with non-imaging variables, suggesting their biological meaningfulness. Already, these findings support the idea that deep learning methods have the potential to generate clinically relevant measurements directly from the medical images themselves. It might be that the errors, or age prediction deltas being measured, are useful health predictors, but most subjects do not have diagnosed illnesses. Fortunately, due to the longitudinal aspect of the UK Biobank study, in the future we will have the ability to explore whether the deltas from models such as this were predictive of any health outcomes.

5

Deep Learning-Based Unlearning of Dataset Bias for MRI Harmonisation and Confound Removal

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

- **Nicola K Dinsdale**, Mark Jenkinson, Ana IL Namburete: *Deep Learning-Based Unlearning of Dataset Bias for MRI Harmonisation and Confound Removal* in Neuroimage, Volume 228, March 2021.

The code used in this chapter is available at:

`github.com/nkdinsdale/Unlearning_for_MRI_harmonisation`.

5.1 Introduction

Although some large-scale neuroimaging projects now exist, for instance the UK Biobank [1], the majority of neuroimaging studies remain small, and many of the larger studies are highly multisite [282], such as ABIDE [98] and ADNI [59]. Therefore, the ability to combine datasets across scanners and protocols is vital to achieve higher statistical power, especially when studying neurological conditions where limited data is available. However, pooling data across scanners and sites leads to an undesirable increase in nonbiological variance [283]. This is the case even when attempts have been made to harmonise acquisition protocols and to use identical phantoms across imaging sites [100, 284–289].

Many studies have confirmed this variation caused by scanner and acquisition differences. In [101] the effect of scanner manufacturer, scanner upgrade, and scanner strength is explored, and it is demonstrated that they have a large impact on the results, alongside the parameters employed in the data processing stages and the effect of gradient non-linearities shown in [104]. The effect of multiple scanners is also explored in [102], where it is shown that, although the inclusion of scanner as a covariate in the analysis can suppress scanner effects, it makes biological differences less likely to be detected. In [103, 290] it is shown that not only can the use of different scanners disguise morphological changes, but also that the use of scanners at different time points, due to scanner drift, can have the same effect, causing difficulties for longitudinal studies. Furthermore, in [291] it is shown that, even when two scanners are nominally identical, consistent inter-site bias exists and in [292] it is shown that the choice of acquisition protocol can significantly bias the group measures of values derived from diffusion MRI. Several works have confirmed that this variability can be identified, by applying the *'Name the Dataset'* game [293] to neuroimaging datasets [294, 295] and

even when effort is taken to harmonise the acquisitions, scans can be reliably clustered by scanner for structural, resting state fMRI [284] and diffusion data [285].

The removal of this scanner-induced variance is therefore vital for many neuroimaging studies. Many existing methods are based on ComBat [296], an empirical Bayes method originally developed to remove 'batch effects' in genetics, which has been applied for harmonising values derived from structural [295, 297, 298], diffusion [299] and functional MRI [100], successfully removing non-biological variance whilst preserving biological associations. ComBat uses multivariate linear mixed effects regression to account for both the biological variables and the scanner, to allow the modelling of the imaging features. For each site, a site-specific scaling factor δ is calculated, yielding a model that adjusts for additive and multiplicative effects. Furthermore, ComBat uses empirical Bayes to learn the model parameters, which assumes that model parameters across features are drawn from the same distribution; this improves the estimation of the parameters where only small sample sizes are available. It has been demonstrated that the use of ComBat to harmonise the data leads to an increase in statistical power compared to random-effects meta-analyses [300]. ComBat has been further developed to include a term explicitly to model for variables of interest to preserve after harmonisation [295], model covariances [301], the incorporation of a generalized additive model (GAM) into the model, extending it to include non-linear variations [298], and longitudinal studies [302]. ComBat, however, is usually applied to the harmonisation of image-derived values and defined associations. It has also been shown to be sensitive to even minor changes in the data pre-processing pipeline [303]. Random forests have also been used to approach the problem similarly, with the approach demonstrated for harmonising structural ROIs [304]. Other models also approach the harmonisation task through finding scaling factors for the raw image values, for instance MICA [305], which estimates non-linear transformations of the voxel intensity values.

Harmonisation has been explored more for diffusion MRI than other modalities [105, 107, 299, 306–312] with many methods using spherical harmonics to harmonise the data [107, 309–312] which greatly limits the ability of these methods to be applied to other modalities. Moyer et al [107] adopt variational autoencoders to produce scanner-invariant representations of the data for diffusion MRI, and [117] also create scanner invariant representations through combining domain adaptation and attention gates for classification. These feature representations can then be used to reconstruct the input images so that they are minimally informative of their original collection site. This work shows that using deep learning techniques to create feature representations that are invariant to scanners presents a strong candidate for MRI data harmonisation.

In addition to the variational autoencoders used in [107], other generative models have been used to harmonise MRI, largely based on deep learning models including UNet [41] based models and cycleGAN [106, 313] based models [314]. These are limited by needing either paired or 'travelling heads' data for training for each site, which is expensive and infeasible to acquire in large numbers, but the methods are hard to evaluate without the paired data [107]. With generative methods, it is also very difficult to validate the generated 'harmonised' images and so there is a risk of unknown errors propagating through pipelines and affecting the results of any completed analysis; furthermore, there is very little exploration of this in the literature. In addition, these methods are data hungry and difficult to train, which leads to many of these methods being implemented in 2D [106] or on a patchwise basis [107]. Such implementations prevent the CNN learning context from adjacent slices or patches and can lead to errors in the reconstructed images, so that optimal performance is usually only achieved when the whole image can be scanned in a single forward pass [85]. Finally, some approaches also try to disentangle the information into scanner information (intensities) and structural information [315, 316]. These decompositions cannot, however, account

for scanner effects which disguise morphological changes.

The key measure of success for harmonisation methods is to be discriminative for the biological variable of interest whilst being indistinguishable with respect to the scanner used to acquire the data. Following the framework introduced by Ben-David et al [110], domain adaptation routinely concerns the scenario where there is access to a source domain with a large set of labels, and a target domain with either no labels or a low number of labels. For this work, only the case where the task to be performed is identical across domains is considered. It would be expected that these two domains would be related but not identical: for instance, the source and target domains could indicate different scanners or different acquisition protocols. The difference between the two domains means that a neural network trained on the source data would have a reduced performance on the target domain [317]; this degree of difference is known as *domain shift* [318]. The smaller the degree of the domain shift, the more likely the domain adaptation is to be successful. Domain adaptation techniques, therefore, attempt to find a feature space that performs a given main task whilst being invariant to the domain of the data. This is achieved by learning features using the source domain labels, but these features are transformed to a representation that should generalise to the target domain by simultaneously adapting the features to have the same distribution for both datasets. Therefore, domain adaptation should be applicable to the task of MRI data harmonisation, creating features that are indiscriminate with respect to scanner, but discriminate with respect to the task of interest.

There have been many methods proposed for domain adaptation, the simplest of which use a divergence-based approach that tries to minimise some divergence criterion between feature distributions for the source and target data distributions, which should provide a domain-invariant representation. This relies on a feature

representation existing, where the classifier is able to perform approximately equally on both tasks. Frequently used divergence measures include maximum mean discrepancy [319], that compares the means of two samples to determine if they are from the same distribution, and correlation alignment [320] that tries to align second-order statistics of the two distributions using a linear transformation. These methods are limited by the assumption that all scanner information can be removed by satisfying a simply definable condition, which is unlikely when dealing with a highly non-linear system such as MRI.

The similarity between domains is generalised further by using adversarial methods, where the features for each domain are forced to be identical for the family of possible non-linear functions that can be produced by a given convolutional neural network. Adversarial domain adaptation was formalised with *Domain Adversarial Neural Networks* (DANNs) [111]. Based on the premise that for effective domain transfer to be achieved, predictions must be based on features that cannot discriminate between domains, they propose a generic framework that jointly optimises the underlying features as well as two discriminative classifiers that operate on these features. The first is a label predictor that predicts the class labels, and the second is a domain classifier that aims to predict the source of the data. The overall aim of the network is then to minimise the loss of the label predictor and to maximise the loss of the domain classifier, such that the learnt features are discriminative for the main task but are completely unable to discriminate between domains. They show that this behaviour can be achieved for feed-forward neural networks simply through the addition of a domain classifier formed of standard convolutional layers and a gradient reversal layer [321]. The gradient reversal layer allows the maximisation of the loss of the domain classifier whilst minimising the loss for the main task. Placed between the feature extractor and domain classifier, the gradient reversal layer acts as an identity function in the forward step, and during the backward pass, it multiplies the gradient function

by $-\lambda$, where λ is a hyperparameter set empirically for a given experimental setup [111]. The extension of domain adaptation to N source domains – key to enable the harmonisation of more than two scanners – was formalised in [110] and demonstrated for adversarial domain adaptation in [314].

Another method for adversarial domain adaptation was introduced in [115] where, rather than using a gradient reversal layer to update the domain classifier in opposition to the task, an iterative training scheme was used. By alternating between learning the best domain classifier for a given feature representation, and then minimising a confusion loss which aims to force the domain predictions to be closer to a uniform distribution, they obtain a domain classifier that is maximally confused [115]. Compared to DANN-style unlearning networks, it is better at ensuring a domain classifier that is equally uninformative across the domains [322] because the confusion loss tries to force all softmax output values to be equal, which is vital for successful data harmonisation, especially as the framework is extended to a larger number of source scanners. This work was then applied in [322] to facial recognition, not for domain adaptation, but for the explicit removal of identified sources of bias in their data, showing that the framework can be extended to 'unlearn' multiple factors simultaneously.

The contributions in this chapter, therefore, are as follows:

- It will be demonstrated that the adversarial framework proposed in [115] can be adapted for use in harmonisation for deep learning tasks. By considering harmonisation to be a multiple source, joint domain adaptation task, it will be shown that shared feature representations can be produced that are invariant to the acquisition scanner while still completing the main task of interest across scanners and acquisition protocols, with negligible performance compromise.

- The framework will be demonstrated for different feed-forward architectures and segmentation tasks – age prediction with a VGG style classification network [22] and tissue segmentation with a UNet style network [41] and, as the framework is applicable to all feed-forward networks and tasks, it should be applicable to the harmonisation of data across a wide variety of neuroimaging studies.

- A range of likely data scenarios will be considered, such as the effect of having limited amounts of training data and the effect of having different distributions of data for different scanners. It will be shown that the training framework is simply adapted to deal with these additional challenges and that it can be expanded to deal with highly multisite datasets.

- Finally, the framework will be extended to remove other additional confounds in addition to harmonising for acquisition scanner, increasing the flexibility for the application of the framework.

5.2 Methods

5.2.1 Data Inputs and Outputs

Before explaining the methods, the notation used throughout will be explained. The network takes as input a set of MRI scans, $\boldsymbol{X}$, and outputs the predicted labels $\hat{\boldsymbol{y}}$, which can take the form of a vector of scalar values (for age regression) or class probability maps (for segmentation). There are also target labels $\boldsymbol{y}$ available for the task of interest (regression or segmentation), which will be made available to the network during training. These data come from N distinct scanners or acquisition protocols, and for each scanner there are S_n subjects, where $n \in \{1..N\}$. Note that images acquired on

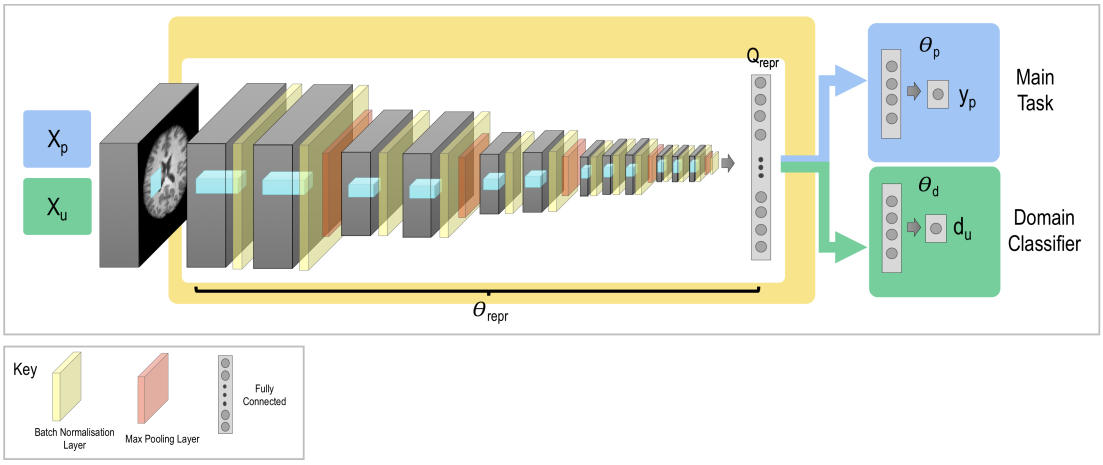


Figure 5.1: General network architecture. The network is formed of three sections: the feature extractor with parameters $\boldsymbol{\Theta}_{repr}$, the label predictor with parameters $\boldsymbol{\Theta}_p$, and the domain classifier with parameters $\boldsymbol{\Theta}_d$. $\boldsymbol{X}_p$ represents the input data used to train the main task with labels $\boldsymbol{y}_p$, and $\boldsymbol{X}_u$ represents the input data used to train the steps involved in unlearning scanner with labels $\boldsymbol{d}$.

the same *physical* scanner but using different acquisition parameters (eg. a change in T1 acquisition protocol) would be treated as if they were acquired on different scanners.

5.2.2 Network Architecture

The network was formed of three parts, illustrated in Fig. 5.1: a *feature extractor* with parameters $\boldsymbol{\Theta_{repr}}$, a *label predictor* with parameters $\boldsymbol{\Theta_p}$, and the *domain classifier* with parameters $\boldsymbol{\Theta_d}$. The first two jointly form the network needed to perform the main task: the feature extractor takes the input image $\boldsymbol{X}$ and outputs a fully connected layer. This fully connected layer is then passed to the label predictor and outputs the main task label, $\boldsymbol{y}$. The domain classifier is also added to the output of the feature extractor to allow us to adversarially remove the scanner information. It therefore takes in the fully connected layer from the feature extractor and outputs softmax values where p_n is the softmax value for the nth scanner. It is assumed that it is possible to obtain the domain label (e.g. scanner or acquisition protocol) for all available samples,

which will be true for nearly all cases. The main task labels, $\boldsymbol{y}$ (e.g. image annotations or subject age), however, may not be available for all samples.

5.2.3 Training Procedure

The training procedure consists of three stages:

1. Optimising the feature extractor and the label predictor for the main task;
2. Optimising the domain classifier to identify the scanner information remaining;
3. Optimising the feature extractor to confuse the domain predictor and remove scanner information.

For each stage, a different loss function is optimised, leading to three consecutive steps (or iterations) in each training batch. These work together to create a feature representation $\boldsymbol{Q}_{repr} = f(\boldsymbol{X}, \boldsymbol{\Theta}_{repr})$ – the activations at the final layer of the feature extractor – where the aim is to make this feature representation invariant to the scanner used for acquisition but discriminative for the main task of interest (e.g. segmentation or regression).

The first stage involves minimising the loss function pertaining to the primary task of interest. For instance, for segmentation it may take the form of the Dice loss, whereas for a classification task it may be categorical cross entropy. This loss function is evaluated separately for each scanner, or dataset acquired with a different protocol, such that the optimisation is not driven by the largest dataset:

$$L_p(\boldsymbol{X}_p, \boldsymbol{y}_p; \boldsymbol{\Theta}_{repr}, \boldsymbol{\Theta}_p) = \sum_{n=1}^{N} \frac{1}{S_n} \sum_{j=1}^{S_n} L_p(\boldsymbol{y}_{j,n}, \hat{\boldsymbol{y}}_{j,n}) \tag{5.1}$$

where $\boldsymbol{X}_p \subset \boldsymbol{X}$ and p selects the subset of data for which main task labels are available, $\boldsymbol{y}_p$. L_p is the primary loss function being evaluated for the subjects acquired on scanner

$n \in \{1..N\}$ – S_n – and $\hat{\boldsymbol{y}}_{j,n}$ is the predicted label for the main task for the jth subject from the nth scanner. The notation $L_p(\boldsymbol{X}_p, \boldsymbol{y}_p; \boldsymbol{\Theta}_{repr}, \boldsymbol{\Theta}_p)$ means that the loss function is evaluated for the data $\{\boldsymbol{X}_p, \boldsymbol{y}_p\}$, which are fixed, given the current values of the parameters $\boldsymbol{\Theta}_{repr}$, and $\boldsymbol{\Theta}_p$ found by the optimisation process (i.e. quantities after the semicolon are ones that can be changed). It can therefore be seen that the optimizer controlling the training relating to this loss function is given access to the parameters in the feature extractor and the label predictor, as these are the values which are indicated to vary. The domain classifier is not involved in this stage.

The scanner information is then 'unlearned' (i.e. removed from the internal representation) using two loss functions in combination. The first is the domain loss that assesses how much scanner information remains in $\boldsymbol{Q}_{repr}$. It simply takes the form of categorical cross-entropy:

$$L_d(\boldsymbol{X}_u, \boldsymbol{d}_u, \boldsymbol{\Theta}_{repr}; \boldsymbol{\Theta}_d) = -\frac{1}{S_u} \sum_{s=1}^{S_u} \sum_{k=1}^{N} \mathbb{1}[d_s = k] log(p_{s,k}) \tag{5.2}$$

where $p_{s,k}$ is the softmax output of the domain classifier for the sth subject and k^{th} possible scanner domain, $\boldsymbol{X}_u \subset \boldsymbol{X}$ and $\boldsymbol{d}_u \subset \boldsymbol{d}$ where u indicates the subset of data to be used for unlearning, and the mean is calculated over all the subjects used for the domain unlearning $S_u = |\boldsymbol{X}_u|$. Note the negative sign, following the standard definition of categorical cross entropy, as the aim is to maximise the performance on the classification task so as to evaluate the scanner information remaining, and so it is necessary to minimise the categorical cross entropy. Although it is assumed that there is access to the domain labels for all of the data points, the full set is not necessarily used for unlearning, depending on the scenario. This will be explored in the later experiments.

For this stage, the optimiser has access only to the weights in the domain classifier and has no influence on $\boldsymbol{\Theta}_{repr}$, which is fixed, as indicated by the loss function

$L_d(\boldsymbol{X}_u, \boldsymbol{d}_u, \boldsymbol{\Theta}_{repr}; \boldsymbol{\Theta}_d)$. Therefore, the best domain classifier is found given the fixed feature representation and, hence, an indication of the amount of scanner information remaining.

The second loss function controlling the unlearning is the confusion loss. Minimising this loss function aims to tune the parameters in the feature extractor $\boldsymbol{\Theta}_{repr}$ such that all scanner information is removed from $\boldsymbol{Q}_{repr}$, thus making it scanner invariant. When this feature representation is entirely scanner invariant, even the best domain classifier will be unable to predict which scanner acquired the data, and so the softmax outputs of the domain classifier will all be equal, corresponding to random chance. The confusion loss therefore has the form:

$$L_{conf}(\boldsymbol{X}_u, \boldsymbol{d}_u, \boldsymbol{\Theta}_d; \boldsymbol{\Theta}_{repr}) = -\frac{1}{S_u} \sum_{s=1}^{S_u} \sum_{k=1}^{N} \frac{1}{N} log(p_{s,k}) \tag{5.3}$$

where $\boldsymbol{X}_u$ and $\boldsymbol{d}_u$ are the same subsets used in the previous loss function. This step only updates the parameters in $\boldsymbol{\Theta}_{repr}$ and depends on the fixed value of $\boldsymbol{\Theta}_d$ as indicated by $L_{conf}(\boldsymbol{X}_u, \boldsymbol{d}_u, \boldsymbol{\Theta}_d; \boldsymbol{\Theta}_{repr})$. Again, note the negative sign, meaning that the loss is minimised when the softmax outputs are all equal in value.

Stages 2 and 3 should be considered to be a unit and therefore the order in which they are updated is fixed. The confusion loss is most effective at removing crucial information to confuse the domain classifier once it has been tuned to find and utilise any domain information. Hence, it is best to update the domain classifier first, to enable it to learn the domain information, prior to using the confusion loss. To this end, equation (5.2) must be updated before equation (5.3).

Therefore, the overall method minimises the total loss function:

$$\begin{aligned} L(\boldsymbol{X}_p, \boldsymbol{X}_u, \boldsymbol{y}_p, \boldsymbol{d}_u, \boldsymbol{\Theta}_{repr}, \boldsymbol{\Theta}_p, \boldsymbol{\Theta}_d) = L_p(\boldsymbol{X}_p, \boldsymbol{y}_p; \boldsymbol{\Theta}_{repr}, \boldsymbol{\Theta}_p) \\ + \alpha L_d(\boldsymbol{X}_u, \boldsymbol{d}_u, \boldsymbol{\Theta}_{repr}; \boldsymbol{\Theta}_d) \\ + \beta L_{conf}(\boldsymbol{X}_u, \boldsymbol{d}_u, \boldsymbol{\Theta}_d; \boldsymbol{\Theta}_{repr}) \end{aligned} \tag{5.4}$$

where α and β represent the weights of the relative contributions of the different loss functions and recalling that L_d and L_{conf} contain negative signs enabling the adversarial training.

Equations (5.2) and (5.3) cannot be optimised in a single step because they act in direct opposition to each other, hence the iterative update scheme and three forward/backward passes per stage for each batch. This is not too computationally expensive as in each stage only a subset of the total parameters are updated.

The domain classifier can only be used to assess how well domain information is being removed if it is capable of accurately predicting the domain *prior* to unlearning. In order to ensure this, the network is pretrained using equations (1) and (2) until the primary task reaches convergence. Provided that the domain classifier is able to accurately predict the scanner label at this stage, optimising the confusion loss in equation (3) such that the domain classifier performs no better than random chance, corresponds to removing the scanner information in $\boldsymbol{Q}_{repr}$. In addition, pretraining will decrease the number of epochs required for unlearning. As epochs with unlearning steps are more computationally expensive, this serves to decrease the overall training time.

As shown by the loss functions, different sets of data can be used to evaluate the different loss functions. This not only means that scanner information can be unlearned for data samples for which the main task labels are not available, but also that scanner information can be unlearned using curated (or 'matched') subsets of

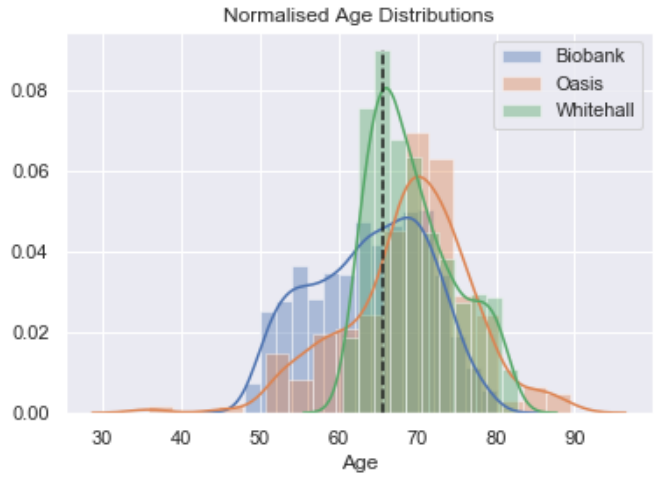


Figure 5.2: Normalised age distributions for the three datasets: UK Biobank, OASIS and Whitehall. The dashed line indicates the mean age of the three datasets.

data samples when potentially problematic biases exist in the data as a whole. These scenarios are explored in the experiments to follow.

5.2.4 Age Prediction Task

First, the task of brain age prediction [217, 225, 323] is considered as an example task to demonstrate the framework: $\boldsymbol{X}$ takes the form of the T1-weighted MRI images and $\boldsymbol{y}$ the true biological age values.

For these experiments, an architecture similar to the VGG-16 [22] network was used. A batch size of 32 was used throughout, with each batch constrained to have at least one example from each dataset. This increases stability during training but requires oversampling of the smaller datasets. The main task loss (equation (5.1)) was the Mean Square Error (MSE) for each scanner. All experiments were completed on a V100 GPU and were implemented in Python (3.6) and Pytorch (1.0.1) [324].

5.2.4.1 Datasets

Three datasets were used for these experiments: UK Biobank [1] (Siemens Skyra 3T) pre-processed using the UK Biobank Pipeline [269] (5508 training, 1377 testing); healthy subjects from the OASIS dataset [4] (Siemens Tesla Vision 1.5T) at multiple longitudinal time points, split into training and testing sets at the subject levels (813 training, 217 testing), and healthy subjects from the Whitehall II study [325] (Siemens Magnetom Verio 3T) (452 training, 51 testing). The model was trained using five-fold cross validation with the training data split as 80% training 20% validation; the results are reported for the hold out test set.

The input images for all the three datasets were resampled, taking account of the original voxel size, to $128 \times 128 \times 128$ voxels, each of size $1 \times 1 \times 1$mm using spline interpolation; then, every fourth slice was retained, leaving 32 slices in the z direction (axial slices). This then maintains the physical size of the object and makes datasets comparable in this way, at the price of some interpolation-related changes to the intensities. Note that such resampling was necessary in order to enable us to use all the datasets within the same model.

Every fourth slice was selected so as to maximise coverage across the whole brain whilst minimising redundancy and allowing us to use a larger batch size and number of filters – for validation of this approach see Table 4.5. The inputs were also normalised to have zero mean and unit standard deviation. The distributions of the data can be seen in Fig. 5.2. The p-values from t-tests on the dataset pairs show that there is only a significant difference between the distribution of the UK Biobank and Whitehall data - UK Biobank and OASIS p=0.22, OASIS and Whitehall p=0.09, UK Biobank and Whitehall p=0.001.

5.2.4.2 Basic Fully Supervised Learning

First, the simplest training scenario was considered, where main task training labels are available for all datasets being used and all the datasets have similar distributions for the main task label. This means that there is not a high degree of correlation between the age and the scanner and therefore it should be possible to remove scanner information from the feature representation without removing information that is discriminative for the age prediction task.

In this scenario, all three loss functions can be evaluated on a single combined dataset $\boldsymbol{X}$ where labels $\boldsymbol{y}$ for the full set are available and the acquisition scanner for all data points $\boldsymbol{d}$ known, meaning that the overall method minimises the total loss function:

$$\begin{aligned} L(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{d}, \boldsymbol{\Theta}_{repr}, \boldsymbol{\Theta}_p, \boldsymbol{\Theta}_d) = & L_p(\boldsymbol{X}, \boldsymbol{y}; \boldsymbol{\Theta}_{repr}, \boldsymbol{\Theta}_p) \\ & + \alpha L_d(\boldsymbol{X}, \boldsymbol{d}, \boldsymbol{\Theta}_{repr}; \boldsymbol{\Theta}_d) \\ & + \beta L_{conf}(\boldsymbol{X}, \boldsymbol{d}, \boldsymbol{\Theta}_d; \boldsymbol{\Theta}_{repr}) \end{aligned} \tag{5.5}$$

To allow comparison, the network is also trained with normal training – training only the feature extractor and label predictor – on the different possible combinations of datasets, and compared to all combinations of datasets. As a form of ablation study, standard training with the main task loss function evaluated as in equation (5.1) is also compared. These comparisons are vital because it is important that the harmonisation process does not significantly degrade the performance on the main task. No comparison with existing harmonisation methods such as ComBat is made because these methods approach harmonisation differently to the proposed method and so are not directly comparable.

5.2.4.3 Biased Domains

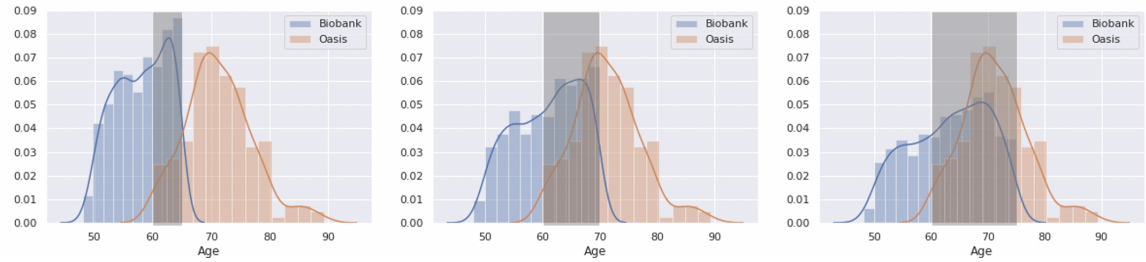


Figure 5.3: The three biased datasets used in this experiment with 5 years, 10 years, and 15 years overlap. Only the shaded overlap region is used for unlearning scanner information; all the data points are used to evaluate the main loss function.

When the distribution of the data for the main task is similar across all data points, unlearning can simply be completed on all of the data points. However, where there exists a large difference between the two domains, such that the main task label is highly indicative of the scanner, it is likely that the unlearning process will also remove information that is important for the main task. This latter scenario could, for instance, be where the age distributions for the two studies are only slightly overlapping or where nearly all subjects with a given condition are collected on one of the scanners.

To reduce this problem, the flexibility of the training framework is utilised, and, whilst evaluating the main task on the whole dataset, the scanner unlearning (equations (5.2) and (5.3)) was performed on only a subset of the data. For the case of age prediction, unlearning is performed only on the overlapping section of the data distributions. If the scenario being considered was where there was data from both subjects and healthy controls and, for instance, most of the subjects had been scanned on one of the two scanners, unlearning could be performed only on the healthy controls rather than the whole dataset. As the main task labels are not needed for unlearning, unlearning could be performed on a separate dataset so long as the scanner and protocol remained identical.

The UK Biobank and OASIS datasets are sub-sampled so as to create three degrees of overlapping datasets: 5 years, 10 years and 15 years (Fig 5.3) and tested on the same test sets as before, spanning the whole age range. Three training options are compared: i) training normally on the full combination of the biased datasets, ii) unlearning on the whole distribution, and iii) unlearning only on the overlapping section of the distributions.

5.2.4.4 Extension to ABIDE Data

Having demonstrated the network on three scanners, the network is then demonstrated on a multi-site dataset, ABIDE [98]. The data is split 90% training and 10% testing across each scanner and processed as described in Section 5.2.4.2. Therefore, there is data from 16 scanners, including Philips, Siemens and GE scanners, each with between 27 and 165 scans per site, for training. This is representative of many neuroimaging studies, with relatively low numbers of scans available for each site. Following the same fully supervised framework, as demonstrated in Section 5.2.4.2, training on the combination of data from all 16 sites is compared to the unlearning approach.

A batch size of 32 was used in training, with the batch constrained to have at least one example from each scanner. To achieve this, the smaller datasets were over-sampled such that the number of batches was limited by the largest dataset. This was performed because it was found that having an example of each dataset in each batch led to higher stability during training. The over-sampled data points were not augmented in order to be sure that improvement was not due to the augmentation, but, in practice, this would be a sensible step.

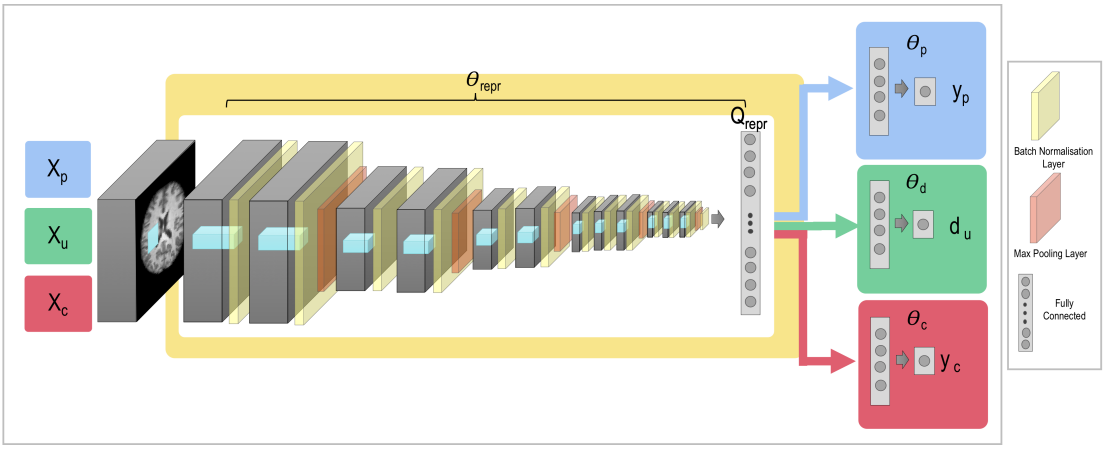


Figure 5.4: The general network architecture can be adapted to also remove other confounds from the data. In addition to the datasets used for the main task and the scanner unlearning, a data pair is added where $\boldsymbol{X}_c$ are the input images used for deconfounding and $\boldsymbol{y}_c$ are the confound labels.

5.2.4.5 Removal of Other Categorical Confounds

In addition to unlearning scanner information to harmonise the data, the framework can also be adapted explicitly to remove other confounds. As shown in Fig 5.4, an additional data pair is added to training and an additional two loss functions are added to the overall loss function per confound. For each confound to be removed, an additional *confound classification loss* is added, which, like the domain classification loss, identifies how much information relating to the confound remains in the feature space, and a *confound confusion loss*, which aims to modify the feature space so as to remove the confound information by penalising deviations from a uniform distribution for the softmax predictions for the given task.

Therefore, the overall method will minimise the loss:

$$
\begin{aligned}
L(\boldsymbol{X}_p, \boldsymbol{X}_u, \boldsymbol{X}_c, \boldsymbol{y}_p, \boldsymbol{d}_u, \boldsymbol{y}_c, \boldsymbol{\Theta}_{repr}, \boldsymbol{\Theta}_p, \boldsymbol{\Theta}_d, \boldsymbol{\Theta}_c) &= L_p(\boldsymbol{X}_p, \boldsymbol{y}_p; \boldsymbol{\Theta}_{repr}, \boldsymbol{\Theta}_p) \\
&+ \alpha L_d(\boldsymbol{X}_u, \boldsymbol{d}_u, \boldsymbol{\Theta}_{repr}; \boldsymbol{\Theta}_d) \\
&+ \beta L_{conf}(\boldsymbol{X}_u, \boldsymbol{d}_u, \boldsymbol{\Theta}_d; \boldsymbol{\Theta}_{repr}) \\
&+ \sum_j^J (\gamma_j L_{c_j}(\boldsymbol{X}_{c_j}, \boldsymbol{y}_{c_j}, \boldsymbol{\Theta}_{repr}; \boldsymbol{\Theta}_{c_j}) + \phi_j L_{Cconf_j}(\boldsymbol{X}_{c_j}, \boldsymbol{y}_{c_j}, \boldsymbol{\Theta}_{c_j}; \boldsymbol{\Theta}_{repr}))
\end{aligned}
\tag{5.6}
$$

where there are J different confounds to be removed, γ_j is the weighting for the classification loss L_{c_j} for the jth confound, and similarly ϕ_j is the weighting of the confound confusion loss L_{Cconf_j} for the jth confound. This is demonstrated with sex as the confound, and so the classification loss can simply be the binary cross-entropy loss. If the confound to be unlearned took the form of a continuous variable, such as age, then following [322] it would need to be split into discrete bins.

Three different scenarios for removing sex as a confound while completing the age prediction task are considered. First, the simplest case is considered, where sex is approximately equally distributed across age and scanner. In this case, all the loss functions can simply be evaluated for the whole set of input images – assuming all labels are available for input images for all tasks.

The scenarios of the confound being correlated with a) the scanner and b) the main task are also explored. For correlation with scanner, datasets are created where 80% of the subjects in the OASIS dataset are male and 80% of the subjects in the UK Biobank dataset are female. For correlation with age, the case where 80% of the subjects under the median age across the two datasets (65) are female and 80% of the subjects over the median age are male is considered. For both scenarios, testing is still completed on the full test set, and comparison is made between normal training, unlearning sex on the whole training set, and unlearning sex on a subset curated to be balanced with

respect to the generated bias. In the last case, the loss functions for the main task and unlearning of scanner are evaluated across the whole datasets and only the equations controlling unlearning the sex confound are evaluated for the subsets.

5.2.4.6 Removal of Continuous Confounds

While the approach for the removal of categorical confounds, such as sex, is clear from the section above, the extension to continuous variables is less clear. The scenario is reframed such that sex prediction is now the main task, and the process of removing age as a continuous variable is considered.

The continuous age labels could be approximately converted into categorical labels by binning the data into single-year bins spanning across the age range. This, however, would not encode the fact that a prediction of 65 for a true label of 66 encodes more true age information than a prediction of 20. Therefore, the true age labels are converted into a softmax label around the true age, normally distributed as a $\mathcal{N}(\mu, \sigma^2)$ where μ was the true age label and σ was set to 10 empirically, allowing us to maintain relative information between bins. The value of 10 was chosen as, when used in normal training, this minimised the mean absolute error (MAE) on the test set. The loss function for the main task then becomes the Kullback-Leibler (KL) divergence between the true softmax distribution and the predicted softmax outputs. Unlearning can then be achieved as before, with the aim still being to make all the softmax outputs equal, so that no information about the age of the subject remains.

5.2.5 Segmentation Task

Having demonstrated the unlearning process on the age prediction task, segmentation is now considered. Specifically, the case where the network architecture takes the form of

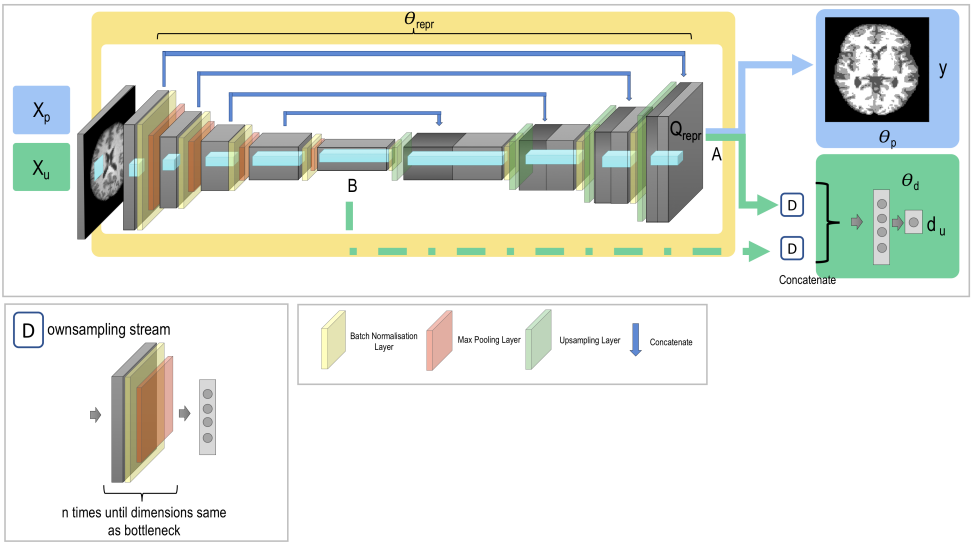


Figure 5.5: Network architecture used for unlearning with segmentation. $\boldsymbol{X}_p$ represents the input images used to evaluate the primary task with $\boldsymbol{y}$ being the main task label segmentations. $\boldsymbol{X}_u$ are the input images used for unlearning scanner information with domain labels d_u. The domain discriminator for unlearning can be attached from A, B, or the two in combination. If it is attached from A and B together, the first fully connected layers (the output of the two downsampling branches D) are concatenated together to produce a single feature representation.

a UNet [41] is considered, where the skip-connections, which form a crucial part of the architecture, have the possibility of increasing the complexity of the unlearning process. Again, the input images $\boldsymbol{X}$ are T1-weighted input images but the labels take the form of tissue segmentations (grey matter/white matter/CSF) produced using FSL FAST [11] as a proxy to manual segmentations, which were converted to one-hot labels for the segmentation. Examples from the UK Biobank [1] (2095 training, 937 testing) and healthy subjects from the OASIS dataset [4] (813 training, 217 testing) were considered. As before, these were resized to $128 \times 128 \times 128$ voxels (each of which is 1 x 1 x 1 mm in size) using spline interpolation and the intensity values were normalised, but then they were split into 2D slices so that a 2D network was trained. The labels were interpolated using trilinear interpolation and then thresholded at 0.5 to create categorical labels. Multi-class Dice Loss was used as the primary task loss function. All experiments were

completed on a V100 GPU and were implemented in Python (3.6) and Pytorch (1.0.1).

5.2.5.1 Basic Fully Supervised Learning

As can be seen in Fig. 5.5, the general structure is identical to that used for age prediction, apart from the location of the domain classifier. In the case of the age prediction task, the unlearning was performed at the first fully connected layer, chosen as it was both a reduced representation of the data and because it was at the end of the network so there were few weights afterwards for the network to re-learn how to exploit any remaining scanner information were it not fully removed by the unlearning process. In the case of the UNet architecture, the most compressed representation of the information is at the bottleneck (B) but there are both several upsampling layers and skip-connections after this point. Therefore, were unlearning only completed at the bottleneck, all scanner information would most likely be re-learned by the subsequent layers in the upsampling branch and so might still influence the output segmentations.

Therefore, unlearning is considered from both the end of the network before the label predictor (A), the bottleneck (B) and the combination of the two, formed by concatenating the first fully-connected layers to form a single domain prediction. Standard training is also completed, providing the benchmark. The simplest regime is utilised, where segmentation labels are available for all of the data from all scanners and we are considering identical segmentation tasks. Therefore the overall loss function takes the same form as equation (5.5) and all of the loss functions are evaluated on all of the data points.

5.2.5.2 Semi-supervised Learning

Finally, the scenario where there are very few or no labels available for one of the scanners is considered: this is a very likely scenario for segmentation where manual

labels are time consuming and difficult to obtain. It is assumed that there is one fully labelled dataset (UK Biobank) - and another dataset for which there are not many labels available. While the unlabelled data points cannot be used to evaluate the main task, they can be used for scanner unlearning.

No changes to the architecture are required; rather, the main task is simply evaluated for those data points for which main task labels are available and all data points are used for unlearning such that the overall method minimises:

$$\begin{aligned} L(\boldsymbol{X}, \boldsymbol{y}_p, \boldsymbol{d}, \boldsymbol{\Theta}_{repr}, \boldsymbol{\Theta}_p, \boldsymbol{\Theta}_d) = L_p(\boldsymbol{X}_p, \boldsymbol{y}_p; \boldsymbol{\Theta}_{repr}, \boldsymbol{\Theta}_p) \\ + \alpha L_d(\boldsymbol{X}, \boldsymbol{d}, \boldsymbol{\Theta}_{repr}; \boldsymbol{\Theta}_d) \\ + \beta L_{conf}(\boldsymbol{X}, \boldsymbol{d}, \boldsymbol{\Theta}_d; \boldsymbol{\Theta}_{repr}) \end{aligned} \tag{5.7}$$

where $\boldsymbol{X}_p$ is the subset of $\boldsymbol{X}$ for which main task labels $\boldsymbol{y}_p$ are available and the full dataset $\boldsymbol{X}$ is used in unlearning scanner information.

The effect of having increasing numbers of data points on the final segmentation is explored, and comparison is made of unlearning to normal training on the combination of the full UK Biobank dataset and OASIS with available segmentations.

5.3 Results

5.3.1 Age Prediction Task

5.3.1.1 Basic Fully Supervised Learning

The results from training with all three datasets separately and the different combinations can be seen in Table 5.1 – p-values can be found in the Appendix C.2.4. The scanner classification accuracy is achieved by taking the frozen feature representation

	Training Data			Biobank MAE	OASIS MAE	Whitehall MAE	Scanner Classification Accuracy %
	B	O	W				
Normal Training							
1.	✓	×	×	3.25 ± 2.36	16.50 ± 6.77	13.81 ± 5.42	-
2.	×	✓	×	5.61 ± 3.52	4.27 ± 3.79	6.73 ± 4.82	-
3.	×	×	✓	5.61 ± 3.65	5.22 ± 4.83	3.15 ± 2.81	-
4.	✓	✓	×	3.30 ± 2.50	4.00 ± 2.78	4.71 ± 3.42	98 (50)
5.	✓	×	✓	3.31 ± 2.49	4.45 ± 3.53	3.05 ± 2.84	100 (50)
6.	×	✓	✓	5.71 ± 3.59	4.05 ± 3.71	3.21 ± 2.94	100 (50)
7.	✓	✓	✓	3.24 ± 2.47	4.19 ± 3.50	2.89 ± 2.70	96 (33)
Normal Training - Eq (5.1)							
8.	✓	✓	×	3.45 ± 2.63	3.99 ± 2.85	4.56 ± 3.38	100 (50)
9.	✓	×	✓	3.42 ± 2.60	4.19 ± 4.01	2.99 ± 2.69	100 (50)
10.	×	✓	✓	4.56 ± 3.05	4.08 ± 3.73	3.18 ± 2.93	100 (50)
11.	✓	✓	✓	3.55 ± 2.68	3.90 ± 3.53	2.62 ± 2.65	98 (33)
Unlearning							
12.	✓	✓	×	3.41 ± 2.04	3.79 ± 2.99	4.60 ± 3.47	48 (50)
13.	✓	×	✓	3.41 ± 2.58	4.07 ± 4.12	2.81 ± 2.57	52 (50)
14.	×	✓	✓	3.38 ± 2.64	3.91 ± 3.53	2.82 ± 2.65	50 (50)
15.	✓	✓	✓	**3.38 ± 2.64**	**3.90 ± 3.53**	**2.56 ± 2.47**	34 (33)

Table 5.1: Results comparing unlearning to training the network in different combinations on the datasets. Mean absolute error is reported in years. Scanner accuracy is the accuracy achieved by a domain classifier given the fixed feature representation at convergence, evaluating only for the datasets the network was trained on. The number in brackets indicates random chance. B = UK Biobank, O = OASIS, W = Whitehall. p values can be found in Appendix C.2.4. Bold indicates the experiment with the best average across the datasets.

$\boldsymbol{Q}_{repr}$ from training, and training a classifier based on this. For the unlearning method, this is the domain classifier that was used during training but fine-tuned to convergence. For normal training, an identical domain classifier is trained using the frozen feature representation as input until convergence. Therefore, the closer the value is to random chance, the less informative the feature representation is about scanner, thus meaning that the influence of the scanner on the prediction is reduced. The results from training on all combinations of datasets, both with standard training, and standard training using our loss function for the main task conditioned on each scanner are shown for comparison.



(a) **(b)**

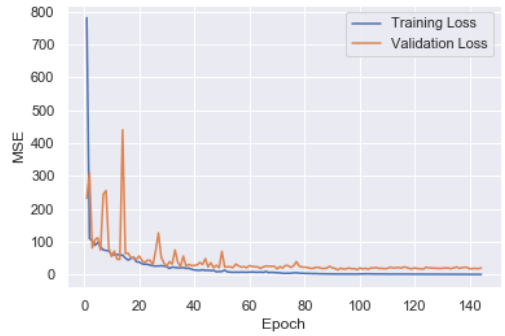
Figure 5.6: a) A t-SNE plot of the activations of the fully connected layer, $\boldsymbol{Q}_{repr}$, before unlearning. It can be seen that the domains can be almost entirely separated, except for two data points grouped incorrectly, showing that data from each scanner has its own distinct distribution. b) A t-SNE plot of the activations of the fully connected layer, $\boldsymbol{Q}_{repr}$, after unlearning. It can be seen that, through the unlearning, the distributions become entirely jointly embedded.

In Fig. 5.6, a t-SNE plot [326] can be seen, allowing visualisation of the activations at the output of the feature extractor, $\boldsymbol{Q}_{repr}$. It can be seen that, before unlearning, the features produced are entirely separable by scanner, but after unlearning the scanner features become jointly embedded and so the feature embedding is not informative of scanner.

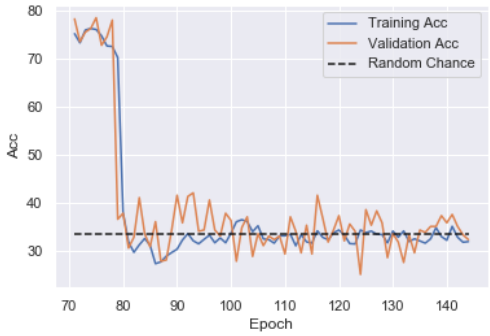
The t-SNE demonstrates that without unlearning, not only are there identifiable scanner effects but that they also affect the age predictions. On the other hand, it can be seen that scanner information can be removed using the unlearning technique, such that the data points for all three scanners share the same embedding. This is confirmed by the scanner classification accuracy being almost random chance after unlearning has been completed.

It can also be seen that unlearning does not decrease substantially the performance on the main task. Graphs showing the different loss functions with training can be found in Fig. 5.7. Only the main task loss (MSE for age prediction) is used during pretraining

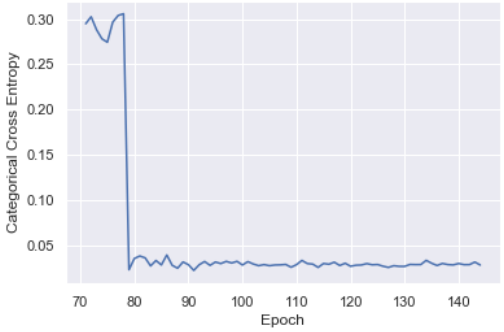
and so the other loss functions begin at epoch 70 when unlearning started. Unlearning began when convergence was achieved on the main task (patience = 15 epochs). It can be seen that beginning unlearning does not reduce the performance on the main task. It can also be seen that all of the loss functions are stable throughout training.



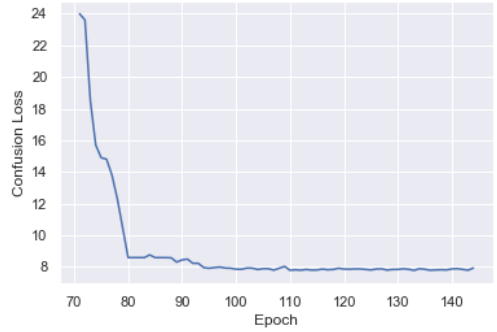
(a) Main Task Loss



(b) Scanner Classification Accuracy



(c) Domain Loss - Training



(d) Confusion Loss - Training

Figure 5.7: Graphs for the different loss functions from training. Only the main task loss is used during pretraining and so all other losses begin at 70 epochs, when convergence on the main task is achieved.

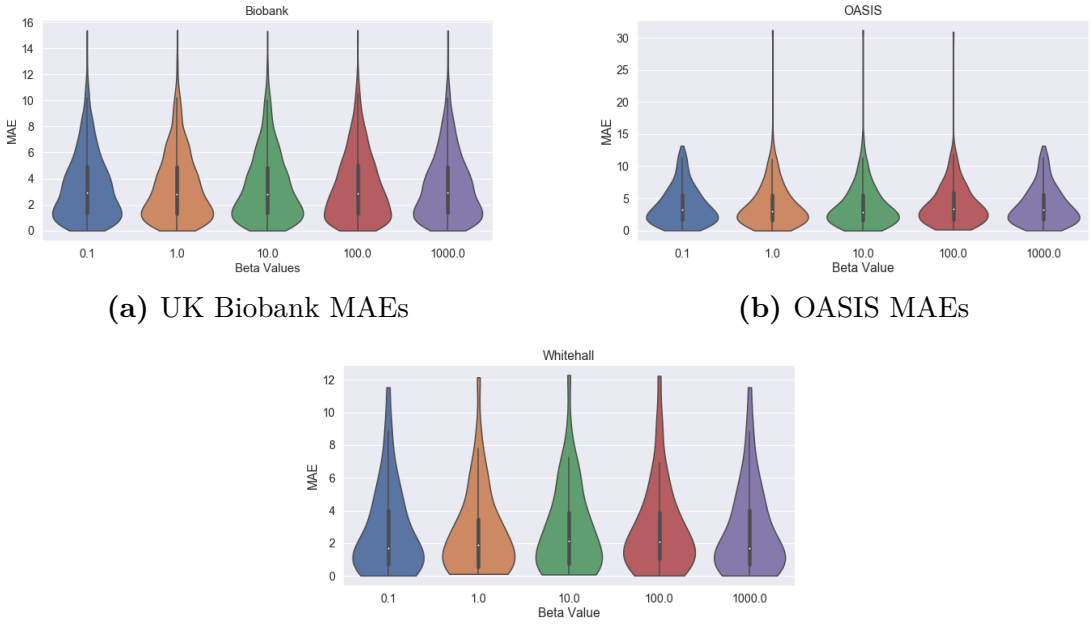
Considering the case of training on all three datasets, there is actually an overall improvement in performance using unlearning across the three datasets (comparing lines 7 and 15 in Tab. 5.1) with the performance on the OASIS and Whitehall datasets substantially improving. The performance on the UK Biobank dataset decreases, but this is as would be expected because the UK Biobank dataset has a much larger size

and so, with normal training, the network is driven by this dataset whereas, with unlearning, all of the datasets drive the performance. Comparing to line 11, where the same loss function was used for the main task, evaluated separately for each dataset, but unlearning was not performed, it can be seen that there was an improved performance when the unlearning was added and so the unlearning itself provides actual improvement on the main task. It can also be seen that the scanner classification accuracy is almost perfect when using normal training (96%), or the loss function evaluated on the datasets separately (98%), and is random chance after unlearning (34%), clearly demonstrating that the unlearning process removes scanner information from the feature space.

It can also be seen that the unlearning process creates a feature space that generalises better to the third unseen dataset. It can be seen, for instance by comparing lines 5 and 13, that not only is there an improved performance on the OASIS and Whitehall datasets used for training, but there is also a significant improvement in the performance on the UK Biobank dataset, which was unseen by the network. This, therefore, suggests that the unlearning process not only creates features that are invariant to the scanners that were present in the training set, but also that these features are more generalisable and so can be more effectively applied to other scanners. It can again be seen that there is an improvement between normal training with our loss function evaluated on each scanner separately, and the performance with unlearning (lines 9 and 13), showing that the increase in generalisability of the features is due to the unlearning process. The same pattern can be seen for training on the other possible pairs of datasets.

The generalisability of the features between datasets and the reduction of the scanner classification accuracy together demonstrate that the unlearning process successfully harmonises the differences between scanners and protocols.

The effect of the hyperparameters on the unlearning process was also explored. For all experiments, the value of α remained set to 1 and no change was required.



(a) UK Biobank MAEs

(b) OASIS MAEs

(c) Whitehall MAEs

Figure 5.8: MAEs for each dataset with increasing values of the constant β used in training, where β is the weight determining the contribution of the confusion loss.

The value of β was varied between experiments, controlling the training stability and the rate of convergence: higher β values provide more stability for the domain unlearning but decrease the rate of convergence for the main task. It should not have any effect on the final age prediction values achieved, given that for all values it is possible to reach convergence, and the value of β was varied between 0.1 and 1000 and the MAEs compared for the three datasets. It can be seen that the value of β has no effect on the achieved MAEs across these values from Fig. 5.8. Therefore, the segmentation performance is robust to the choice of β and the value chosen can be selected to maximise the stability of training without impacting on the final prediction values. Stability could also be controlled by using different learning rates for each stage; however, this was found to be harder to tune in practice.

Finally, the effect of the batch size was explored. For training, the largest batch that

could fit into the memory was used to obtain the results in this chapter; here, the effect of smaller batches is explored. Figure 5.9 shows the validation loss during training for batch sizes of 3, 8 and 32 (other batch sizes are not shown for the purpose of clarity) and the pattern was the same across the different loss functions. It can be seen that the training was more stable with the smaller batch sizes. However, it can be seen from Fig. 5.10 that the overall MAEs achieved across the datasets were better with larger batch sizes. This is most likely because larger batch sizes force the network to unlearn scanner information more thoroughly than, for instance, with a batch size of 3, where given the constraint on the batches, there can only be one example per scanner. Given the result from varying the value of β above, the batch size should be chosen so as to maximise the performance on the main task, varying the β value to control the stability of training.

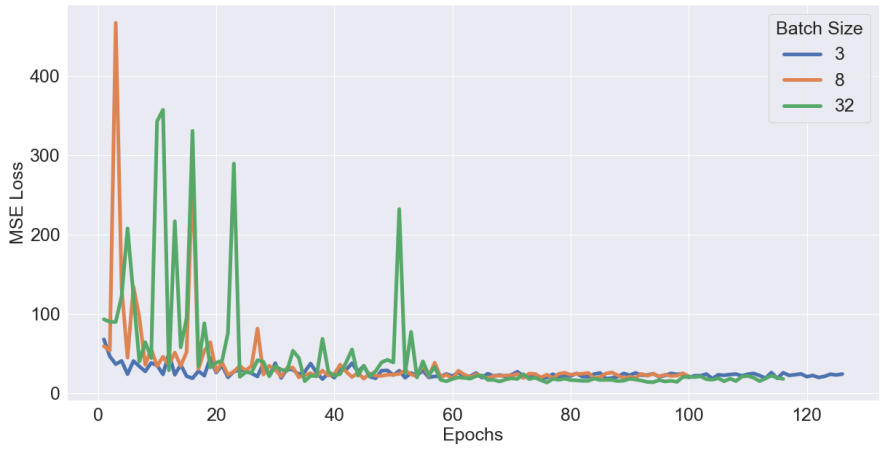
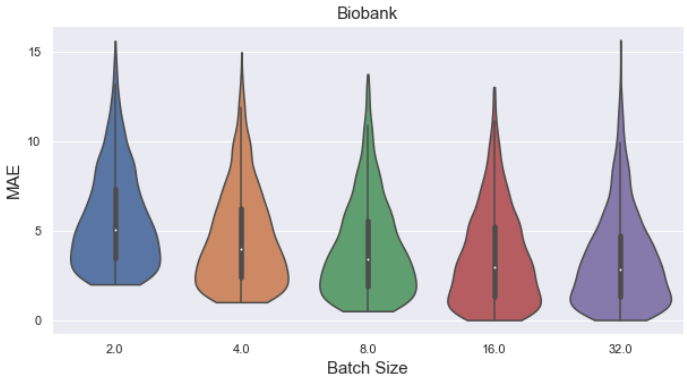
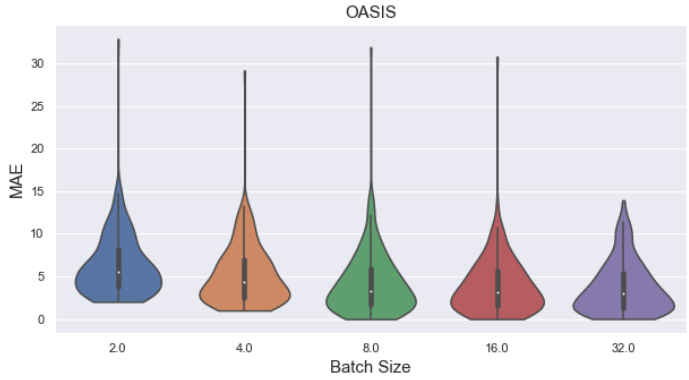


Figure 5.9: Validation loss with epochs for batch sizes of 3, 8 and 32.

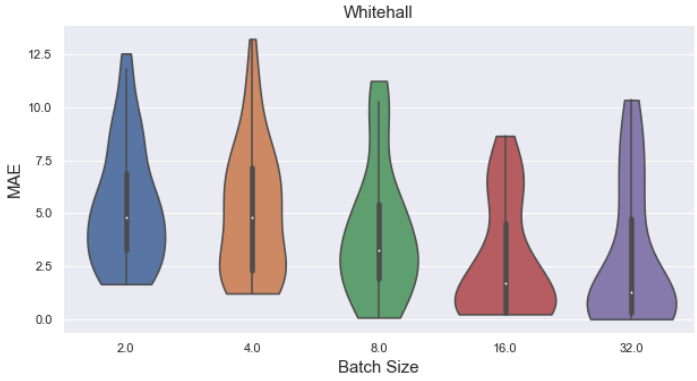
It was considered if it was better to use the full imbalanced datasets to maintain the number of data points or if this hindered the unlearning process. Therefore, the UK Biobank and OASIS datasets were randomly downsampled to have the same number of examples as the Whitehall dataset (452 subjects). Testing was evaluated



(a) UK Biobank MAEs



(b) OASIS MAEs



(c) Whitehall MAEs

Figure 5.10: MAEs for each dataset with increasing batch sizes used during training.

on the same test dataset as in the previous experiments. The results can be seen in Table 5.2. For both normal training and unlearning, there was a large decrease in

Training Method	Testing Dataset (MAEs)			
	UK Biobank	OASIS	Whitehall	Scanner Classification Accuracy
Full Dataset				
Normal Training	3.24 ± 2.47	4.19 ± 3.50	2.89 ± 2.70	100 (33)
Unlearning	**3.38 ± 2.64**	**3.90 ± 3.53**	**2.56 ± 2.47**	34 (33)
Balanced Dataset				
Normal Training	5.31 ± 3.82	5.80 ± 3.93	3.50 ± 3.09	98 (33)
Unlearning	4.11 ± 3.23	4.10 ± 3.68	3.11 ± 2.99	34 (33)

Table 5.2: Comparison of the effect of training with the full datasets with imbalanced numbers from each scanner versus training with balanced numbers of subjects from each scanner (452 per scanner, randomly selected compared to 5508 UK Biobank, 813 OASIS and 452 Whitehall), comparing both normal training and unlearning. The method giving the best average across datasets is highlighted.

performance using the balanced datasets, but the decrease was less pronounced when using unlearning than when using normal training. Using balanced datasets made no difference to the ability to classify the scanner: with normal training the classifier was still able fully to distinguish between scanners; with unlearning, the classification was almost random chance. Therefore, any advantage gained from having balanced datasets is outweighed by the reduction in performance from having reduced numbers of training examples; thus, it is better to train on the whole dataset. Given that the method is able to remove nearly all information with significant imbalance in this case (UK Biobank 5508 training data points compared to Whitehall's 452) it is likely that unlearning would be sufficient in most cases.

5.3.1.2 Biased Datasets

The performance of the network training with biased datasets curated from subsets of the UK Biobank and OASIS datasets was assessed. The cases of 5-year, 10-year and 15-year overlaps were considered, where the smaller the overlap, the harder the task. The trained networks were tested on the same test sets used above, with subjects across the whole age range included.

Method	UK Biobank MAE	OASIS MAE	Scanner Classification Accuracy (%)
5 Years Overlap			
1. Standard	16.5 ± 5.94	15.5 ± 6.95	100 (50)
2. Naïve Unlearning	6.11 ± 3.99	4.44 ± 4.20	58 (50)
3. Unlearning on Overlap	**5.49 ± 3.67**	**4.37 ± 4.05**	53 (50)
10 Years Overlap			
4. Standard	9.66 ± 5.83	13.6 ± 6.58	100 (50)
5. Naïve Unlearning	4.20 ± 2.90	4.29 ± 4.01	56 (50)
6. Unlearning on Overlap	**3.93 ± 2.81**	**4.04 ± 3.86**	52 (50)
15 Years Overlap			
7. Standard	8.91 ± 5.31	10.4 ± 5.55	100 (50)
8. Naïve Unlearning	3.82 ± 2.84	4.39 ± 4.07	57(50)
9. Unlearning on Overlap	**3.75 ± 2.78**	**3.99 ± 3.52**	50 (50)

Table 5.3: MAE results for UK Biobank and OASIS data from training with datasets with varying degrees of overlap as shown in figure 5.3. Standard refers to training normally and naïve unlearning refers to unlearning on the whole of the datasets. Scanner accuracy is calculated by training a domain classifier on the fixed feature representation. Random chance is given in brackets. The method with the best average across the datasets for each degree of overlap is highlighted.

The results can be seen in Table 5.3 where normal training is compared to naïve unlearning on the whole dataset, and unlearning on only the overlapping age range. It can be seen that for all three degrees of overlap, standard training leads to much larger errors than unlearning and that unlearning on only the overlap gives a lower degree of error than unlearning on all data points. Plots of the MAEs with age can be seen in Fig. 5.11 for the 10-year overlap case. It can be seen that the overall lowest MAEs are achieved across the age range when unlearning is performed only on the overlapping subjects.

As expected, it can be seen that the normal training regime produces large errors, especially outside of the range of the UK Biobank training data, as the learnt weights are very largely driven by the much larger size of the UK Biobank training data. With

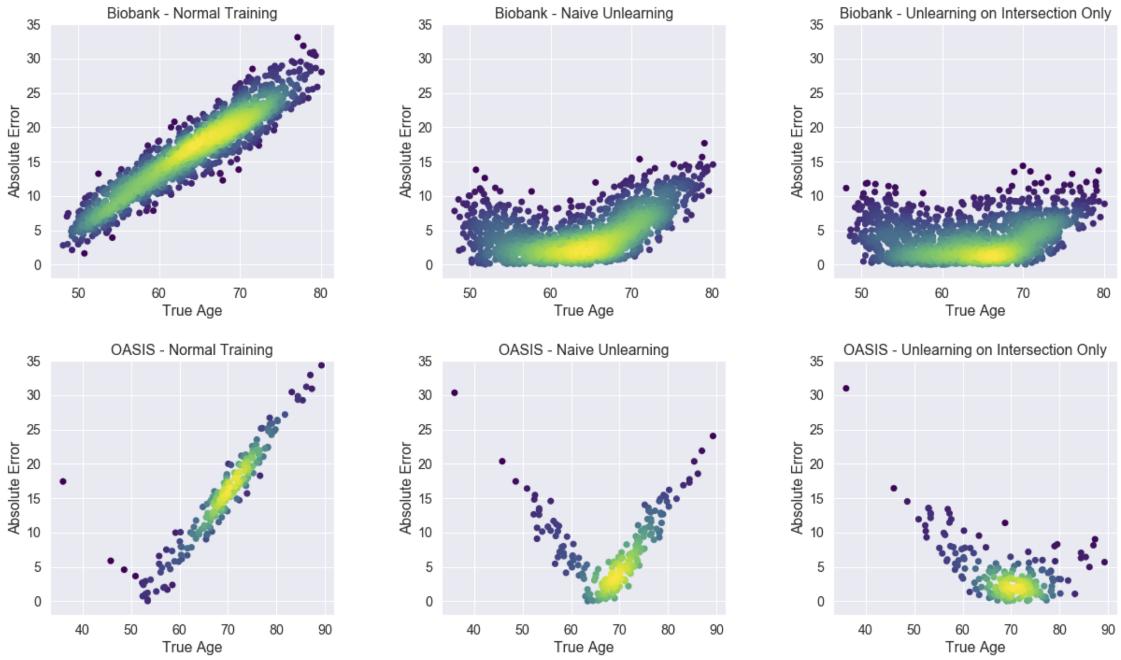


Figure 5.11: Density plots showing the absolute errors for the three different training regimes: standard training, naïve unlearning and unlearning only on the overlap data for 10-year overlap. It can be seen that unlearning only on the overlap dataset leads to substantially lower losses across both datasets.

naïve unlearning, the network is not able to correct for both scanners and the results for the OASIS data are poor, whereas by unlearning on just the overlapping subjects, the error is reduced on both of the datasets during testing. The only area where there is a reduction in performance is the lower end of the OASIS dataset, possibly because when the network was being driven by the UK Biobank data, the network generalised well to the OASIS subjects from the same age range. Naïve unlearning also performs slightly less well at removing scanner information on the testing data, probably indicating that the features removed also encode some age information and so generalise less well across the whole age range.

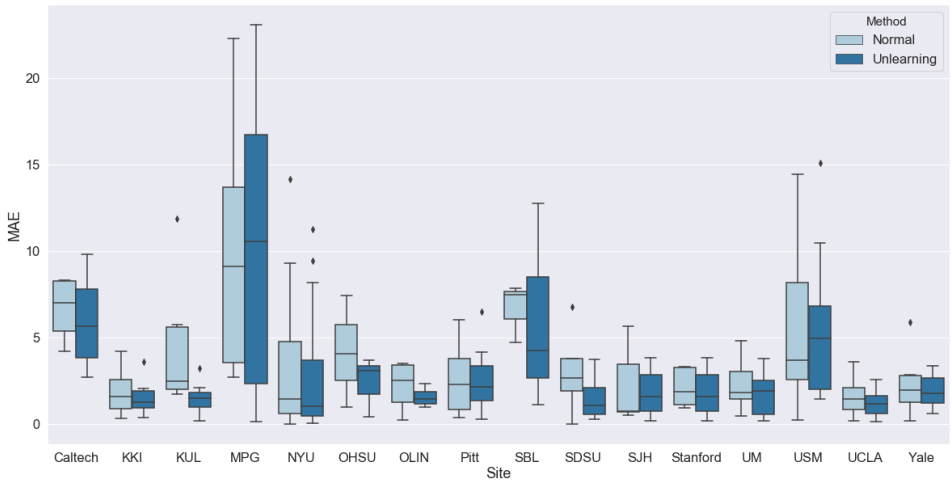


Figure 5.12: MAE results broken down by site for the ABIDE data, comparing normal training to unlearning.

5.3.1.3 Extension to ABIDE Data

Figure 5.12 shows the results of applying the unlearning process to the ABIDE data, comparing normal training to unlearning. It can be seen that unlearning improves the MAE across all sites apart from MPG, which contains significantly older subjects compared to the other sites. The scanner classification accuracy before unlearning was 55.9% and after unlearning was reduced to 6.42%, where random chance was 6.25%. Therefore, it can be seen that scanner information is present in the feature embeddings, even when harmonised protocols are used.

These results therefore show that the framework can be applied to many sites with no changes needed. It also shows that the framework is applicable to lower numbers of subjects, with some of the sites having as few as 27 subjects for training. Thus, the framework should be applicable to many studies.

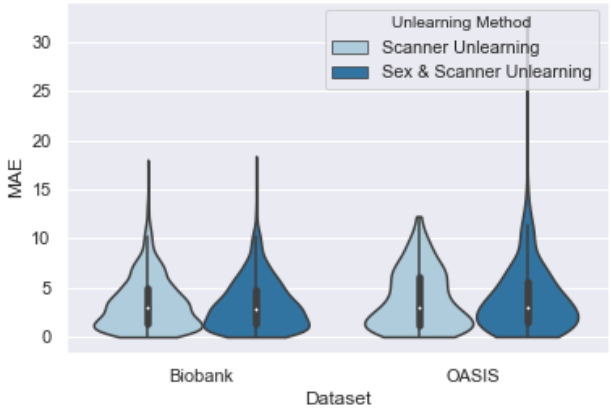


Figure 5.13: Results comparing unlearning scanner only to unlearning scanner and sex information.

5.3.1.4 Removal of Other Categorical Confounds

The effect of removing sex information as an additional confound in addition to harmonising for scanner was investigated. In Fig. 5.13 it can be seen that there is no significant effect on the MAE results when removing sex information in addition to scanner information. Unlearning sex information had no substantive effect on the ability to remove scanner information with the scanner classification accuracy being 48% and 49%, respectively. The sex classification accuracy was 96% before unlearning and 54% after unlearning. Therefore, multiple pieces of information can be removed simultaneously and, so long as the information that needs to be unlearned does not correlate with the main task, it can be done without significantly reducing the performance on the main task.

Next, the case where sex information was correlated with the acquisition scanner was considered, for the case where 80% of the subjects in the OASIS dataset are male and 80% of the subjects in the UK Biobank dataset are female. Table 5.4 shows the comparison of normal training on these datasets compared to unlearning sex on all

Training Method	UK Biobank MAE	OASIS MAE	Scanner Classification Accuracy	Sex Classification Accuracy
Normal Training	3.44 ± 2.56	3.98 ± 3.33	100 (50)	98 (50)
Unlearning All	3.41 ± 2.60	4.05 ± 3.66	52 (50)	57(50)
Unlearning Balanced	3.43 ± 2.60	4.04 ± 3.60	49 (50)	56 (50)

Table 5.4: MAE results for unlearning scanner and sex when sex is highly correlated with scanner.

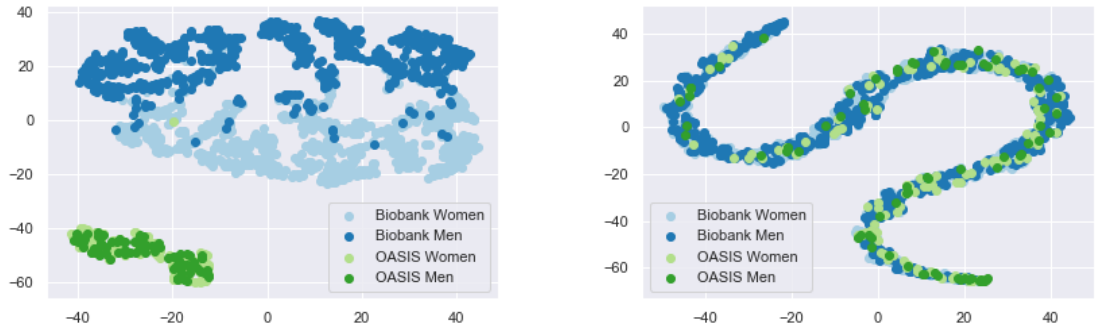
data points and unlearning sex on a subset with balanced numbers of each sex for each scanner. The full testing set was still used.

Finally, the case where sex – the confound to be unlearned – is highly correlated with age – the main task – was considered, with 80% of the subjects below the median age being female and 80% of the subjects over the median age being male. Again, normal training was compared to unlearning on the whole dataset and to unlearning on a curated subset, with uniform distributions of sex with age for each scanner. Testing was again performed on the full testing datasets.

It was found that it was not possible to unlearn on the whole distribution as it caused the training to become unstable almost immediately (after approximately 5 epochs), with the training and validation loss exploding before sex and scanner information was unlearned. The results from unlearning scanner can be seen in Table 5.4 where it can be seen that by unlearning only on a subset of data points with equal sample numbers, both scanner and sex information could be almost entirely unlearned. Figure 5.14 shows a t-SNE [326] of $\boldsymbol{Q}_{repr}$ where it can be seen that, before unlearning, the data was largely separable by scanner and sex, and that, after unlearning, these are indistinguishable.

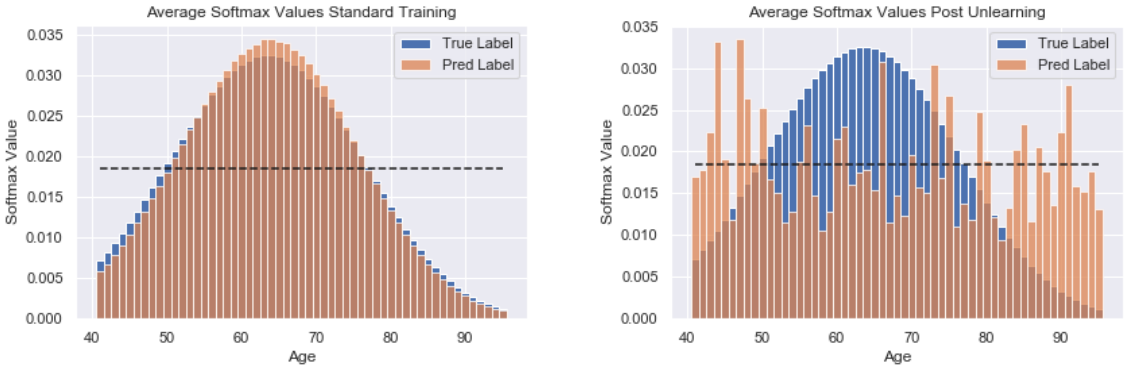
5.3.1.5 Removal of Continuous Confounds

For the removal of age as a confound, sex prediction was used as the main task. An average of 96.3% was achieved on the sex prediction task before unlearning and 95.9%



(a) **(b)**

Figure 5.14: a) t-SNE plot of $\boldsymbol{Q}_{repr}$ from training without unlearning, having trained on a dataset where sex correlates highly with age. It can be seen that the scanners can still be almost entirely separated, except for two data points grouped incorrectly (located within the cluster of light blue data points), showing that data from each scanner has its own distinct distribution and that the data points can also be entirely split by sex for the UK Biobank data. b) t-SNE plot of $\boldsymbol{Q}_{repr}$ after unlearning. It can be seen that, through the unlearning, the distributions become entirely jointly embedded in terms of both scanner and sex.



(a) **(b)**

Figure 5.15: Softmax labels for the age prediction task, averaged across all examples for the UK Biobank dataset showing both the true values converted into Gaussian distributions and the predicted values. The dashed line corresponds to the value if all of the softmax values were equal. a) Predictions with standard learning b) Predictions after unlearning. It can be seen that after unlearning the softmax values are much closer to the dashed line and there is no trend towards the true maximum.

after unlearning. As before, it was possible to unlearn scanner information, reducing the scanner classification accuracy from 100% to 53%. Figure 5.15a shows the averaged

softmax labels from the age prediction task with normal training, where it can be seen that there is a large degree of agreement between the true labels and the predicted labels, showing that it is possible to learn age using the continuous labels and KL divergence as the loss function. An MAE value of 3.26 ± 2.47 was achieved for UK Biobank and 4.09 ± 3.46 for OASIS with normal training in this manner.

Figure 5.15b shows the softmax labels after unlearning. It can be seen that the predicted labels no longer follow the same distribution as the true labels and that they are distributed around the random chance value indicated by the dotted line. It can also be seen that there is no trend towards the true age value, indicating that a large amount of the age information has been removed. After unlearning, the MAEs for the age prediction task were 17.54 ± 7.65 for UK Biobank and 22.86 ± 9.40 for OASIS. The average KL divergence has increased from 0.0022 with standard learning to 0.251 after unlearning. This therefore demonstrates that, by adapting the labels, the same training regime can be used to remove continuous information.

5.3.2 Segmentation Task

5.3.2.1 Supervised Learning

Location of Domain Classifier	UK Biobank Dice	OASIS Dice	Scanner Classification Accuracy
Final Convolution (A)	0.910 ± 0.023	0.916 ± 0.021	51 (50)
Bottleneck (B)	0.871 ± 0.046	0.882 ± 0.030	100 (50)
Both (A+B)	0.903 ± 0.025	0.912 ± 0.021	55 (50)

Table 5.5: Dice scores comparing different locations for attaching the domain classifier in the network (as indicated in Fig. 5.5) A) at the final convolutional layer, B) at the bottleneck, and A+B) the combination of the two locations. The scanner classification accuracy was the accuracy achieved by a separate domain classifier using the fixed feature representation at the final convolutional layer. Random chance is given in brackets.

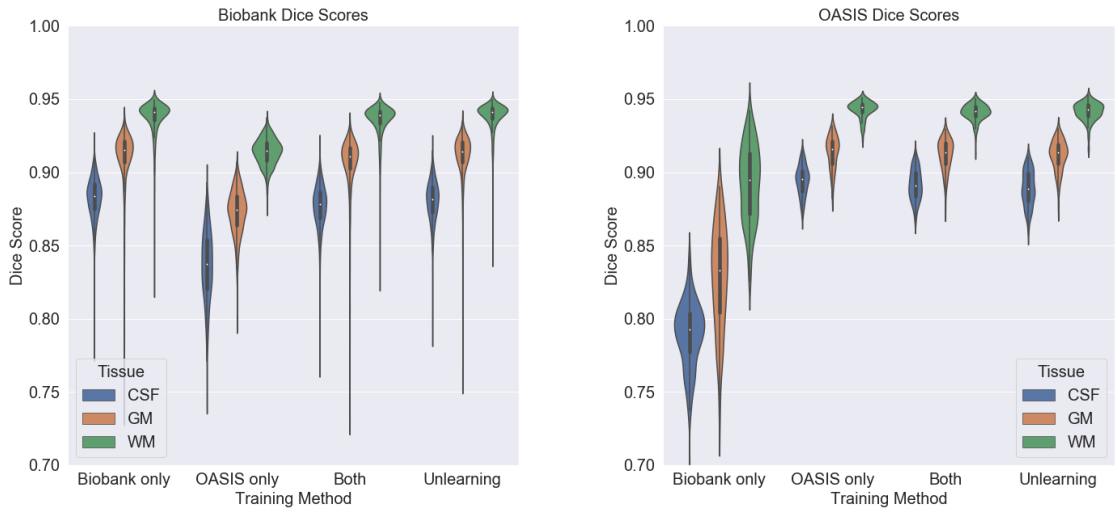
Training Data	UK Biobank Dice	OASIS Dice	Scanner Classification Accuracy (%)
UK Biobank Only	0.910 ± 0.022	0.836 ± 0.043	-
OASIS Only	0.874 ± 0.032	0.917 ± 0.020	-
Both (Normal Training)	0.906 ± 0.024	0.915 ± 0.020	100 (50)
Both (Unlearning)	0.910 ± 0.023	0.916 ± 0.021	51 (50)

Table 5.6: Dice scores comparing unlearning to training the network on different combinations of the datasets, averaged across the tissue types. Scanner accuracy is the accuracy achieved by a domain classifier given the fixed feature representation, with random chance given in brackets.

The results from comparing the location of the domain classifier can be seen in Table 5.5. Dice scores and scanner classification accuracy are compared. It can be seen that the best results for both Dice and the scanner classification accuracy were achieved with the domain classifier simply connected at the final convolution. Training with the domain classifier attached to the bottleneck was also much less stable. Given this result, the location of the domain classifier is fixed to the final convolutional layer for the rest of the experiments that follow.

This finding goes against the expectation from the literature [113] which suggests a mixture of locations is best for domain adaptation. This is most likely due to the prioritisation of the removal of the scanner information to ensure harmonisation through unlearning: at the bottleneck (B) it is the most compressed representation of the data and it is probable that by unlearning at that location the features are constrained too highly and they are unable to find as successful a representation of the data.

The results for training on the combination of datasets with normal training and unlearning can be seen in Table 5.6, averaged across tissue type, and in Fig. 5.16 by tissue type for each scanner. It can be seen that the unlearning process does not reduce the Dice score achieved across tissue types and it is possible to remove scanner information almost entirely.



(a) **(b)**

Figure 5.16: Dice scores for the two datasets for each method broken down by tissue type. CSF = Cerebrospinal Fluid, WM = White Matter, GM = Grey Matter

5.3.2.2 Semi-supervised Segmentation

To allow the exploration of the effect of training the network with low numbers of labelled training points for one of the scanners, the network was trained with both normal learning and with unlearning for different numbers of OASIS data points with labels. Only the examples with labels were used for evaluating the main task but all of the data points were still used for unlearning scanner as these loss functions do not require main task labels to be evaluated. The results can be seen in Fig. 5.17: for all cases, the scanner classification accuracy for normal training was 100% and between 50% and 55% for unlearning. It can be seen that unlearning gives a large improvement in the segmentations with low numbers of data points, not only in terms of the mean value but also the consistency of the segmentations, including when the network is unsupervised with regards to the OASIS dataset and so has no training labels. Even with large numbers of training examples, there is never a disadvantage to using unlearning.

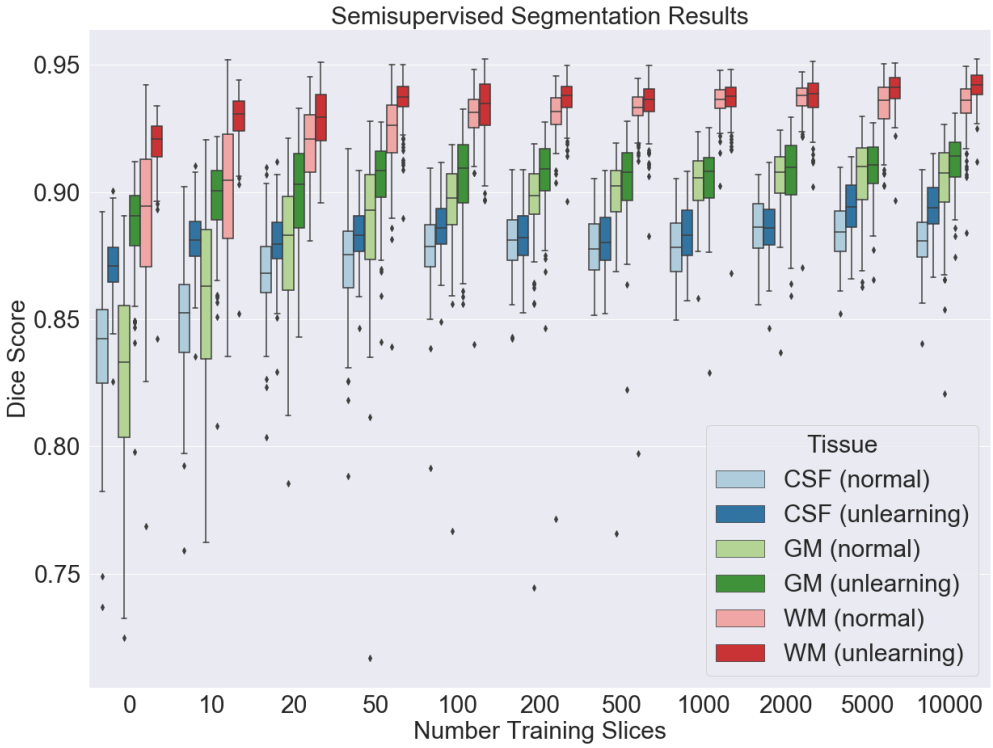


Figure 5.17: Dice scores for the three different tissue types for the OASIS data with increasing numbers of OASIS training slices, comparing both normal training and unlearning with the full UK Biobank dataset used throughout. For clarity, the x axis is not plotted to scale.

5.4 Discussion

In this work, it has been demonstrated that a domain adaptation technique can be adapted to harmonise MRI data for a given task using CNNs, by creating a feature space that is invariant to the acquisition scanner. Removal of scanner information from the feature space prevents the information from driving the prediction. The cost of the harmonisation is an approximately two fold increase in training time (from 0.88 seconds per batch to 1.53 seconds on a V100) but there is no increase in prediction time, and so, once the model is trained, there is no disadvantage to having unlearned scanner

information. The duration of a batch increase is not the three times that might be expected, with the increase from one-forward and backward pass to three forward and backward passes per epoch, because the second two loss functions only update subsets of the parameters, reducing the number of calculations that must be completed.

The technique has been demonstrated for both regression and segmentation. The applicability to a wide range of networks, even those with skip connections such as the UNet, provides a high level of flexibility. It does, however, only harmonise the features for a given task and therefore a separate network must be trained for each task. Consequently, an area for future research would be to harmonise across both scanners and tasks.

As a comparison, a DANN-style network [111], with the same architecture as our network but an additional gradient reversal layer inserted for the domain classifier branch, was also trained, but it was found that when training with only two datasets, the domain classifier was unstable and predicted the domains with 7.2% accuracy (results can be seen in Appendix C.2.2). This was a local minimum but clearly meant that the scanner information was not successfully removed. This instability in adversarial tasks posed as a minmax optimisation has also been experienced for other applications across the literature [327, 328]. By contrast, the confusion loss used in our iterative approach forces the logits towards random chance and the training of the domain unlearning branch is much more stable. It was not possible to compare to existing harmonisation methods, such as ComBat, because the different approaches are not easily compared, especially for the segmentation task, and focus on different stages of the pipeline.

A potential limitation of the work is that the framework is not generative and therefore cannot easily be used in conjunction with tools such as VBM or Freesurfer, although it could be added into deep learning equivalents such as Fastsurfer [82], especially as more tools gain CNN-based equivalents. The decision not to make the

framework generative was due to generative CNN methods needing large amounts of data to train or even, in some examples, paired data, meaning that they are not applicable to most real life neuroimage studies, whereas this framework is applicable to even highly multi-site datasets with limited examples from each site, such as the ABIDE data. Furthermore, due to the data-hungry nature of generative CNNs, most of these methods reduce 3D volumes to 2D slices or even patches. This then leads to errors when reconstructing the outputs, such as inconsistencies between slices. Were these images then to be used in further downstream analysis, it is very hard to anticipate how these errors might impact on obtained results and to know how to account for them. Although harmonised images are not produced, the image-derived values can be used in downstream analysis, for instance tissue volumes or segmentation maps can be used with a GLM to explore group differences. The approach, therefore, is limited to within CNNs but this is a rapidly growing area of study within neuroimaging and the flexible nature of the framework means it should be applicable across feed-forward architectures and different neuroimaging tasks.

The ablation study on the age prediction task shows that not only is it possible to unlearn the scanner information but also that there is an increase in performance that is not just due to the main task loss function. This indicates that the learned features generalise better across datasets, which is confirmed by the performance on the third dataset when only training on two. This corroborates the scanner classification accuracy and t-SNE plots, showing that scanner information has largely been removed and that the features that remain encode age information that transfers across datasets. Even for the segmentation task, where there is no significant improvement in the performance, including the unlearning gives us the assurance that the output values are not being driven by scanner. This could be important for downstream analysis, for instance comparing white matter volumes, allowing the data to be pooled across

sites. This result is also robust to the choice of the weighting factors and so the unlearning is not sensitive to the hyper-parameters.

The results also show that the unlearning scheme can be used even when there is a strong relationship between the main task label and the acquisition scanner. This could be especially useful when combining data between studies with different designs and with very different numbers of subjects from each group. As unlearning can be performed on a different dataset to the main task, there is the flexibility to apply the method to a range of scenarios.

At the limit where there exists no overlap between the datasets' distributions, unlearning scanner information would be highly likely to remove information relating to the main task. This limitation of learning invariant representation for MRI harmonisation is discussed in [329], where it is discussed that if a real label value is highly informative about the source domain it cannot be accurately predicted by an invariant predictor. Using the proposed framework, it should be possible to mitigate for this problem by having an additional dataset for each scanner, acquired with the same protocol, and so this represents a potential limitation of the method. The extension to the ABIDE dataset, however, shows the applicability of the method to many small datasets, and therefore this may be solvable using a set of small datasets which together span the whole distribution. It also shows that the framework is able to harmonise across scanner manufacturers as the data was collected on a range of GE, Siemens and Philips scanners.

The approach can be extended to allow the removal of additional confounds, and a way has been that demonstrated enabling the framework to also be extended to allow the removal of continuous confounds such as age. For each confound to be removed, two additional loss functions are required, and two additional forward and backward passes. Therefore, the training time will increase with each additional confound, presenting a

potential limitation. Were many confounds to be removed, it might also be necessary to increase the number of times the passes for the main task are performed to prevent the performance on the main task suffering from too much degradation due to the feature space being optimised for multiple tasks. It has, however, also been shown how confounds can be removed even when they correlate with the main task. By carefully selecting a subset of the data with which to unlearn the scanner, confounds can be removed, including, for instance, where sex is correlated with age, which is a case where both the main task and confound to be removed are associated with structural changes.

The approach can also be applied when no labels are available for one or more of the domains. The technique has been demonstrated for the segmentation task, and can be used effectively when labels are only available for one domain and it would be expected that this would extend to multiple unlabelled datasets.

A method for MRI harmonisation and confound removal has been presented that should be applicable across many tasks for neuroimaging and data scenarios. It has been shown that the framework can easily be applied to segmentation, classification and regression tasks and, given the highly flexible nature of the framework, it should be applicable to any feed-forward network. The ability to remove scanner bias influencing the predictions of CNNs should enable the combination of small datasets and the exploration of problems for which there are no single-scanner datasets of adequate size to consider.

6

Simultaneous Model Pruning and Training for Low Data Regimes in Medical Image Segmentation

The code used in this chapter will be released on acceptance for publication at: `github.com/nkdinsdale/STAMP`.

6.1 Introduction

Semantic segmentation in medical imaging is vital for the understanding and monitoring of the progression of disease. For instance, the accurate segmentation of the

hippocampus is essential for volumetric and morphological assessment, as hippocampal atrophy, observed through MRI, is one of the most validated biomarkers of Alzheimer's disease [330]. Manual segmentation, however, is time consuming and difficult and many structures have ambiguous boundaries, making consistent delineation of the structures between raters hard to achieve. Therefore, there is a need for automatic segmentation methods, capable of reliably providing accurate segmentations of the structures of interest.

Deep learning-based methods have become the state of the art for medical image segmentation, with most methods being based on the UNet architecture [41–43], a version of which can be seen in Fig. 6.3. This is formed of an encoder branch that takes the image and generates a high-dimensional feature vector, aggregating features at multiple levels to create a full representation of the variation of the input images, and a decoder that takes the high-dimensional feature representation and, from it, generates a segmentation mask, decoding the features that were aggregated by the encoder at multiple levels. It also has skip connections between representations at the same levels of abstraction. Thus, the encoder captures context, and its symmetric expanding path decoder enables precise localisation, with the skip connections allowing the transfer of spatial information into the decoder. This promotes the flattening of the optimisation space, aiding the optimisation of the loss function and thus making it easier to train the network [44]. Methods either using UNets directly, or methods highly inspired by the UNet, have been applied to the spectrum of medical imaging segmentation tasks across a range of modalities, for example: in MRI [35, 50, 169, 331], in CT [50, 166, 332–334], in X-Ray [50, 335] and in ultrasound [166, 181, 336, 337].

A major limitation of these methods, and with deep learning-based methods in general, is the large amount of labelled training data required to train the models. In medical imaging, manual delineation by domain experts is considered to be the 'gold

standard' for labels. Therefore, to produce a dataset large enough to train a network for segmentation is expensive, requiring large amounts of expert time to curate, and expert domain knowledge. Furthermore, the labels required often do not form part of standard clinical practice, and therefore have to be produced specifically for the network training. Thus, there is a need to develop methods that can work in data domains where low numbers of labelled data points are available.

One of the causes of the need for large dataset sizes for training when using deep neural networks is that the models contain large numbers of parameters, and many more than the amount of training examples available. Many methods have been proposed to help train models in low data regimes, as was discussed in the CNN background chapter, with the most frequently used approaches being model pretraining or transfer learning, and data augmentation. Model pretraining uses a related large dataset to initialise the model weights, allowing the optimisation to begin from a more informed place. Data augmentation applies random transformations to the data during training, increasing the size and variation of the dataset seen, and so artificially creating more data points to train the model. Here utility of model pruning will be investigated, reducing the number of parameters in the model and so the model complexity, and thus potentially reducing the number of data points required to train the model.

Model pruning for neural networks was first explored in [182], where they introduced the standard pruning framework of training a model until it performs acceptably well on a given task and then removing the weights with small saliency - that is, the weights that have the least effect on the error, then finetuning this pruned network. The process is repeated iteratively until a model of the desired size or performance is obtained. This general format is replicated across the literature; however, the details of each stage vary between the study and the applications.

First, calculation of which weights have the least effect on the model error is computationally highly expensive [183] and so an approximation is desirable. Various metrics have been proposed to approximate this saliency, including the second derivative of the gradient [182, 338, 339], the magnitude of the weights or activations [340–343], and the product of the activation and gradients [183]. All of these methods aim to remove the unimportant parts of the network, but the approximations they use to calculate those that are unimportant differ.

Methods also differ as to the mechanism of the pruning. Some methods prune individual model weights or connections (*unstructured pruning*) [182, 338, 341, 344, 345] whereas others prune whole convolutional filters (*structured pruning*) [183, 339, 340, 342, 343, 346, 347]. Unstructured pruning results in networks that are sparse and it is hard to accelerate these methods without specialised libraries, which makes them difficult to apply using general-purpose hardware [183]; thus, this chapter will focus on structured pruning, removing whole convolutional filters through the pruning.

Methods also differ as to what is done to the pruned components of the networks, with some being set to zero and frozen [182, 338], some reset to the original initialisation value [341] and others entirely removed from the network architecture [340]. This work will focus on this final example, and so not only create a final model that has fewer parameters, but also has a *progressively* reduced parameter count during the training process.

The methods across the literature prune the network after training to convergence, then finetune the resulting model. This, however, requires the model to be trained to convergence – or at least to good performance by some metric [182] – before we are able to prune the network, leading to long training times and the methods not being applicable if one cannot train the full-sized network to an acceptable performance in the first place. Few methods attempt to prune the model during training: [348] learn

a sparse representation during training; [349] complete channelwise pruning during training, but set the weights to the final value they held before pruning and all are present at test time. Therefore, the final model does not reduce in size or complexity, and much of the capacity to overfit to the data remains.

The obvious question, given that it is often possible to prune networks to less than 50% of the original size of the model, is why we do not just train smaller models directly, removing the need for the pruning process. It has, however, been shown by several works in the literature that it is not normally possible to train the smaller pruned networks from random initialisation and achieve the same performance [340, 343]. The Lottery ticket hypothesis [341] demonstrates that there exists a subnetwork of the original network – a pruned network that, taken in isolation, with the rest of the network pruned away – that can be trained to the same level of task performance as the original network, given that the starting point is a dense network pruned using unstructured pruning, were a suitable initialisation found. These networks are very sensitive to initialisation and smaller networks may be achievable through maintaining the weights of the pruned networks, due to coadapted layers from earlier in the pruning process maintaining the complex relationships that were learned. The sensitivity to the initialisation when pruning convolutional filters as opposed to individual parameters is demonstrated in [350].

Other methods exist that have been applied to medical imaging, which aim to create networks with inherently reduced numbers of parameters, such as model distillation [178, 181, 351] and the use of separable convolutions [181, 352–356]. Model distillation focuses on teaching a smaller student network to mimic a larger teacher network, and so, sufficient data is needed to train a reasonable teacher network. Separable convolutions [352] and depthwise separable convolutions [353] break down large convolutional filters into a larger number of smaller kernels (e.g. replacing one $3 \times 3 \times 3$ kernel with

three kernels with sizes $3 \times 1 \times 1$, $1 \times 3 \times 1$, and $1 \times 1 \times 3$, then using the outer product of these), reducing the number of parameters in the network. Separable convolutions severely limit the range of filters that can be learned by the network, which is overcome by depthwise separable convolutions, but these smaller networks can easily be underpowered for some more difficult tasks. It will be demonstrated that there is a benefit to beginning with a larger network and pruning the network to become smaller, and giving the network the ability to determine where the parameters are needed, rather than reducing the number of parameters throughout the network.

It is therefore proposed to explore the utility of model pruning for model training in low data regimes. An algorithm will be proposed to explore simultaneously training and pruning a UNet architecture. Existing pruning methods have been validated in high data regimes, primarily on large computer vision classification datasets such as MNIST [39] – [182, 341, 342, 344] – CIFAR10 [63] – [341, 343] – and ImageNet [62] – [340, 342, 344]. Datasets of this size do not exist in medical imaging, where manual annotations are expensive and time consuming to acquire, and datasets are more commonly of the order of hundreds of labelled examples. In [357] model pruning is explored with a training set of 700 examples, but for most medical imaging applications this would still be regarded as quite a large dataset, and their dataset shows less heterogeneity than is common. Furthermore, their model is trained in 2D and so has significantly fewer parameters than the 3D models commonly used in medical imaging. Therefore, the utility of pruning in very low data regimes is explored here for a range of segmentation tasks and datasets for medical imaging.

The contributions, therefore, are as follow:

- An algorithm is proposed to simultaneously train and prune a UNet. It is, to

the best of our knowledge, the first exploration of pruning applied to a UNet for medical image segmentation, with existing approaches focusing on image level classification or object detection. The method removes the pruned values as in [340] but prunes whole convolutional filters. Therefore, it produces a model that becomes progressively smaller with training. The technique is demonstrated on several open access datasets and segmentation tasks.

- A channelwise targeted dropout is developed, based on that proposed in [358], but adapted to work channelwise and without requiring information about the filters that are most likely to be removed to be available before pruning. This is similar to the method proposed in [359], where channelwise weighted dropout is used to regularise the training of a model, but the normalised filter magnitudes calculated for pruning are used to modulate the dropout, and the same value is applied to all filters at a given depth, preventing the information from the lowest ranking filters simply being transferred to another at the same depth. If, through removing filters, overfitting is reduced, the information that is removed is not needed for the task and ideally it should not be redistributed to other filters. It is demonstrated that this improves the performance and stability of the models during training, by making them more robust to the pruning process.

- Finally, model pruning is explored for very low data regimes, and it is shown that, through training and pruning the model simultaneously, it is possible to achieve significantly higher performance than the baseline UNet models.

6.2 Methods

Our aim is to train a deep neural network to segment regions of interest from our input images while limiting the degree of overfitting, by pruning the model during training. Thus, the trained model should requires less training data points to produce good quality segmentations, due to the reduction in parameters that must be optimised. As this work uses medical images, where labels are expensive to acquire, it must function in a low data regime.

To allow the network to be trained and pruned simultaneously, consider the scenario where there is access to a training dataset $\mathcal{D} = \{\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2 ... \boldsymbol{x}_N\}, \boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2 ... \boldsymbol{y}_N\}\}$ where $\boldsymbol{x}_n \in \mathcal{R}^{W \times H \times D \times 1}$ represents an input image and $\boldsymbol{y}_n \in \mathcal{R}^{W \times H \times D \times C}$ is its corresponding target segmentation, and where C corresponds to the number of classes in the target segmentation. A network is trained to predict the target segmentations $\boldsymbol{y}_n$ from the input images $\boldsymbol{x}_n$ and the network is parametrised by $\mathcal{W} = \{(\boldsymbol{w}_1^1, b_1^1), (\boldsymbol{w}_1^2, b_1^2), ... (\boldsymbol{w}_L^{C_L}, b_L^{C_L})\}$ where $\boldsymbol{w}_m^n$ corresponds to the weight kernel for the nth filter of the mth layer, b_n^m is the corresponding bias, and C_L is the number of channels in layer L. The weights are first randomly initialised and are optimised during training by minimising a loss function $\mathcal{L}(\mathcal{D}|\mathcal{W})$ that aims to optimise the performance on the segmentation task and is not chosen because of the pruning process.

During the pruning process, the aim is to refine the parameters of the network $\mathcal{W}$ to a smaller subset of the parameters $\mathcal{W}'$ such that $\mathcal{L}(\mathcal{D}|\mathcal{W}') \approx \mathcal{L}(\mathcal{D}|\mathcal{W})$. To achieve this, the observation that the parameters of the smallest magnitude have the least impact on the final segmentation produced by the network, and can be removed with the least impact on model performance, is used.

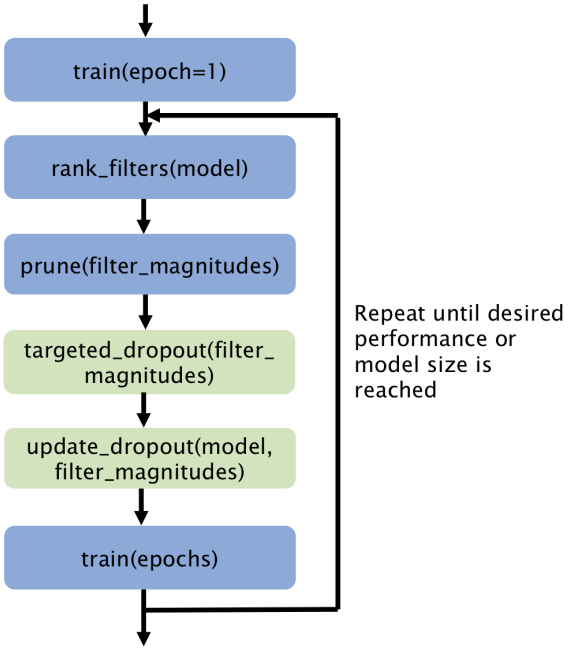


Figure 6.1: Overall training procedure for training and pruning a model simultaneously. The targeted dropout algorithm is presented in Algorithm 1. Steps in green are removed in ablation experiments.

6.2.1 Training Procedure

The overall training procedure is shown in Fig. 6.1. The model is initialised with Xavier initialization [360] and the model is trained for a single epoch before pruning begins, so that the filters contain some information about the task and not just the initialised weights.

Our pruning removes whole convolutional kernels as shown in Fig. 6.2, meaning that we do not end up with a sparse representation and are able to use standard libraries and hardware. When removing whole convolutional filters, we remove the filter from the kernel matrix $\boldsymbol{w}_i$, the corresponding bias, and the filter from the weight kernel matrix $\boldsymbol{w}_{i+1}$. To decide which filters to remove, the feature activation maps are considered, denoted by $\boldsymbol{z}_l^{(k)} \in \mathcal{R}^{W_l \times H_l \times D_l \times C_l}$ where $l \in \{1, L\}$, is the current layer, and C_l is the number of channels at layer depth l. The activation maps are

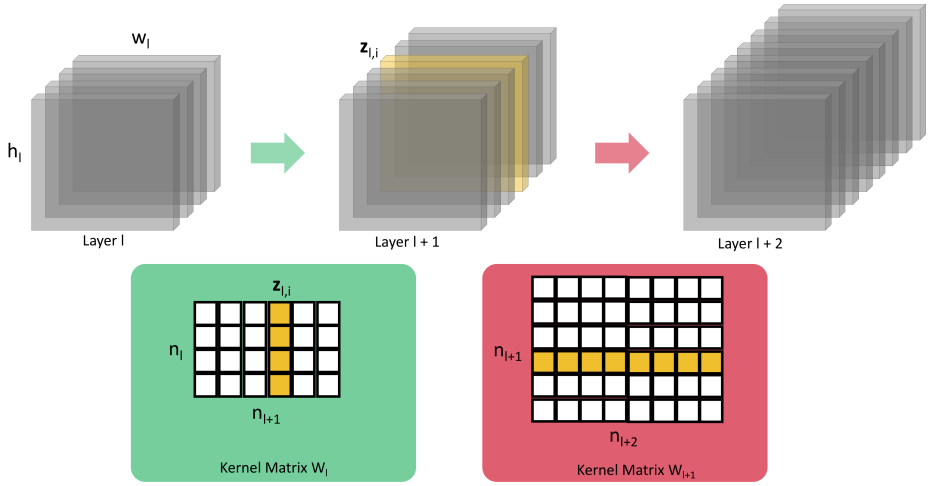


Figure 6.2: Whole convolutional filters are pruned, based on the magnitude of the corresponding feature maps $\boldsymbol{z}_{l,i}$. This requires the weights to be pruned from the kernel matrix $\boldsymbol{w}_l$ and $\boldsymbol{w}_{l+1}$. The corresponding biases are also pruned but are not shown.

related to the kernel weights by:

$$\boldsymbol{z}_l^{(k)} = ReLU(\boldsymbol{z}_{(l-1)} * \boldsymbol{w}_l^{(k)} + b_l^{(k)}) \tag{6.1}$$

where ReLU is the activation function used throughout our network, other than for the final convolution. The use of ReLU means that our network contains a substantial number of zero-valued parameters, creating redundancy that can be pruned.

The observation that weights with the smallest impact on the final prediction are those with the smallest magnitudes [182] is used, and so the weights, $\boldsymbol{w}_l^{(k)}$, and biases $b_l^{(k)}$, corresponding to the smallest magnitude filter activations, $\boldsymbol{z}_l^{(k)}$, are removed. To assess the magnitude of the filters, the $L2$ norm averaged across all training data points (calculated from an additional forward pass) is considered:

$$\Theta_{L_2}(\boldsymbol{z}_l^{(k)}) = \frac{1}{N}\sum_{i=1}^{N} || \boldsymbol{z}_{l,i}^{(k)} ||_2 \tag{6.2}$$

The $L2$ norm is used as it is computationally simple and provides stable performance,

but other metrics could be used to evaluate the magnitude of the filters: it was found that the performance was robust to the choice of metric used – see Section 6.3.1.

These values must then be normalised across the layer depth, because the values are at different scales at different depths [183]. Therefore, a simple $L2$ normalisation across the values at each layer is employed.

$$\hat{\Theta}_{L_2}(\boldsymbol{z}_l^{(k)}) = \frac{\Theta_{L_2}(\boldsymbol{z}_l^{(k)})}{\sqrt{\sum_j (\Theta_{L_2}(\boldsymbol{z}_j^{(k)}))^2}} \tag{6.3}$$

To prune the filter, it is removed entirely from the model architecture, not just set to zero, and so the model architecture reduces in the number of parameters as the training continues. Practically, this is achieved by creating a smaller model with the selected filter removed and then reloading all of the weights apart from those corresponding to the pruned filter.

6.2.2 Adaptive Channelwise Targeted Dropout

Algorithm 1: Adaptive Targeted Dropout Algorithm. The *index* function returns the index (or position) of the kth filter for layer l within the full sorted list, so that the average index at each layer depth can be calculated.

Input: Normalised filter magnitudes for every filter at every depth, output from pruning algorithm: $\boldsymbol{\Theta} = [\hat{\Theta}_{L_2}(\boldsymbol{z}_l^{(k)}), [l, k]] \; \forall \; l \in \{1, L\}, k \in \{1, C_l\}$
Result: Targeted dropout values: $\boldsymbol{p}_l \; \forall \; l \in L$

$\boldsymbol{\Theta}_{sorted} = sort_descending(\boldsymbol{\Theta})$ ▷ according to $\hat{\Theta}_{L_2}(\boldsymbol{z}_l^{(k)})$ values
for $l \leftarrow 1$ *to* L **do**
  $\boldsymbol{\mu}_l = \frac{1}{C_l} \sum_{k=1}^{C_l} index(\boldsymbol{\Theta}_{sorted}, [l, k])$ ▷ Find index for $[l, k]$
end

$\boldsymbol{\mu} = \boldsymbol{\mu} / max(\boldsymbol{\mu})$
$\boldsymbol{p} = \boldsymbol{\mu} \times b_{drop}$ ▷ Multiply by base dropout value

To make the model more robust to the pruning, adaptive channelwise targeted

dropout, based on [358] and [359], is introduced. In [358] dropout is applied to units which are *a priori* believed to be the least useful, thus encouraging the network to learn a representation that is more robust to *post hoc* sparsification. In [359] weighted channelwise dropout is applied during training to regularise the model training. Global Average Pooling is used to determine the importance of individual filters, and a different dropout value is applied to each filter. Here the same dropout value is applied to all of the filters at a given depth such that the information carried by the pruned filters is removed, and not redistributed to other filters at that depth, otherwise the model will continue to overfit to the training data. Like [359] the dropout is applied channelwise, rather than parameter-wise [358], so that the network is prepared for the removal of whole convolutional filters; thus, channelwise or *spatial* dropout [361] is added to our model.

It is not known which filters are the most likely to be pruned *a priori* and, due to not having a pretrained model which is then pruned, the filter magnitudes cannot be used as a basis to predict these; so, unlike [358], the targeted dropout values cannot be based on a known probability distribution. The dropout values are therefore changed adaptively during training, based on the calculated filter magnitudes, $\hat{\Theta}_{L_2}(\boldsymbol{z}_l^{(k)})$. For the first iteration, all dropout values are set equal to the base dropout value b_{drop} (the maximum dropout value applied) and then the values are recalculated after each pruning iteration, as the new magnitudes $\hat{\Theta}_{L_2}(\boldsymbol{z}_l^{(k)})$ are calculated.

Algorithm 1 shows the procedure used to calculate the adaptive targeted dropout values. It should be noted that the indices of the ordered normalised filter magnitudes (i.e. position in the sorted list) are used to calculate the dropout values rather than the $\hat{\Theta}_{L_2}(\boldsymbol{z}_l^{(k)})$ values directly, as the discrete values better encode the relative likelihoods of the filter being pruned from the model. When the new dropout values, $\boldsymbol{p}$, are calculated, the model architecture is updated with the new spatial dropout values,

with the same value being applied to all filters in a layer.

6.2.3 Model Training

After pruning and updating the targeted dropout values, the model is then trained using standard backpropagation. The number of epochs between filter prunings is a hyperparameter that needs to be decided. It is shown that 2 epochs between the pruning of filters worked across the experiments, but this could be increased if the network was unable to learn how to segment the regions of interest. The effect of this hyperparameter on the segmentation performance will be explored.

In the experiments, the network is trained until it cannot be pruned any more – a model cannot be pruned any further when only a single filter remains at each depth and so to prune further would be to break the model – to allow exploration of the model's performance, even as the model becomes very small. In practice, the model could be trained with standard early stopping, to save the best performing model.

6.2.4 Network Architecture

Across the experiments, a standard UNet architecture is explored (3D or 2D depending on dataset) [41, 42] (Fig. 6.3) with two blocks of convolutional layers at each depth, each with a batch normalisation layer [26]. The UNet architecture is considered as it is so widely used throughout medical image segmentation and many winning architectures in segmentation challenges are based on the UNet architecture.

ReLU activations are used throughout, apart from the final layer, where a sigmoid activation function is used to create the segmentations. ReLU is very commonly used in network architectures, and, when we are considering pruning, comes with the added advantage that it encourages the network parameters to become sparse,

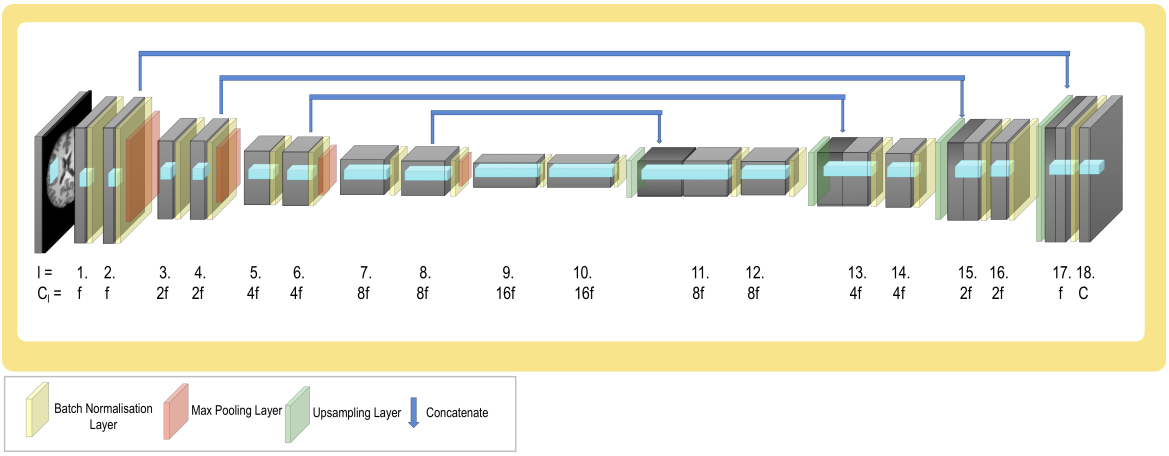


Figure 6.3: UNet model architecture - it follows the standard pattern of halving resolution and doubling filters at each depth. l corresponds to the layer depth, C_l is the number of channels in that layer and C is the number of classes in the output segmentation. f is the initial number of filters, and is varied across experiments, but is 4 unless otherwise stated.

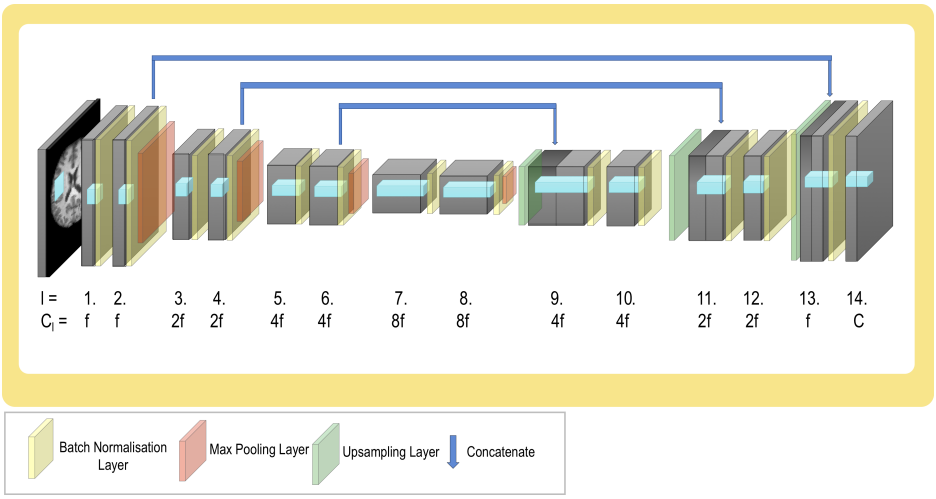


Figure 6.4: Shallower UNet model architecture - it follows the standard pattern of halving resolution and doubling filters at each depth but has less depth than the UNet architecture previously considered. l corresponds to the layer depth, C_l is the number of channels in that layer and C is the number of classes in the output segmentation.

so filter values on average become lower. Spatial dropout is applied to each convolutional block during training, with the values determined by the Targeted Dropout algorithm explained above.

The final layer ($l = 18$) contains the same number of filters as output classes C

and so clearly cannot be pruned – in all cases the background class is considered as an additional class. Therefore, only filters in layers $1 - 17$ will be considered as candidates to be pruned and therefore for Eq. (1) is evaluated over $l \in \{1, L\}$ where $L = 17$ and the final layer is not considered.

The architecture follows the conventional pattern of halving resolution and doubling features at each depth. The value of f, which determines the number of filters in each layer, is varied in the experiments, allowing exploration of the effect of the initial model size on the pruning process and the model's performance, but will be set to 4 unless stated otherwise. The maximum size of the network architecture explored was determined by the available GPU memory.

6.2.4.1 Shallower Network

To allow further exploration of the effect of the initial size of the model, pruning a shallower original model is also considered. A downsampling and an upsampling layer are removed, so the original model is smaller, with fewer parameters, but still having the ability to learn the same number of features at the layers of abstraction which remain. A shallower network is considered, rather than a deeper one, as it allows the exploration of another way in which the model could be smaller in terms of parameters. This architecture can be seen in Fig. 6.4. All other parameters are held the same as in the experiments with the full-sized model and the value of f will again be varied.

6.2.5 Datasets and Implementation Details

In all of the experiments, 2 epochs of training were used between filter prunings, with a batch size of 16 and an Adam optimiser [30] with a learning rate of 0.01. Five-fold cross-validation – the training data was split 80% training, 20% validation – was used for

all experiments and all reported values use a further held-out test set for each dataset.

For the various experiments, two variations of the proposed method and two benchmarks are considered, which will be referred to as follows:

- *TrainPrune* - The proposed method, training and pruning the network simultaneously, without the green blocks shown in Fig. 6.1;
- *TrainPrune + Targeted Dropout* (TrainPruneTD) - The proposed method, including the targeted dropout – all of the blocks in Fig. 6.1. These first two methods allow exploration of the impact of the pruning and the targeted dropout separately;
- *Standard UNet* - The first baseline, the UNet model trained to convergence, without pruning, with the same hyperparameters and the patience value for early fitting set to 25 epochs, representing the standard approach;
- *PruneFinetune* - The second baseline, a UNet trained to convergence, without pruning, is used as the initialisation, then the model is pruned using the same framework as for our proposed method. A patience of 10 epochs was used during the fine-tuning stages. This is the classic pruning regime, allowing exploration of the advantage of pruning and training simultaneously.

The same data splits were used for evaluating all methods. No augmentation was applied to the data at any stage, so that any increase in generalisability is due to the pruning process and is not a function of any augmentation applied. Dice scores are reported throughout, but it can be seen in Appendix C.3.1 that the result is consistent across metrics. All reported statistics are paired t-tests. When a single value is reported for the performance of the pruning methods, the best model is chosen on the basis of results obtained from the validation data. The results are then reported for the

Dataset	Modality	2D/3D	Task	Image Size	Training/ Testing Subjects
HarP [168]	T1 MRI	3D	Hippocampal Segmentation	$64 \times 64 \times 64$	200/70
IXI - HH	T1 MRI	3D	Brain Extraction	$128 \times 128 \times 128$	166/19
IXI - Guys	T1 MRI	3D	Subcortical Segmentation	$128 \times 128 \times 128$	257/65
Medical Decathlon - Cardiac [362]	Mono-modal MRI	2D	Right Atrium	320×320	15/5 (split into slices)
Medical Decathlon - Spleen [363]	CT	2D	Whole Spleen	512×512	28/10 (split into slices)
Medical Decathlon - Prostate [364]	Multimodal MR (T2 & ADC)	2D	Whole Prostate	320×320	28/10 (split into slices)
INTERGROWTH-21st Fetal Ultrasound [365]	Ultrasound	3D	Fetal Brain	$160 \times 160 \times 160$	822 / 274

Table 6.1: Details of the datasets used in pruning experiments.

held out testing data. All graphs show the mean results and interquartile range. The base dropout value b_{drop} was set to 0.05 in all cases.

The code was implemented using Python 3.5.2 and PyTorch 1.0.1.post2 [324]. All models were trained using a V100 GPU.

6.2.5.1 Datasets

A range of medical imaging datasets were considered for these experiments. Basic details are listed in Table 6.1. The datasets were chosen so as to span a range of tasks and imaging modalities.

HarP: A single-site dataset with high quality manual segmentations. $64 \times 64 \times 64$ voxel patches are extracted around each hippocampus, providing 200 volumes – left and right hippocampi combined – for training and 70 for testing. T1 images are used as the input. $C = 2$ (foreground and background).

IXI[1] **- HH**: Hammersmith hospital data from a 3T MRI scanner. The data was processed using the FSL Anat pipeline. The task considered was brain extraction and the labels were automatically generated using the FSL Anat tool, which non-linearly

[1]Data downloaded from: `https://brain-development.org/ixi-dataset`

registers an atlas to the target image. $C = 2$ (foreground and background).

IXI - Guy's: Guy's hospital data from a 1.5T MRI scanner. The data was processed using the FSL Anat pipeline. The task considered was subcortical segmentation of three deep grey matter structures: caudate, putamen and thalamus, with the labels automatically generated using FSL FIRST [12] run on the bias field corrected and registered images output from the FSL Anat pipeline. $C = 4$ (caudate, putamen and thalamus, and background).

Medical Decathlon - Cardiac: 20 MRI datasets covering the entire heart which were acquired during a single cardiac phase (free breathing with respiratory and ECG gating). This represents a small dataset with large variability, where the task was the segmentation of the right atrium. $C = 2$ (right atrium and background).

Medical Decathlon - Spleen: 41 CT scans with automatic segmentations that were manually corrected. This demonstrates the method on a different imaging modality with the dataset having a large range of foreground sizes; the task was the segmentation of the spleen. The dataset was originally reported in [363]. $C = 2$ (spleen and background).

Medical Decathlon - Prostate: 38 Multimodal MRI datasets with labels of the whole prostate. This demonstrates that the method can be applied to multichannel input data; the dataset requires the segmentation of two touching regions with large intersubject variation. The dataset was originally reported in [364]. $C = 3$ (two prostate regions and background).

INTERGROWTH-21st: 1096 Ultrasound scans which were unaligned with varying fields of view. The data were collected as part of the INTERGROWTH-21st study and represent a healthy fetal cohort, with gestational ages between 14 and 31 weeks. In each scan, the brain occupied $> 50\%$ of the image space. The images had a range of sizes and were resampled, then cropped to create images of the

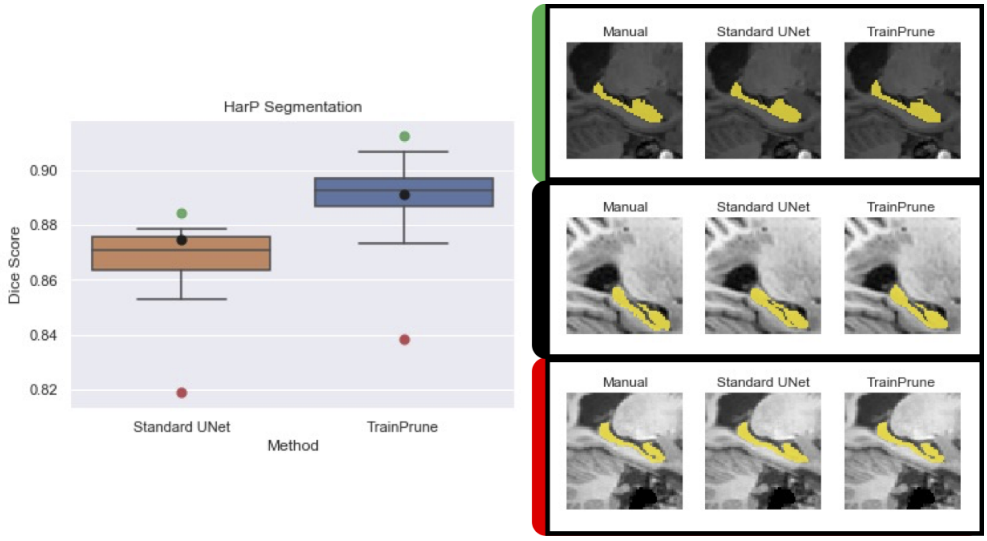


Figure 6.5: Representative segmentations from the HarP test set. The colour coded boxes correspond to the same colour point on the box plot. The result shown in green (top) is the best segmentation, in red (bottom) the worst, and black (middle) represents a standard result.

same size, with the image centred. The brain mask labels were created by manually aligning (similarity transform) the age-matching fetal MRI atlas from [366] to each US image. $C = 2$ (fetal brain and background).

6.3 Results

6.3.1 HarP Data - Evaluation of the Algorithm

The HarP dataset was first used to allow the validation of the different aspects of the approach, using a dataset with high quality, consistent manual labels. Once the different aspects of the pruning algorithm are understood, the effect of the available training data can be explored. Therefore, this section explores the effect of:

- the choice of pruning metric,
- the number of filters,

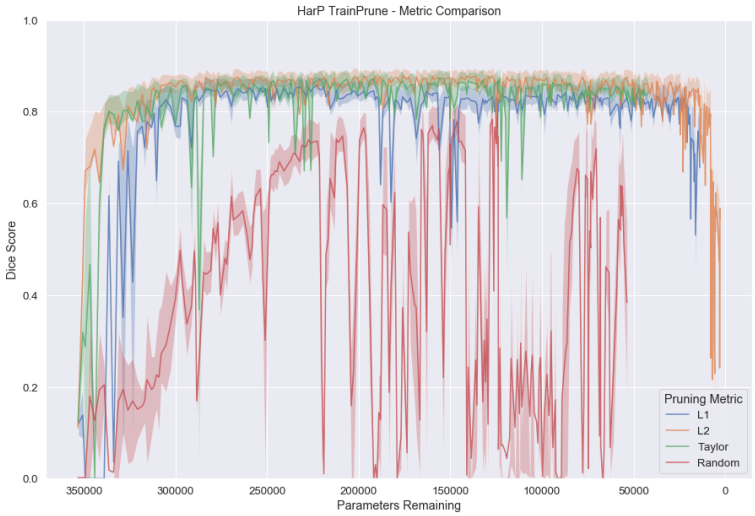


Figure 6.6: The performance the *TrainPruneTD* algorithm with different choices of pruning metric using the HarP data, with $f = 4$. It can be seen that the three metrics performed similarly, with no significant difference between L2 and Taylor metrics. All three metrics were better than random pruning. Note that the results are plotted against the remaining parameters rather than the pruning iteration.

- the use of the targeted dropout,
- the base value of the targeted dropout out, b_{drop},
- the effect of the number of recovery epochs between filter pruning
- the initial depth of the UNet architecture.

Example segmentations from these experiments can be seen in Fig. 6.5.

6.3.1.1 Choice of Metric

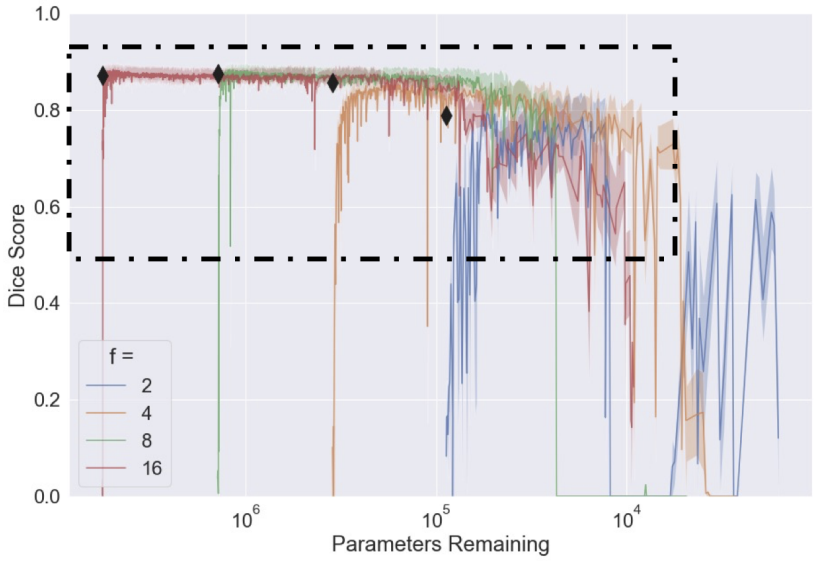
First, the choice of metric used to determine which filters to prune is explored. As stated in the methods, the L2 norm was chosen as the pruning metric and the results shown in Fig. 6.6 form the basis of that decision. The L2 norm is compared with:

- L1 norm [343]: $\Theta_{L_1}(\boldsymbol{z}_l^{(k)}) = \frac{1}{N}\sum_{i=1}^{N} || \boldsymbol{z}_{l,i}^{(k)} ||_1$

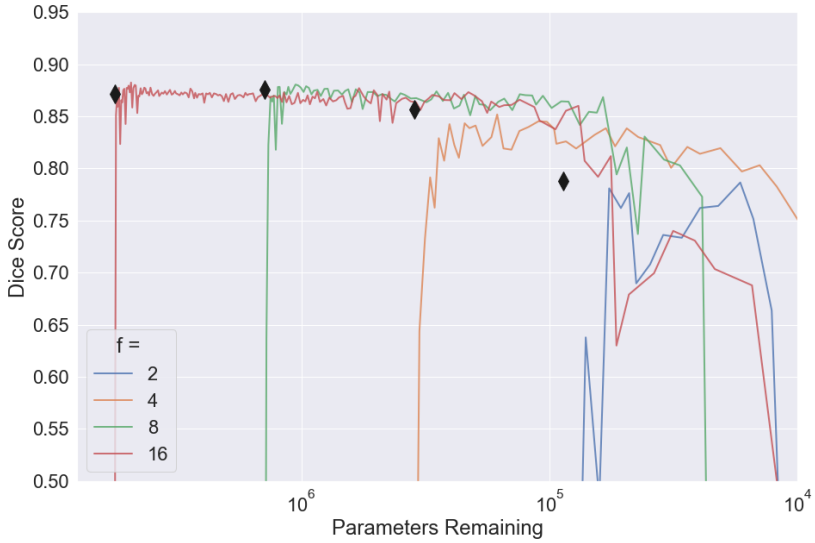
- Taylor – proposed in [183], which utilises the product of the activation and the gradient of the loss function: $\Theta_{TE}(\boldsymbol{z}_l^{(k)}) = ||\frac{1}{N}\sum_{i=1}^{N}\frac{\delta C}{\delta \boldsymbol{z}_{l,i}^{(k)}}\boldsymbol{z}_{l,i}^{(k)}||_1$
- Random: $\Theta_{Rand} \sim Uniform(0, 1)$

where all the metrics but Random are normalised according to Equation 6.3. This normalisation ensures that the last filter at any depth is not pruned. With random pruning, the condition that the last kernel at a given depth cannot be pruned is explicitly coded. Figure 6.6 shows the average Dice scores on the test dataset for each pruning iteration for models trained using *TrainPruneTD* with each metric in turn. For each metric the mean value across the test set is shown as the solid line, and the shaded region indicates the interquartile range. The results are plotted against the remaining number of parameters in the network after each pruning iteration. This is done because different pruning iterations remove different numbers of parameters, depending on from where in the model the filter is removed: for instance, if it also leads to filters being removed across the skip connection.

It can first be seen that all three metrics perform better than randomly pruning channels, as would clearly be expected. As the models start from random initialisation, the performance initially for all four metrics is poor, then improves rapidly for all metrics but the random one, as the model training continues. It can then be seen that all three of the metrics perform comparably on this task, with no significant difference between the performance of the L2 and Taylor metrics (L2 - Taylor: $p = 0.07$, L2 - L1: $p = 0.003$, L2 - Random: $p = 5.7 \times 10^{-11}$). As the L2 norm is computationally more efficient than the Taylor metric, it will be used throughout the remaining experiments.



(a)



(b)

Figure 6.7: a) Results for pruning models of varying model sizes – the value of f – while training to segment the hippocampus on the HarP data. The black diamond corresponds to a model of the same size being trained to convergence on the same data, with all parameters held the same except for the absence of pruning the network. b) The same result, zoomed in and subsampled (only every 10th data point shown and no bounds, only the mean, no interquartile range shown) to allow the result to be seen more clearly. The box on a) indicates the zoomed in region.

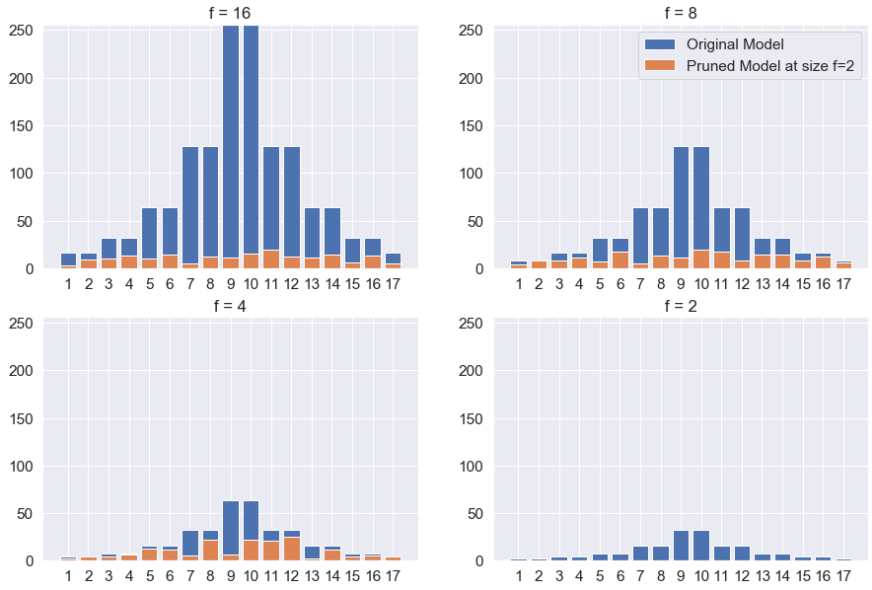


Figure 6.8: The number of filters in each layer for the models trained in figure 6.7. The distribution of the number of filters shown in blue corresponds to the model's original filter distribution, and the one shown in orange is the distribution when the model has the same total number of parameters as the original model with $f = 2$.

6.3.1.2 Varying Numbers of Filters f

Having validated the use of the L2 norm as the pruning metric, the *TrainPruneTD* algorithm was tested on the HarP data for models with varying numbers of filters - $f = [2, 4, 8, 16]$. Each model is simultaneously pruned and trained until the model is unable to be pruned any further, with only one filter remaining in each layer. Figure 6.7a shows the performance of the model, the performance is evaluated on the test set for each iteration of the pruned model; the mean value (the solid line) and the interquartile range bounds (shaded) are shown. The result is shown for four different initial sizes of model, plotted against the number of parameters remaining in the model. For comparison, the correspondingly sized *Standard UNet* models were also trained until convergence. The same result can be seen in Fig. 6.7b but is zoomed in to allow the result to be seen more clearly.

First, it can be seen that the model can be pruned and trained simultaneously, such that the same network performance is reached as the model trained to convergence; therefore, the pruned network is sufficiently powerful to be able to represent the variation in the data. It is also evident from Fig. 6.7 that it is possible to prune the models to a fraction of their original size without reducing the network's performance. It can also be seen that the smallest model, when $f = 2$, is not able to perform the segmentation well (maximum Dice score $= 0.788 \pm 0.052$ for the standard UNet and 0.799 ± 0.050 for the pruned model), but the larger networks can be pruned to the same number of parameters as the $f = 2$ model – and smaller – and still perform well on the segmentation task. This indicates that by pruning a larger model we are able to learn a better arrangement of parameters than when naïvely creating a model of that size.

Figure 6.8 shows the distribution of filters of the differently sized networks at initialisation and at the point in pruning where they have the same number of parameters as the $f = 2$ model. These filter distributions correspond to the $f = 2$ model trained to convergence having a Dice score of 0.788 ± 0.052 compared to the pruned models' performances of 0.833 ± 0.041, 0.859 ± 0.047, and 0.861 ± 0.028 for $f = 4, 8, 16$ respectively, meaning that all of the larger pruned models perform significantly better than the *Standard UNet* model of that size (the best performing model was selected using the validation data). It can be seen that the filter distributions learned through pruning the models are very different to the standard UNet architecture of doubling filters but instead have a far flatter distribution, with a similar number of features at each depth being maintained.

Furthermore, in a separate experiment, the weights of the pruned model with both $f = 16$ and $f = 8$ are reset when they have the same number of parameters as the full $f = 2$ model using Xavier initialisation [360]. It was not possible to train them to perform significantly better than the $f = 2$ model, indicating that, through pruning a



Figure 6.9: Comparison of pruning the HarP data with $f = 4$ for our proposed method of *TrainPrune with Targeted Dropout*, *PruneFinetune* and *TrainPrune* without targeted dropout - green blocks from Fig. 6.1 removed. The black diamond indicates the performance of the *Standard UNet* trained to convergence before the pruning began - $f = 4$.

larger model to that size, complex co-adapted layers and representations are learned that cannot be learned with the smaller model alone.

These results show that the proposed technique is able to complete the segmentation tasks, and indicate that key features and relationships that are learned in the larger networks are preserved when pruning the network. These relationships are, however, not easily learnt when directly training smaller sized models to convergence from random initialisations.

6.3.1.3 Method Comparison

In Fig. 6.9 the results achieved using different pruning methods are compared: *TrainPruneTD*, *PruneFinetune* and *TrainPrune*. First, it can be seen, by comparing *TrainPrune* with and without targeted dropout, that the targeted dropout both improves the model's performance (from 0.844 ± 0.03 to $0.879 \pm 0.02, p = 2.9 \times 10^{-18}$)



Figure 6.10: The adaptive targeted dropout values for every 30 epochs of training on the HarP data – $b_{drop} = 0.05$ – for $f = 8$. It can be seen that the dropout values are quite consistent across pruning iterations until the network becomes very small in size. When a layer has a single filter at that depth, the dropout value becomes 0, guaranteeing that information can be propagated through the model.

and additionally makes the training more stable. This shows that the targeted dropout successfully makes the model more robust to being pruned, even with this relatively simple task.

Comparing *TrainPrune* to *PruneFinetune*, we can see that pruning the already converged model leads to more stable training than *TrainPrune* (Fig. 6.9). However, the removal of a single filter at a time is very conservative compared to the normal approaches taken in the literature, where a percentage of filters, or all filters under a threshold value are removed, which we would then expect to be less stable. Furthermore, the recovery time between filter prunings is much longer, as the model is allowed to

recover entirely before additional pruning is completed. Finally, the original had to be trained to convergence before the network could be pruned, and so *PruneFinetune* has a much larger computational cost while performing less well than the proposed method of *TrainPruneTD* ($0.860 \pm 0.03, 0.879 \pm 0.02, p = 1.3 \times 10^{-9}$). Therefore, the proposed method allows better performance to be achieved without having to train the original model.

6.3.1.4 Filter Magnitudes

Remembering that the UNet architecture conventionally has doubling features with increasing model depth and halving resolution, Fig. 6.10 shows the adaptive targeted dropout values through training, for a model trained with $f = 8$, with every 30 iterations being reported. The x value corresponds with the location model, as indicated by the model architecture (Fig. 6.3). It can be seen that until the model becomes very small (highly pruned, iterations higher than 600) the distribution of the calculated values is very similar between iterations, with the tendency towards the bottleneck of the model having the highest dropout value. This is to be expected, as this is where most filters in the model are located and, thus, where there is likely to be the greatest level of redundancy. It can also be seen that towards the end of training the dropout values for some of the depths are 0. These correspond to the depth where only a single filter remains, thus to prune – or to apply dropout here – would break the model and information would be unable to flow. This is coded into the algorithm for setting the targeted dropout values.

Figure 6.11a then shows the distribution of the filters at each depth as the network is pruned. The darker the section of the bar, representing the filters at a given layer, the longer they remained in the model architecture. As would be expected, because

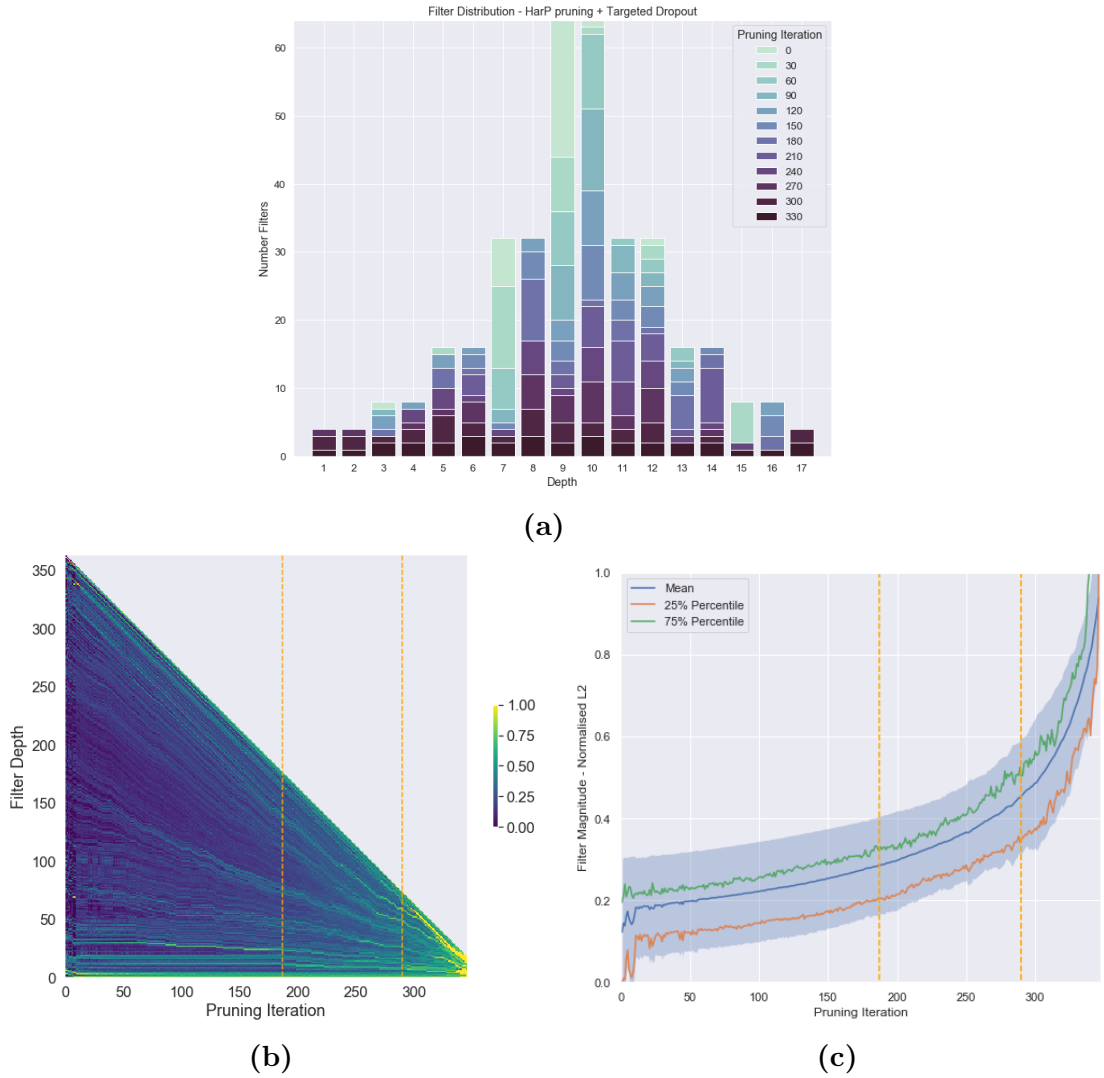


Figure 6.11: a) Shows the distribution of the filters with network depth as the model is pruned, meaning that the darker the shade of the filter block, the longer the filters at that depth were maintained in the models. b) Shows the magnitudes of the activations, averaged across the training data, maintained in the model as it is pruned. It can be seen that the lower magnitude activations are pruned first and the average value of the activation increases as the model is pruned. The first vertical dashed line at 182 filters shows the distribution of filters which gave the best performance on the testing data, and the second dashed line indicates the distribution of the filters for the smallest model that was able to complete the segmentation successfully, with no substantial difference in performance from the best performing model. c) Shows the average filter magnitude, and the lower and upper quartile bounds with pruning iteration. It can be seen that the average value increases consistently with pruning iteration.

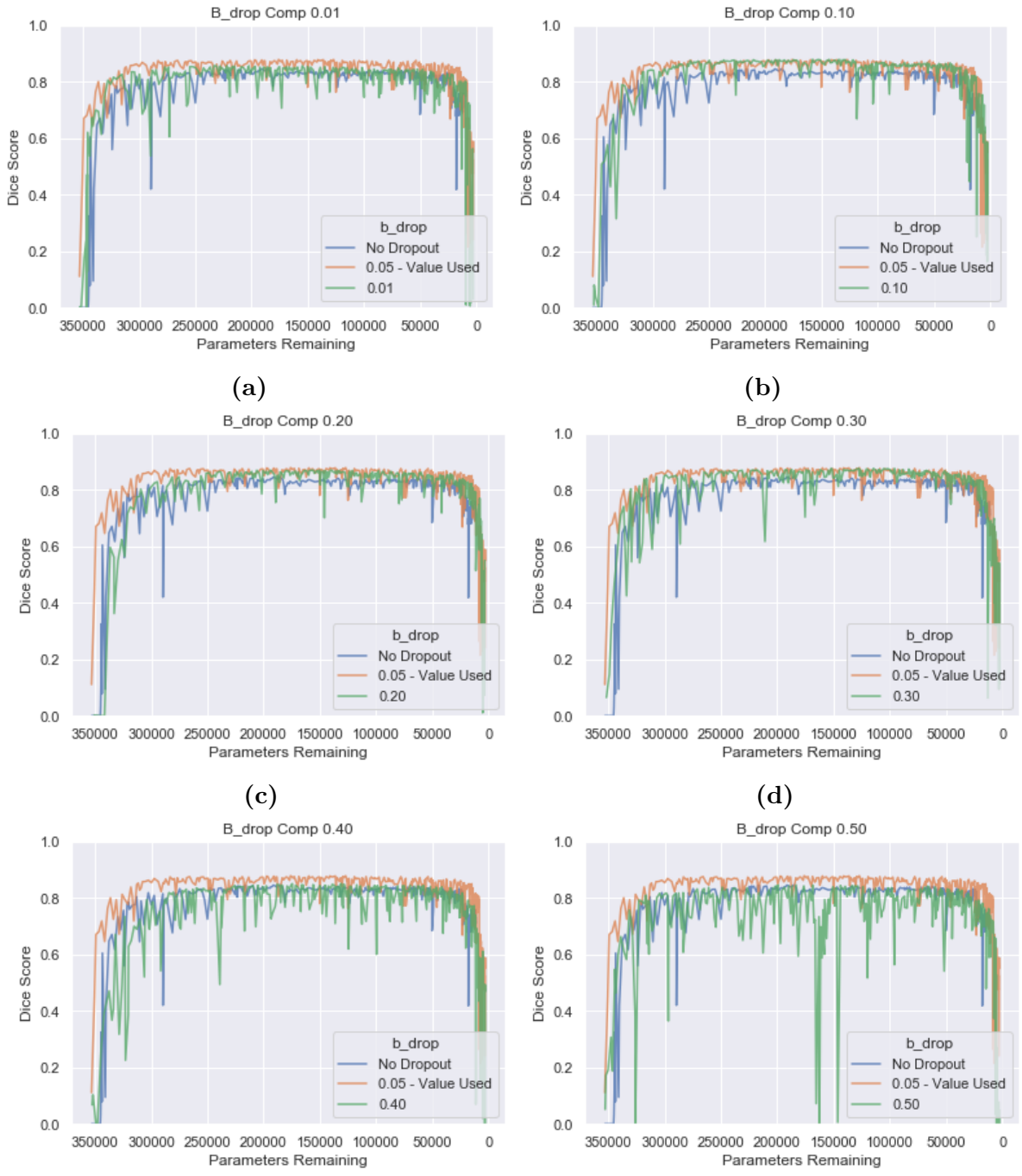


Figure 6.12: Here, the effect of changing the value of b_{drop} is explored. For clarity, three versions are compared in each panel: no dropout (blue), the base dropout value we used across the experiments (0.05 - in orange), and the base dropout value of interest for that panel (green).

they are based on the activation magnitude values, the filters in the bottleneck of the network are pruned aggressively first, corresponding with the high dropout values. It can also be seen across the network, that the first layer within a pair of layers at a given depth, is pruned more quickly than the second, for both the encoder and the decoder.

Figure 6.11b shows the magnitudes of the filters - $\hat{\Theta}_{L_2}(\boldsymbol{z}_l^{(k)})$ averaged across the testing dataset as the model is gradually pruned from left to right. It can be seen that as the model is gradually pruned, the average value of the remaining filters increases and the high magnitude filters remain throughout the pruning. The first dashed vertical line at iteration 182 represents the point in the pruning where the model had the best performance on the testing data (49% of the filters of the original model). It can be seen that the majority of the low magnitude filters – and so assumed uninformative – have been removed, showing that more than half of the filters in the model can be removed without having a negative impact on the segmentation. This demonstrates that it is possible to prune in a low-data regime when the network architecture is a UNet. The second dashed line represents the smallest network (20% of the filters of the original model) for which the performance was not significantly worse than the best performing model. This indicates that the network can be reduced to a small number of filters, where all of the activations are playing a more important role in producing the outputs.

These results clearly demonstrate that the proposed method can be utilised to simultaneously train and prune a UNet, while working in a low-data regime, as is common in medical imaging. The results also indicate the benefit of the addition of the targeted dropout, showing an increased performance even on this relatively easy task.

6.3.1.5 Base Dropout Value b_{drop}

The addition of the targeted dropout does, however, introduce a hyperparameter: the base dropout value b_{drop}. Figure 6.12 explores the effect of changing the value of b_{drop}

on the proposed *TrainPruneTD* method. Across the experiments, a base value of 0.05 was used, meaning that the maximum dropout probability that can be applied to a layer is 0.05. This is a lot lower than the dropout probability used with parameterwise dropout [87] where values as high as 0.5 can be used; however, spatial or channelwise dropout is much more aggressive, and as shown in Fig. 6.9 the addition of targeted dropout with $b_{drop} = 0.05$ made a significant ($p < 0.001$) improvement to the model performance. It can be seen in Fig. 6.12 that very low values (for instance 0.01 shown in Fig. 6.12a) are too low to give the benefit of the addition of the targeted dropout. Very similar performance was then observed for a large range of dropout values $0.05 - 0.30$, showing that the method is not too dependent on the value chosen. As the b_{drop} value is increased to very high values, $0.40 - 0.50$, it can be seen, as would be expected, that the model becomes unstable and the performance degrades. Therefore, the results are sensitive to the choice of the b_{drop} value, but there is a large range of values for which good performance is achieved and there is clear benefit from its addition. The value 0.05 is used across the experiments shown in this chapter; it was found that this produced good results in all cases.

6.3.1.6 Number of Recovery Epochs

The number of recovery epochs the network is allowed to train for between filter pruning is then explored. Figure 6.13 shows the performance of the network with pruning for increasing numbers of recovery epochs. It can be seen that the maximum performance reached was very similar in all cases (1 epoch = 0.870, 2 epochs = 0.862, 5 epochs = 0.876, 10 epochs = 0.874, with the best performing model selected using the validation data). The difference with the recovery epochs is the stability of the training, with the stability in performance between pruning iterations increasing with an increasing number of recovery epochs. Therefore, the setting of the recovery epochs

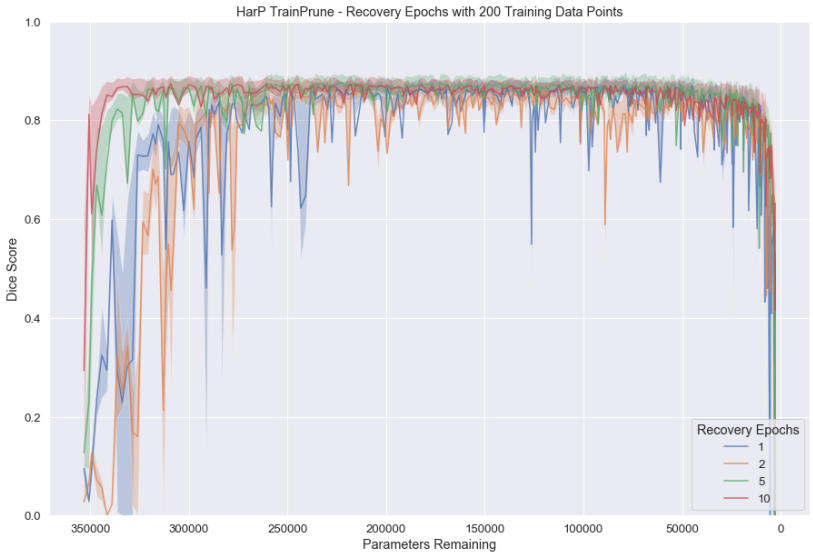


Figure 6.13: Performance of the *TrainPruneTD* framework with increasing numbers of recovery epochs between filter prunings.

is a trade off between the stability and the duration of the training.

6.3.1.7 Network Architecture Depth

In this section the case where the initial model architecture is shallower is considered, where this contains fewer parameters in total while having the same number of features at each depth that remains, as shown in Fig. 6.4. Given that this model has fewer parameters to begin with for the same richness of filters at each depth, the overfitting to the training data might be expected to be less.

First, the results obtained using the HarP data were investigated, for different values of $f = [2, 4, 8, 16]$. Figure 6.14 shows the results, where it can be seen that the results with the shallower UNet follow the same pattern as was seen with the *standard UNet* (Fig. 6.7). Again, it is clearly visible that a better performance on the segmentation task is achieved if we begin with a larger model and prune it to be the same size as the smaller model, rather than originally training the smaller model to

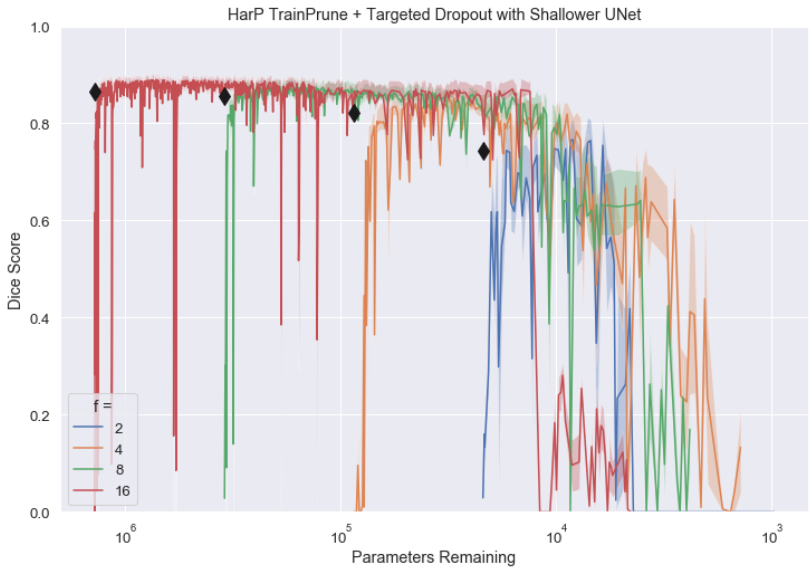
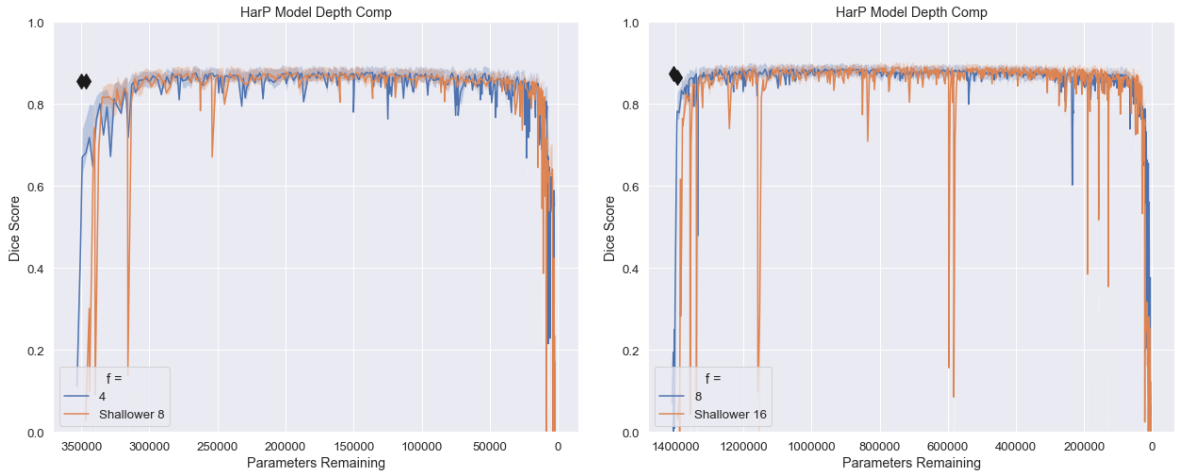


Figure 6.14: Dice scores on the HarP data using the shallower UNet architecture for $f = [2, 4, 8, 16]$, with interquartile bounds shown, using *TrainPruneTD*. Diamonds represent a standard UNet of that number of parameters trained to convergence.

convergence. The pruning was, however, less stable, showing that the model training was less robust to filters being removed than when the model was deeper.

Fig. 6.15 compares the results from the shallower UNet to the original UNet when the original models were matched in terms of parameters, where Fig. 6.15a compares an original depth UNet with $f = 4$ to a shallower UNet with $f = 8$, and Fig. 6.15b compares an original depth UNet with $f = 8$ to a shallower UNet with $f = 16$. The black diamonds represent the original *Standard UNets* trained to convergence; they perform very similarly on the data. As they are pruned, the two similarly sized networks have very similar performance on the main task, indicating that the exact arrangement of the filters is not vital for the model's performance. This therefore suggests that the algorithm is reasonably robust to the initial model configuration and the results are not due to the initial choice of model architecture. This is also demonstrated by Fig. 6.16, which shows the distribution of filters during training for the shallower UNet when training with the HarP data. It can be seen that, as is evident in the



(a) **(b)**

Figure 6.15: Direct comparison of pruning the original UNet considered – Fig. 6.3 – and the shallower UNet – Fig. 6.4 – for the case where the two models have similar numbers of parameters originally: a) compares the UNet with $f = 4$ to the shallower UNet with $f = 8$. b) compares the UNet with $f = 8$ to the shallower UNet with $f = 16$. The black diamonds represent the original *Standard UNets* trained to convergence.

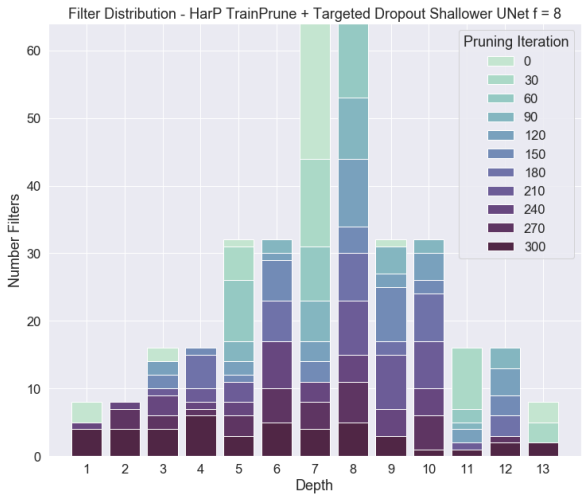


Figure 6.16: The distribution of the filters with network depth as the shallower model is pruned for hippocampal segmentation on the HarP data, meaning that the darker the shade of the filter block the longer the filters at that depth were maintained in the model.

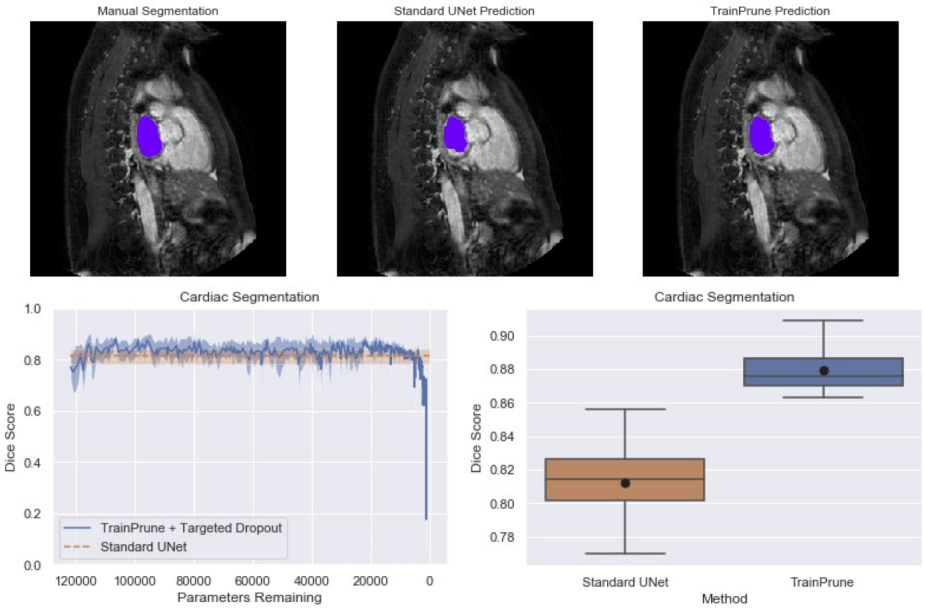


Figure 6.17: Example segmentation of the right atrium and comparison of the performance of the *Standard UNet* and the proposed *TrainPruneTD*. The segmentation shown is indicated by the point on the boxplot.

other examples shown, the bottleneck of the model is pruned aggressively early in the pruning, and the first layer at each depth is pruned first – the result shown is for $f = 8$ but the pattern was the same for all values of f considered.

6.3.2 Application to Other Datasets

To test the robustness of the chosen hyperparameter values, the same settings were applied to a range of datasets and tasks. Therefore, the following results were all run with the hyperparameters set as:

- Number of recovery epochs = 2
- $b_{drop} = 0.05$
- Learning Rate = 1×10^{-2}
- $f = 4$

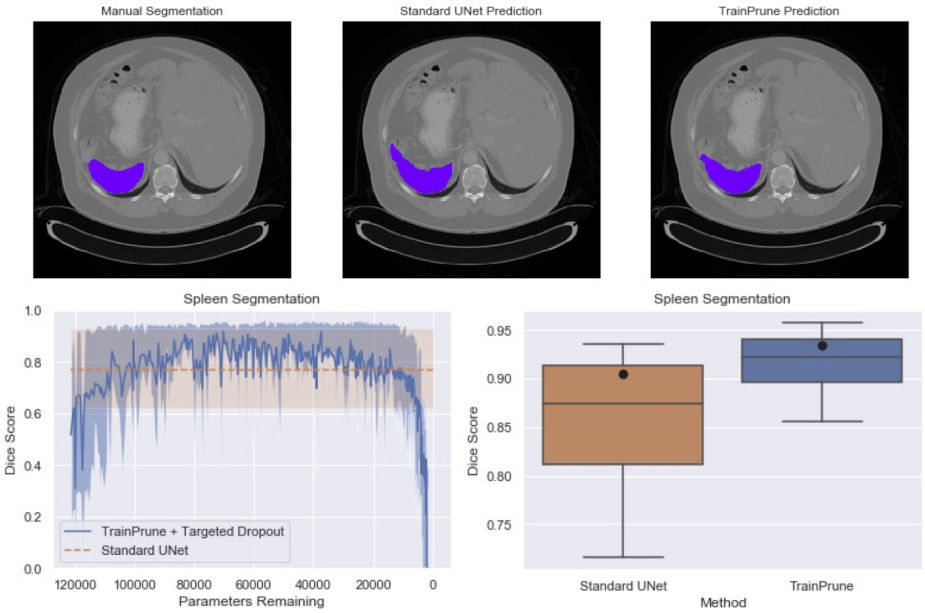


Figure 6.18: Example segmentation of the spleen from the CT dataset and comparison of the performance of the *Standard UNet* and the proposed *TrainPruneTD*. The segmentation shown is indicated by the point on the boxplot.

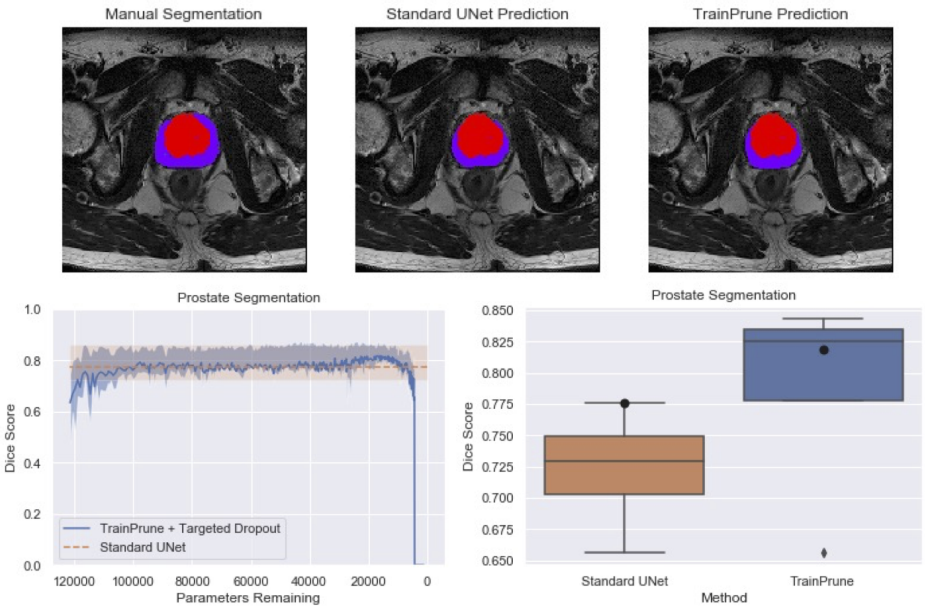


Figure 6.19: Example segmentation of the prostate from the multimodal MRI dataset and comparison of the performance of the *Standard UNet* and the proposed *TrainPruneTD*. The segmentation shown is indicated by the point on the boxplot.

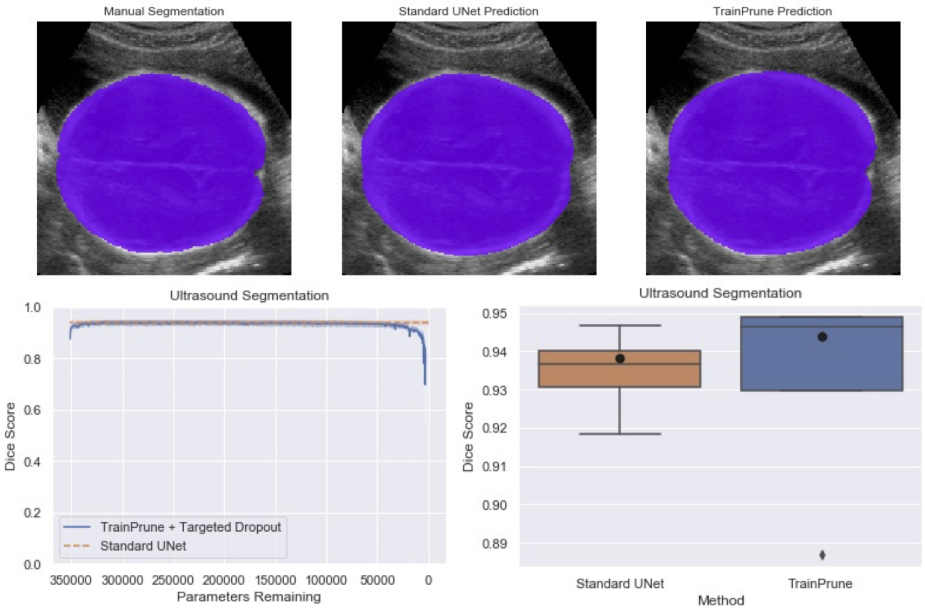


Figure 6.20: Example segmentation of the fetal brain from ultrasound data and comparison of the performance of the *Standard UNet* and the proposed *TrainPruneTD*. The segmentation shown is indicated by the point on the boxplot.

The results can be seen in Figs 6.17 – 6.20, where the ground truth segmentations and predictions, and the segmentation performance with pruning for each task can be seen. The results for the datasets from the Medical Decathlon are the values for the reconstructed 3D volumes, so that they are comparable with the results reported in the Medical Decathlon Challenge. The predictions from *TrainPruneTD* are the results on the held out test data from the best model, selected on the performance on the validation data. It can be seen that despite the hyperparameter values that are used not being optimised for these tasks, the proposed approach is able successfully to train and prune the network simultaneously, producing networks that match the performance of, or out-perform, the *Standard UNet*. This shows that the chosen hyperparameters are robust across tasks, dataset size and image modality, and so indicates that the approach should be applicable for a wide range of medical imaging tasks.

6.3.3 Application to the IXI Subcortical Segmentation - Ablation Study

The task of segmenting three deep grey matter structures was also considered: the caudate, putamen and thalamus. Without data augmentation, it was not possible to train a *Standard UNet* to successfully segment all three regions; however, it was possible to train the *TrainPruneTD* method to segment all three regions. Therefore, an ablation study was performed to explore the effect of the pruning and the targeted dropout on the performance of the model. Therefore, the following methods are tested: *Standard UNet*, *Standard UNet* with channelwise dropout, *Standard UNet* with targeted dropout, *TrainPrune*, *TrainPrune* with channelwise dropout, and *TrainPruneTD*. For the models with spatial dropout, the dropout probability was set to the average value of the targeted dropout values.

First, the training distributions for the three pruning studies can be seen in Fig. 6.21. This shows that the filter distributions as the models are pruned are very similar. This demonstrates that the targeted dropout is not driving the pruning process, with all methods showing the same pattern, and so the pruning pattern is being driven by the data. The improvement seen through the addition of the targeted dropout is due to the dropout making the model more robust to the removal of filters.

Figures 6.22 and 6.23 show the results of the ablation study, with Fig. 6.22 showing a subject from the test set, and Fig. 6.23 showing the box plots for each site, averaged over the three subcortical regions. It can be seen from the example segmentations that the *Standard UNet* fails to segment all of the regions and even the addition of the targeted dropout, whilst it leads to a small amount of the third region being segmented, does not lead to a substantial improvement in the performance of the network. Therefore, in this low-data regime for a relatively difficult task, where the *Standard UNet* fails

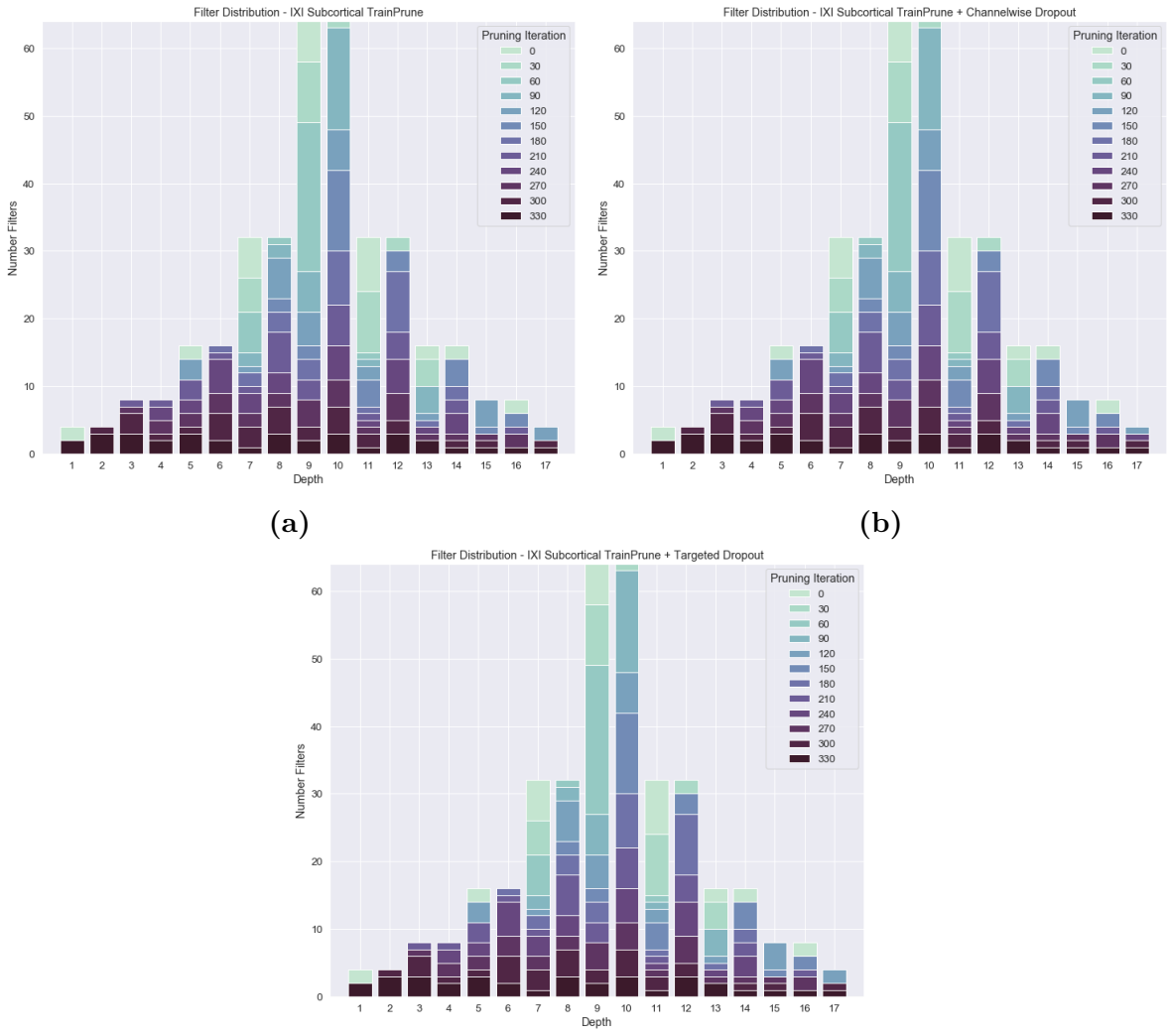


Figure 6.21: (a) *TrainPrune*, (b) *TrainPrune + Channelwise dropout out* (c) *TrainPruneTD* for subcortical segmentation on the IXI data. The darker the filter, the longer it was maintained in the model. It can be seen that all the methods lead to a very similar pattern of pruning of the filters, showing that the adaptive dropout does not drive which filters are pruned, merely makes the model more robust to being pruned.

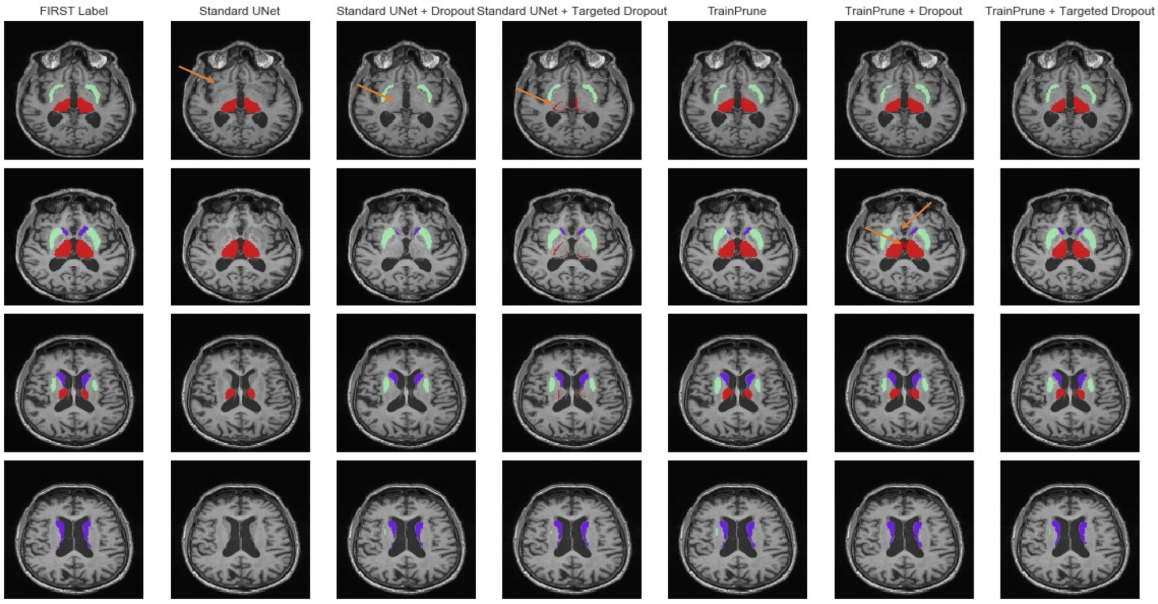


Figure 6.22: Example segmentations for the caudate, thalamus and putamen trained using the Guy's data from the IXI dataset for, left to right, starting at the second column: *Standard UNet*, *Standard UNet* with channelwise dropout, *Standard UNet* with targeted dropout, *TrainPrune*, *TrainPrune* with channelwise dropout, and *TrainPruneTD*. The labels generated using the FSL FIRST tool (shown in the first column) are used as a proxy for manual labels. Arrows indicate regions with clear differences between approaches.

to complete the segmentation to a standard that would be acceptable, the pruning process alone creates a network that is significantly better at segmenting the three regions. Across all folds, the *Standard UNet* found a local minimum which led to one of the regions being segmented very well at the cost of another not being segmented at all. Table 6.2, which explores the effect of training with reduced amounts of training examples, shows that this is also the mode of failure of the pruned network when the amount of training data is reduced too much. The region that is not segmented does, however, vary between folds and is not necessarily the smallest region.

It can also be seen that while the improvement provided by the targeted dropout is relatively small, it is significant ($p = 0.0007$) and, in addition, increases the stability of the *TrainPruneTD* procedure. There is also a significant improvement compared

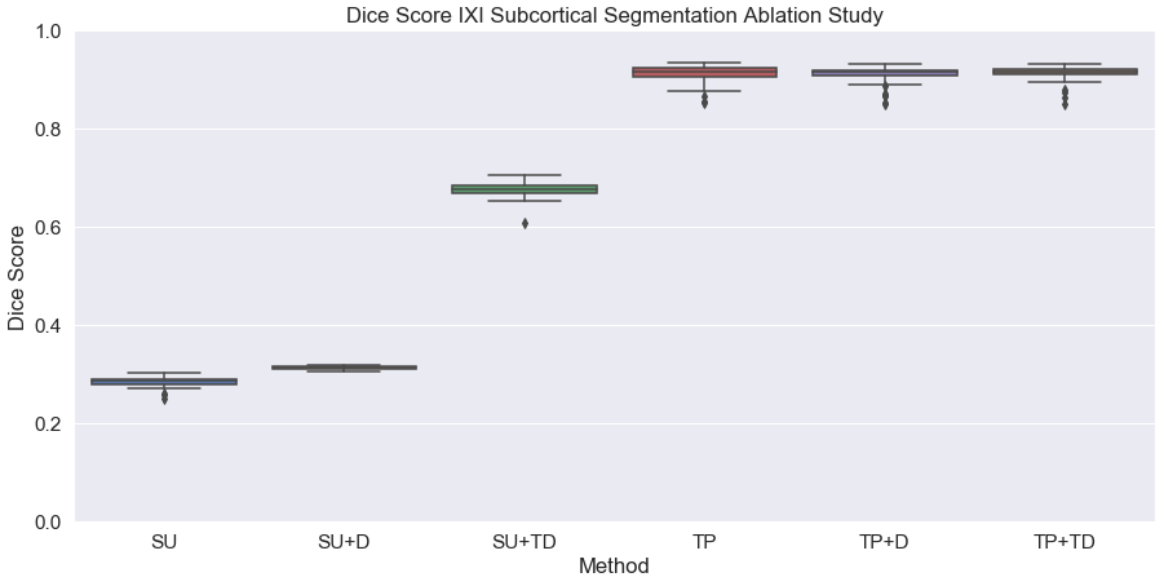


Figure 6.23: Dice results for the IXI subcortical segmentation ablation study, comparing *Standard UNet* (SU) , *Standard UNet* with channelwise dropout (SU+D), *Standard UNet* with targeted dropout (SU+TD), *TrainPrune* (TP), *TrainPrune* with channelwise dropout (TP+D) and *TrainPruneTD*(TP+TD). It can be seen that the simple pruned model outperforms all the models trained without pruning, which were unable to segment all three labels, as seen in Fig. 6.22.

to using the same spatial dropout across the networks ($p = 1.2 \times 10^{-8}$): this might be expected as this does not encode information about which filters are the most likely to be removed, so there is a chance that filters that are very important to the performance would be dropped, especially as the pruning progresses, hindering the model training. The value of spatial dropout used was very low (0.037) and with a higher value the performance of the network was hindered further.

6.3.3.1 Amount of Training Data

The poor performance of the *Standard UNet* on this task is almost certainly due to the shortage of training data available for the task. Working in a low data regime means how much more data is required to improve the performance cannot be explored; however, reducing the amount of training data to explore the performance of the



(a)

Figure 6.24: Dice scores averaged across the three subcortical regions for increasing numbers of available training images for the IXI data. 250 represents the full training set.

Number of Training Images	Standard UNet			TrainPruneTD		
	Caudate	Putamen	Thalamus	Caudate	Putamen	Thalamus
50	0.28 ± 0.07	0.19 ± 0.03	0.14 ± 0.04	0.78 ± 0.04	0.75 ± 0.04	0.0 ± 0.0
100	0.54 ± 0.12	0.65 ± 0.08	0.75 ± 0.06	0.86 ± 0.02	0.89 ± 0.04	0.93 ± 0.02
150	0.46 ± 0.06	0.78 ± 0.06	0.21 ± 0.07	0.88 ± 0.02	0.90 ± 0.03	0.92 ± 0.02
200	0.032 ± 0.01	0.0 ± 0.0	0.82 ± 0.02	0.90 ± 0.02	0.91 ± 0.03	0.94 ± 0.01
250	0.90 ± 0.03	0.080 ± 0.05	0.0 ± 0.0	0.89 ± 0.02	0.92 ± 0.03	0.94 ± 0.01

Table 6.2: Table showing the Dice scores of the different subcortical regions with different amounts of training data for the testing data coming the Guy's site, comparing *Standard UNet* and *TrainPruneTD*. The pattern of results was the same for the other two sites.

TrainPruneTD method can be explored. Figure 7.16 shows the Dice scores on the training set, averaged across the three subcortical regions. Table 6.2 then shows the corresponding Dice scores broken down by subcortical region for the testing data, with the performance of the *Standard UNet* provided as a benchmark. Both sets of results show that the pruned network can be successfully trained with substantially less data than the *Standard UNet* and, in all cases, provides a better performance. This is not just a function of the reduced number of parameters, as we were unable to train a

Standard UNet with $f = 2$ on the data. Some of the increase in performance might be due to the network learning coadapted relationships between layers and overfitting less to the very small amount of training data available. Therefore, *TrainPrune* clearly provides advantages for segmentation tasks when working in low-data regimes, as is very common in medical imaging applications.

6.3.4 Training in Low Data Regimes

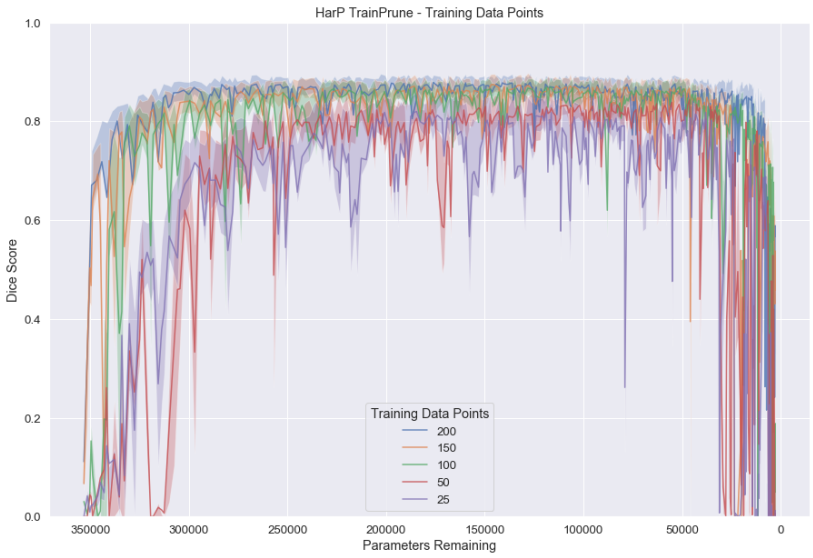
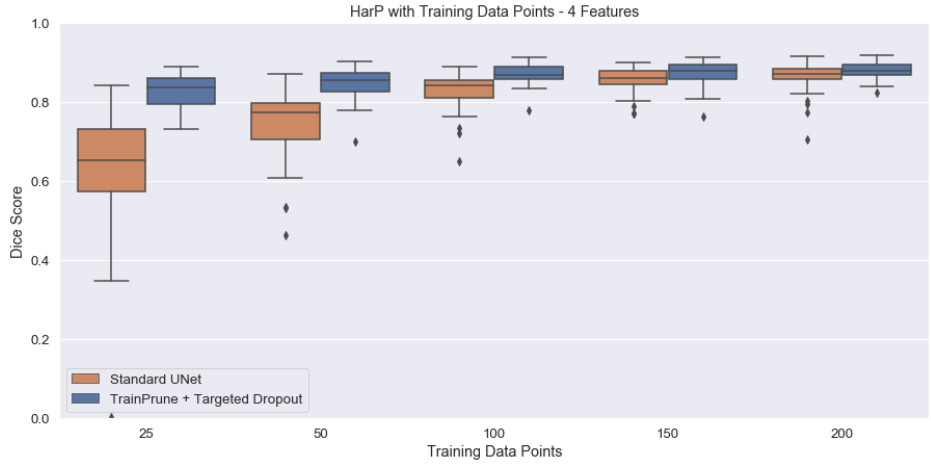


Figure 6.25: Dice scores for the HarP hippocampal segmentation for increasing amounts of training subjects.

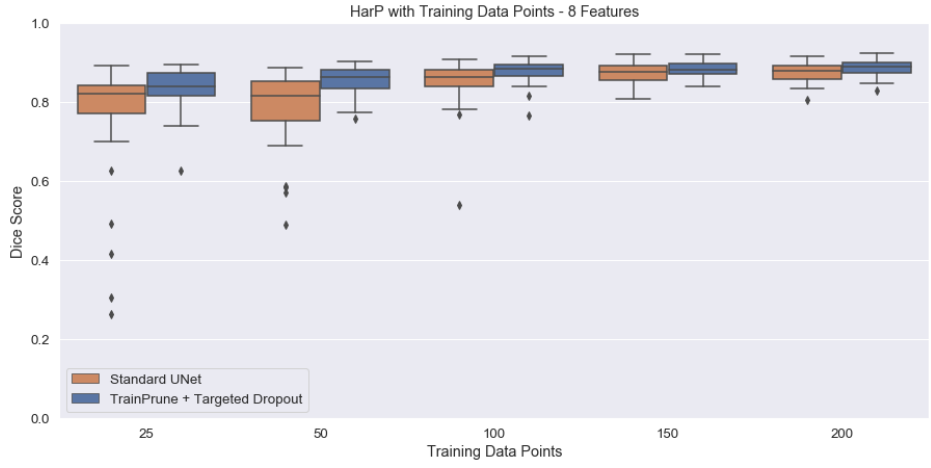
Having shown that the *TrainPruneTD* algorithm has worked in a low data regime when the *Standard UNet* was not able to create acceptable segmentations, training in low data regimes was now more systematically explored. Clearly, as would be expected, the performance of the method improves with increasing amounts of training data, as shown in Fig. 6.25, with the maximum dice score increasing and the stability of the score between pruning iterations increasing. The network is able to produce dice scores of over 0.82 even when presented with only 25 training examples.



(a)



(b)



(c)

Figure 6.26: Dice scores for the HarP hippocampal segmentation for increasing amounts of training subjects, comparing *TrainPruneTD* and *Standard UNet*, for $f = \{2, 4, 8\}$.

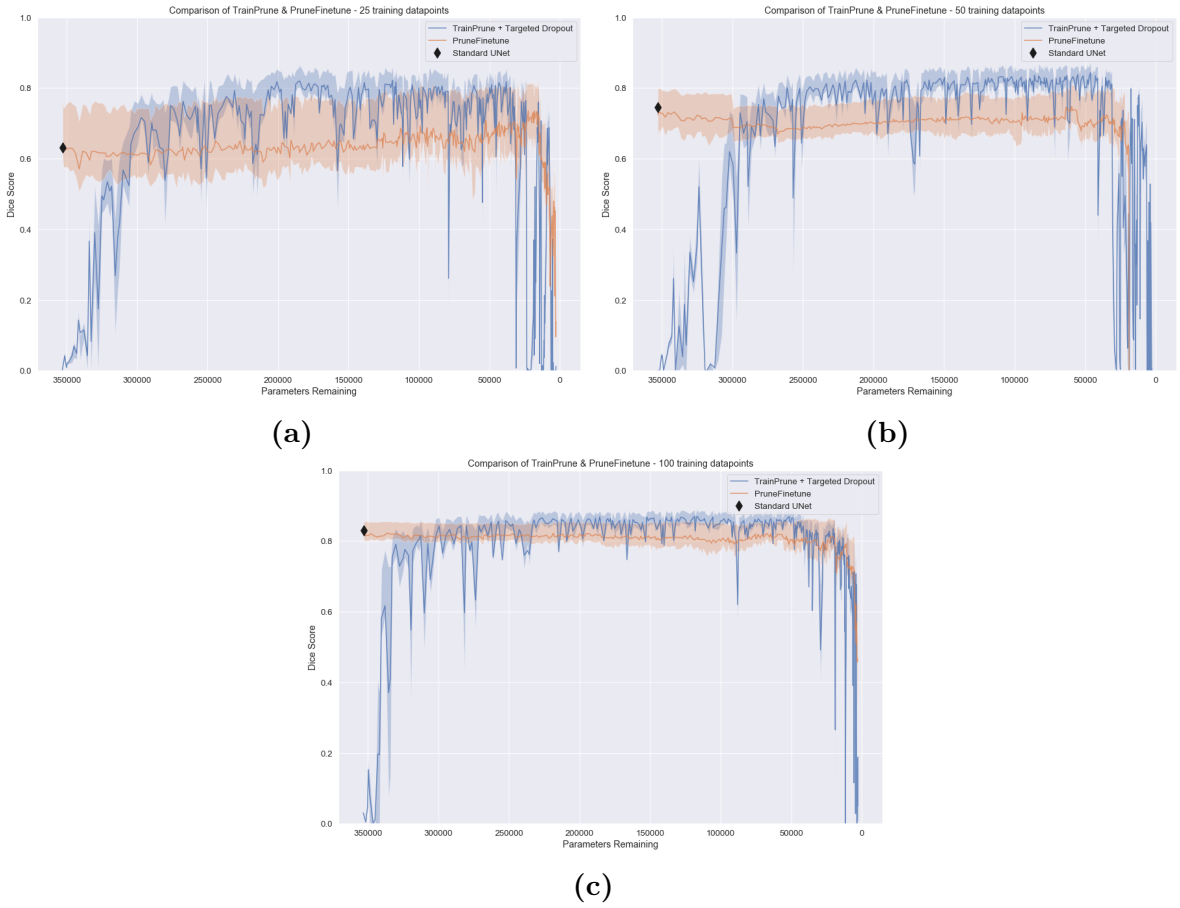
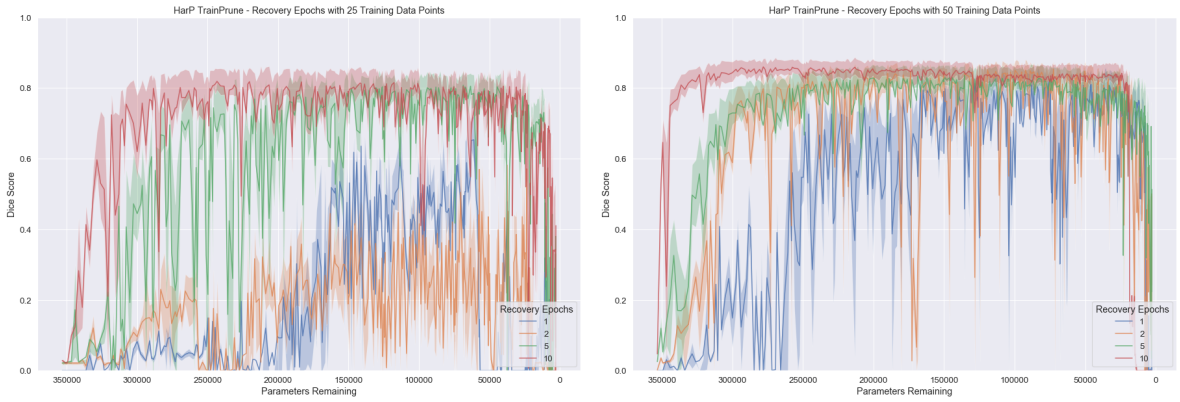


Figure 6.27: Dice scores for the HarP hippocampal segmentation comparing *TrainPruneTD* and *PruneFinetune* for 25, 50 and 100 training subjects.

6.3.4.1 Comparison to PruneFinetune

Figure 6.26 then compares *TrainPruneTD* to *Standard UNet* for increasing numbers of training subjects, for models with $f = \{2, 4, 8\}$. It can be seen that for all sizes of model, and for all amounts of data, the *TrainPruneTD* model outperforms the *Standard UNet*, demonstrating its utility in low data regimes.

TrainPruneTD is then compared to *PruneFinetune* for increasing numbers of training subjects. In Fig. 6.27 the comparison can be seen for $25, 50$ and 100 training subjects. It can be seen that when the number of training subjects is very low – for instance, consider 25 training subjects – the original *Standard UNet*



(a) **(b)**

Figure 6.28: Dice scores for the HarP hippocampal segmentation for exploring the effect of the number of recovery epochs for 25 and 50 training subjects.

model performed badly on the data. As this model is then the initialisation for the *PruneFinetune* pruning and there are still very few subjects available for training, the model struggles to perform well from the poor starting point. The *TrainPruneTD* model has noisier performance between pruning iterations, but substantially outperforms the *PruneFinetune* approach. Even as the number of training subjects increases, *TrainPruneTD* outperforms *PruneFinetune*.

6.3.4.2 Number of Recovery Epochs

It was hypothesised that *TrainPruneTD* might be more sensitive to the number of recovery epochs between pruning iterations when working in a low data regime. Therefore, the experiment exploring the effect of the number of recovery epochs was repeated for 25 and 50 training subjects.

The results can be seen in Fig. 6.28. It can be seen that when working in a low data regime, the number of recovery epochs between pruning iterations needs to be increased to allow the model to reach good segmentation performance. Therefore, the number of recovery epochs needs to be increased at a cost to the amount of time



Figure 6.29: Dice scores for the heart segmentation task for low numbers of subjects comparing *TrainPruneTD* and *Standard UNet*. Note that the*Standard UNet* was unable to complete the segmentation for the two lowest numbers of data points.

required to train the model. Despite this, training of the *TrainPruneTD* method still takes many fewer training epochs than *PruneFinetune*.

6.3.4.3 Training in Low Data Regimes – Application to Other Datasets

Finally, training in low data regimes was explored for other datasets. Based on the previous experiment, the number of recovery epochs was set to 5 and the other hyperparameters remained the same as in previous experiments. The results can be seen in Figs. 6.29 - 6.32. It can be seen that for all the tasks, *TrainPruneTD* outperforms the *Standard UNet* for the low numbers of training subjects, showing the power of the pruning method for working in low data regimes, as are common across medical imaging.

6.4 Conclusion

In this chapter a method has been proposed to enable the simultaneous training and pruning of the UNet architecture for medical image segmentation. While pruning is well explored for large classification datasets, to the best of our knowledge it has not been

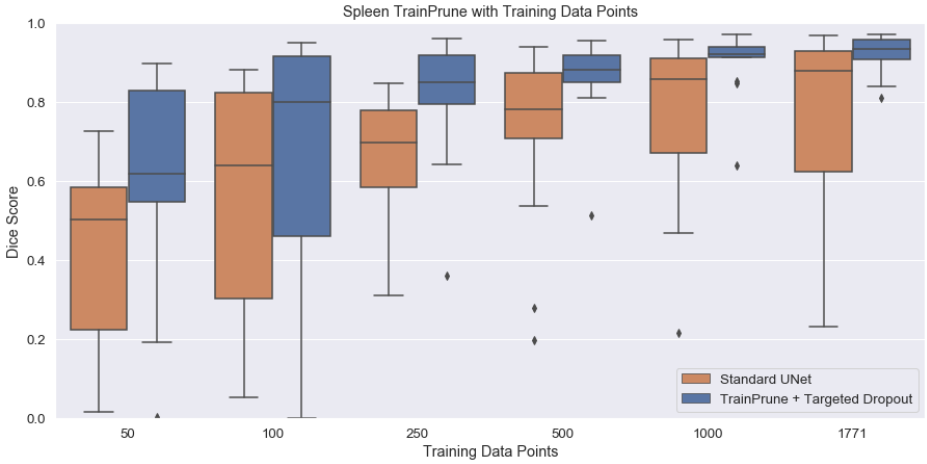


Figure 6.30: Dice scores for the spleen segmentation task for low numbers of subjects comparing *TrainPruneTD* and *Standard UNet*.



Figure 6.31: Dice scores for the prostate segmentation task for low numbers of subjects comparing *TrainPruneTD* and *Standard UNet*.

explored for medical image segmentation, where different architectures and generally small training datasets create a different challenge for pruning algorithms. Therefore, the *TrainPrune* method and the addition of Targeted Channelwise Dropout has been explored for a range of segmentation datasets and tasks; it has been shown that not only is it possible to create smaller models with as good or increased performance on the task of interest, but also, that through pruning, better performing models

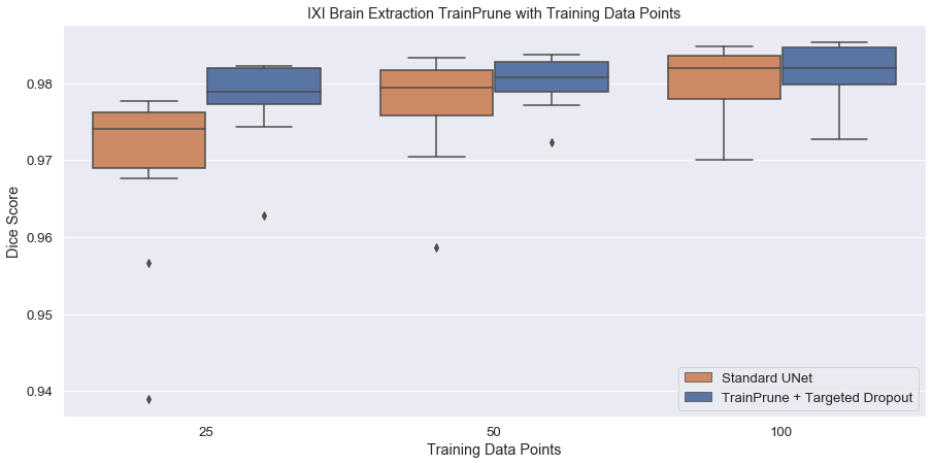


Figure 6.32: Dice scores for the IXI brain extraction task for low numbers of subjects comparing *TrainPruneTD* and *Standard UNet*.

can be produced in low data regimes.

Using the HarP data, the method has been validated on data with manual segmentation labels and the effects of the design decisions have been explored. The results indicate that better performance is achieved by initially choosing a large model and pruning the model to be smaller, rather than beginning with a small model, and that different distributions of the filters are kept by the pruning process than might be naïvely chosen. The experiments also show that the addition of the Targeted Dropout leads to an improvement in performance and that these results show that this improvement is present for a relatively wide range of b_{drop} values – 0.05 to 0.30.

It has been demonstrated that the hyperparameters used are robust across a range of tasks and imaging modalities and for 3D and 2D problems, showing that the method is applicable across a range of medical imaging tasks, and should be able to be used in many cases without the need for extensive parameter searching or optimisation: the default parameters led to the same or improved performance as the *Standard UNet* comparisons for all cases. Further, the increase in performance in low data regimes has been demonstrated across these different datasets and tasks. The ablation study on

the IXI dataset indicates that the improvement is not just a function of the pruning or the targeted dropout, but a combination of the two aspects.

Whilst care has been taken to try to test the method across modalities and segmentation tasks, the labels used for the IXI dataset were generated using automated tools. The two tools used – FSL FIRST and ANAT – use differing model assumptions so the results should be valid for comparison between methods; however, it does potentially limit the maximum achievable performance due to the imperfections in the labels.

The method has also only been explored for the standard UNet architecture. This decision was made as the UNet is the most popular architecture for medical image segmentation, and the majority of methods either use the UNet architecture or derivatives thereof. It is expected that the results would generalise to other similar networks, but this has not been explored explicitly within this work.

Finally, a potential limitation of the method is that the performance between pruning iterations is unstable, especially compared to the *PruneFinetune* method. Across this work, the best model was evaluated using the validation data, and this has corresponded well with good performance on the testing data – although it does not always achieve the highest performance on the testing data. This has been true across the datasets explored here but would not necessarily be true if the method were to be applied to other datasets. The use of the targeted dropout helps to reduce this by increasing the stability between iterations, but if this were not seen to be the case for a given dataset, it may be necessary to increase the number of recovery epochs to increase the stability of the model training. It is probable that a lot of the instability is due to working in the low data regime, as the stability visibly increases as the number of training subjects is increased.

In the next chapter, the method will be applied to multisite MRI data, to explore model generalisability.

7

Model Pruning for Generalisability Across Imaging Sites

7.1 Introduction

As explored in the previous chapter, deep learning based segmentation methods are able to produce high quality automatic segmentations across segmentation tasks and modalities. A major limitation of these methods, and with deep learning-based methods in general, is that they suffer significant degradation in performance when applied to data that is drawn from a different distribution to the data used during training [111, 367], for instance, when the data was collected on a different scanner or with

a different acquisition protocol. This problem is exacerbated in medical imaging by the lack of labelled training data available, especially when multisite data is collected; collecting data spanning the variation of devices and acquisition protocols whilst also obtaining labels for all of these is not viable. Therefore, methods are needed to tackle this lack of generalisability – not just between the training and testing sets, as is often considered in computer vision, but also for potential datasets that have not yet even been collected, to ensure that segmentation methods can be confidently used and will produce reliable segmentations.

One of the causes of this lack of generalisability is the model overfitting to the training data, due to the models containing very high numbers of parameters. Several methods have been proposed to try to reduce the overfitting to the training data, including node dropout [87, 368] and data augmentation [40, 369, 370]. Dropout removes nodes at random with a given probability, forcing the model not to be dependent on certain nodes, and reducing the dependency on individual nodes. Alternatively, augmentation applies random transformations to the data during training, increasing the size and variation of the dataset seen, so that the trained model is more robust to variation between the training and test sets.

Here, the utility of model pruning, specifically the *TrainPrune + Targeted Dropout* (TrainPruneTD) algorithm previously introduced, for increasing model generalisability in medical imaging applications is explored. In [182] an increased generalisability between the training and testing data was observed. These datasets were, however, drawn from the same distribution; this work instead considers what happens when the testing data has a different distribution, for instance when it has been acquired with a different scanner or protocol, which is rarely explored within computer vision [293]. As previously, a UNet will be used as the architecture being explored.

The contributions are as follows:

- The method is demonstrated on multi-site data and it is shown that, not only is the same performance as the *Standard UNet* achieved on the testing data from the site seen during training, but a significant improvement in performance is achieved on the unseen site. The design decisions explored previously are re-explored for this new data scenario.
- Methods for exploring the degree of the model generalisability are explored, with PCA being used to create reasonable variations to the model in a data driven manner, and augmentations being used to generate plausible variations.
- The *TrainPruneTD* algorithm is compared to the addition of dropout and augmentation, to allow comparison with the standard approaches.

7.2 Methods

7.2.1 Implementation Details

The method and network architectures will be as detailed in Section 6.2. The hyperparameters across the experiments were set as follows, unless otherwise stated:

- learning rate $= 1 \times 10^{-2}$
- base dropout $b_{drop} = 0.05$
- recovery epochs $= 2$
- batch size $= 32$
- features scaled by $f = 4$

7.2.2 Datasets

The aim was to explore the effect of pruning on deep neural networks' ability to generalise to other datasets beyond the one on which the model was trained. To achieve this, ideally there would be access to multiple labelled datasets for the same tasks, collected on different MRI scanners, with similar populations. Acquiring data with consistent manual labels with dataset sizes large enough to train deep neural networks is rendered infeasible due to the expense, both in terms of time and expertise, demonstrating the need for automated methods to produce labels. Therefore, in the previous chapter, the method has first been demonstrated on a single-site with manual labels, allowing validation of the performance. Now, the method has been applied to multisite data, where labels have been automatically generated as a proxy for manual labels, using a range of tools to generate the labels. This potentially inflates the possible performance of our method and can be interpreted as a form of upper bound, but still allows for comparison between approaches. This is, therefore, a limitation of the work, but the method has previously been validated on manually labelled datasets.

Multisite datasets are considered here, as the data between sites within such a dataset is acquired to be as similar as possible, and therefore the domain shift is the least, although standard models still experience significant performance degradation when applied to the unseen sites in situations like this. In this scenario, there is a reasonable expectation that a model trained on one of the datasets should perform well when applied to the other datasets, but in reality, domain shift is still experienced. This could, for instance, represent the scenario where the scanner is upgraded during a study and a user wants to be able to continue to use their trained segmentation model on the data from the upgraded scanner. Therefore, for each dataset, one site is used for training and the other two sites are only seen at test time.

Dataset	Modality	2D/3D	Task	Image Size	Training/ Testing Subjects
ABIDE - **UM**, Yale, MPG	T1 MRI	3D	Tissue Segmentation	$128 \times 128 \times 128$	**80/15**,56,57.
IXI - **HH**, Guy's, IOP	T1 MRI	3D	Brain Extraction	$128 \times 128 \times 128$	**166/19**,65,74
IXI - **Guy's**, HH, IOP	T1 MRI	3D	Subcortical Segmentation	$128 \times 128 \times 128$	**257/65**,74,185

Table 7.1: Details of the datasets used in pruning experiments.

The datasets are shown in Table 7.1. The site used in training is shown in bold.

ABIDE [98] - UM, Yale, MPG: These three sites were chosen because UM had the largest number of samples available, Yale had a very similar age profile to UM, and MPG had a very different age profile – distributions can be seen in the appendix in Fig. C.13. Tissue segmentation (grey matter, white matter, CSF) is then used as the segmentation task, meaning that MPG, with a number of subjects in the higher age range compared to the age range of the UM training data, will have visibly different amounts of CSF in the subjects. The data was pre-processed using the FSL Anat pipeline, for the bias field and registration stages, and then FSL FAST [11] labels were used as proxy manual labels. The T1 brain-extracted images were used as the input images, and the number of classes, $C = 4$.

IXI[1] **- HH**, IOP, Guy's: Hammersmith hospital data from a 3T MRI scanner, IOP and Guy's was acquired on a 1.5T machine. The data was pre-processed using the FSL Anat pipeline [2]. The task considered was brain extraction and the labels were automatically generated using the FSL Anat tool, which non-linearly registers an atlas to the target image. The number of classes $C = 2$ (foreground and background).

IXI - Guy's, HH, IOP: Guy's hospital data is from a 1.5T MRI scanner, and the two unseen sites were acquired on a 1.5T (IOP) and a 3T (HH). The data was pre-processed

[1]Data downloaded from: `https://brain-development.org/ixi-dataset`
[2]For details see: `http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/fsl_anat`

using the FSL Anat pipeline. The task considered was subcortical segmentation of three deep grey matter structures: caudate, putamen and thalamus, with the labels automatically generated using FSL FIRST [12] run on the bias field corrected and registered images, which were output from the FSL Anat pipeline. The number of classes considered was $C = 4$ (caudate, putamen and thalamus, and background).

7.2.3 Generalisability Across Sites: ABIDE - Tissue Segmentation

The approach is first used on the ABIDE data. As it is expected that a lot of the lack of generalisability of the models is due to the model overfitting to the training data, it is expected that model pruning will lead to improved performance on the unseen sites. Example segmentations from the three sites can be seen in Fig. 7.1. It can be seen that both models perform well on the UM test data (the same site as seen during training), but that the pruned model performs best on the two unseen sites. The segmentation performance is worst for the Yale data, for example in the cortical folds, as would be expected due to the lower image quality of the data from the Yale scanner compared to the other two sites.

Figure 7.2 shows our pruned model's performance as the model is gradually pruned and trained, until no further filters can be removed. As a comparison, the *Standard UNet* is also shown. The model was trained on the data from the UM site, and the data from the Yale and MPG sites were entirely held out. Considering the results on the testing data from UM, it can be seen that the pruned model is able to achieve the same performance as the original full-sized model ($0.899 \pm 0.01, 0.900 \pm 0.01, p = 0.199$). The same model then performs significantly better than the original UNet model on the two unseen sites.

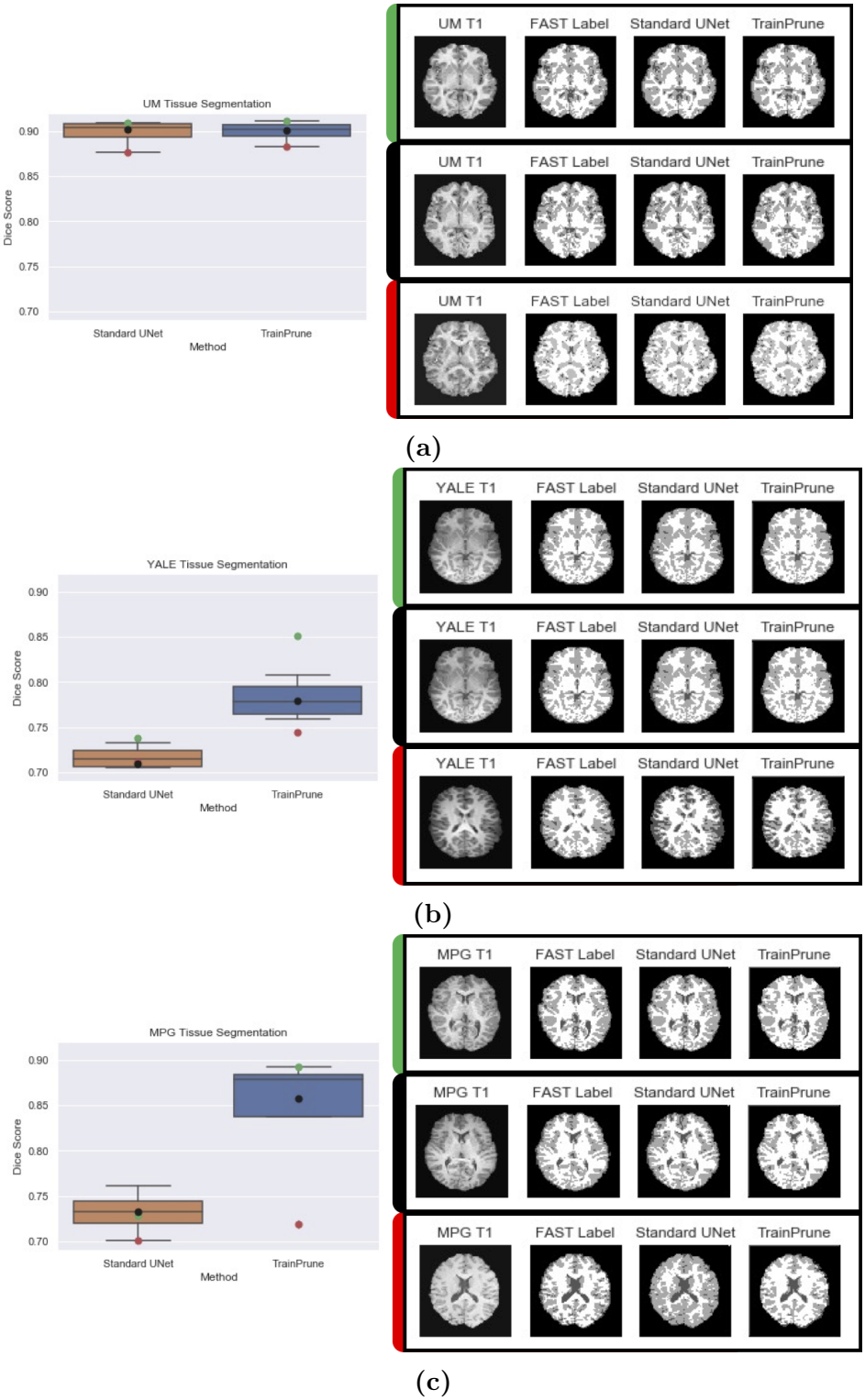


Figure 7.1: Representative segmentations from the ABIDE test set. Data from the UM site were seen during training, the other two sites were unseen. The colour coded boxes correspond to the same colour points on the box plot (best in green, typical in black, worst in red).

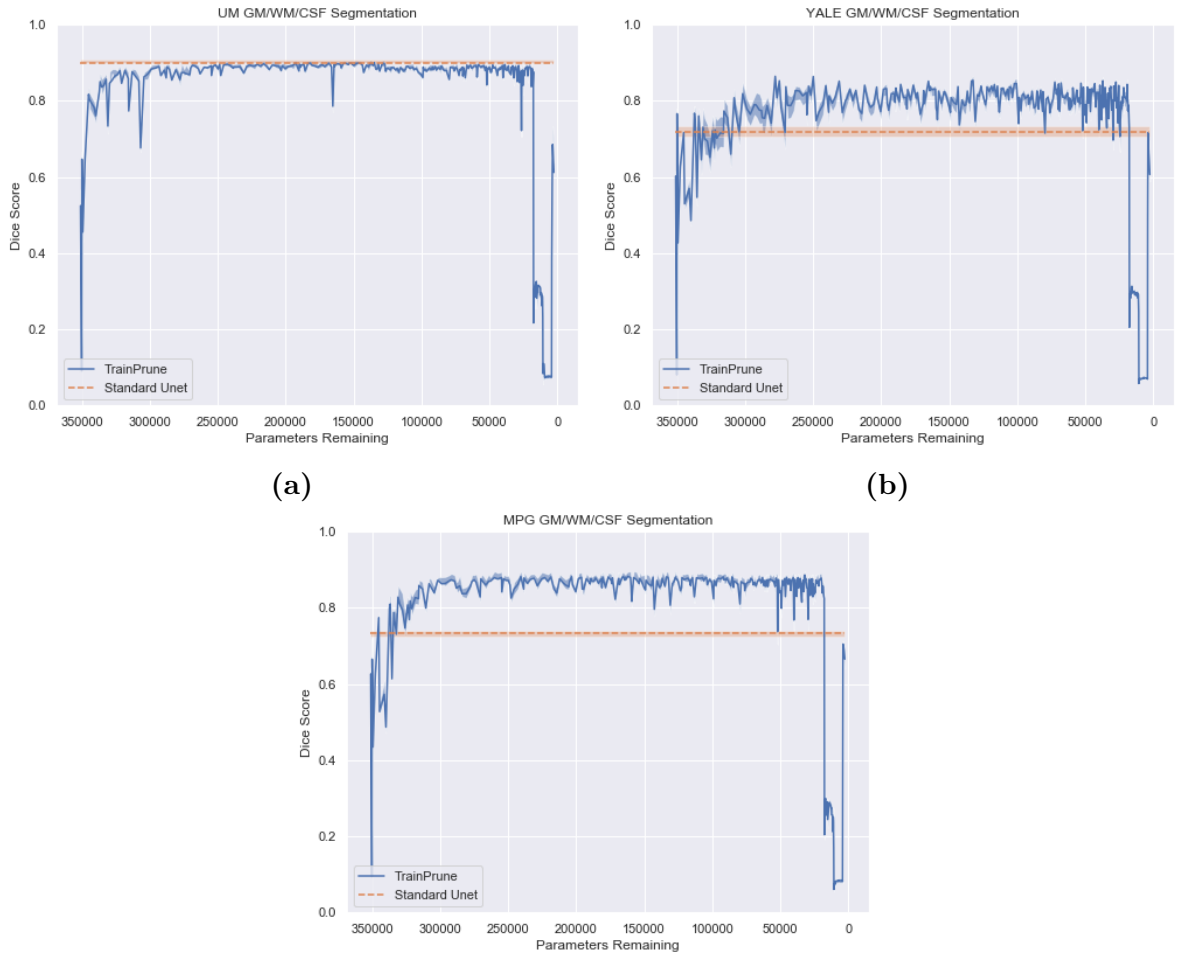


Figure 7.2: Dice results averaged across the three tissue types, comparing the Standard UNet to the same model trained and pruned simultaneously with targeted channelwise dropout. The model was trained using the data from UM and is tested on a hold out set from UM and two unseen sites, Yale and MPG. The same results evaluated with different metrics can be seen in Fig. C.15, where it can be seen that the pattern is the same across the metrics.

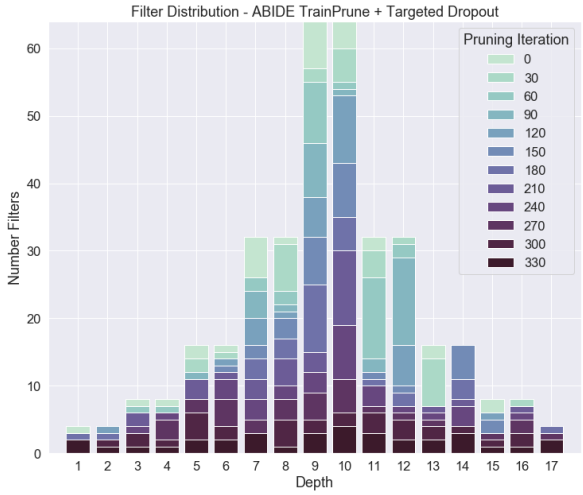


Figure 7.3: The distribution of the filters with network depth as the model is pruned for tissue segmentation on the UM site of the ABIDE data. The darker the shade of the filter block, the longer the filters at that depth were maintained in the model.

The pruned models have a fraction of the parameters of the original model, which indicates that the lack of generalisability of the original models is due to their overfitting of the training data. By pruning the models, we are able to create a representation that learns features that generalise better to the unseen data, suggesting that the additional parameters learned information about the specific dataset that were not necessary to complete the task of interest. The pruned model achieved a Dice score of 0.864 ± 0.02 on Yale compared to the *Standard UNet* 0.721 ± 0.06 ($p = 9.2 \times 10^{-14}$), and 0.886 ± 0.01 on MPG compared to 0.732 ± 0.05 ($p = 2.5 \times 10^{-10}$). Fig. C.15 in the Appendix shows that the result is consistent across the metrics regularly used to assess the performance of segmentation models; therefore only Dice scores [371] will be reported for the rest of the experiments, as it is the most commonly used metric for assessing segmentation performance with known behaviour [31].

Figure 7.3 shows the filter distribution as the model is pruned. It can be seen that the pruning pattern is very similar to that which we saw for the HarP data in the previous chapter, with the first layer at each depth being pruned earlier than the

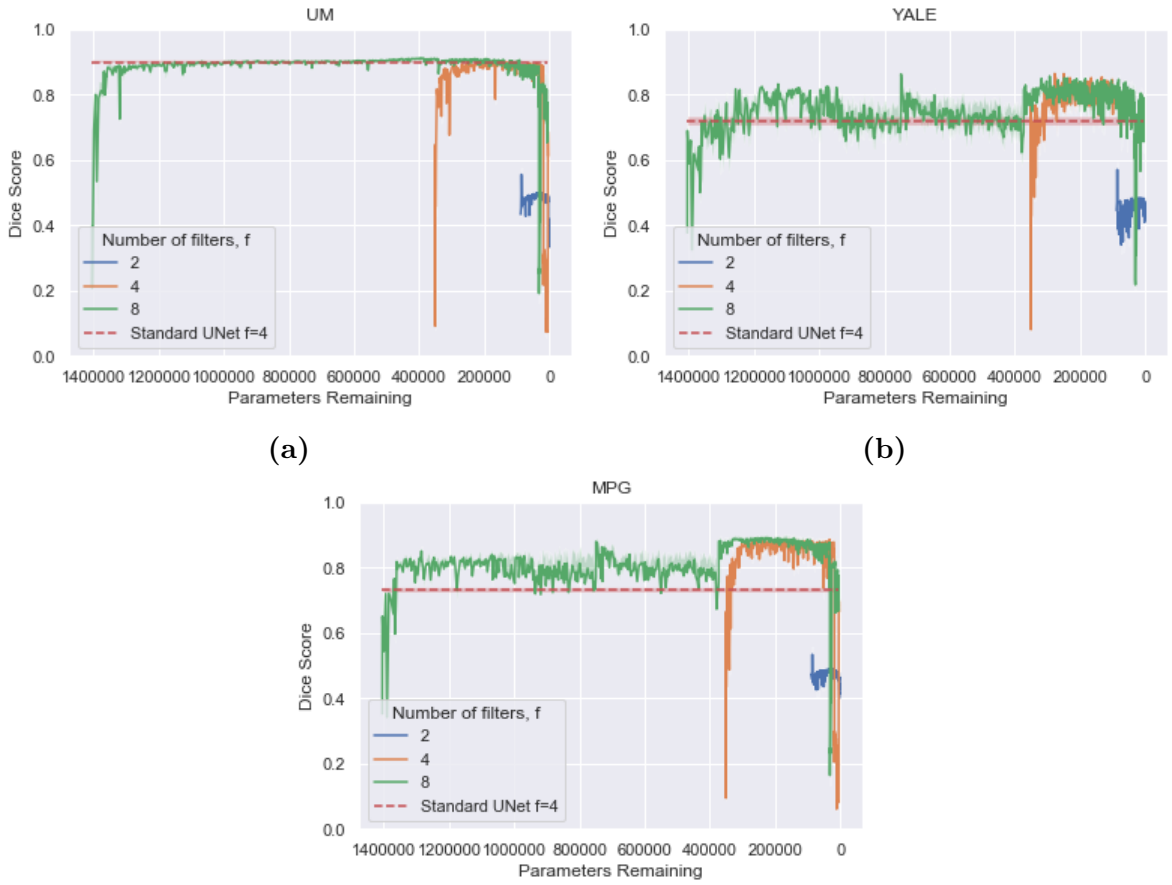


Figure 7.4: The effect of model size on the model performance is explored, for $f = [2, 4, 8]$. It can be seen that the smallest model - $f = 2$ was unable to perform the segmentation task well despite being able to prune various other networks, initially with larger sizes, down to the same size as the $f = 2$ network, which can then complete the segmentation.

second, and the bottleneck being aggressively pruned early on.

The fact that it is possible to prune the networks to a fraction of their original size indicates that that the initial models have more parameters than are necessary to complete the task. Figure 7.4, however, shows that this cannot be solved by simply naïvely halving the size of the model: despite the fact that the pruned model is able to perform the segmentation task well while being in the parameter range of the model with $f = 2$, it was not possible to train such a small model to complete the task

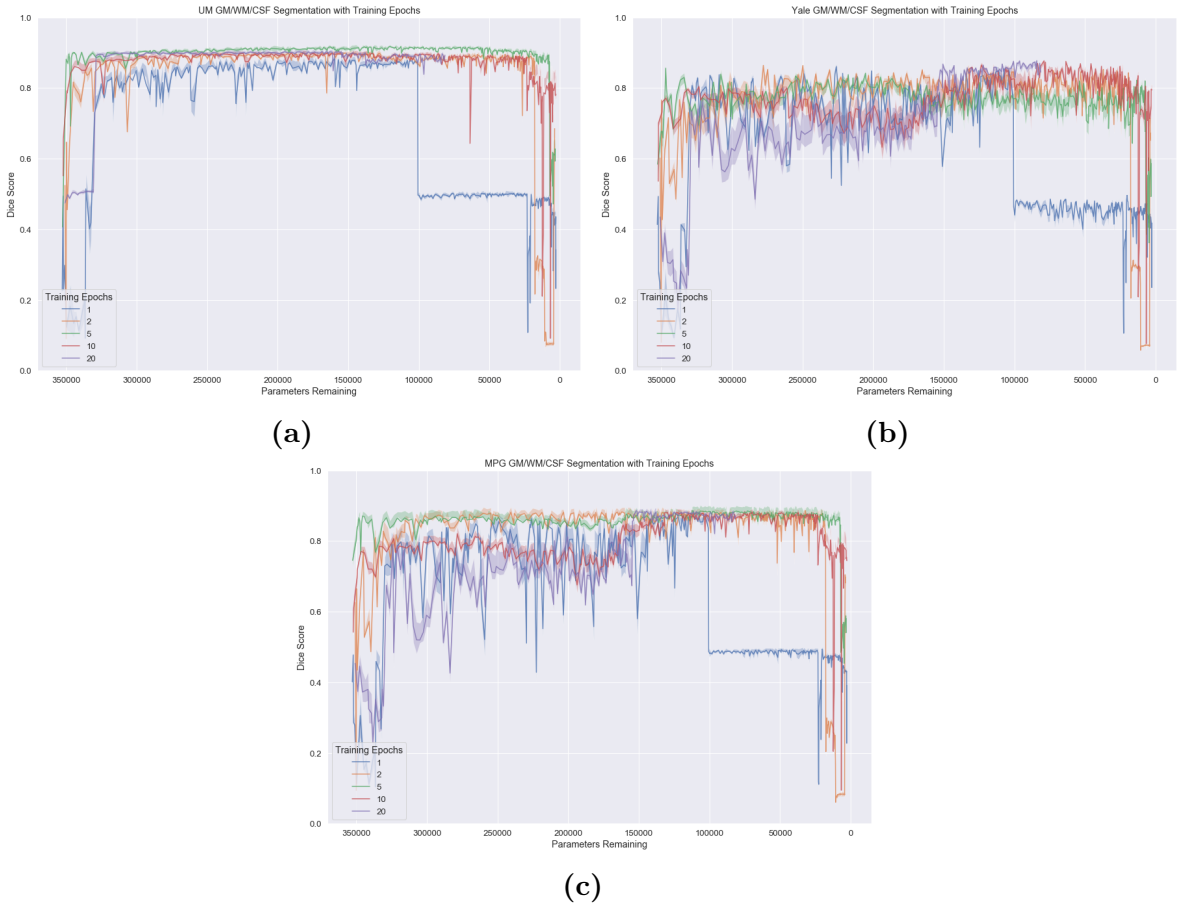


Figure 7.5: The *TrainPruneTD* method, comparing different numbers of training epochs, for tissue segmentation on the ABIDE data. The model was trained using the data from UM and is tested on a hold out set from UM site and two unseen sites, Yale and MPG. It can be seen that the model is robust to the number of training epochs as long as the value chosen is greater than one. At very high values – 20 – the model continues to overfit until the model is pruned heavily to a much smaller size, suggesting that the smallest number of training epochs to successfully train the model should be used.

successfully. This suggests that by pruning a larger model to be smaller, we are able to learn parameters that are coadapted, and that rely on once having been part of a larger model, and that these coadapted features generalise well to the unseen sites.

7.2.3.1 Number of Recovery Epochs

The effect of the number of epochs of training completed between each pruning iteration was also investigated. In all of the experiments, a value of 2 training epochs was used, but this is a hyperparameter in the training, and the effect of this on the results should be understood. Figure 7.5 shows the results on the tissue segmentation task with the ABIDE data as the number of training epochs between pruning iterations is varied between 1 and 20. It can be seen that, for this data scenario, 1 training epoch between pruning iterations is not sufficient to provide good performance on the main task: the performance on the data from UM never reaches the same performance that can be achieved by training the full normal model, with the performance collapsing very early on. Therefore, a single epoch is not sufficient to allow the model both to recover from the pruning and to learn more information about the segmentation task. It can also be seen that when the number of epochs is very large – such as 20 epochs – that although the model performs very well on the UM site it saw during training, it performs less well on the unseen sites until the model becomes very small. This indicates that when the number of epochs for recovery is allowed to become very large, it enables the model to continue to overfit, despite the pruning, to the site seen during training. This would not be too problematic for this situation as the model can be pruned to be small enough that there are insufficient parameters for the model to overfit and the effect is then removed, but it is desirable to avoid this effect so that there is no requirement to have to prune the model to a very small size. Therefore, the choice of the number of training epochs between pruning iterations needs to be selected such that there are sufficient epochs for the model to learn, but not so long that the model overfits to the training data. All of the other results shown in this chapter were for 2 training epochs, which worked across the range of data scenarios

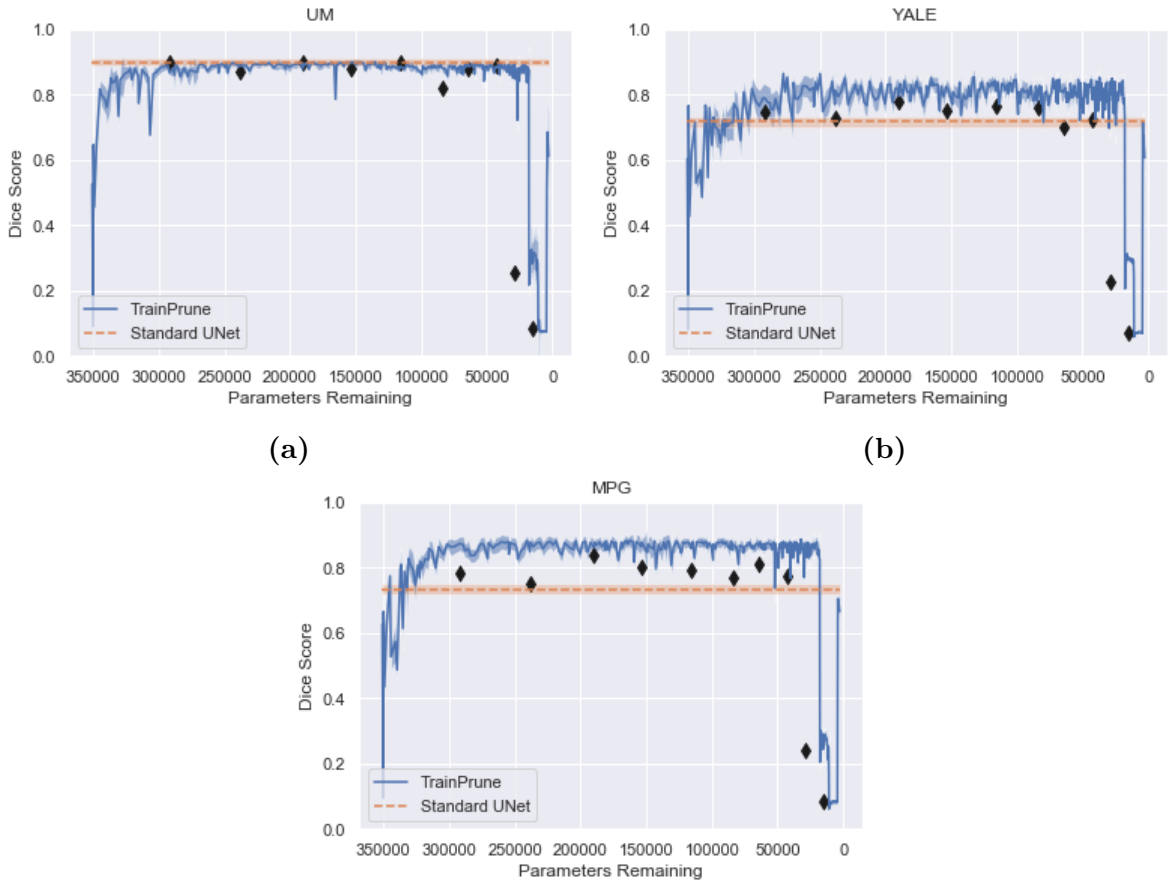


Figure 7.6: Comparison of (i) the performance when randomly initialising and training to convergence the pruned models (black diamonds) to (ii) the performance of the *TrainPrune* models (blue line. Results for the three different sites are shown (training always using UM); only every 30th model was chosen for random initialisation and re-training, due to the computational cost of training all the models.

presented here, showing that the technique is not very sensitive to this choice.

7.2.3.2 Exploring Potential Co-Adaptation

In Fig. 7.6 the effect of taking the pruned models and reinitialising the weights in the models using Xavier initialisation is explored. These models were then trained to convergence using standard training. In [347] they found that fine-tuning a pruned model generally gave comparable or worse performance than training that model with

randomly initialised weights, which went against common findings in the literature [340, 343]. Our results from the UM data show that training smaller networks from random initialisation leads to a network that performs no better or worse than the pruned model. Additionally, the models could be pruned to a smaller size and obtain better performance than when training the small models from scratch. This indicates that complicated relationships between layers can be learned through pruning that are difficult to learn directly using the randomly initialised network. Furthermore, considering the application of the network to the two unseen sites, it can be seen that the randomly initialised models do not generalise well to the two unseen datasets, indicating that our increased generalisability is not just due to the new, smaller architecture but is also a function of the pruning process. It could be that the performance improves because of the co-adaptation of layers and the relationships between layers that are learned when the model is larger. This may be surprising: dropout aims to reduce co-adaptation to reduce overfitting [87] but a degree of co-adaptation is, potentially, beneficial to help learn complex relationships to complete the segmentation tasks, although not so much that they overfit to extraneous information in the dataset.

7.2.3.3 Method Comparison

The results of comparing the *Standard UNet* to *PruneFinetune* are shown in Fig. 7.7. Comparison is made to pruning a single filter at a time and fine-tuning this model until convergence. It can be seen that the two methods have very similar performance on the UM data and when generalising to the two unseen sites of data. The *PruneFinetune* method, however, requires many more training epochs and removing a single filter at a time is much more conservative than would typically be seen in the literature, where usually all filters with magnitudes under a given threshold, or a given percentage or filters, are pruned simultaneously [182]. Figure 7.7d shows the effect on the results

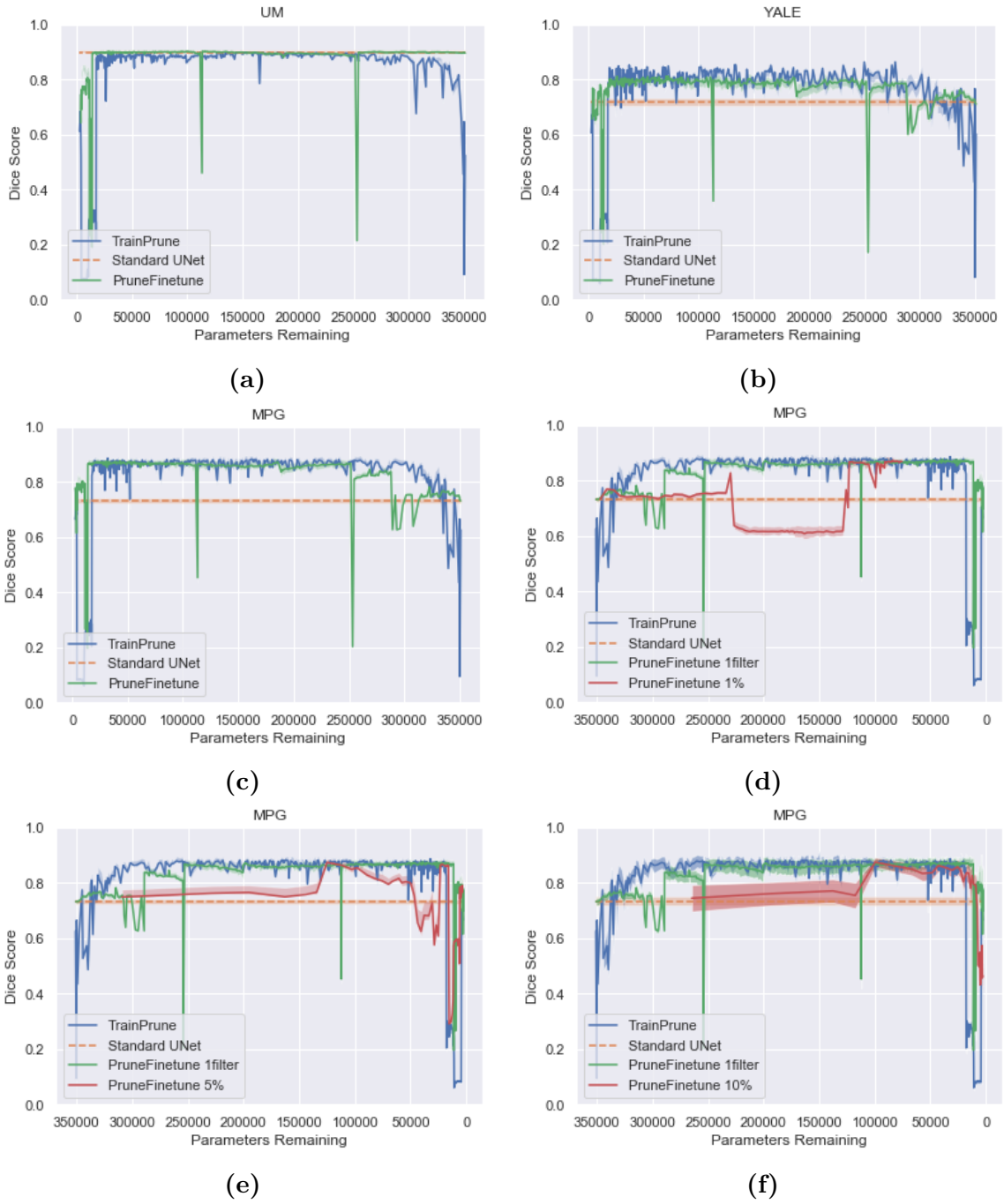


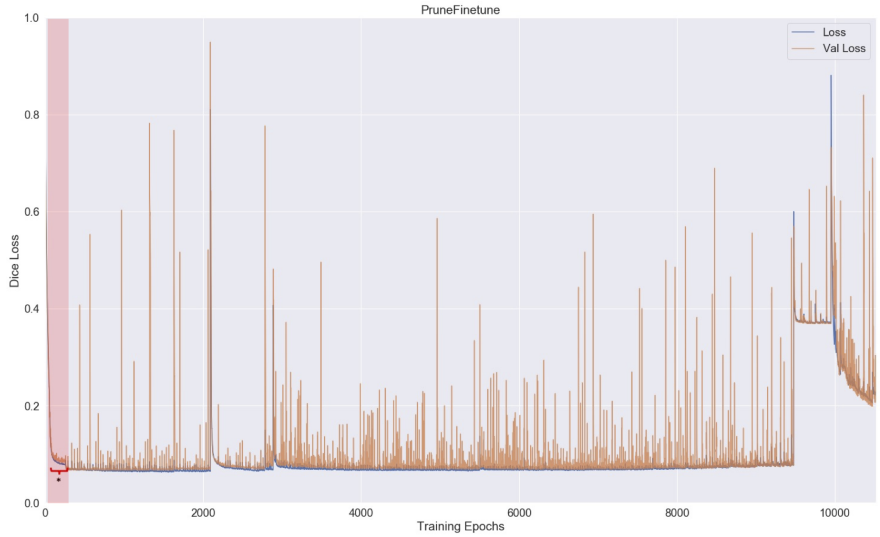
Figure 7.7: The *TrainpruneTD* method compared with training a model trained to convergence and pruning a single filter before finetuning - *PruneFinetune*. The model was trained using the data from UM and is tested on a hold out set from UM and two unseen sites, (b) Yale and (c) MPG. Panel (d) shows the performance on the MPG site data if 1% of filters are removed at a time rather than removing a single filter at a time, showing that the method's generalisability performance is not stable and degrades quickly as larger numbers of filters are removed. Panels e) and f) show the same as d) but for 5% and 10% of filters removed at a time, respectively. The results for the other sites can be found in Appendix C.3.4 and show the same pattern.

on the unseen MPG site if 1% of filters are removed at a time and the model is given time to recover, rather than a single filter (the results for the other sites can be found in the Appendix C.3.4 showing the same pattern). This model had the same performance as when a single filter was removed on the UM site data, but it can be seen that the model has a significantly reduced performance on the unseen MPG data, with the model having to become very small before the performance reaches the same level. This indicates that the amount of training required to recover the model's performance (epochs between pruning iterations), when more parameters at a time are removed, leads to the model overfitting to the training data, despite the pruning. The instability on the unseen sites is observed as the pruning percentage is increased – 5% and 10% can be seen in Fig. 7.7e and f.

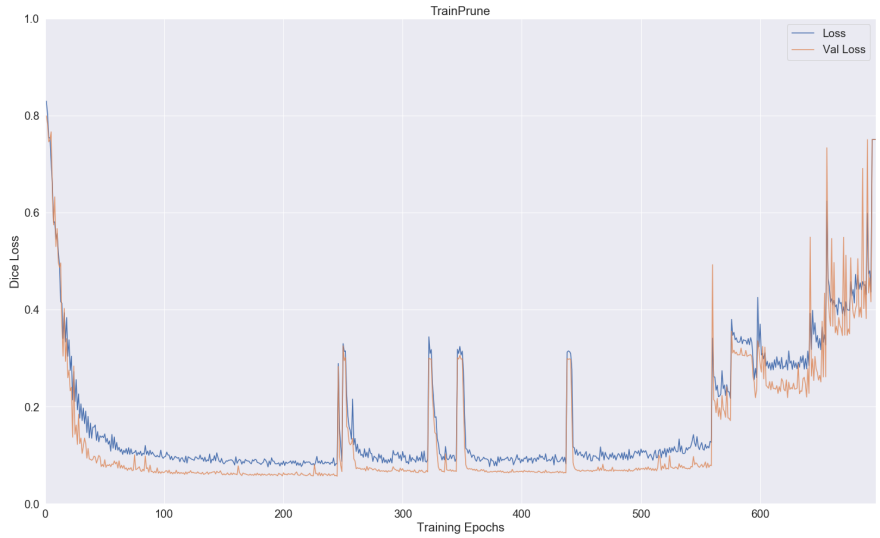
Representative example training graphs can be seen in Fig. 7.8, with *PruneFinetune* taking in the order of thousands of epochs compared to *TrainPrune* taking hundreds. The region indicated in red on Fig. 7.8 represents the epochs used to train the model to convergence before pruning began. This clearly represents a very small amount of the computational cost of the procedure; the bulk of the cost is the pruning and fine tuning, even though the patience was set to 10 epochs. In addition, as explained above, the *TrainPrune* method takes a fraction of the training time and has far more stable validation loss when pruning.

7.2.3.4 Comparison to Dropout

The method was also compared to adding parameter-wise dropout [87] to the *Standard UNet* during training. Dropout is regularly added to models during training to try to prevent the model from overfitting. The value of the dropout was varied between 0 (*Standard UNet*) and 0.5 as can be seen in Fig. 7.9. It can, however, clearly be seen that the addition of even low levels of pruning leads to a reduction in the performance



(a)



(b)

Figure 7.8: Loss graphs with epoch for a) *PruneFinetune* and b) *TrainPrune* against epochs of training. The section indicated in red in a) is the training to convergence of the *Standard UNet*. A patience of 10 epochs was used during training for *PruneFinetune*.

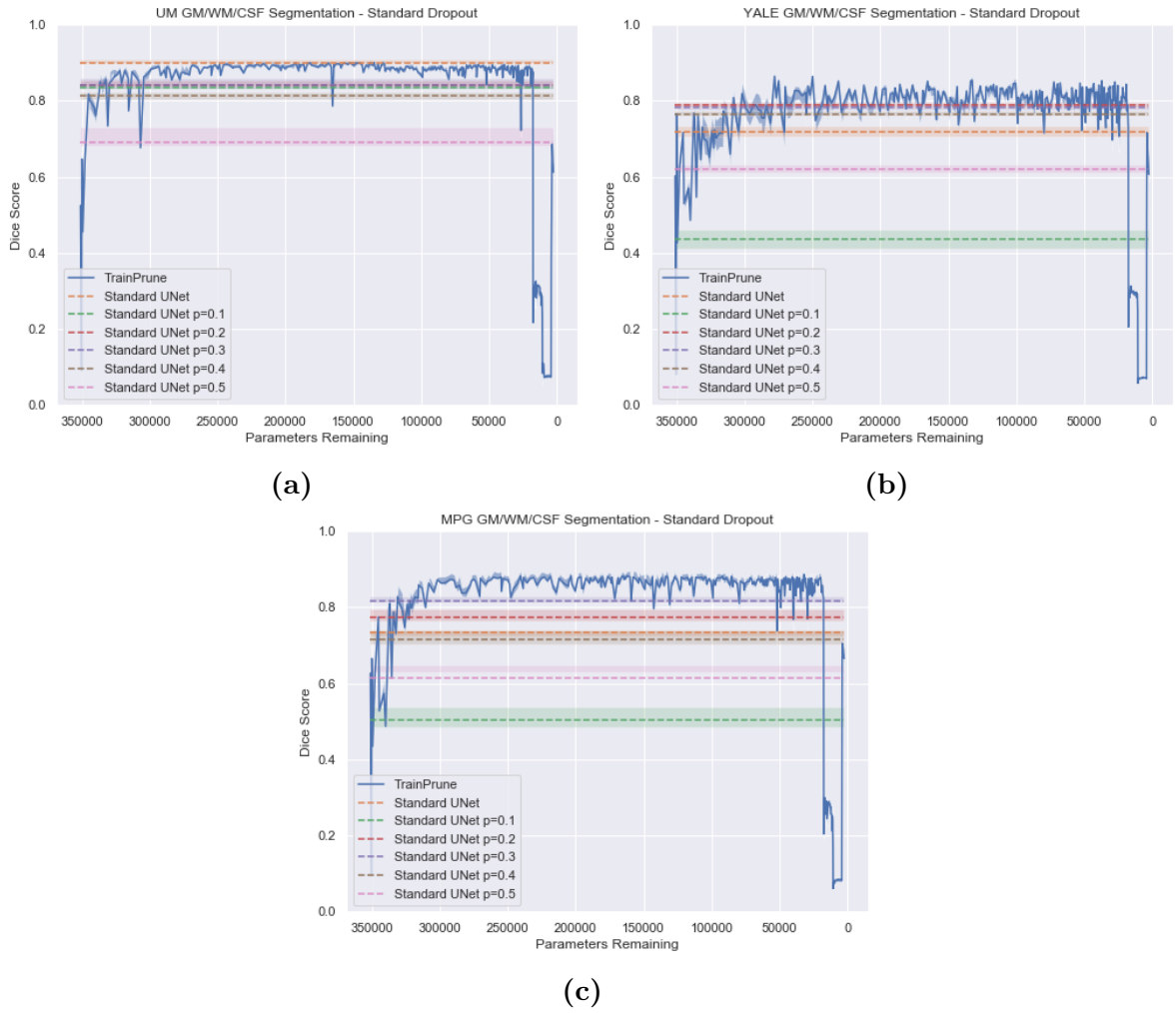


Figure 7.9: Comparison of *TrainPrune* method to *Standard UNet* with increasing amounts of parameterwise dropout [87] added for all three sites. Data from the UM site was seen during training and the other two sites were entirely held out during training.

on the testing data for the UM site data used for training. However, adding low levels of dropout does give an increase in performance over the *Standard UNet* baseline on the two unseen sites, Yale and MPG. In addition, it can be clearly be seen that *TrainPrune* not only maintains better the performance on the UM training site, but also has better generalisability performance on the two unseen sites and so, for this scenario, acts as a better means of preventing overfitting than using dropout. Using a higher level of dropout would not lead to better generalisation performance since,

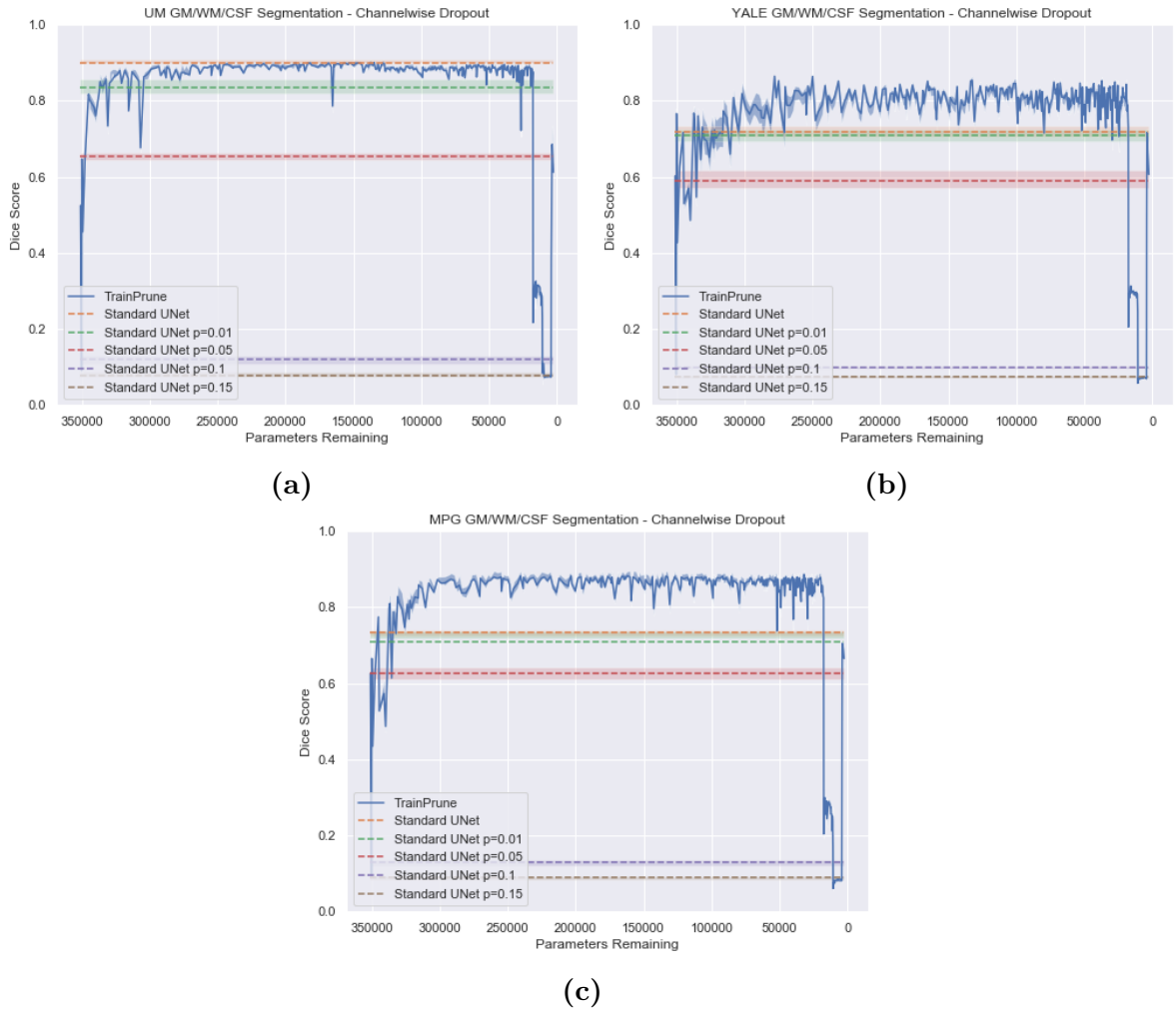


Figure 7.10: Comparison of *TrainPrune* method to *Standard UNet* with increasing amounts of channelwise dropout [361] added for all three sites. Data from the UM site was seen during training and the other two sites were entirely held out during training.

as shown by Fig. 7.9a, the model performed badly with a dropout probability of 0.5, and the performance became worse as the dropout probability was increased. It can be seen from the other two sites that this degree of model performance also leads to poor segmentations on the two unseen sites: as the dropout value becomes too high, the performance on these sites decreases. The experiment was then repeated with channelwise or spatial dropout [361] and in Fig. 7.10 it can be seen that adding spatial dropout does not improve the generalisability performance of the network. Note that

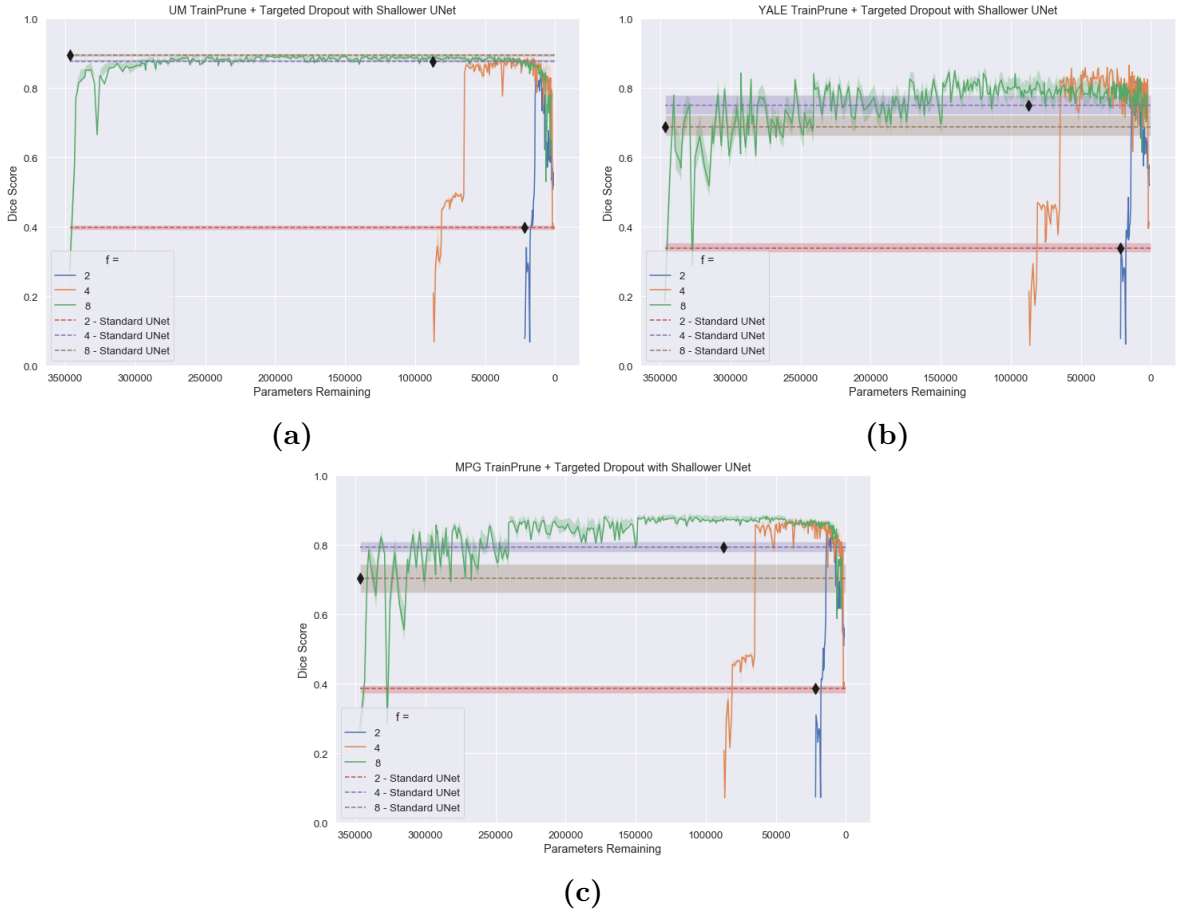


Figure 7.11: Dice results averaged across the three tissue types, comparing the *Standard UNet* and *TrainPruneTD* for models with different numbers of filters - $f = [2, 4, 8]$. The model was trained using data from UM and tested on hold out data from UM and two unseen sites, Yale and MPG.

much smaller values of the dropout were used than for parameterwise dropout, as the effect of channelwise dropout is much more aggressive.

7.2.3.5 Network Architecture Depth

The experiment exploring the use of the shallower network architecture, shown in Fig. 6.4 , was also repeated for the ABIDE dataset. This model has fewer parameters while still having the richness of filters at each depth, so it might be expected to suffer less from overfitting. First, f is varied, testing $f = [2, 4, 8]$, training both the

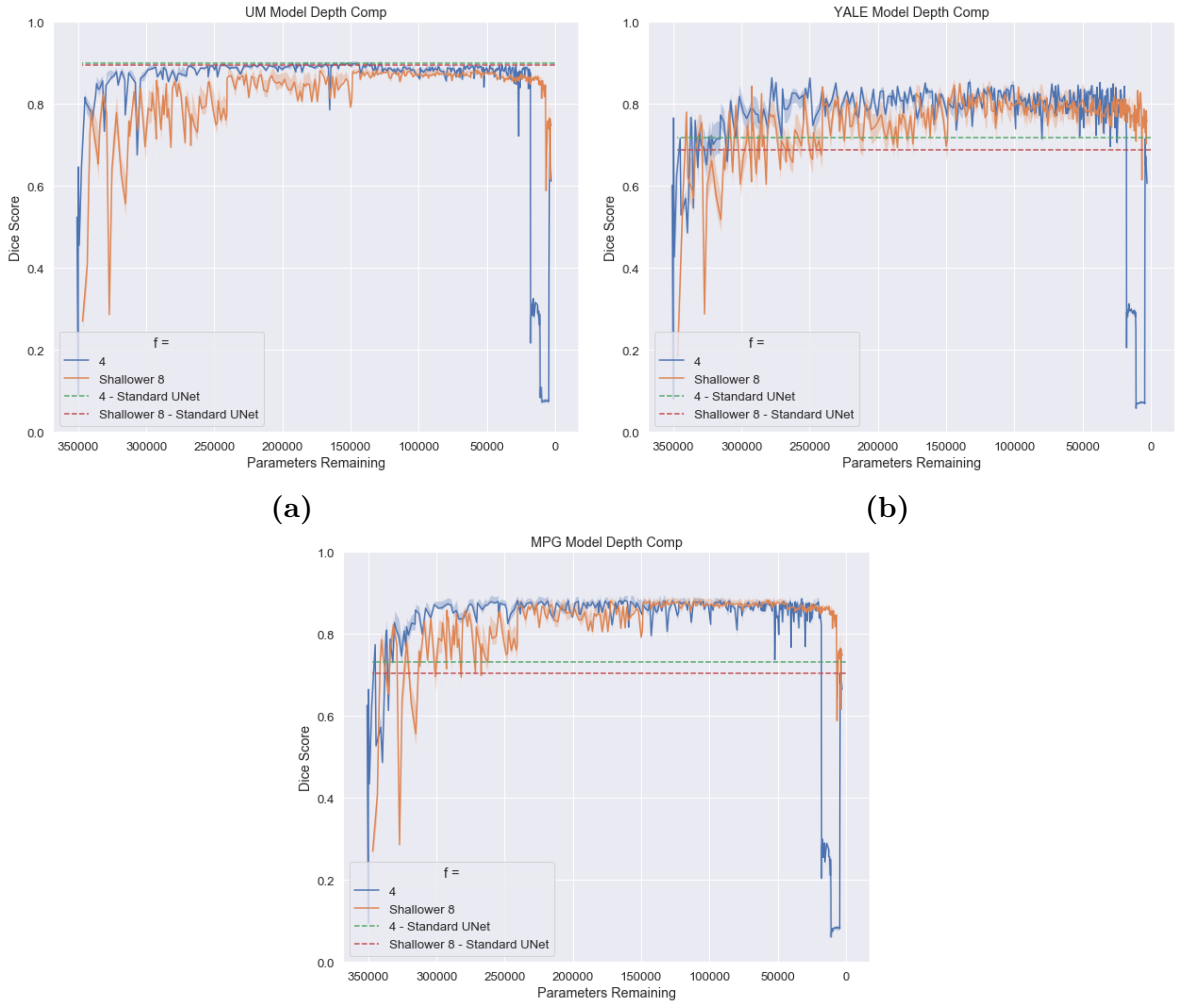


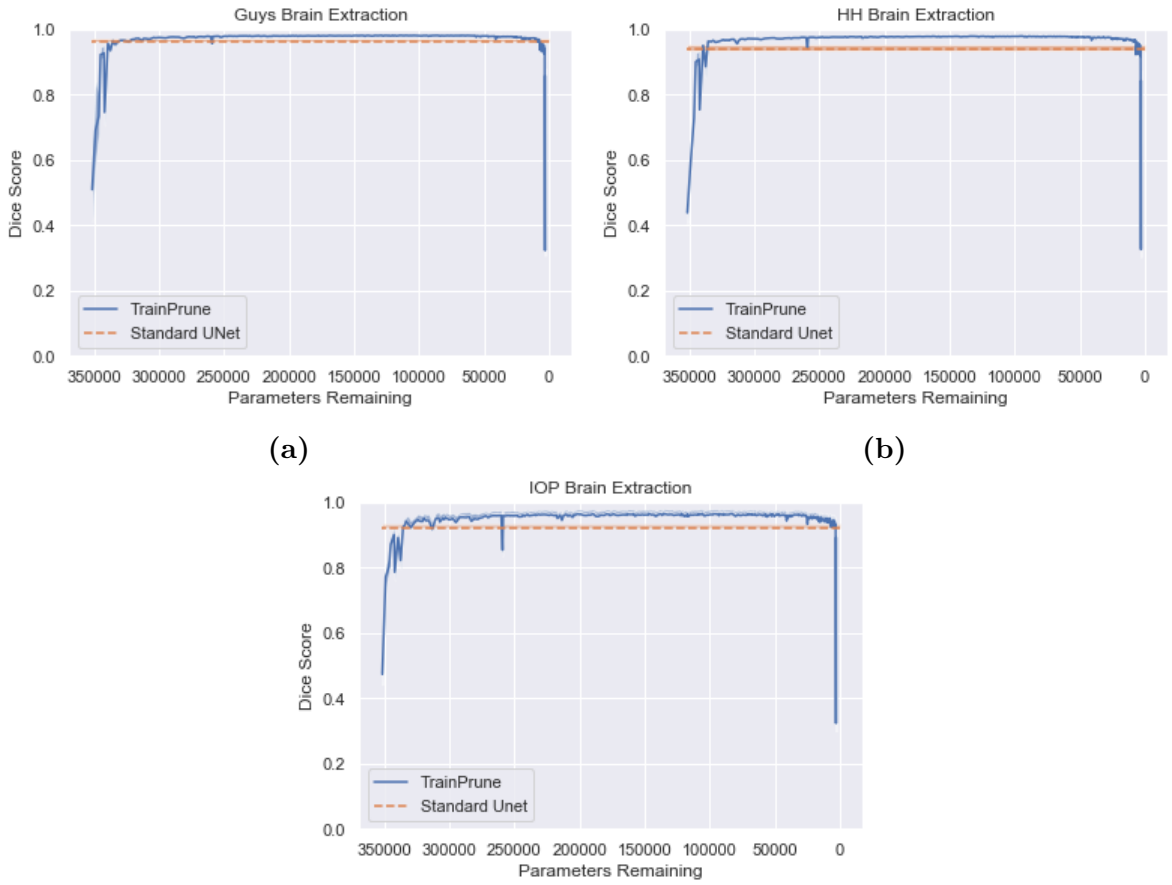
Figure 7.12: Comparison of the results with the original UNet considered with $f = 4$ – Fig. 6.3 – and the Shallower UNet with $f = 8$ – Fig. 6.4. These two models had very similar numbers of parameters at initialisation. The results are shown for both *TrainPruneTD* and *Standard UNet*, with training performed using the data from the UM dataset, and results shown for the hold out data from the UM site and from the held out sites, Yale and MPG.

TrainPruneTD and the *Standard UNet*. From the results in Fig. 7.11 it can be seen that the same pattern is the same seen as with the original sized UNet – Fig. 7.4: the pruned models are able to achieve the same performance as the *Standard UNet* for the training site, UM, and significantly out-perform the *Standard UNet* on the two unseen sites. It can also be seen that out of the Standard UNets, the one with $f = 8$ performs best on the UM site seen during training but generalises worse than the one with $f = 4$ on the unseen sites, clearly demonstrating the degree to which these models overfit to the training data. It was not possible to train a Standard UNet with $f = 2$ to an acceptable standard, but it was possible to prune the larger models to this size with high performance.

Fig. 7.12 shows a comparison of the original size UNet ($f = 4$) and the Shallower UNet ($f = 8$) chosen so that they had approximately the same number of parameters at initialisation. The performance of the two *Standard UNets* was the same but the shallower network overfit the training data more and suffered greater performance degradation when applied to the two unseen sites. The two pruned networks, however, had very similar performance, with significant improvement over the two baseline networks. Therefore, the increased generalisability is not dependent on our initial choice of the network depth and the performance is robust to the initial filter distribution.

7.2.4 Generalisability Across Sites: IXI - Brain Extraction

The TrainPrune method was then applied to the task of brain extraction on the IXI dataset. It was expected that the increased generalisation seen on the ABIDE data should also be seen when working with the IXI data, especially as the data from the Guy's site was obtained with a 1.5T scanner and the data from HH with a 3T scanner; therefore, a visible difference would be expected due to scanner field strength.



(a) **(b)**

(c)

Figure 7.13: Dice results for the brain extraction task, comparing *Standard UNet* to *TrainPruneTD*. The model was trained using the data from Guy's and was tested on a hold out set from Guy's and two unseen sites, HH and IOP.

Figure 7.13 shows the results for the brain extraction task with data from the HH site being used for training and Guy's and IOP being entirely held out. It can be seen that our pruned model is able to successfully complete the segmentation task and that there is an improvement in performance on all three sites. (Guy's: 2.20×10^{-15}, HH: 2.74×10^{-76}, IOP: $p = 5.17 \times 10^{-47}$). It is a relatively simple task and so there is not a large performance degradation on the two unseen sites, but it is still possible to prune the models to a fraction of their original size with a slight improvement in performance.

Figure 7.14 then shows the distribution of the filters as the model is pruned and

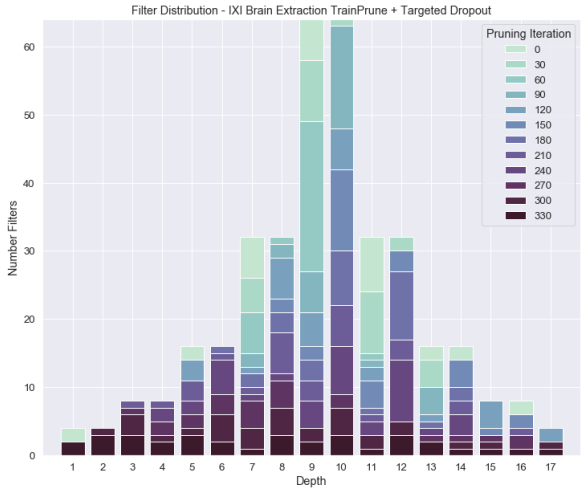


Figure 7.14: The distribution of the filters with network depth as the model is pruned for brain extraction on the IXI data. The darker the shade of the filter block, the longer the filters at that depth were maintained in the models.

trained simultaneously. Again, the distribution of filters as the model is pruned is similar to those from the HarP data (Fig. 6.11a) and the ABIDE data (Fig. 7.3), indicating that the regions which are pruned first can afford to be pruned for a range of tasks, and suggesting that this observation could potentially be used to influence model architectures when the number of parameters needs to be reduced.

7.2.5 Generalisability Across Sites: IXI - Subcortical Segmentation

Finally, the task of segmenting three deep grey matter structures is considered: the caudate, putamen and thalamus. The ablation study explored in the previous chapter is repeated for all three sites. Therefore, the following methods are tested: *Standard UNet*, *Standard UNet* with channelwise dropout, *Standard UNet* with targeted dropout, *TrainPrune*, *TrainPrune* with channelwise dropout, and *TrainPruneTD*. For the models with spatial dropout, the dropout probability was set to the average value of the

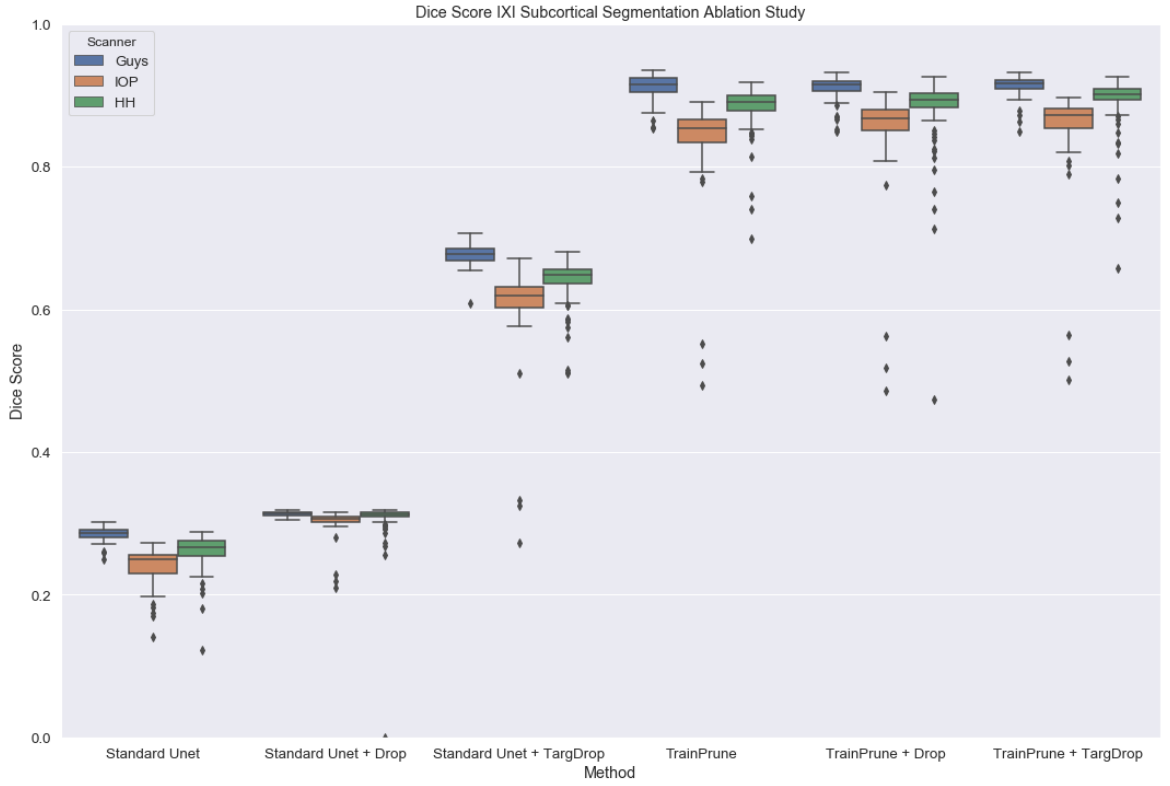


Figure 7.15: Dice results for the IXI subcortical segmentation ablation study, comparing *Standard UNet*, *Standard UNet* with channelwise dropout, *Standard UNet* with targeted dropout, *TrainPrune*, *TrainPrune* with channelwise dropout and *TrainPruneTD*. It can be seen that the simple pruned model outperforms all the models trained without pruning, which were unable to segment all three labels, as seen in Fig. 6.22.

targeted dropout values.

Figure 7.15 shows the results of the ablation study, showing the box plots for each site, averaged over the three subcortical regions. It can be seen that the results are consistent across the three sites.

It can also be seen that while the improvement provided by the targeted dropout is relatively small, it is significant for all three sites (Guy's: $p = 0.0007$, HH: $p = 0.02$, IOP: $p = 4.1 \times 10^{-7}$) and, in addition, increases the stability of the *TrainPrune* procedure. There is also a significant improvement compared to using the same spatial dropout across the networks (Guy's: $p = 1.2 \times 10^{-8}$, HH: $p = 1.2 \times 10^{-12}$, IOP: $p = 3.8 \times 10^{-12}$).

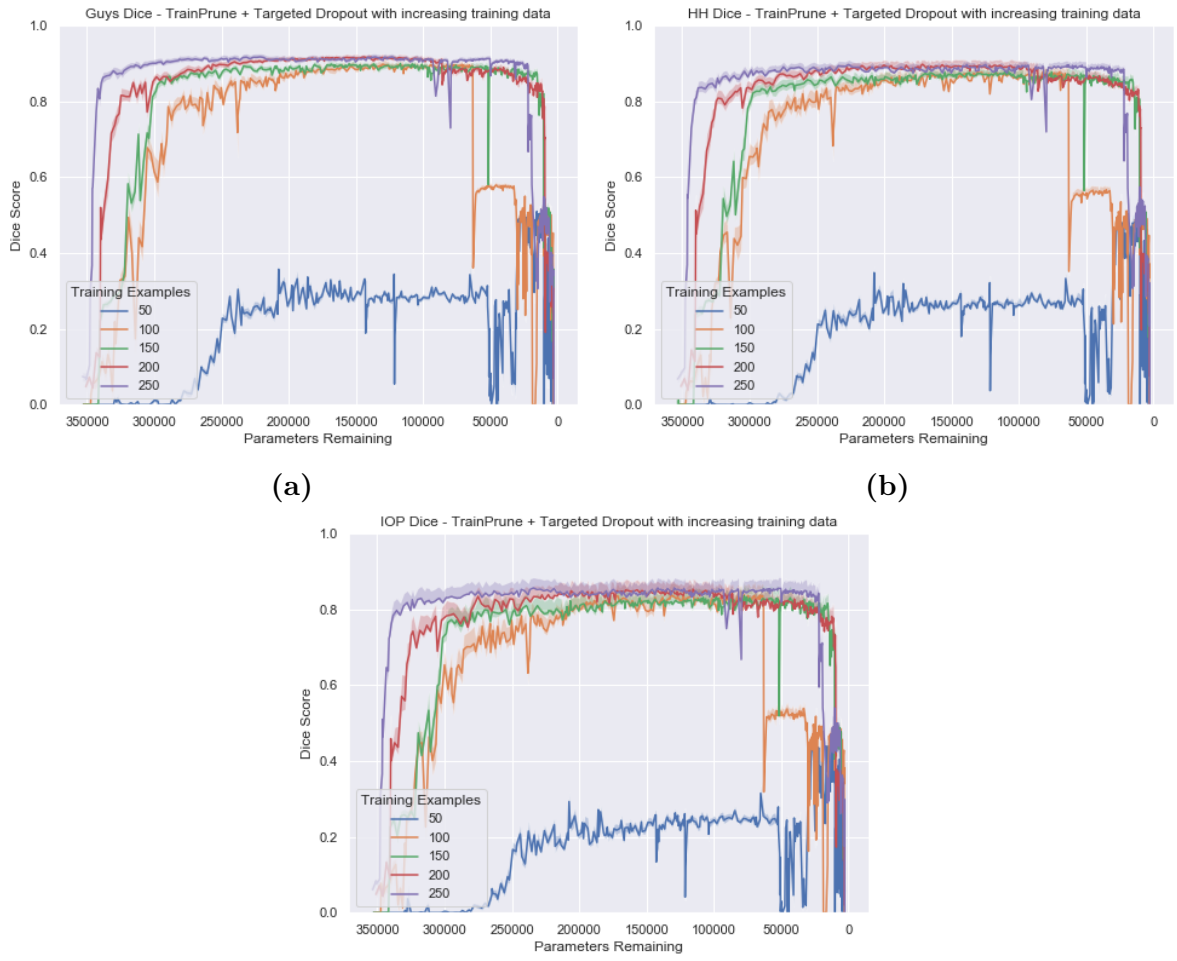


Figure 7.16: Dice scores averaged across the three subcortical regions for increasing numbers of available training images for the IXI data. 250 represents the full training set.

This might be expected as this does not encode information about which filters are the most likely to be removed, so there is a chance that filters that are very important to the performance would be dropped, especially as the pruning progresses, hindering the model training. The value of spatial dropout used was very low (0.037) and with a higher value the performance of the network was hindered further.

Having explored the performance with low data for this data set in the previous chapter, here it is repeated with the predictions for the two unseen datasets. Figure

7.16 shows the Dice scores on the training set for the three sites separately, averaged across the three subcortical regions. The results show that the pruned network can be successfully trained with substantially less data than the *Standard UNet* and, in all cases, provides a better performance; the trained models generalise better to the two unseen sites. This clearly demonstrates an advantage of the pruning approach when working in very low data regimes as is common across medical imaging.

7.2.6 Robustness of Pruned Model

It was hypothesised that, if the pruned model has overfit less to the data, then it should be more robust to reasonable perturbations to the testing data. To test this, the main modes of variation of the data were identified using principal component analysis (PCA). PCA was run on the T1 images from the UM site of the ABIDE data. The latent space of transformed values was then systematically perturbed, to different degrees, for a single component at a time and new, reconstructed images were produced. These were then passed through the trained models for *Standard UNet* and *TrainPruneTD*. The results for the first three components can be seen in Figs. 7.17 - 7.19.

From the figures, it can be seen that the first component appears to be largely due to bias field; the second appears to be due to rotation or registration errors; the third appears to be largely due to ventricular changes that are highly correlated with the age of the subject. It is not possible to quantitatively assess the model performance, due to the lack of ground truth labels for the perturbed images, but qualitative assessment clearly shows that the performance of the *Standard UNet* breaks down earlier than the proposed *TrainPruneTD* model. This can be seen for all three of the components considered, which explain 57% of the variation of the dataset. This indicates that the pruned model is more likely to produce usable segmentations for a wider range

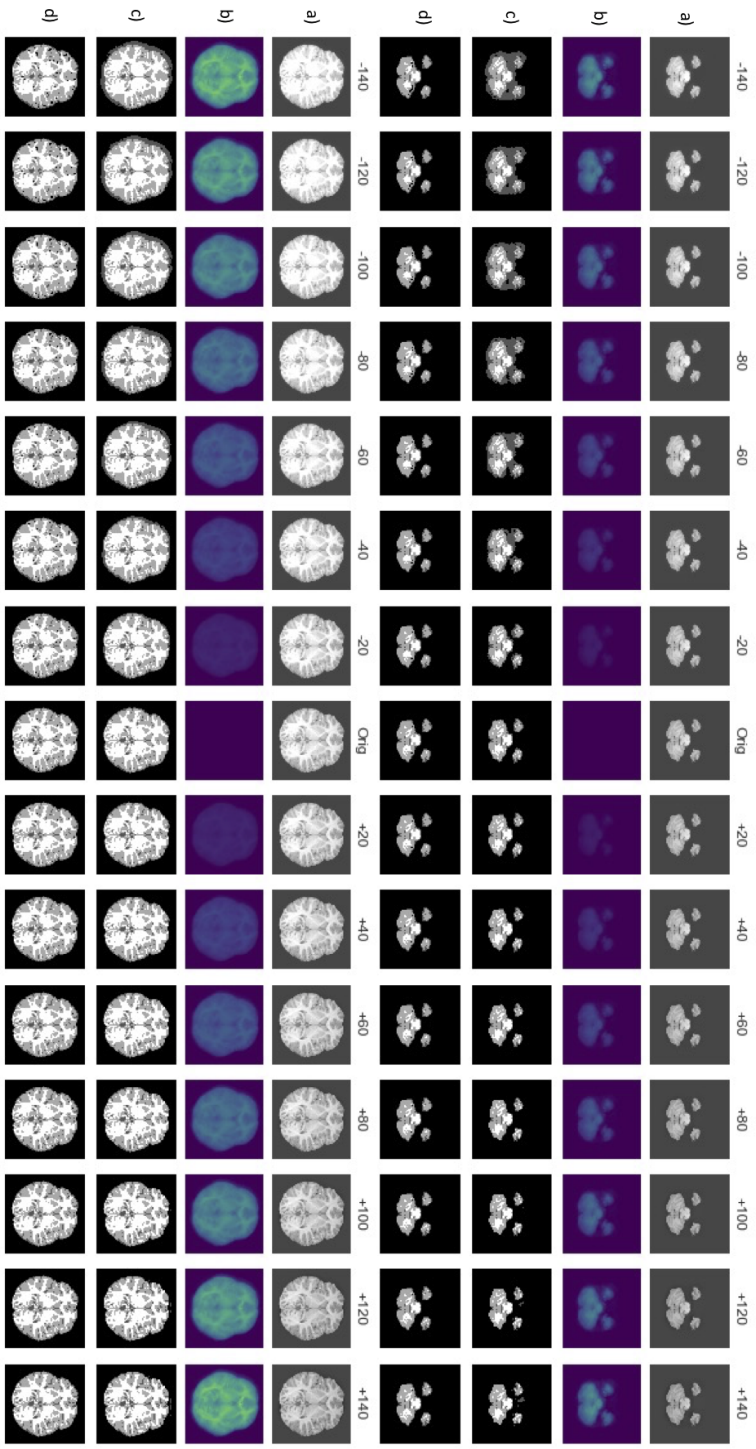


Figure 7.17: Exploration of the first PCA component with two slices for the same subject shown. a) PCA reconstruction perturbed by the specified amount (see percentage change value at top of column) from the original for the first PCA component. b) The absolute difference map between the reconstructed image and the original reconstructed image, showing the changes made by perturbing the component, c) The model prediction from the *Standard UNet*, d) The model prediction from *TrainPruneTD*.

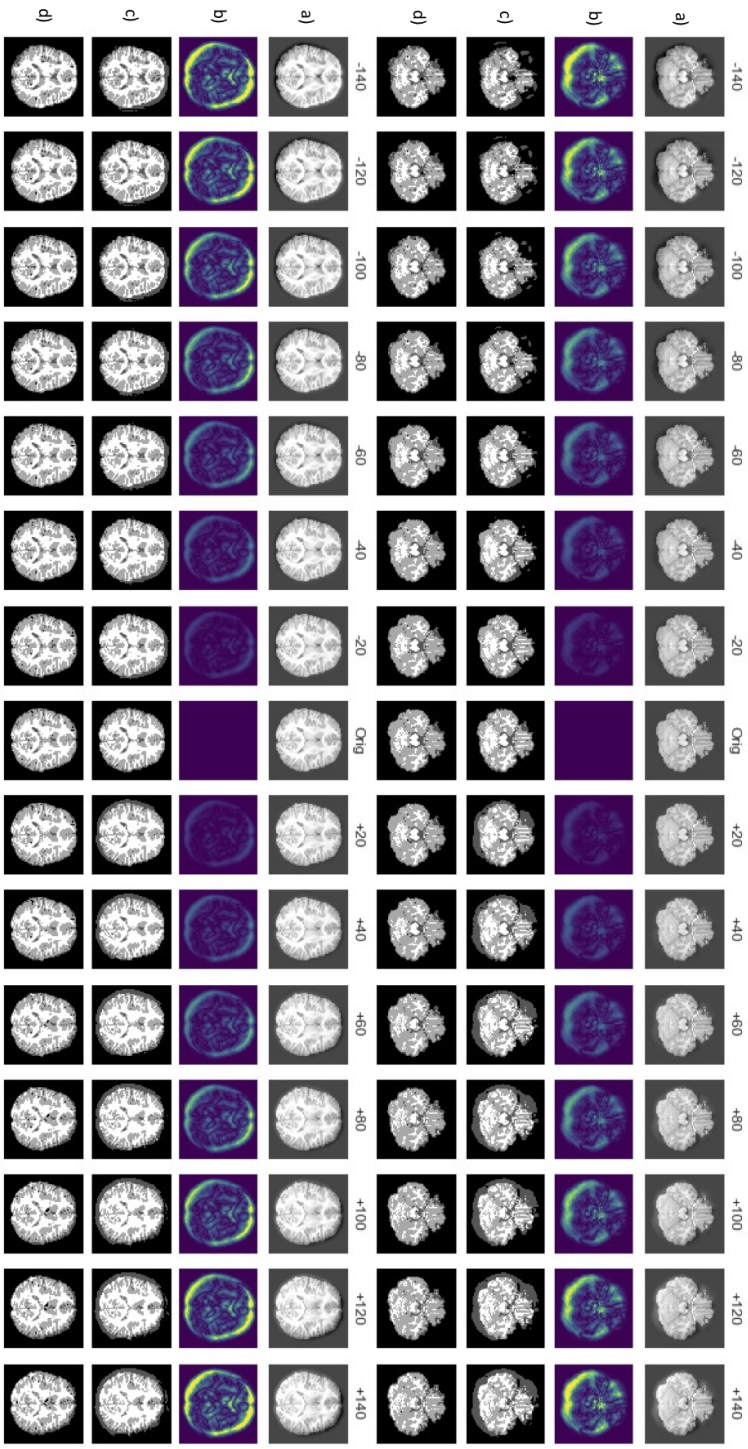


Figure 7.18: Exploration of the second PCA component with two slices for the same subject shown. a) PCA reconstruction perturbed by the specified amount (see percentage change value at top of column) from the original for the second PCA component. b) The absolute difference map between the reconstructed image and the original reconstructed image, showing the changes made by perturbing the component, c) The model prediction from the *Standard UNet*, d) The model prediction from *TrainPruneTD*.

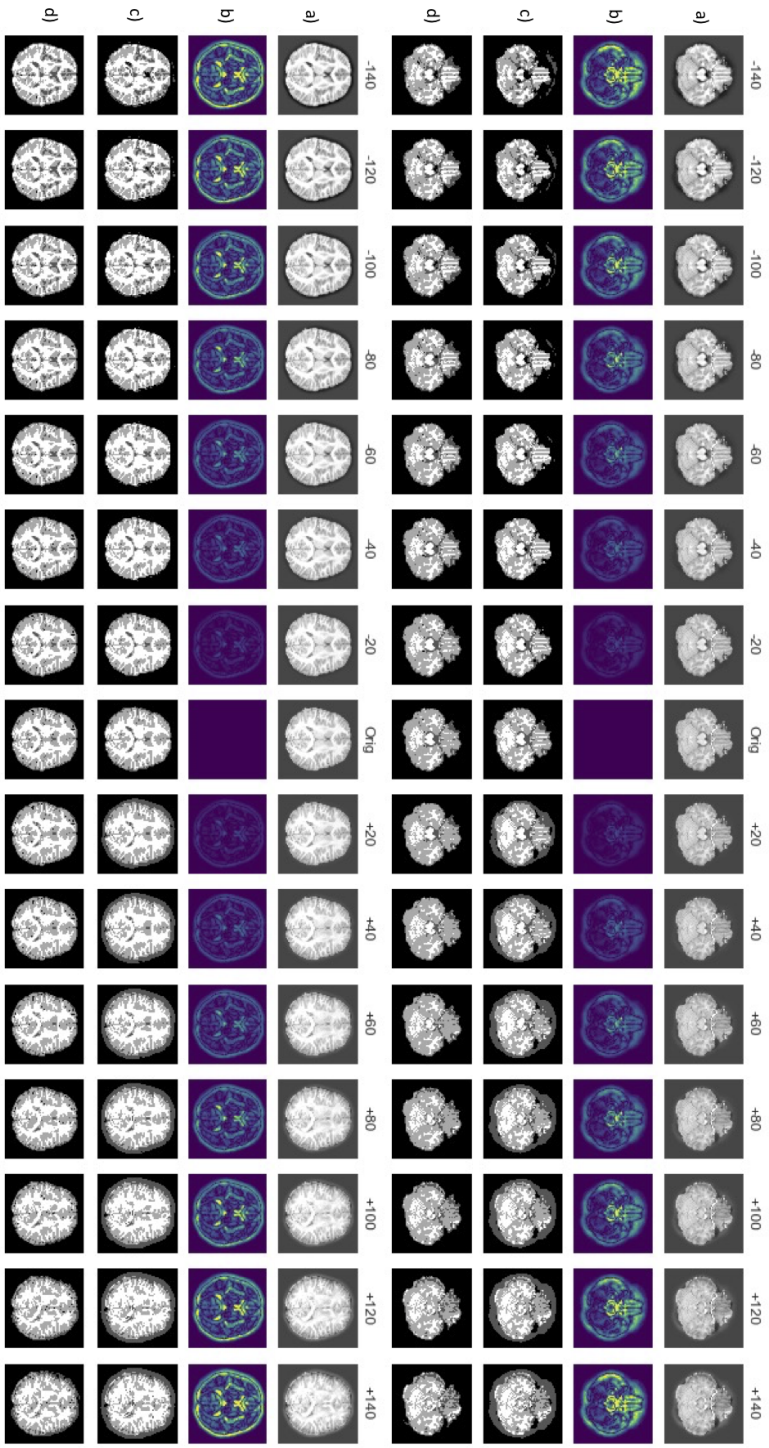


Figure 7.19: Exploration for the third PCA component with two slices for the same subject shown. a) PCA reconstruction perturbed by the specified amount (see percentage change value at top of column) from the original for the third PCA component. b) The absolute difference map between the reconstructed image and the original reconstructed image, showing the changes made by perturbing the component, c) The model prediction from the *Standard UNet*, d) The model prediction from *TrainPruneTD*.

of variation in the data. It can also be seen that both models cope well with the structural changes, as the ventricles grow or shrink in component 3, suggesting that both models are more robust to structural changes than intensity changes.

For further investigation, augmentation was applied at test time to represent changes or artefacts likely to be experienced when considering data collected on different scanners. Therefore, intensity changes were explored, as these seem to be the changes the *Standard UNet* model is most sensitive to, alongside noise in the form of Gaussian blurring, and the effect of anisotropic images, as are common in clinical research. It is not surprising that the model is sensitive to intensity changes as CNNs commonly learn many features based on intensity information. Augmentations were applied to the data using the TorchIO Python package [372] with only one type of augmentation used at any time, and the dataset was augmented 5 times for each type of applied augmentation, creating 100 test samples. The applied augmentations were:

- Simulated Bias Field: the bias field (or RF receive inhomogeneity) is modelled as a linear combination of polynomial basis functions and the maximum magnitude of the polynomial coefficients is varied across the experiment where the coefficient is drawn from $n \sim Uniform(0, maximum)$. The order of the polynomial was set to the default value.

- Contrast change: the intensity values are changed by raising the values to the power of $\gamma = e^{\beta}$. The range from which β is sampled – $\beta \sim Uniform(-val, val)$ is varied throughout the experiments.

- Gaussian blurring: the image is blurred using a Gaussian filter where the maximum value of the standard deviation of the Gaussian is varied across experiments and is drawn from $\sigma \sim Uniform(0, val)$.

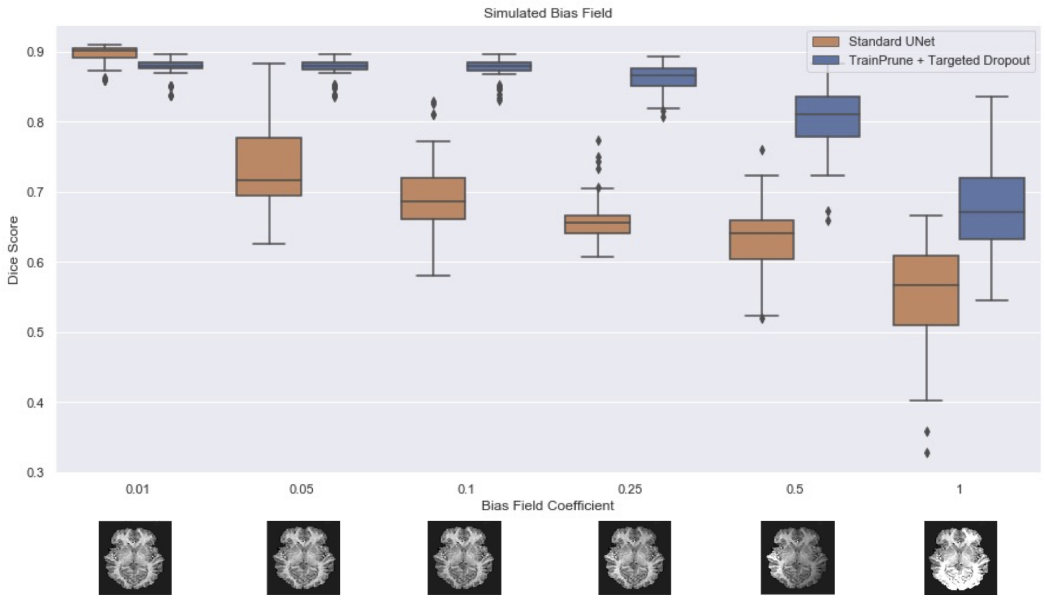


Figure 7.20: Results from applying simulated bias field to the images at test time, comparing the *Standard UNet* to *TrainPruneTD.*

- Anisotropic sampling: the image is downsampled along an axis chosen at random and then upsampled back to the initial space with the downsampling factor varied across the experiments.

It is assumed that the ground truth labels remain the same despite the augmentation. This is potentially not true, especially for Gaussian blurring and anisotropic sampling, but it allows for comparison between the two methods.

The results from applying these augmentations at test time can be seen in Figs. 7.20 - 7.23. Considering the two intensity changes first, simulated bias field and contrast change (Fig. 7.20 and Fig. 7.21), it can be seen that the pruned model maintains higher performance values than the *Standard UNet*, which suffers significant performance degradation even for very low levels of intensity change. This confirms the results of the PCA investigation that the *Standard UNet* is more sensitive to intensity values and that this causes the network to overfit to the training data set, as intensity values and the relative intensity values are known to vary across scanners.

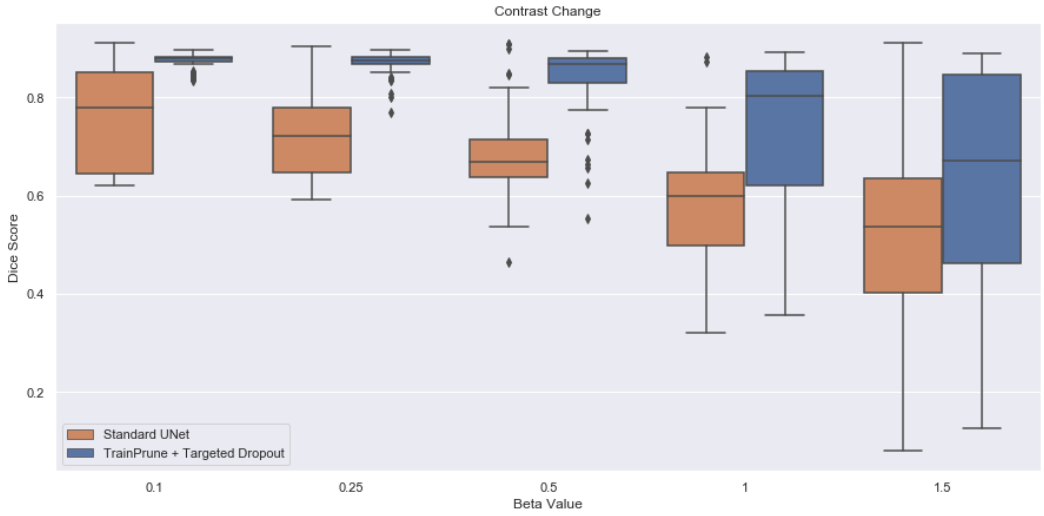


Figure 7.21: Results from applying intensity changes to alter the image contrast of the data at test time, comparing the *Standard UNet* to *TrainPruneTD.*

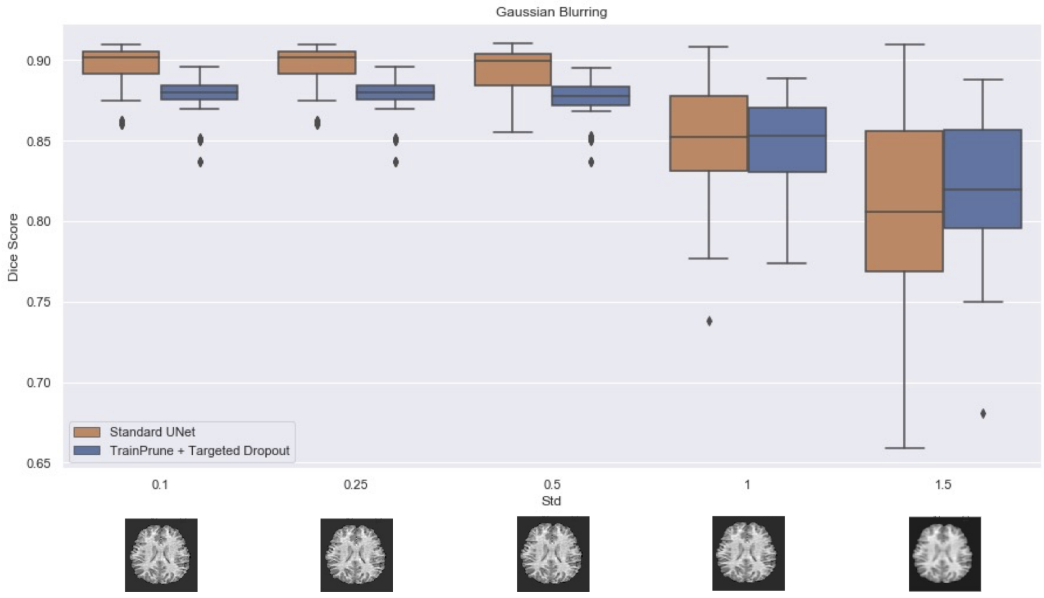


Figure 7.22: Results from applying Gaussian blurring augmentation to the data at test time, comparing the *Standard UNet* to *TrainPruneTD.*

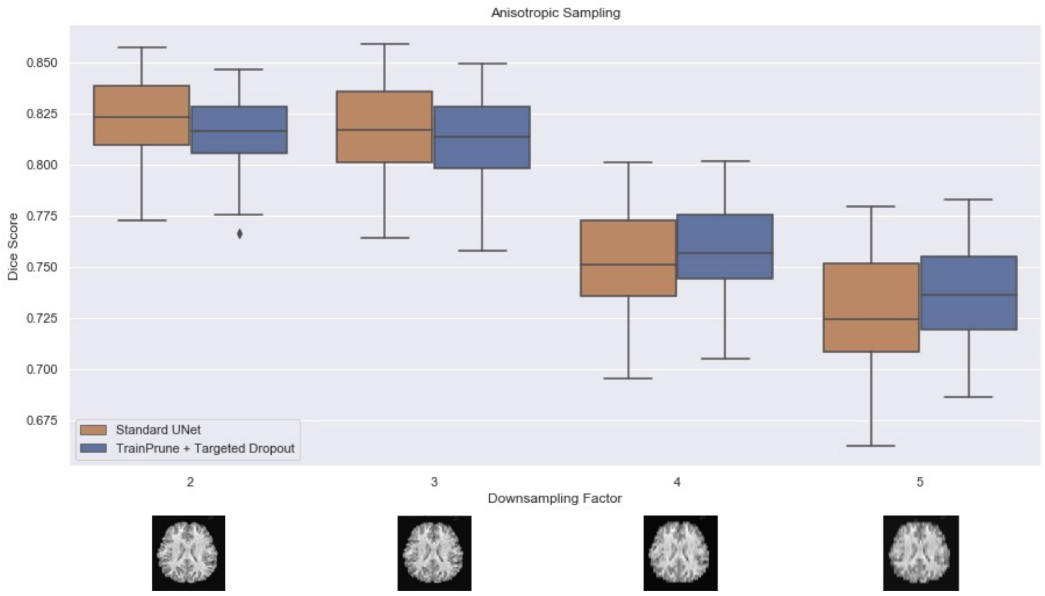


Figure 7.23: Results from downsampling an axis chosen at random to simulate the effect of anisotropic voxels in the data at test time, comparing the *Standard UNet* to *TrainPruneTD*.

Considering Fig. 7.22 and Fig. 7.23, it can be seen that the two methods perform very similarly when presented with blurred images or anisotropic voxels. Therefore, if these types of variation were expected to be seen in the datasets, then it would be helpful for them to be applied as augmentations during training since the pruning approach does not improve the robustness to these changes.

To test the relative benefits of data augmentation and pruning, intensity-augmented data was applied to the training data and used to train the *Standard UNet*. The *TrainPruneTD* network was not retrained, and so did not see any of these augmentations at training time. The data was augmented 5 times and the original data included, using the same values that were explored in the testing. The result can be seen in Fig. 7.24. It can be seen that although the addition of the augmentations led to a reduction in performance on the UM site seen during training, it led to an increase in performance on the two unseen sites. The pruning alone still gave an increased generalisability on the two unseen sites. This indicates that although the generalisability may also

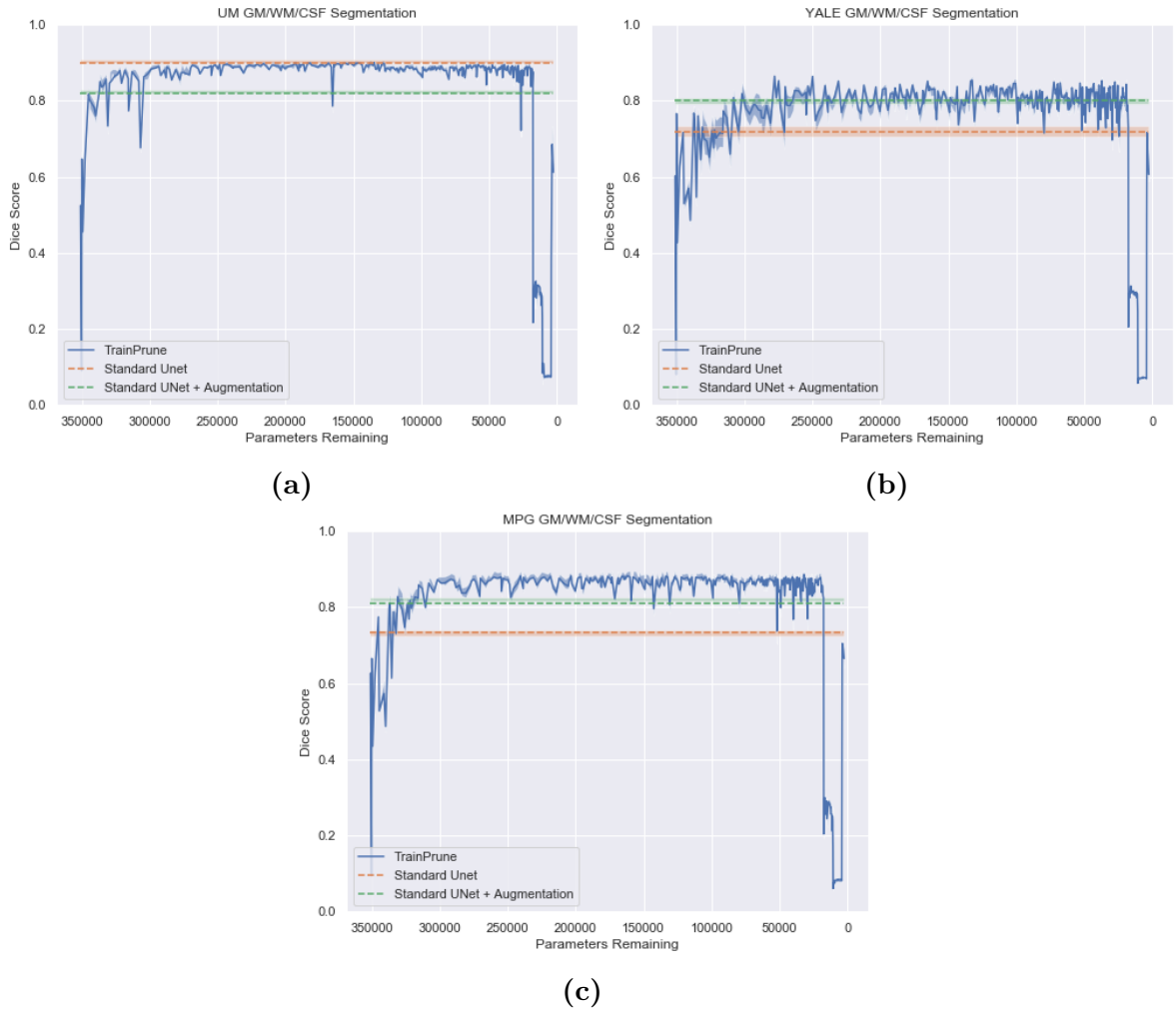


Figure 7.24: Results on the ABIDE data from training *Standard UNet* with intensity-based augmentation and comparing to *TrainPruneTD* and *Standard UNet*, both trained without the augmentation.

be achievable through augmentation, the use of pruning allows us to achieve the generalisability without having to identify the set of augmentations required.

7.3 Conclusion

The *TrainPrune + Targeted Dropout* (TrainPruneTD) algorithm introduced in chapter 6 has been applied to multisite data, and has been shown to lead to an increase in

generalisability across unseen sites. The lack of generalisability of deep networks when applied to datasets drawn from different distributions to the training data – such as data acquired on different MRI scanners – represents a major limitation for the deployment of deep learning methods. While some of the problem can be removed by pooling datasets and applying data augmentation to increase the variability of the training data, the lack of manually labelled segmentation data available in medical imaging means that deep learning methods have a strong tendency to overfit the training data. The results presented clearly show that the *TrainPrune* approach can be used to limit this overfitting. It has also been shown that this is not the case when using traditional pruning approaches – eg. *PruneFinetune* – or through training smaller models in the first place, indicating that the improvement is not just due to the reduction in parameters but also due to the training method reducing the degree of overfitting to the data. Comparison with the use of dropout also indicates that the method leads to better generalisability and better performance on the dataset seen during training.

Clearly, a limitation of the work is that simulated labels were used for the training and testing labels, so that it was possible to explore multisite datasets. These labels are clearly not gold standard manual labels, but the three types of automated labels were generated under different assumptions so it should be that the result is not a function of the method used to generate the labels. The use of the generated labels is likely to impact the maximum performance that we can achieve but allows comparison between the methods, especially given that the method has previously been validated on data with manual labels.

The robustness of the model and therefore the limits of the generalisability have also been explored. Perturbation using the three strongest components from a PCA analysis indicates that the major types of variation are intensity, head position and morphological changes as seen in ageing. Both the *Standard UNet* and *TrainPrune* were

robust to the ventricular changes as is often seen in ageing, but the *Standard UNet* was much more affected by the intensity changes. This was confirmed by the application of augmentation at test time, where the performance of the pruned model was much more robust to the intensity-based augmentation. Training the *Standard UNet* with these intensity-based augmentations gives an improvement in generalisability compared to the *Standard UNet* baseline but not to the same extent that was achieved by the pruned model. This indicates that a benefit of the pruning approach is not having to identify the full set of transformations that would need to be applied during training to lead to the same degree of generalisability. The results on the blurred data or the anisotropic voxels indicate, however, that if this sort of variation were to be expected across the sites then it would be helpful to apply this variation as an augmentation during training, since pruning alone does not improve the results. The results are limited by the labels which were generated automatically rather than being manual labels, so it is not certain what effect the augmentations would have on the ground truth, although they can be used to compare between methods with an understanding of this limitation. Therefore, further investigation is needed on the interplay between pruning and augmentation for different scenarios. The results also indicate that the variation that will cause models to fail can be begun to be explored, but needs to be investigated further when assessing the performance of models.

The results here are limited to multisite datasets; future developments should focus on whether the use of model pruning can help segmentation networks generalise to other datasets not acquired as part of the same study or protocol.

8

Conclusion

8.1 Contributions

This thesis has explored three distinct problems relating to the analysis of MRI of the brain. For all three contributions, the problem has been introduced, a method developed, and then the results explored and compared to appropriate alternative methods. This chapter summarises the main contributions of this thesis and discusses the limitations of the present work. Finally, opportunities for future research directions based on this work will be discussed.

- **Model of brain ageing using data from the UK Biobank**

  The first contribution was the development of a model to explore brain ageing.

A model was trained on subjects from the UK Biobank to predict age from T1-weighted MRI scans, stratified by sex, and the resulting $\Delta_{BrainAge}$ values were correlated with the IDPs and non-imaging variables from the UK Biobank. The approach was compared with an elastic net method trained on the IDPs. The results were also consistent when the data were screened for the health of the subjects. Finally, the effect of the registration in the pre-processing pipeline was explored: non-linear registration is a standard part of many pre-processing pipelines, due to its requirement in many group analyses. It was shown that the use of non-linear registration when using the CNNs to predict brain age may lead to the model being driven by artefacts of the registration process and missing subtle changes which occur away from the ventricles, meaning that the predictions may not appropriately reflect deviations from 'normal' ageing.

- **Development of an 'unlearning' domain adaptation based approach for harmonisation**

  A method for harmonisation was then explored. The data harmonisation problem in MRI is the fact that data collected on different scanners or with different acquisition protocols encode information about this acquisition, such that, when the data across sites is combined, there is an increase in variance due to nonbiological factors. It was observed that the harmonisation problem can be modelled as a domain adaptation challenge, and so domain adaptation methods should be applicable, and an 'unlearning' domain adaptation approach is adopted for harmonisation. It was shown that the method is able to remove scanner information while maintaining the performance on the task of interest for regression, classification and segmentation tasks, and for a range of data scenarios. This indicates that the framework should be applicable to a range of

neuroimaging tasks, especially as the framework can be used with the range of feedforward architectures.

- **Simultaneous model pruning and training for low data regimes in medical image segmentation**

  A method for simultaneous model training and channelwise pruning was developed, alongside an adaptive channelwise targeted dropout algorithm, to encourage the model to be robust to the pruning. The method is applied for segmentation across modalities and tasks. It was demonstrated that this method led to improved performance compared to standard training of the UNet model and the effect of the design decisions were explored. It was then demonstrated that the use of the pruned models yielded improved performance in low data regimes, which are common scenarios across medical imaging tasks.

- **Model pruning for Generalisability Across Imaging Sites**

  Finally, the pruning algorithm was applied to multisite MRI data where significant degradation in performance is experienced when applying a model trained on one site to the remaining sites. The pruning algorithm developed in the previous chapter was applied to this multisite data and it was shown that the decrease in overfitting due to the pruning leads to an increase in generalisability across the sites. Further, the data were perturbed, both through augmentation and through exploration of the major PCA components, to allow exploration of the extent of the increase in generalisability.

8.2 Limitations

The limitations of each chapter have been discussed in the respective concluding sections, but here, the major limitations of each project will be discussed, which provide the opportunity for future research to improve on the performance, applicability and impact of the proposed work.

- **Use of the UK Biobank data:** The major limitation of the brain age model is the use of the UK Biobank data. The UK Biobank is the largest imaging study to date and the availability of the health data makes it a powerful resource; however, there are limitations of the dataset, and, therefore, of any model trained based on these data. One limitation is the selection bias among people who are able to volunteer, with the UK Biobank population not representing well the UK population as a whole, in terms of education, socio-economic status and health, leading to a lower frequency of diseases within the participant cohort than is present in the general British population. Another limitation of the dataset is that the study participants are overwhelming white and so any findings resulting from using this data cannot necessarily be expected to generalise beyond this specific population. Additionally, some of the categories in the biological health data used rely upon self reporting, such as diet and lifestyle, and so there may be bias introduced depending on how the questionnaires were completed. These limitations are likely to be experienced across studies of this type and so motivate the need to be able to combine data across scanners, thus, the solving of the harmonisation problem, to enable diverse populations to be incorporated into a model. Generally, this needs to be considered when interpreting the results.

- **Brain age paradigm:** Although the use of the brain age paradigm (where a

CNN is trained to minimise the difference between the true age and the predicted age for a given metric) has been widely explored e.g. [217, 220, 258, 373] it would be reasonable to question the validity of minimising the value that is of interest, the $\Delta_{BrainAge}$ value. A large $\Delta_{BrainAge}$ value could either be an indicator of deviation from normal ageing, an indication that the testing data has been drawn from a different distribution, or that the model has been trained poorly and so is performing badly on the testing data. While effort has been taken to explore some subjects with large $\Delta_{BrainAge}$ values, to try to explore the validity of the values, and correlations have been found which match our understanding of the ageing brain, it is hard to understand the effect of the decisions made in training the model on the results. Further, the validation of the model would need to be completed for every model of its kind; the result cannot be assumed to generalise.

- **Harmonised images:** The proposed approach for harmonisation is potentially limited by the fact that it is not able to create harmonised images, but rather only to limit the influence of the scanner information on the predictions of the model. For use in traditional processing and analysis pipelines, harmonised images would be desirable. Adversarial approaches are unlikely to be appropriate for generating harmonised images, as the process removed information, and so the generated images are unlikely to truly resemble 'normal' images, and would be likely to be visually blurred. Validation of the results obtained with the generative models would be difficult without large quantities of 'travelling heads' data, and would need to be validated for every downstream tool, to ensure that they function as expected with the generated images. Many generative methods for harmonisation are based on disentangled representations [315, 374] or style

transfer [314, 375] which can only account for contrast changes and so cannot account for systematic morphological differences, for instance due to gradient non-linearities. Therefore, while generative solutions to the harmonisation problem are desirable, the development of an approach that can deal with the range of effects that can be caused through the use of different acquisition scanners is necessary.

- **Validation of the pruned model's robustness:** The model's robustness to variation was evaluated both using qualitative examination of the predictions from the PCA (as no ground truth was available and the effect of the perturbation on the model could not be quantified), and through applying augmentations and evaluating the model's performance on the augmented data. The best policy as to whether or not to apply applying the augmentations to the labels – and what effect they would have – is not clear, and although the results are sufficient for direct comparison between methods, a future direction for studying model robustness and generalisability more widely could consider a principled way to investigate this question. Ideally, images and labels spanning the expected variation would be acquired, but that is infeasible and so a method for generating high-quality image and label pairs to simulate this would be desirable. Further, no labelled multisite datasets were available, and therefore the results are based on simulated labels using different tools. The three tools use different models to generate the labels; however, these labels are not perfect and may potentially be adding systematic bias to the results. Without the availability of manual labels for multiple datasets, it is hard to remove this risk.

- **Interplay between augmentation and model pruning:** Both data augmentation and the proposed *TrainPrune* method aim to improve the model's

generalisability and robustness to expected variation in the data. Although the effect of training the *Standard UNet* model with augmentation was compared to the pruned model, the interplay between data augmentation during training and the model pruning is yet to be explored.

8.3 Future Directions

Many developments based on this thesis and related problems could be investigated. Here, just a few possible directions for further research will be discussed.

- **The interplay between harmonisation and model pruning:** As has been shown in chapter 7, the use of model pruning reduces the sensitivity of the model to changes in intensity due to the scanner rather than tissue. As these changes are a major component of the variation that must be removed during the harmonisation process, it is probable that pruning the network being used in the harmonisation process might aid the harmonisation process. If the bulk of the overfitting to the intensity information were removed through the pruning process, then the features would be able to specialise on the removal of information relating, for instance, to the systematic morphological changes, which are harder to account for. The pruning algorithm would, however, first need to be validated for model architectures other than the UNet to allow for application across tasks.

- **Pruning other model architectures:** The *TrainPrune* algorithm has only been applied to a basic UNet architecture. While it is expected that the results should generalise to other network architectures – and many segmentation architectures are based upon the UNet architectures – the results have not been validated for any other architecture. This would be required to ascertain whether

the seen generalisability and reduction in model overfitting would apply across architectures. Further work therefore, could systematically explore architectures and tasks to see where the most benefit can be gained from the use of model pruning.

- **Removal of known confounders:** Model interpretability in deep learning is an unsolved problem and a problem of high importance in medical imaging. None of the contributions in this thesis have contributed to solving that challenge; however, it has been demonstrated that the harmonisation framework can also be used for the removal of additional confounds from the data. Therefore, the framework could be applied to remove information from models that should not influence the model prediction in any way. Through explicitly removing the information from the features used to make the model predictions, a level of confidence could be provided that this did not influence the decision making process; so, whilst this approach could not give information about what *did* drive the prediction, it could give information about what did *not* drive the model prediction. For instance, for conditions unrelated to sex, sex information could be removed to ensure that this did not influence the predictions. Whenever bias exists in datasets, this could be a tool towards enabling models to be applied with some confidence that the bias is not unduly affecting model predictions.

A

Dataset Details

Multiple datasets were used throughout this thesis, all of which are either openly available or available to researchers on application. No new data was acquired for this study. Listed below are the datasets used, with discussion of the dataset aims and the ethics statement for the study.

A.1 UK Biobank

The UK Biobank [1] is a large-scale prospective epidemiological study with all data accessible to researchers world wide. The included imaging modalities are T1, T2 FLAIR, swMRI, dMRI, rfMRI, tfMRI, and health and lifestyle information is also available for the subjects. The data used in this study had been processed using the UK

Biobank pipeline [269]. The T1 images were acquired with $1mm$ isotropic resolution.

A list of the IDPs can be found at `https://www.fmrib.ox.ac.uk/ukbiobank/gwaspaper/10k_IDPinfo.txt`. They represent summary statistics generated from the different imaging modalities. Generally, the values from the structural images represent the volumes of different structures: from the fMRI they are generally the median or 90th percentile values of the BOLD effect or z-stat values, and from the dMRI the mean of a diffusion derived characteristic (such as MD) across a predefined tract. More details about the IDPs can be found at `https://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=100`.

The data has been used in the following ways in this thesis:

- Brain age prediction: T1 data and age information used for training data, correlations made with all modalities and health and lifestyle information;
- Harmonisation: T1 data and age information used as training data.

The UK Biobank has approval from the North West Multi-centre Research Ethics Committee (MREC) to obtain and disseminate data and samples from the participants, and these ethical regulations cover the work in this study. Written informed consent was obtained from all participants. Details can be found at `www.ukbiobank.ac.uk/ethics`. This research has been conducted in part using the UK Biobank Resource under Application Number 8107. We are grateful to UK Biobank for making the data available, and to all UK Biobank study participants, who generously donated their time to make this resource possible. Analysis was carried out on the clusters at the Oxford Biomedical Research Computing (BMRC) facility and FMRIB (part of the Wellcome Centre for Integrative Neuroimaging).

A.2 OASIS

The Open Access Series of Imaging Studies (OASIS) [4] is a project aimed at making neuroimaging data sets of the brain freely available to the scientific community. The data is freely available on application. OASIS-3 was used in this study which is a longitudinal neuroimaging, clinical, cognitive, and biomarker dataset for normal ageing and Alzheimer's Disease (AD). Only healthy subjects were used in this thesis, screened for both AD and other conditions known to impact the brain: brain tumours, benign neoplasms of the brain, hydrocephalus/congenital malformations of the nervous system, stroke/cerebrovascular disease, all neurological diagnoses and all mental and behavioural disorders, diabetes and hypertension. Further, only the image acquired on the 3T scanner were kept, and these images were all acquired on a single scanner.

The data has been used in the following ways in this thesis:

- Harmonisation: T1 data and age information used as training data.

The OASIS data was previously collected under several study protocols at Washington University. All studies were approved by the University's Institutional Review Board (IRB). All subjects gave written informed consent at the time of study participation. The University's IRB also provided explicit approval for open sharing of the anonymized data. More information relating to this can be found in [4].

A.3 Whitehall II

The Whitehall II [325] study was established to investigate the causes of social inequalities in health. A cohort of 10,308 participants aged 35-55, of whom 3,413 were women and 6,895 men, was recruited from the British Civil Service in 1985. The study is now a study on ageing, with the 30 year old follow-up allowing research on

multimorbidity, functional decline, frailty, disability, and dementia. Imaging data for this follow-up was used in this study, screened to keep only healthy subjects.

The data has been used in the following ways in this thesis:

- Harmonisation: T1 data and age information used as training data.

Ethical approval was granted generically for the "Protocol for non-invasive magnetic resonance investigations in healthy volunteers"
(MSD/IDREC/2010/P17.2) by the University of Oxford Central University / Medical Science Division Interdisciplinary Research Ethics Committee (CUREC/MSD-IDREC), who also approved the specific protocol: "Predicting MRI abnormalities with longitudinal data of the Whitehall II sub-study" (MSD-IDREC-C1-2011-71).

A.4 ABIDE

The Autism Brain Imaging Data Exchange I (ABIDE I) represents the first ABIDE initiative. It involves 17 imaging sites, sharing resting state functional magnetic resonance imaging (rfMRI), anatomical and phenotypic datasets made available for data sharing with the broader scientific community. This effort yielded 1112 subjects, including 539 from individuals with autism spectrum disorder and 573 from typical controls (ages 7-64 years, median 14.7 years across groups).

The data has been used in the following ways in this thesis:

- Harmonisation: T1 data and age information used as training data;
- Pruning: T1 data from three sites (UM, Yale, MPG) used as training data.

Prior to data contribution, sites are required to confirm that their local Institutional Review Board (IRB) or ethics committee have approved both the initial data collection

and the retrospective sharing of a fully de-identified version of the datasets (i.e., after removal of the 18 protected health information identifiers including facial information from structural images as identified by the Health Insurance Portable and Accountability Act [HIPAA]). More details can be found in [98].

A.5 IXI

The IXI dataset is a collection of nearly 600 MR images from normal, healthy subjects, with T1, T2, PD, magnetic resonance angiography (MRA) and diffusion weighted images available for each subject, across three scanners. The data and demographic information is freely available from the webstite: `https://brain-development.org/ixi-dataset/`.

The data has been used in the following ways in this thesis:

- Pruning: T1 data used as training data.

The data was collected as part of the project: IXI – Information eXtraction from Images (EPSRC GR/S21533/02) and was used in accordance with their data access declaration.

A.6 HarP

The EADC-ADNI Harmonized Hippocampal Protocol (HarP) [168] was created to be the standard segmentation method for hippocampal volumetry in diagnostic studies, clinical trials, and algorithm validation efforts. The imaging dataset consists of 135 ADNI images segmented in AC-PC, spanning cognitively normal, mildly cognitively impaired, and AD subjects; details can be found in [168].

The data has been used in the following ways in this thesis:

- Pruning: T1 data and manual labels used as training data.

The MR images used were part of the ADNI [59]. As per ADNI protocols, all procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. More details can be found at adni.loni.usc.edu. The data was used in correspondence with the usage agreement for the HarP data.

A.7 Medical Decathlon

Data from the medical decathlon [376] is used in this thesis. Details about the data is given in the chapter. The data has been used in the following ways in this thesis:

- Pruning: Imaging data and manual labels used as training data from heart, spleen and prostate tasks.

A.8 INTERGROWTH-21st Study

The data were collected as part of the INTERGROWTH-21st study and represent a healthy fetal cohort - see [365] for more details. Thanks is given to INTERGROWTH-21st for access to the dataset.

- Pruning: Imaging data was used as training data

B

Appendix: Additional Background

B.1 Classification Networks

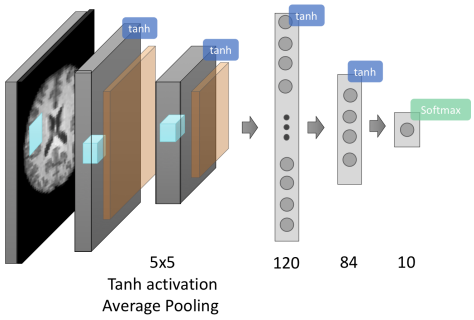


Figure B.1: LeNet architecture

The development of deep neural networks has progressed rapidly as the availability

of both processing power and large-scale, labelled datasets has increased, building on the previous work which lay the foundations of convolutional neural networks [21], back propagation [38] and the network architecture LeNet [39] (Fig. B.1). LeNet demonstrated that image features are distributed across the entire image, and convolutions with learnable parameters are an effective way to extract similar features at multiple locations with relatively few parameters (compared to MLP).

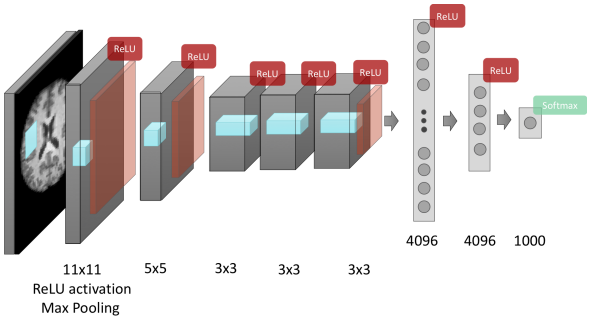


Figure B.2: AlexNet architecture

The first major breakthrough for deep learning was for classification tasks, with AlexNet [40] outperforming the other methods on the ImageNet dataset, leading to their extensive use across a multitude of tasks – see Fig. B.2 for the model architecture. AlexNet was a much deeper CNN than LeNet. The successful deployment of convolutional layers in AlexNet [40] on the ImageNet dataset has led to their extensive use across a multitude of tasks. The use of ReLU [23] activation functions was demonstrated to significantly speed up training time compared to the tanh functions that had been used previously. They also employed two methods to reduce overfitting: data augmentation - see section 3.4.1.4 and dropout [87].

The VGG-16 [22] network – Fig. B.3 – showed a large increase in model depth, with that increase leading to a corresponding increase in model performance. The network combined 13 convolutional layers, 3 fully connected layers, and ReLU activation

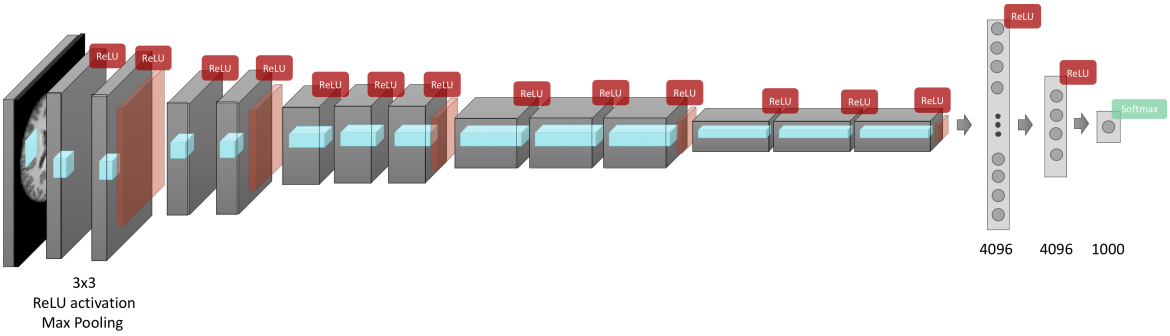


Figure B.3: VGG-16 architecture

functions. It was also demonstrated in [22] that it is better in terms of performance to use two layers with kernels of 3×3 than one layer with a larger kernel size (5×5) with the same receptive field, largely due to the increase in the number of non-linear activation layers and, thus, the discriminative power of the network. This has led to many models being developed with 3×3 kernel sizes throughout.

Inception modules [377] were then introduced which parallelise different convolutional layers, using 1×1 filters, this time to reduce the computational expense. This arrangement allowed the network to learn which features were the most useful at a given depth from the set of options. This module architecture was a novel development, and enabled the network to identify the useful features at different depths.

As the networks being trained became deeper – AlexNet [40] was formed of 8 layers whereas VGG-16 [22] contains 16 trainable layers – the training became more difficult, potentially due to the vanishing gradients problem preventing the propagation of important information through the network. ResNet [378] was designed to overcome this problem through the insertion of skip connections which allow the network to bypass two layers so that the network learns the difference (residual) between the representations. This empirically led to a large increase in performance on datasets such as ImageNet, and ResNet has been the basis of many developments since. The ResNet architecture

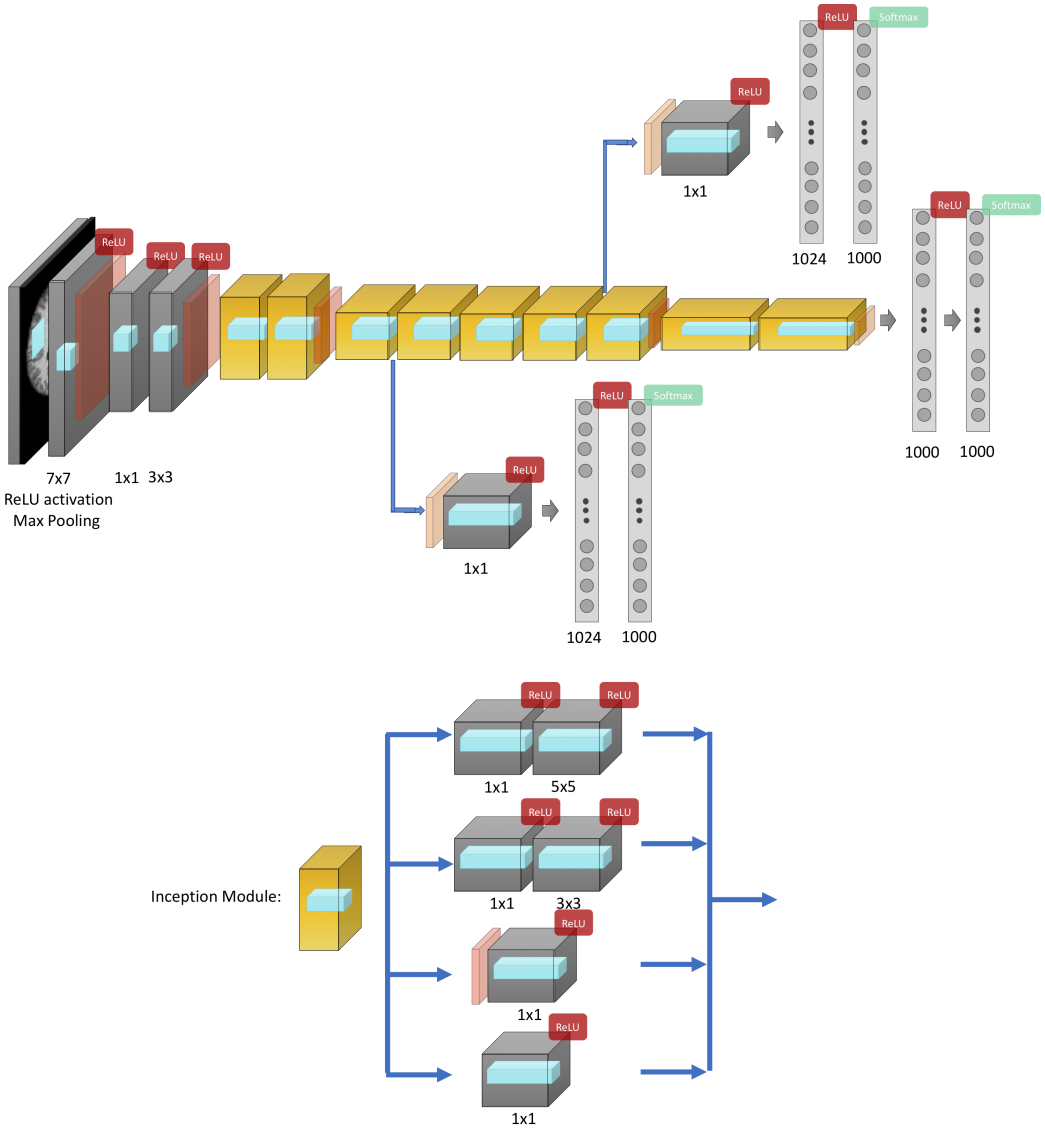


Figure B.4: Inception V1 architecture

was also one of the first architectures to adopt the use of batch normalisation.

DenseNet then built on the idea developed in ResNet, also focusing on solving the problem of vanishing gradients with increasing model depth. To try to ensure maximum information sharing, and so maximum gradient flow, every layer is connected directly to each other as shown in the bottom panel of Fig. B.6. Counter-intuitively, this can lead to DenseNets requiring fewer parameters than equivalent traditional

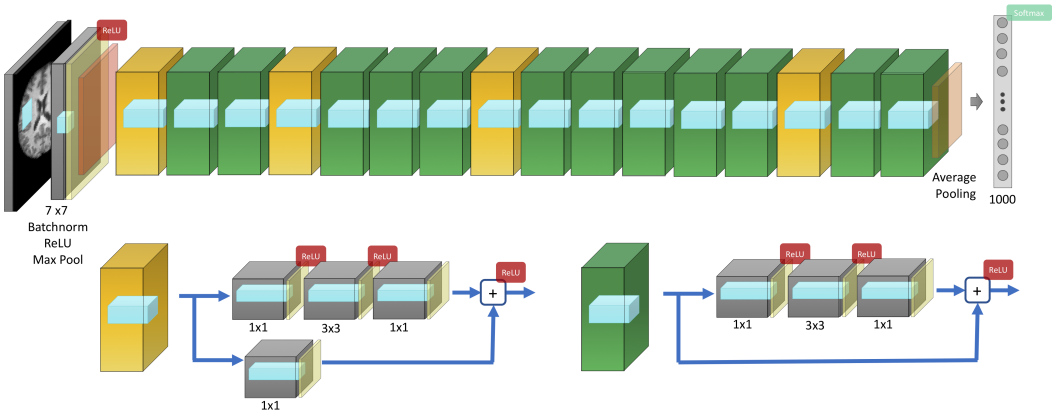


Figure B.5: ResNet-50 architecture

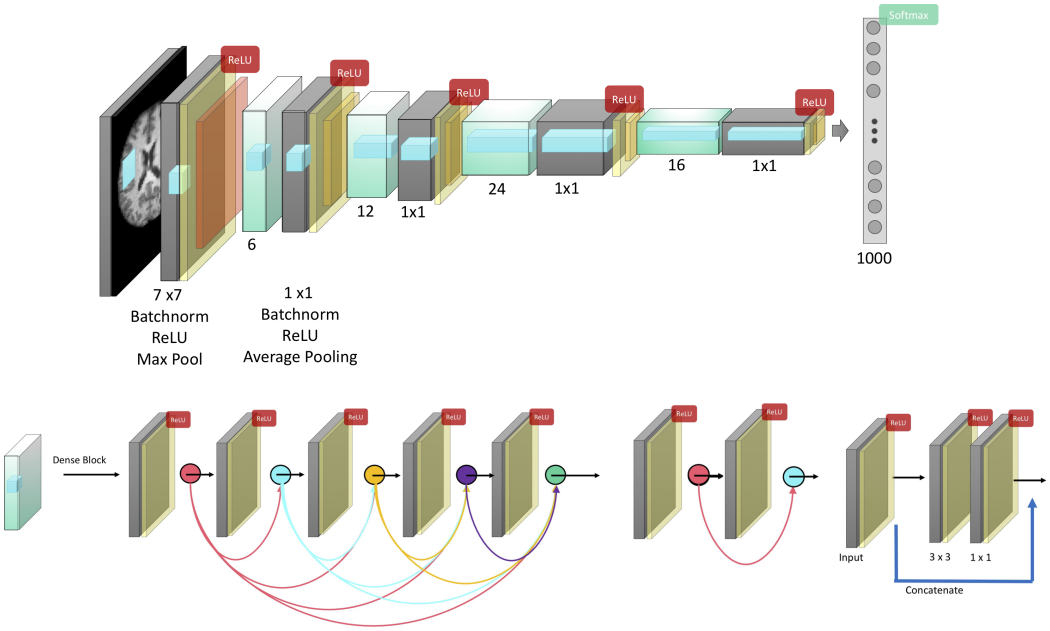


Figure B.6: DenseNet-121 architecture

CNNs, as fewer redundant feature maps are learned, with very narrow architectures achieving the same performance as wider ResNet architectures.

These architectures, and the design decisions that they informed, form the foundation of the vast majority of the networks that are used for classification or regression tasks in medical imaging. Classification tasks for MRI focus on classification of patients from controls [379], distinguishing between pathologies [380], and disease trajectories

[381]. Regression tasks include, for instance, brain age prediction [217].

B.2 Segmentation Networks

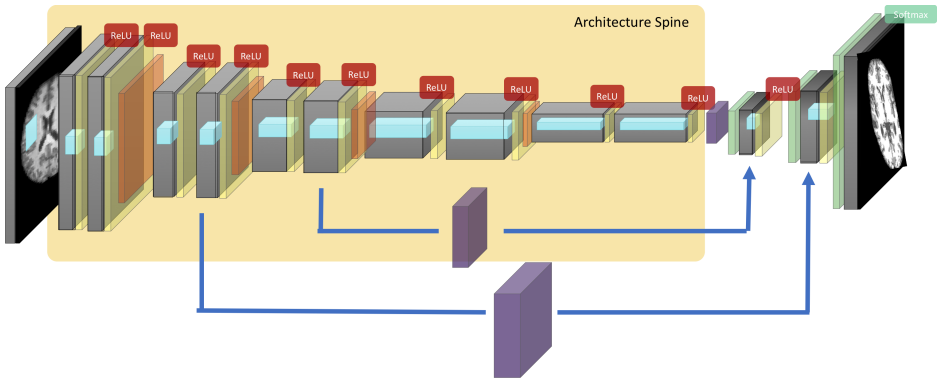


Figure B.7: Fully Convolutional Network (FCN) architecture. The architecture spine can take many different forms, commonly the classification networks discussed in the classification section.

The fully convolutional network (FCN) [382] has been one of the most successful architectures for segmentation, an example schematic for which can be seen in Fig. B.7. This network takes the image classification networks, such as AlexNet, to serve as the encoder, then appends a decoder with transpose convolution layers (upsampling layers where the upsampling is learned by the network) to upsample the coarse feature maps into the full resolution segmentation map. Rather than upsampling in one stage, the upsampling is completed gradually, and features from the earlier layers are summed with the upsampled features using skip connections, helping to produce more fine-grained detail in the predicted segmentations than the original encoder-decoder architectures.

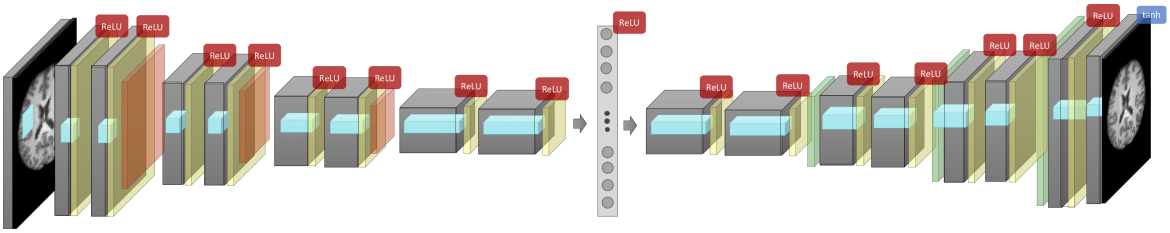


Figure B.8: Example architecture for an autoencoder. A similar architecture could be used for a variational autoencoder, where the fully connected layer uses the reparameterisation trick to draw samples from the learned distribution [51].

B.3 Image Generation or Reconstruction

Generative methods aim to find underlying patterns in the data to either compress data to reduced dimensions or create new data points from the same distribution. An *autoencoder* is an unsupervised learning technique for representation learning, where the model is trained to reconstruct the input by minimising some loss $L_{recon}(\boldsymbol{x}, \hat{\boldsymbol{x}})$. As can be seen from Fig. B.8, the basic form of an autoencoder is an encoder branch which compresses the image information down to a fully connected latent space, and a decoder which projects this compressed information back up into the image space. The latent space is an information bottleneck, which forces the network to learn a compressed representation of the original input. Any autoencoder model has to balance two factors: being sensitive enough to the input data so that it is able to accurately create a valid reconstruction, whilst being insensitive enough to the data that the model does not simply memorise, or overfit, to the data.

Variational autoencoders [51] are a development of autoencoders, which encode a probabilistic description of the data in the latent space. Rather than building an encoder which outputs a single value to describe each latent state attribute, the encoder is formulated to describe a probability distribution for each latent attribute. When decoding, a sample is drawn randomly from the latent state distribution to generate a vector as input for the decoder model. For any of these samples from the

latent distribution, the decoder should be able to reconstruct the input accurately; thus, values which are nearby one another in the latent space should correspond with very similar reconstructions, creating a latent space which is often more meaningful than just using an autoencoder.

Autoencoders can be used for a range of tasks, such as classification via clustering in the latent space [52], and anomaly detection, through observation of large reconstruction errors for subjects with unexpected pathologies [53], and unsupervised segmentation [54].

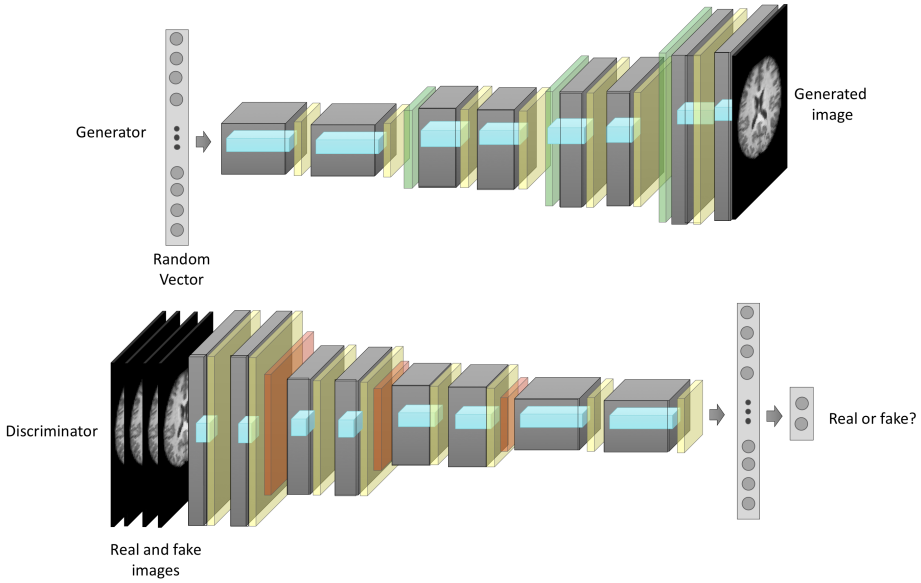


Figure B.9: Example architecture for a Generative Adversarial Network (GAN) architecture.

Generative adversarial networks (GANs) [55] aim to automatically discover and learn patterns in the input data in such a way that the model can be used to generate or output new examples that plausibly could have been drawn from the original dataset distribution. As can be seen in Fig. B.9 they are formed of two models: the *generator* model is trained to generate new examples from a random vector initialisation, and a *discriminator* model is trained to classify the examples either as real data samples drawn from the data distribution, or fake ones produced by the generator. GANs were then

further developed in conditional GAN (cGAN) [92], where the input to the generator is conditioned on some other piece of information, for instance gender or an example input image. CycleGAN [95] combines two GAN models for image-to-image translation without requiring paired examples. The models are trained in an unsupervised manner, using a collection of images from two domains that do not have to be related in any way. They enable style transfer, for instance across imaging modalities.

GANs have been used in medical imaging for a range of tasks, such as data augmentation [56], generating MR images from CT [57] and segmentation [58]. For a review, see, for instance, [383].

B.4 Brain Age Prediction Studies

The following table lists Brain Age prediction papers, focusing on MRI and lists the training data used, model type, the age of the subjects, the resulting MAE and the weighted MAE. Weighted MAE is the MAE normalised by the range of the ages in the training data, as proposed by [256] to allow a degree of comparison between methods. The search terms used were Brain Age Prediction. Studies focusing only on patient populations were not included, as our study focused on healthy populations. The work published from this study is shown in the table in bold.

Study	**Training Data**	**Model**	**Age Range & Mean**	**MAE**	**Weighted MAE**
Ashburner (2007) [224]	T1: GM	RVR	17–79 (38.1)	-	-
Franke et al (2010) [225]	T1: GM	RVR	20-86 (48.2)	4.61	0.070
Dosenbach et al (2010) [245]	fMRI	SVR	7-30	-	-
Wang & Pham (2011) [231]	T1	HMM	50-86	6.2	0.172
Brown et al (2012) [236]	T1	RVR	3-20	1.71	0.101
Franke et al (2012) [257]	T1	RVR	5-19 (10.7)	1.2	0.086
Groves et al (2012) [252]	T1, dMRI	Smoothing Spline	8-85	-	-
Wang et al (2012) [247]	fMRI	SVR	8-79	7.5	0.106

Konukoglu et al (2013) [384]	T1	NAF	18-86	-	-
Koutsouleris et al (2013) [226]	T1	SVR	18-65 (36.8)	4.6	0.098
Mwangi et al (2013) [251]	dMRI	RVR	4-85	6.94	0.086
Su et al (2013) [227]	T1	RVR	18-91 (43.3)	5.7	0.078
Han et al (2014) [250]	dMRI	CBR	4-85	7.66	0.095
Irimia et al (2014) [253]	T1, dMRI	MLR	18-61 (33.3)	-	-
Wang et al (2014) [237]	T1	RVR	20-82 (47.0)	4.57	0.074
Cole et al (2015) [232]	T1	GPR	18-90 (41.8)	6.2	0.086
Erus et al (2015) [385]	T1, dMRI	SVR	8-22 (15.1)	1.22	0.087
Fujimoto et al (2015) [234]	T1	RVR	20-80	4.48	0.075
Khundrakpam et al (2015) [238]	T1	Elastic Net	5-21	1.68	0.105
Kondo et al (2015) [233]	T1	RVR	20-75 (45.6)	4.50	0.082
Cherubini et al (2016) [212]	T1, dMRI	MLR	20-74 (42)	-	-
Lin et al (2016) [249]	dMRI	ANN	50-79 (65.5)	4.29	0.148
Tsvetanov et al (2016) [246]	rfMRI	MLR	18-88	-	-
Cole et al (2017) [217]	T1	GPR	18-90 (37.0)	5.01	0.070
Cole et al (2017) [217]	T1	CNN	18-90 (37.0)	4.65	0.065
Valizadeh et al (2017) [221]	T1	NN	18 - 96 (split)	4.97	0.160
Huang et al (2017) [373]	T1	CNN	20-80	4.0	0.067
Liem et al (2017) [280]	T1	NN	19-82 (58.7)	4.83	0.077
Guggenmos et al (2017) [243]	T1	Ridge Regression	21-65 (43.7)	6.9	0.157
Aycheh et al (2018) [242]	T1	LASSO + GPR	49-91	4.05	0.088
Al Zhoubi et al (2018) [386]	EEG	Ensemble	18-58	7.01	0.175
Bagarinao et al (2018) [244]	T1	Ridge Regression	21-86	7.18	0.110
Eavani et al (2018) [229]	T1	SVR	50-96 (72.5)	4.41	0.096
Guitierrez Becker et al (2018) [239]	T1	GPR	6-92	3.86	0.045
Lancaster et al (2018) [228]	T1	SVR	16-90 (36.5)	5.08	0.069
Li et al (2018) [248]	fMRI	CNN	8-22	2.15	0.154
Madan & Kensinger (2018) [241]	T1	RVR	18-94	-	-
Monté-Rubio et al (2018) [235]	T1	GPR	20-69 (46.9)	5.0	0.075
Pardoe & Kuzniecky (2018) [240]	T1	RVR	8-69 (24.8)	7.2	0.118
Varikuti et al (2018) [230]	T1	LASSO	18-81	4.9	0.078
Jonsson et al (2019) [258]	T1	CNN	18-75	4.006	0.070
Smith et al (2019) [266]	Multiple	Elastic Net	45-80	-	-
Wang et al (2019) [387]	T1: GM	CNN	66	4.45	-
Anaturk et al (2020) [254]	T1 dMRI	Regression	60-83 (69.8)	3.11	0.135
Cole (2020) [255]	Multiple	LASSO	45-80	3.55	0.101

Hong et al (2020) [259]	T1	CNN	0-5	0.185	0.037
Jiang et al (2020) [260]	T1	CNN	18-90 (39.5)	5.55	0.060
Rao et al (2020) [261]	T1	CNN	12-30	1.96	0.10
Baecker et al (2021) [388]	T1	RVR	47-80 (61.3)	3.7	0.103
Bellantuono et al (2021) [389]	T1	Complex Network	7-64	2.19	0.04
Dinsdale et al (2021) [79]	T1: Preprocessed	CNN	44-80	2.71	0.074
Hwang et al (2021) [390]	T2	CNN	19-88 (56.4)	4.22	0.061
Peng et al (2021) [220]	T1, GM, WM	CNN	44-80 (52.7)	2.14	0.059

Table B.1: Brain age prediction studies. ANN = Artificial Neural Network; AD = Axial Diffusivity; CNN = Convolutional Neural Network; CBR = Correlation-based regression; ENet = Elastic Net; FA = Fractional Anisotropy; GM = Grey matter; GPR = Gaussian Processes Regression; MD = Mean Diffusivity; MO = Mode of Anisotropy; MLR = Multiple Linear Regression; MVNL = Multivariate Non-Linear Regression; RD = Radial Diffusivity; RF = Random Forests; RVR = Relevance Vector Regression; SVR = Support Vector Regression

B.5 Harmonisation Studies

Study	Modality	Method	Harmonised Output	Paired
Mirzaalian et al (2016) [310]	dMRI	RISH Features	Image	N
Fortin et al (2017) [297]	T1	ComBat [296]	IDPs	N
Fortin et al (2017) [299]	dMRI	ComBat	IDPs	N
Karayumak et al (2018) [306]	dMRI	RISH Features	Images	N
Mirzaalian et al (2018) [311]	dMRI	RISH Features	Image	N
Yu et al (2018) [100]	fMRI	ComBat	IDPs	N
Blumberg et al (2019) [307]	dMRI	MTL	Image	N
Dewey et al (2019) [106]	T1	UNet	Image	Y
Huynh et al (2019) [308]	dMRI	Method of Moments	Image	N
Karayumak et al (2019) [105]	dMRI	RISH Features	Image	N
Koppers et al (2019) [391]	dMRI	RISH Features	Image	Y
Meyer et al (2019) [392]	T1	RVM	IDP	N
Nath et al (2019) [393]	dMRI	Null Space DL	Spherical Harmonics	Y
Yamashita et al (2019) [394]	fMRI	L2 Regularisation	IDPs	Y
Zhao et al (2019) [314]	Cortical Surface	CycleGAN	Image	N
Beer et al (2020) [302]	T1	ComBat + Longitudal Effects	IDPs	N
Chen et al (2020) [301]	T1	ComBat + Covariance	IDPs	N
Dewey et al (2020) [315]	T1	Disentangled Representations	Image	N
Garcia-Dias (2020) [304]	T1	Random Forest	IDPs	N
Karayumak et al (2020) [303]	dMRI	ComBat	IDPs	N
Modanwal et al (2020) [395]	DCE-MRI	CycleGAN	Image	N
Moyer et al (2020) [107]	dMRI	Autoencoder	Image	N
Pomponio et al (2020) [298]	T1	ComBat-GAM	IDPs	N
Radua et al (2020) [300]	T1	ComBat	IDPs	N
Robinson et al (2020) [316]	T1	ISTNs	Image	N
St-Jean et al (2020) [312]	dMRI	Dictionary Learning	Image	N
Wrobel et al (2020) [305]	dMRI	MICA	Image	N
Zhong et al (2020) [396]	dMRI	GANs	Image	N
Dinsdale et al (2021) [108]	T1	Domain Adaptation	IDPs	N
Guan et al (2021) [117]	T1	Domain Adaptation	Classification	N
Wachinger et al (2021) [295]	T1	ComBat + sex and age regressors	IDPs	N
Zuo et al (2021) [374]	T1	Disentangled Representations	Image	N

Table B.2: Harmonisation Studies. Where a study has been published both as a conference proceeding and a journal paper, only the journal paper is included. Search term: MRI Harmonisation or MRI Harmonization

C

Additional Results

C.1 Brain Age Prediction

C.1.1 Deconfounding for Age

Most brain age literature shows an underestimation of brain age for old subjects and an over estimation of brain age for young subjects, resulting from several factors. This results in an age dependence in the estimated deltas, which will be problematic when computing association with nonimaging variables, as they will be driven by the true age rather than the age delta. Therefore, the process described by Smith et al [266] to correct for both linear and quadratic associations with age is followed.

The results can be described as:

$$Y_B = Y + \delta = F(X) \tag{C.1}$$

where Y are the true age values, Y_B the estimated 'biased' age values, X the array of brain imaging measurements, which in our case take the form of the T1-weighted MRI images, and F(.) the model mapping from X to Y_B, which in our case takes the form of the CNN discussed in section 4.2.2.

First, we consider only the linear dependence with age, which is simply removed with a second step. Considering the scatter plots such as those shown in Fig. 4.4 before the removal of the age dependence, it can be seen that ideally the deltas would be distributed around 0 with no overall slope and thus, no overall dependence on true age Y. If there is the presence of an overall slope, then a straight line can be simply fitted to the full dataset and subtracted from the deltas. Therefore:

$$\delta_1 = Y\beta_2 + \delta_2 \tag{C.2}$$

where β_2 is the regressor to model the linear dependence and δ_2 are the residuals from this fitting, which are orthogonal to age with the biases removed. Therefore, the predicted brain age for this step is now:

$$Y_{\beta 2} = Y + \delta_2 \tag{C.3}$$

This therefore provides us with an estimate for the deltas with linear relationships to age removed.

This now needs to be extended to include non-linear relationships with age, especially because the acceleration of the effects of ageing in older age seems likely, especially

with disease. Therefore the model is simply adapted to include an additive non-linear term in Y. A quadratic term is added as the most natural extension and so:

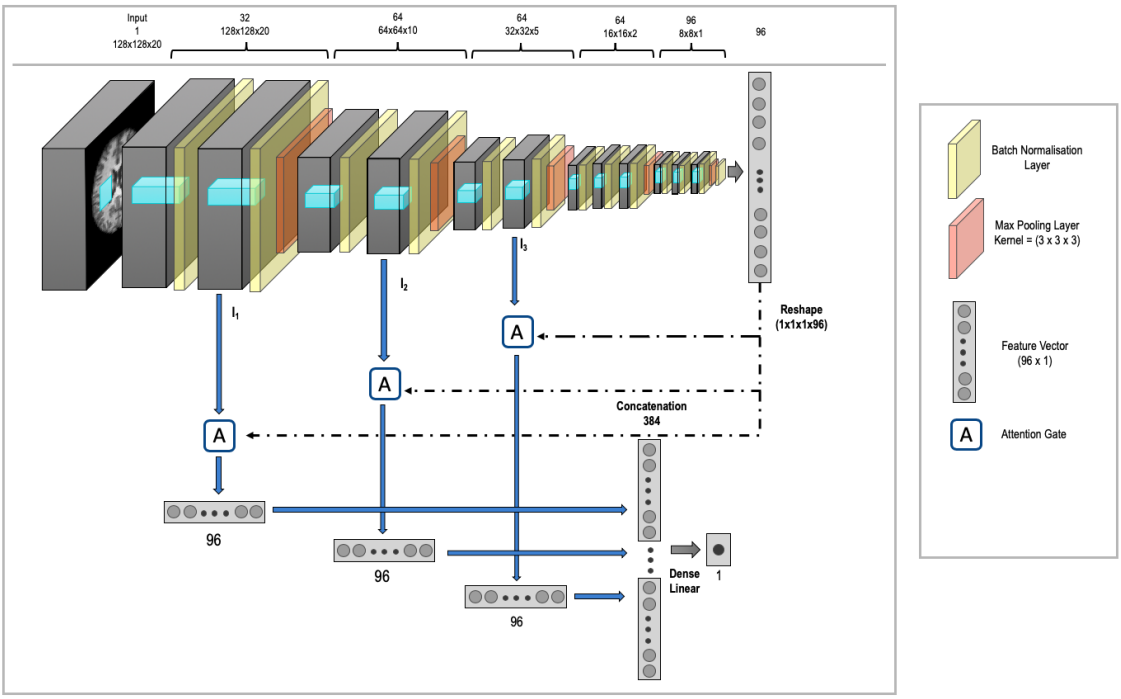
$$Y_B = Y + Y^2\alpha + \delta = X\beta \tag{C.4}$$

and

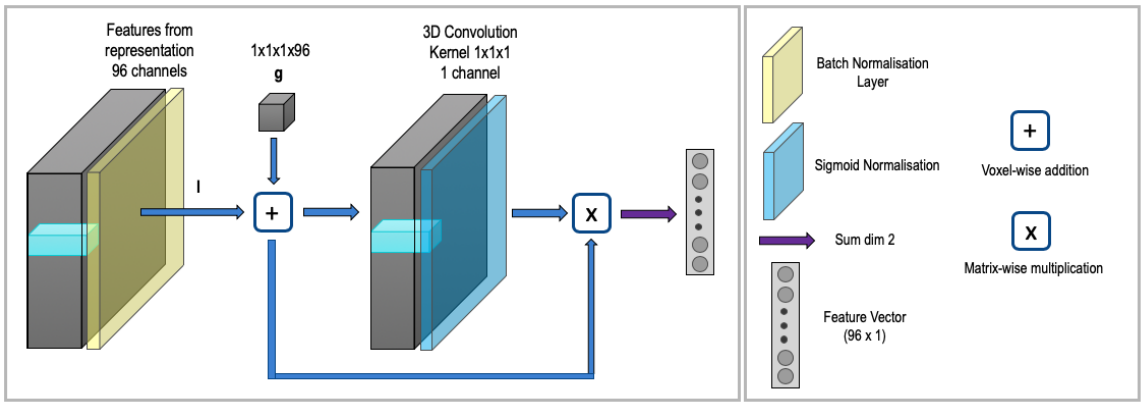
$$\delta_1 = [Y\ Y^2]B_{2q} + \delta_{2q} \tag{C.5}$$

where β_{2q} has two regressors covering the linear and quadratic regressors. Therefore, the linear and quadratic relationships with age from the deltas can be removed through multiple regression.

C.1.2 Attention Network Design



(a) Network with attention gates added. The spine of the network is the same as is used throughout this work.



(b) Attention gate used. Where necessary, the features were first passed through a projection block to ensure the number of features corresponded.

Figure C.1: Network used to create the attention maps for the non-linearly registered data.

C.1.3 Training on Healthy Subjects

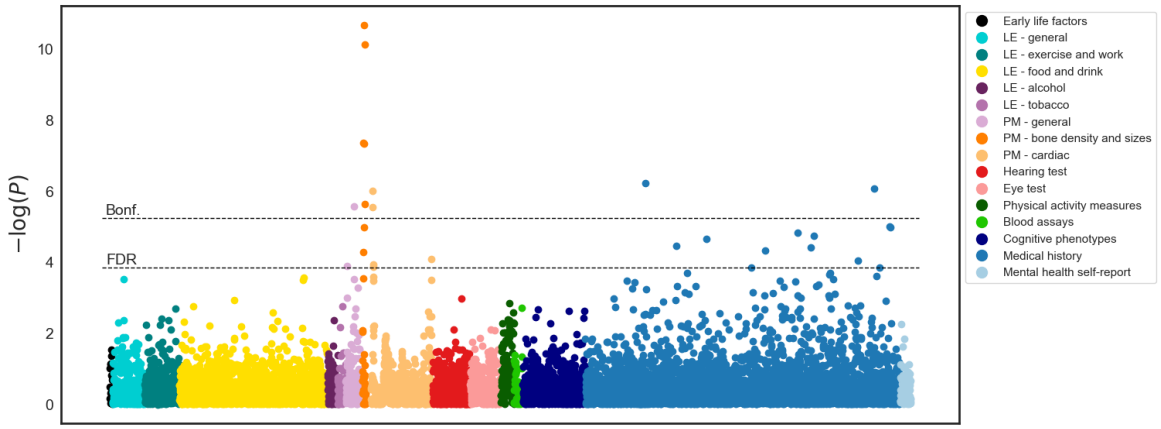
Not all of the subjects in the UK Biobank are healthy, so models were also trained on only the healthy subjects to see what effect this had on the predictions. Not all patients with reported ICD9/10 codes or self-reported conditions were removed, as this left very few subjects. Rather, following [255] subjects were removed with conditions that were likely to impact on the brain. Subjects were filtered out who had either ICD9/10 codes or self-reported conditions which were likely to influence the brain. Subjects were removed who reported brain tumours, benign neoplasms of the brain, hydrocephalus/congenital malformations of the nervous system, stroke/cerebrovascular disease, all neurological diagnoses and all mental and behavioural disorders, leaving around 14k subjects in total. The same approach was then taken as explored in the methods but only for the non-linearly registered images, as this was sufficient for comparison.

The MAE results and correlations can be seen below – Table C.1 and Figs. C.2 and C.3.

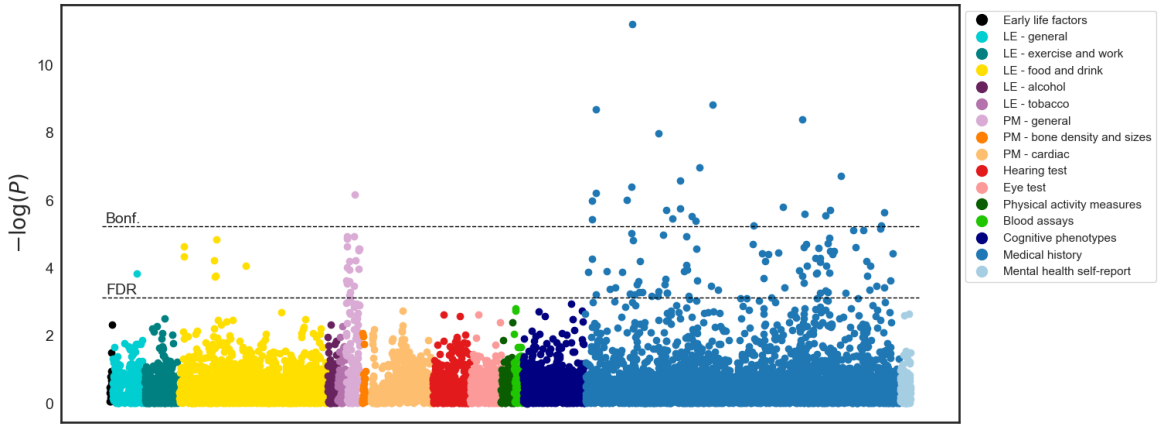
	Non-linear Registered Input	
	Female	Male
MAE	2.94±2.22	3.03±2.29
MSE	13.56±19.1	14.44±21.1
R	0.866	0.864

Table C.1: Ensemble network results split by sex for only the healthy subjects.

The effect of also removing conditions that have been identified by other works in the literature [266] and by our method to correlate with the $\Delta_{brainage}$ values – diabetes and hypertension – was also explored. The removal of these subjects left around 10k subjects in total. Again, the same approach was taken as explored in the methods, but the results are only shown for the non-linearly registered images as a comparison. The MAE results and correlations are shown – Table C.2 and Figs. C.4 and C.5.



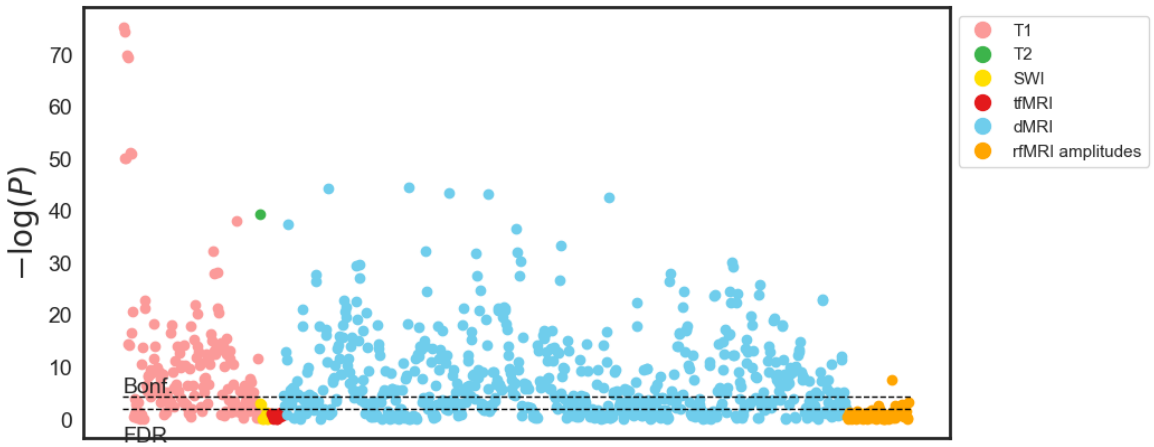
(a) Female Correlations (n=2360)



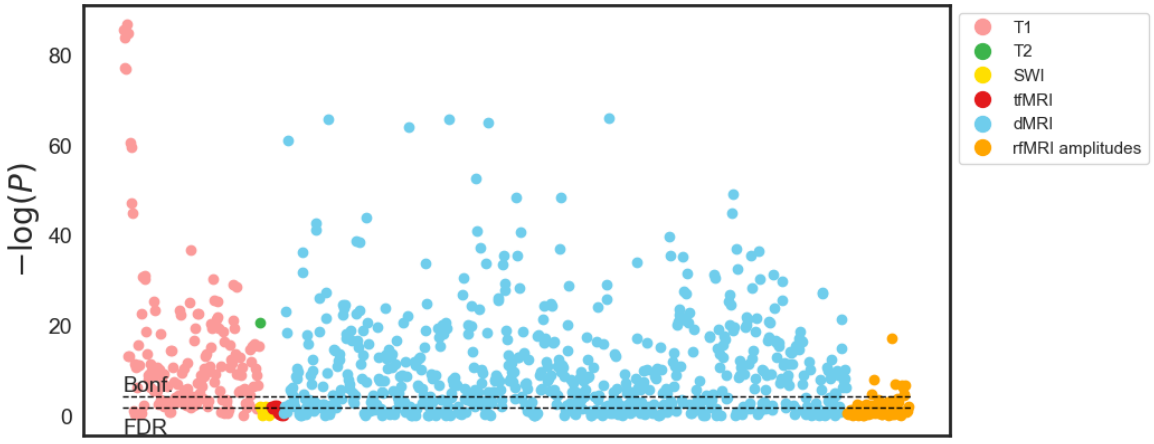
(b) Male Correlations (n=2366)

Figure C.2: Manhattan plot relating each of the non-imaging UK Biobank variables with the delta measurements from the CNN (with non-linearly registered images as input) on female and male subjects, training and testing on only the healthy subjects. Each dot represents the statistical significance of the correlation between one UK Biobank variable and the deltas.

Finally, the model was tested on the whole test dataset – healthy and unhealthy – using the model trained on the healthy subjects, so that subjects with correlating conditions were also removed. The results can be seen in Figs. C.6 and C.7. It can be seen that the results remain consistent. The highest correlations for women still include bone density, blood pressure, height, weight and the presence of MS, and for men diabetes, hypertension, weight and height.

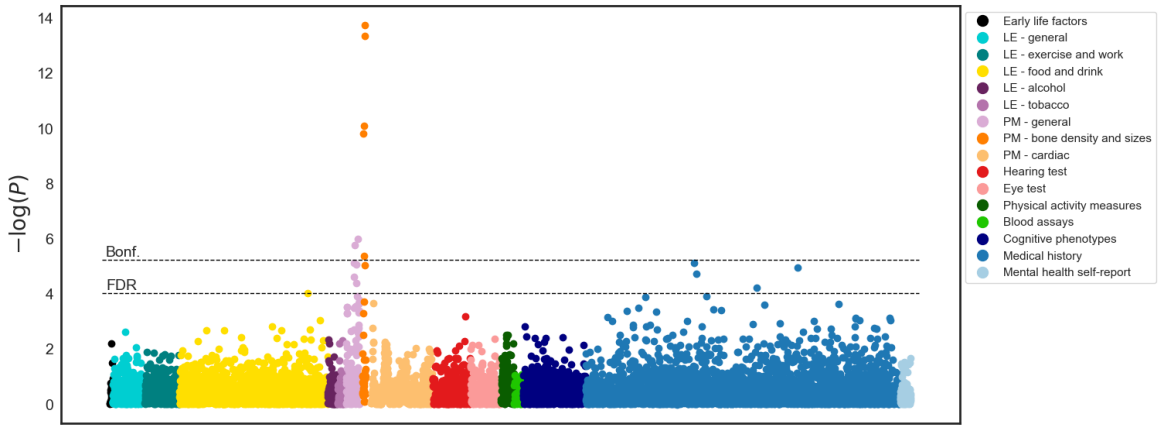


(a) Female Correlations (n=2360)

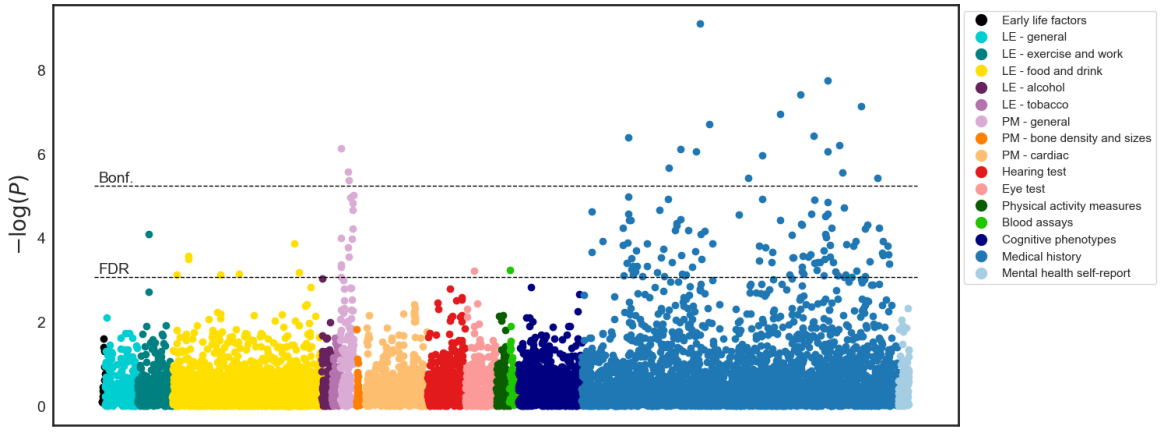


(b) Male Correlations (n=2366)

Figure C.3: Manhattan plot relating each of the IDPs with the delta measurements from the CNN (with non-linearly registered images as input) on female and male subjects, training and testing on only the healthy subjects. Each dot represents the statistical significance of the correlation between one UK Biobank variable and the deltas.

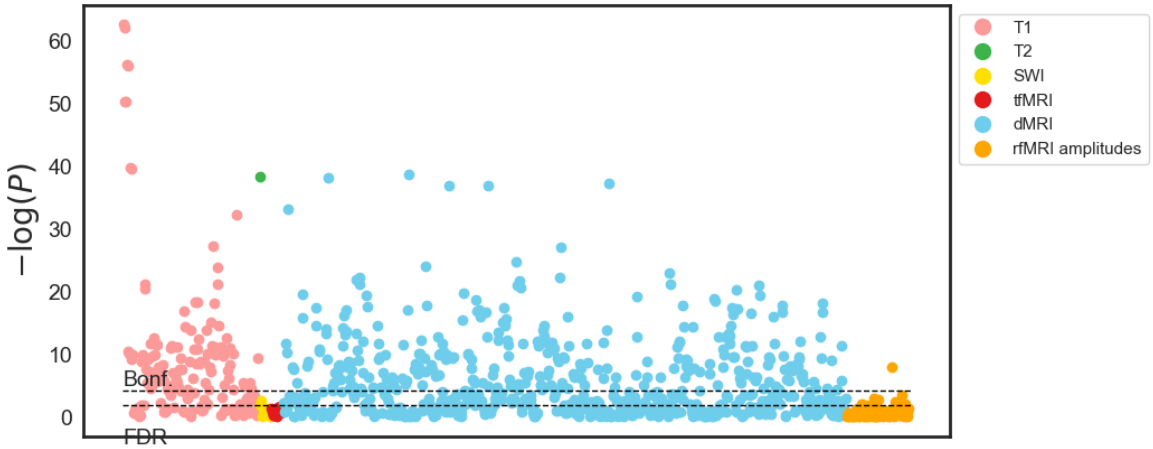


(a) Female Correlations (n=1848)

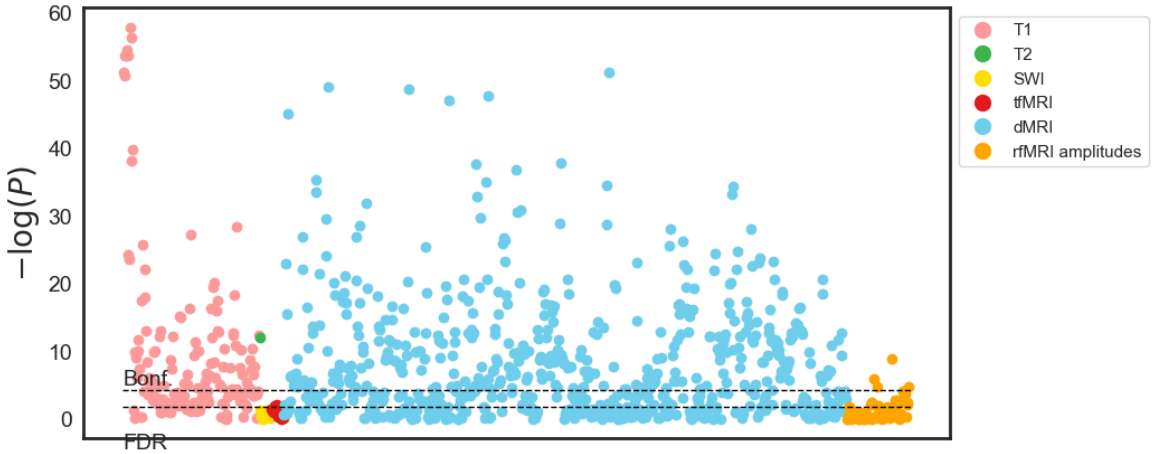


(b) Male Correlations (n=1594)

Figure C.4: Manhattan plot relating each of the non-imaging UK Biobank variables with the delta measurements from the CNN (with non-linearly registered images as input) on female and male subjects training and testing on only the healthy subjects, with correlating conditions removed. Each dot represents the statistical significance of the correlation between one UK Biobank variable and the deltas.

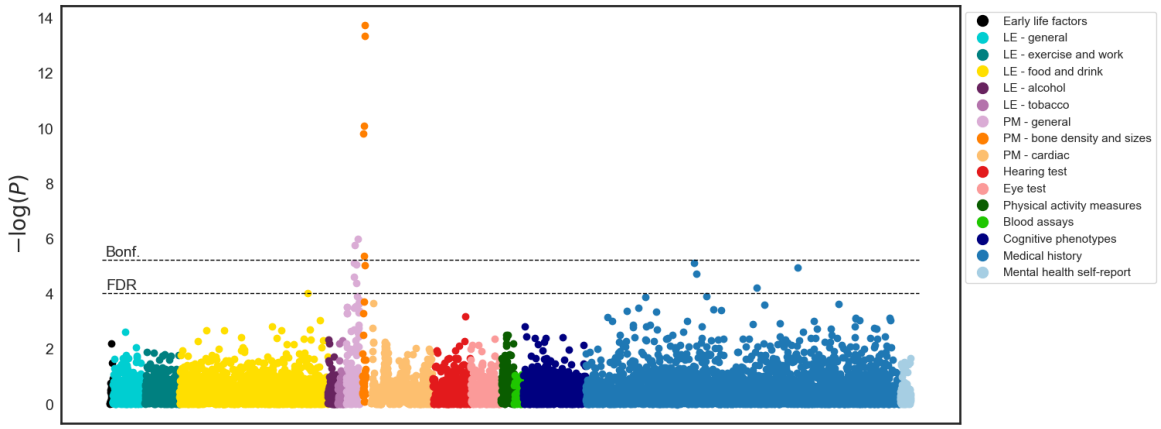


(a) Female Correlations (n=1848)

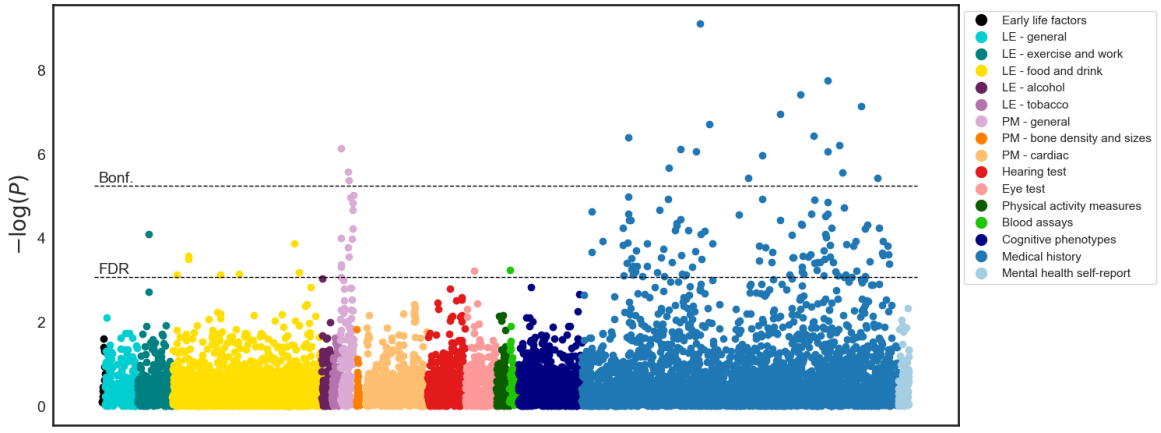


(b) Male Correlations (n=1594)

Figure C.5: Manhattan plot relating each of the IDPs with the delta measurements from the CNN (with non-linearly registered images as input) on female and male subjects training and testing on only the healthy subjects, with correlating conditions removed. Each dot represents the statistical significance of the correlation between one UK Biobank variable and the deltas.

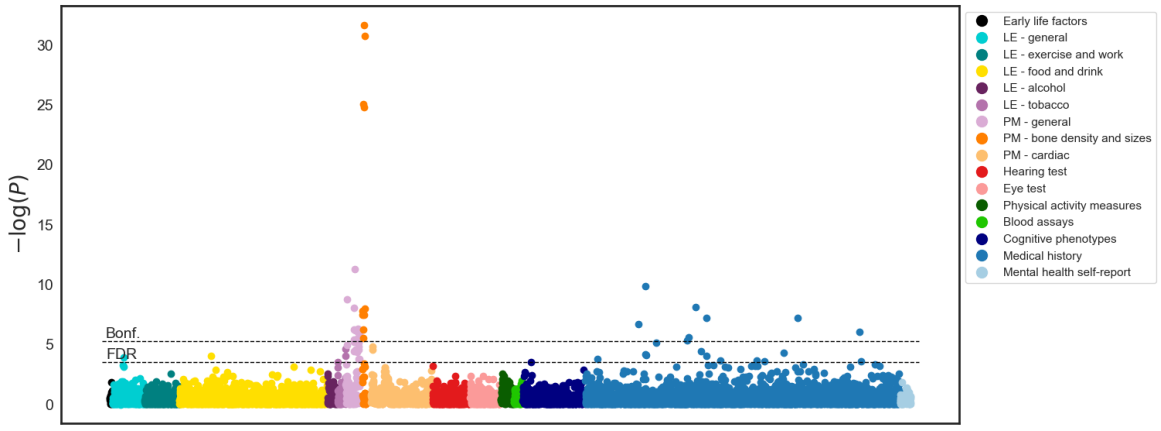


(a) Female Correlations (n=3654)

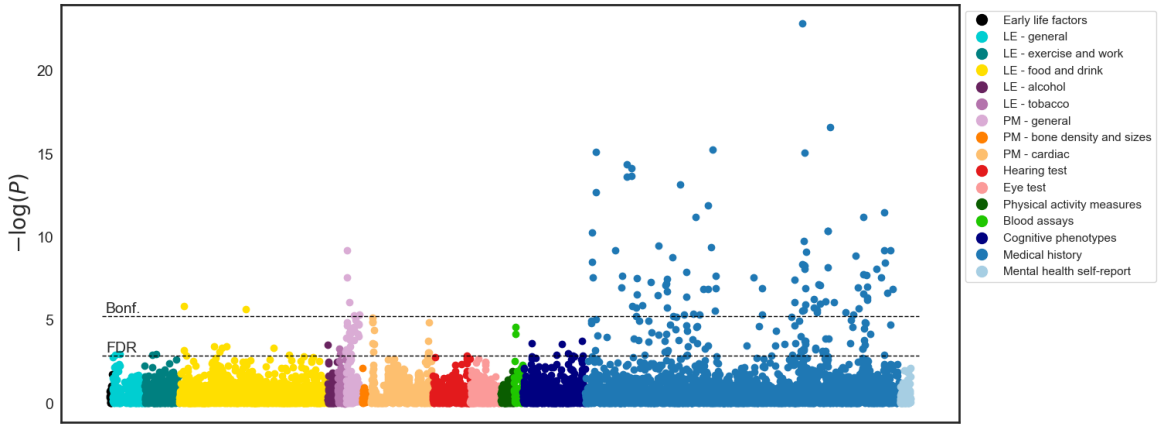


(b) Male Correlations (n=3231)

Figure C.6: Manhattan plot relating each of the non-imaging UK Biobank variables with the delta measurements from the CNN (with non-linearly registered images as input) on female and male subjects training on only the healthy subjects, with correlating conditions removed and testing on all subjects. Each dot represents the statistical significance of the correlation between one UK Biobank variable and the deltas.



(a) Female Correlations (n=3654)



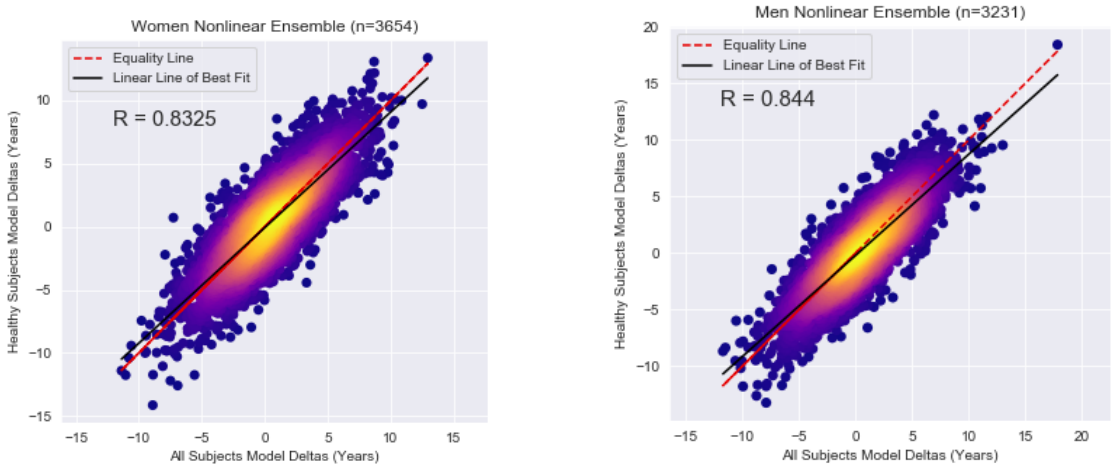
(b) Male Correlations (n=3231)

Figure C.7: Manhattan plot relating each of the IDPs with the delta measurements from the CNN (with non-linearly registered images as input) on female and male subjects, training on only the healthy subjects with correlating conditions removed and testing on all subjects. Each dot represents the statistical significance of the correlation between one UK Biobank variable and the deltas.

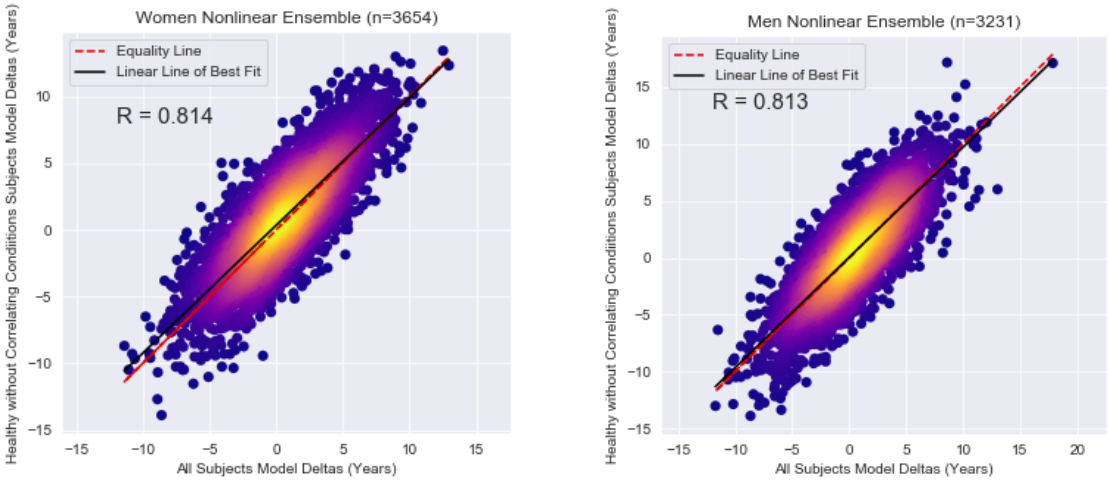
	Non-linear Registered Input	
	Female	Male
MAE	3.20±2.43	3.48±2.60
MSE	16.14±22.9	18.91±25.67
R	0.840	0.833

Table C.2: Ensemble Network results split by sex for just the healthy subjects with correlating conditions removed.

Figure C.8 shows the correlations between the deltas from the models trained on the two different healthy cohorts with the model trained on all subjects, where testing was performed using all subjects. It can be seen that there is a high degree of correlation between the deltas produced by the different models, indicating that incorporating the unhealthy subjects into training has little effect on the model predictions. It can also be seen that, while the degree may differ, the models are consistent with the subjects they predict to have accelerated or reduced ageing, meaning that the deltas should be providing meaningful indications of the progression of ageing. Finally, it can be seen that there are no subjects that have a large negative delta with one model and large positive delta with the other.



(a) **(b)**



(c) **(d)**

Figure C.8: Density plot of the model deltas (true age - predicted age) with the model trained on all subjects versus healthy subjects, testing on all subjects (top row) or healthy subjects minus correlating conditions, testing on all subjects (bottom row) separated by sex.

C.1.4 Significant IDPS and Variables

For each experiment, the significant correlations are presented – those which passed Bonferroni correction. As too many IDPS passed the threshold reasonably to report them all, the top 10 for each modality for IDPs and the top 50 for biological variables are shown. The number of the correlation indicates their ranking of significance, with 1 being the most significant.

Linear Training Data Women:

IDPS:

T1:

1.(r = -0.3), IDP T1 SIENAX grey normalised volume

2.(r = -0.3), IDP T1 SIENAX grey unnormalised volume

3.(r = -0.29), IDP T1 SIENAX peripheral grey normalised volume

4.(r = -0.29), IDP T1 SIENAX peripheral grey unnormalised volume

5.(r = -0.28), IDP T1 SIENAX brain-normalised volume

6.(r = -0.28), IDP T1 SIENAX brain-unnormalised volume

7.(r = 0.27), IDP T1 SIENAX CSF normalised volume

8.(r = 0.27), IDP T1 SIENAX CSF unnormalised volume

14.(r = -0.25), IDP T1 FAST ROIs brain stem

14.(r = -0.25), IDP T1 FAST ROIs brain stem

DMRI:

9.(r = 0.26), IDP dMRI TBSS MD Fornix

10.(r = 0.26), IDP dMRI TBSS L1 Fornix

11.(r = 0.26), IDP dMRI TBSS L2 Fornix

12.(r = 0.26), IDP dMRI TBSS L3 Fornix

13.(r = 0.26), IDP dMRI TBSS ISOVF Fornix

15.(r = -0.25), IDP dMRI TBSS FA Fornix

17.(r = -0.23), IDP dMRI TBSS ICVF Superior fronto-occipital fasciculus L

18.(r = 0.23), IDP dMRI ProbtrackX MD atr r

19.(r = 0.23), IDP dMRI TBSS L2 Fornix cres+Stria terminalis R

20.(r = 0.23), IDP dMRI TBSS L3 Fornix cres+Stria terminalis R

T2:

16.(r = 0.24), IDP T2 FLAIR BIANCA WMH volume

rfMRI:

353.(r = -0.1), rfMRI amplitudes (ICA100 node 38)

Linear Training Data Women:

Non Imaging Variables:

1. (r =-0.12) ['Heel broadband ultrasound attenuation (right) (2.0)']

2. (r =-0.1) ['Treatment/medication code (1140926934 - tamsulosin)']

3. (r =-0.1) ['External causes (W319 - W31.9 Unspecified place)']

4. (r =-0.1) ['External causes (W319 - W31.9 Unspecified place)']

5. (r =-0.1) ['Heel bone mineral density (BMD) T-score, automated (right) (2.0)']

6. (r =-0.1) ['Heel broadband ultrasound attenuation (left) (2.0)']

7. (r =0.09) ['Diastolic blood pressure, automated reading (0.1)']

8. (r =-0.09) ['Heel bone mineral density (BMD) T-score, automated (left) (2.0)']

9. (r =-0.08) ['Weight (2.0)']

10. (r =0.08) ['Diagnoses - main ICD10 (G35 - G35 Multiple sclerosis)']

11. (r =-0.08) ['Sitting height (2.0)']

12. (r =-0.08) ['Sitting height (0.0)']
13. (r =0.08) ['Diastolic blood pressure, automated reading (0.0)']
14. (r =-0.08) ['Hip circumference (2.0)']

Linear Training Data Men:

IDPS:

T1:

1.(r = -0.37), IDP T1 SIENAX grey normalised volume
2.(r = -0.37), IDP T1 SIENAX grey unnormalised volume
3.(r = -0.36), IDP T1 SIENAX peripheral grey normalised volume
4.(r = -0.36), IDP T1 SIENAX peripheral grey unnormalised volume
5.(r = 0.34), IDP T1 SIENAX CSF normalised volume
6.(r = 0.34), IDP T1 SIENAX CSF unnormalised volume
7.(r = -0.33), IDP T1 SIENAX brain-normalised volume
8.(r = -0.33), IDP T1 SIENAX brain-unnormalised volume
17.(r = -0.29), IDP T1 FIRST left thalamus volume
23.(r = -0.28), IDP T1 FIRST right thalamus volume

dMRI:

9.(r = 0.33), IDP dMRI ProbtrackX L1 str r
10.(r = 0.31), IDP dMRI TBSS L2 Fornix cres+Stria terminalis R
11.(r = 0.31), IDP dMRI ProbtrackX L1 str l
12.(r = -0.3), IDP dMRI TBSS ICVF Superior fronto-occipital fasciculus L
13.(r = 0.3), IDP dMRI TBSS L3 Fornix cres+Stria terminalis R
14.(r = 0.3), IDP dMRI TBSS ISOVF Fornix
15.(r = -0.3), IDP dMRI TBSS ICVF Superior fronto-occipital fasciculus R
16.(r = 0.3), IDP dMRI TBSS L2 Fornix
18.(r = -0.3), IDP dMRI TBSS FA Fornix
19.(r = 0.29), IDP dMRI TBSS L3 Fornix

T2:

134.(r = 0.2), IDP T2 FLAIR BIANCA WMH volume

rfMRI:

483.(r = -0.11), rfMRI amplitudes (ICA100 node 17)
500.(r = -0.1), rfMRI amplitudes (ICA100 node 55)
508.(r = -0.1), rfMRI amplitudes (ICA100 node 51)
548.(r = -0.09), rfMRI amplitudes (ICA100 node 54)
567.(r = -0.08), rfMRI amplitudes (ICA100 node 43)
573.(r = -0.08), rfMRI amplitudes (ICA100 node 52)
600.(r = -0.08), rfMRI amplitudes (ICA100 node 21)
611.(r = -0.07), rfMRI amplitudes (ICA100 node 15)

Linear Training Data Men:

Non Imaging Variables:

1. (r =0.13) ['Diabetes diagnosed by doctor (0.0)']
2. (r =0.12) ['Diabetes diagnosed by doctor (2.0)']
3. (r =0.12) ['Non-cancer illness code, self-reported (1220 - diabetes)']
4. (r =0.11) ['Treatment/medication code (1140884600 - metformin)']
5. (r =-0.1) ['Diagnoses - secondary ICD10 (F402 - F40.2 Specific (isolated) phobias)']
6. (r =0.1) ['Non-cancer illness code, self-reported (1065 - hypertension)']
7. (r =0.1) ['Medication for cholesterol, blood pressure or diabetes (2.0)']

9. (r =0.1) [‘Diagnoses - main ICD10 (O60 - O60 Preterm delivery)’]
10. (r=0.1) [‘Diagnoses - secondary ICD10 (E119 - E11.9 Without complications)’]
11. (r=-0.1) [‘Non-cancer illness code, self-reported (1221 - gestational diabetes)’]
12. (r =-0.09) [‘Non-cancer illness code, self-reported (1073 - gestational hypertension/pre-eclampsia)’]
13. (r =0.1) [‘Number of treatments/medications taken (2.0)’]
14. (r =-0.09) [‘Diagnoses - main ICD10 (O16 - O16 Unspecified maternal hypertension)’]
15. (r =-0.09) [‘Diagnoses - secondary ICD10 (Z369 - Z36.9 Antenatal screening, unspecified)’]
16. (r =0.09) [‘Diagnoses - secondary ICD10 (O691 - O69.1 Labour and delivery complicated by cord around neck, with compression)’]
17. (r =0.09) [‘Number of treatments/medications taken (0.0)’]
18. (r =-0.09) [‘Hip circumference (0.0)’]
19. (r =-0.09) [‘Standing height (0.0)’]
20. (r =-0.09) [‘Sitting height (2.0)’]
21. (r =0.08) [‘Treatment/medication code (1140868226 - aspirin)’]
22. (r =-0.08) [‘Cereal intake (2.0)’]
23. (r =0.08) [‘Operation code (1453 - splenectomy)’]
24. (r =-0.09) [‘Standing height (2.0)’]
25. (r =0.08) [‘Treatment/medication code (1140856342 - syndol tablet)’]
26. (r =0.08) [‘Number of self-reported non-cancer illnesses (0.0)’]
27. (r =-0.08) [‘Diagnoses - secondary ICD10 (M255 - M25.5 Pain in joint)’]
28. (r =0.08) [‘Non-cancer illness code, self-reported (1223 - type 2 diabetes)’]

Non-linear Training Data Women:

IDPS:

T1:

1.(r = -0.33), IDP T1 SIENAX peripheral grey normalised volume
2.(r = -0.33), IDP T1 SIENAX grey normalised volume
3.(r = -0.33), IDP T1 SIENAX peripheral grey unnormalised volume
4.(r = -0.33), IDP T1 SIENAX grey unnormalised volume
6.(r = -0.28), IDP T1 SIENAX brain-normalised volume
7.(r = -0.28), IDP T1 SIENAX brain-unnormalised volume
20.(r = 0.25), IDP T1 SIENAX CSF normalised volume
21.(r = 0.25), IDP T1 SIENAX CSF unnormalised volume
22.(r = -0.24), IDP T1 FAST ROIs brain stem
51.(r = -0.23), IDP T1 FAST ROIs L cent operc cortex

T2:

5.(r = 0.3), IDP T2 FLAIR BIANCA WMH volume
5.(r = 0.3), IDP T2 FLAIR BIANCA WMH volume

dMRI:

8.(r = 0.29), IDP dMRI ProbtrackX MD atr r
9.(r = 0.28), IDP dMRI ProbtrackX L3 atr r
10.(r = 0.27), IDP dMRI ProbtrackX MD atr l
11.(r = 0.27), IDP dMRI TBSS L3 Fornix cres+Stria terminalis R
12.(r = -0.27), IDP dMRI TBSS ICVF Superior fronto-occipital fasciculus L
13.(r = 0.27), IDP dMRI ProbtrackX L3 atr l
14.(r = 0.27), IDP dMRI ProbtrackX L2 atr r
15.(r = 0.26), IDP dMRI ProbtrackX L1 atr r
16.(r = 0.26), IDP dMRI ProbtrackX L2 atr l
17.(r = 0.25), IDP dMRI TBSS L3 Fornix cres+Stria terminalis L

rfMRI:

321.(r = -0.14), rfMRI amplitudes (ICA100 node 38)
511.(r = -0.09), rfMRI amplitudes (ICA100 node 55)
516.(r = -0.09), rfMRI amplitudes (ICA100 node 43)
523.(r = -0.08), rfMRI amplitudes (ICA100 node 47)
547.(r = -0.08), rfMRI amplitudes (ICA100 node 17)
578.(r = -0.07), rfMRI amplitudes (ICA100 node 23)
580.(r = -0.07), rfMRI amplitudes (ICA100 node 44)

SWI:
535.(r = 0.08), IDP SWI T2star right thalamus

Non-linear Training Data Women:
Non Imaging Variables
1. (r =-0.16) ['Heel broadband ultrasound attenuation (right) (2.0)']
2. (r =-0.14) ['Heel bone mineral density (BMD) T-score, automated (right) (2.0)']
3. (r =-0.13) ['Heel broadband ultrasound attenuation (left) (2.0)']
4. (r =-0.12) ['Heel bone mineral density (BMD) T-score, automated (left) (2.0)']
5. (r =-0.08) ['Weight (2.0)']
6. (r =0.08) ['Treatment/medication code (1140879802 - amlodipine)']
7. (r =0.08) ['Diastolic blood pressure, automated reading (0.1)']
8. (r =-0.08) ['Sitting height (0.0)']
9. (r =-0.1) ['Heel Broadband ultrasound attenuation, direct entry (0.0)']
10. (r =-0.1) ['Heel quantitative ultrasound index (QUI), direct entry (0.0)']

Non-linear Training Data Men:
IDPS:
T1:
1.(r = -0.38), IDP T1 SIENAX grey normalised volume
2.(r = -0.38), IDP T1 SIENAX grey unnormalised volume
3.(r = -0.38), IDP T1 SIENAX peripheral grey normalised volume
4.(r = -0.37), IDP T1 SIENAX peripheral grey unnormalised volume
5.(r = 0.33), IDP T1 SIENAX CSF normalised volume
6.(r = 0.33), IDP T1 SIENAX CSF unnormalised volume
7.(r = -0.33), IDP T1 SIENAX brain-normalised volume
8.(r = -0.33), IDP T1 SIENAX brain-unnormalised volume
34.(r = -0.27), IDP T1 FIRST left thalamus volume
38.(r = -0.27), IDP T1 FIRST right thalamus volume

dMRI:
9.(r = 0.32), IDP dMRI ProbtrackX L1 str l
10.(r = 0.32), IDP dMRI ProbtrackX L1 str r
11.(r = 0.32), IDP dMRI TBSS L2 Fornix cres+Stria terminalis R
12.(r = -0.32), IDP dMRI TBSS ICVF Superior fronto-occipital fasciculus L
13.(r = 0.31), IDP dMRI TBSS L3 Fornix cres+Stria terminalis R
14.(r = 0.31), IDP dMRI TBSS L2 Fornix
15.(r = 0.31), IDP dMRI TBSS L3 Fornix
16.(r = -0.31), IDP dMRI TBSS FA Fornix
17.(r = 0.31), IDP dMRI TBSS MD Fornix
18.(r = 0.31), IDP dMRI TBSS ISOVF Fornix

T2:
130.(r = 0.21), IDP T2 FLAIR BIANCA WMH volume

rfMRI:

234.(r = -0.17), rfMRI amplitudes (ICA100 node 38)

428.(r = -0.12), rfMRI amplitudes (ICA100 node 17)

484.(r = -0.11), rfMRI amplitudes (ICA100 node 55)

503.(r = -0.11), rfMRI amplitudes (ICA100 node 15)

511.(r = -0.11), rfMRI amplitudes (ICA100 node 43)

522.(r = -0.1), rfMRI amplitudes (ICA100 node 51)

527.(r = -0.1), rfMRI amplitudes (ICA100 node 21)

551.(r = -0.09), rfMRI amplitudes (ICA100 node 54)

566.(r = -0.09), rfMRI amplitudes (ICA100 node 52)

580.(r = -0.08), rfMRI amplitudes (ICA100 node 37)

Non-linear Training Data Men:

Non Imaging Variables

1. (r =0.14) ['Diabetes diagnosed by doctor (2.0)']
2. (r =0.14) ['Non-cancer illness code, self-reported (1220 - diabetes)']
3. (r =0.14) ['Treatment/medication code (1140884600 - metformin)']
4. (r =0.12) ['Diabetes diagnosed by doctor (0.0)']
5. (r =0.1) ['Non-cancer illness code, self-reported (1065 - hypertension)']
6. (r =-0.1) ['Standing height (0.0)']
7. (r =0.1) ['Treatment/medication code (1140874744 - gliclazide)']
8. (r =-0.11) ['Standing height (2.0)']
9. (r =0.1) ['Treatment/medication code (1140868226 - aspirin)']
10. (r =0.1) ['Medication for cholesterol, blood pressure or diabetes (2.0)']
11. (r =0.11) ['Number of treatments/medications taken (2.0)']
12. (r =0.1) ['Diagnoses - secondary ICD10 (E119 - E11.9 Without complications)']
13. (r =-0.1) ['Diagnoses - secondary ICD10 (F402 - F40.2 Specific (isolated) phobias)']
14. (r =-0.1) ['Diagnoses - secondary ICD10 (M255 - M25.5 Pain in joint)']
15. (r =-0.1) ['Trunk fat-free mass (0.0)']
16. (r =-0.1) ['Arm fat-free mass (left) (0.0)']
17. (r =-0.1) ['Arm predicted mass (left) (0.0)']
18. (r =0.09) ['Diagnoses - main ICD10 (O60 - O60 Preterm delivery)']
19. (r =-0.1) ['Sitting height (2.0)']
20. (r =-0.09) ['Non-cancer illness code, self-reported (1221 - gestational diabetes)']
21. (r =-0.09) ['Basal metabolic rate (0.0)']
22. (r =0.09) ['Number of treatments/medications taken (0.0)']
23. (r =-0.09) ['Arm predicted mass (right) (0.0)']
24. (r =-0.09) ['Diagnoses - main ICD10 (O16 - O16 Unspecified maternal hypertension)']
25. (r =0.09) ['Non-cancer illness code, self-reported (1223 - type 2 diabetes)']
26. (r =-0.09) ['Weight (2.0)']
27. (r =0.09) ['Diagnoses - main ICD10 (O034 - O03.4 Incomplete, without complication)']
28. (r =0.08) ['Treatment/medication code (1140860806 - ramipril)']
29. (r =0.19) ['Diabetes diagnosed by doctor (1.0)']
30. (r =0.08) ['Medication for cholesterol, blood pressure or diabetes (0.0)']
31. (r =-0.08) ['Cereal intake (2.0)']
32. (r =0.08) ['Treatment/medication code (1140883066 - insulin product)']
33. (r =0.08) ['Treatment/medication code (1140879760 - bisoprolol)']
34. (r =-0.08) ['Diagnoses - main ICD10 (Z131 - Z13.1 Special screening examination for diabetes mellitus)']
35. (r =0.08) ['Non-cancer illness code, self-reported (1470 - anorexia/bulimia/other eating disorder)']
36. (r =-0.08) ['Weight (0.0)']

IDP Based Method Women:

IDPS:

T1:

1.(r = -0.56), IDP T1 SIENAX grey unnormalised volume

2.(r = -0.55), IDP T1 SIENAX grey normalised volume

3.(r = -0.52), IDP T1 SIENAX peripheral grey unnormalised volume

4.(r = -0.52), IDP T1 SIENAX peripheral grey normalised volume

5.(r = -0.48), IDP T1 SIENAX brain-unnormalised volume

6.(r = -0.48), IDP T1 SIENAX brain-normalised volume

23.(r = 0.38), IDP T1 SIENAX CSF normalised volume

24.(r = 0.38), IDP T1 SIENAX CSF unnormalised volume

44.(r = -0.33), IDP T1 FIRST right thalamus volume

49.(r = -0.32), IDP T1 FAST ROIs brain stem

dMRI:

7.(r = 0.44), IDP dMRI ProbtrackX MD atr l

8.(r = 0.43), IDP dMRI ProbtrackX MD atr r

9.(r = 0.43), IDP dMRI ProbtrackX L1 atr r

10.(r = 0.42), IDP dMRI ProbtrackX L1 atr l

12.(r = 0.41), IDP dMRI ProbtrackX L3 atr l

13.(r = 0.4), IDP dMRI TBSS L2 Fornix cres+Stria terminalis R

14.(r = 0.4), IDP dMRI TBSS L1 Fornix

15.(r = 0.4), IDP dMRI ProbtrackX L2 atr l

16.(r = 0.4), IDP dMRI TBSS MD Fornix

17.(r = 0.4), IDP dMRI ProbtrackX L3 atr r

T2:

11.(r = 0.41), IDP T2 FLAIR BIANCA WMH volume

rfMRI:

91.(r = -0.29), rfMRI amplitudes (ICA100 node 38)

135.(r = -0.26), rfMRI amplitudes (ICA100 node 51)

185.(r = -0.24), rfMRI amplitudes (ICA100 node 15)

191.(r = -0.24), rfMRI amplitudes (ICA100 node 52)

212.(r = -0.23), rfMRI amplitudes (ICA100 node 21)

248.(r = -0.22), rfMRI amplitudes (ICA100 node 8)

253.(r = -0.21), rfMRI amplitudes (ICA100 node 2)

255.(r = -0.21), rfMRI amplitudes (ICA100 node 55)

263.(r = -0.21), rfMRI amplitudes (ICA100 node 44)

265.(r = -0.21), rfMRI amplitudes (ICA100 node 54)

tfMRI: 736.(r = -0.08), IDP tfMRI median zstat shapes

739.(r = -0.08), IDP tfMRI median zstat faces

IDP Based Method Women:

Non Imaging Variables:

1. (r =0.23) ['Non-cancer illness code, self-reported (1507 - haemochromatosis)']

2. (r =0.22) ['Diagnoses - main ICD10 (E119 - E11.9 Without complications)']

3. (r =-0.22) ['Diagnoses - main ICD10 (N433 - N43.3 Hydrocele, unspecified)']

4. (r =0.21) ['Diagnoses - secondary ICD10 (M1099 - M10.99 Gout, unspecified (Site unspecified))']

5. (r =-0.2) ['Operation code (1453 - splenectomy)']

6. (r =0.2) ['Diagnoses - secondary ICD10 (I460 - I46.0 Cardiac arrest with successful resuscitation)']

7. (r =0.2) ['Diagnoses - secondary ICD10 (I120 - I12.0 Hypertensive renal disease with renal failure)']
8. (r =0.18) ['Diagnoses - main ICD10 (N47 - N47 Redundant prepuce, phimosis and paraphimosis)']
9. (r =0.18) ['Diagnoses - secondary ICD10 (I248 - I24.8 Other forms of acute ischaemic heart disease)']
10. (r =0.18) ['Diagnoses - main ICD10 (M2551 - M25.51 Pain in joint (Shoulder region))']
11. (r =0.18) ['External causes (W319 - W31.9 Unspecified place)']
12. (r =0.18) ['External causes (W319 - W31.9 Unspecified place)']
13. (r =0.17) ['Diagnoses - main ICD10 (K403 - K40.3 Unilateral or unspecified inguinal hernia, with obstruction, without gangrene)']
14. (r =0.17) ['Non-cancer illness code, self-reported (1519 - kidney nephropathy)']
15. (r =-0.17) ['Diagnoses - main ICD10 (K081 - K08.1 Loss of teeth due to accident, extraction or local periodontal disease)']
16. (r =0.17) ['External causes (Y400 - Y40.0 Penicillins)']
17. (r =-0.17) ['Diagnoses - main ICD10 (H810 - H81.0 Meniere's disease)"]
18. (r =0.17) ['Diagnoses - secondary ICD10 (M109 - M10.9 Gout, unspecified)']
19. (r =0.16) ['Treatment/medication code (1140874686 - glucophage 500mg tablet)']
20. (r =-0.16) ['Operation code (1526 - stapedectomy)']
21. (r =-0.15) ['Diagnoses - main ICD10 (I840 - I84.0 Internal thrombosed haemorrhoids)']
22. (r =0.15) ['Diagnoses - secondary ICD10 (N210 - N21.0 Calculus in bladder)']
23. (r =-0.15) ['Non-cancer illness code, self-reported (1629 - fracture face / orbit / eye socket)']
24. (r =0.15) ['Treatment/medication code (1140888594 - fluvastatin)']
25. (r =0.15) ['Treatment/medication code (1141187810 - tadalafil)']
26. (r =0.15) ['Diagnoses - main ICD10 (N210 - N21.0 Calculus in bladder)']
27. (r =0.14) ['Diagnoses - main ICD10 (S430 - S43.0 Dislocation of shoulder joint)']
28. (r =0.14) ['Treatment/medication code (1140881938 - beclomethasone dipropionate+salbutamol)']
29. (r =0.14) ['Diagnoses - main ICD10 (N481 - N48.1 Balanoposthitis)']
30. (r =-0.14) ['Treatment/medication code (1140882498 - penicillin)']
31. (r =0.14) ['Histology of cancer tumour (9061 - Seminoma, NOS)']
32. (r =0.14) ['Treatment/medication code (1140875408 - allopurinol)']
33. (r =-0.13) ['Diagnoses - main ICD10 (S4200 - S42.00 Fracture of clavicle (closed))']
34. (r =0.13) ['Diagnoses - secondary ICD10 (E831 - E83.1 Disorders of iron metabolism)']
35. (r =0.13) ['Diagnoses - main ICD10 (F100 - F10.0 Acute intoxication)']
36. (r =0.12) ['Diagnoses - secondary ICD10 (Z572 - Z57.2 Occupational exposure to dust)']
37. (r =-0.13) ['Hip circumference (2.0)']
38. (r =-0.13) ['Heel broadband ultrasound attenuation (right) (2.0)']
39. (r =-0.12) ['Type of cancer: ICD10 (D046 - D04.6 Skin of upper limb, including shoulder)']
40. (r =0.12) ['Treatment/medication code (1141179974 - cozaar 25mg tablet)']
41. (r =-0.13) ['Heel broadband ultrasound attenuation (left) (2.0)']
42. (r =-0.12) ['Body mass index (BMI) (2.0)']
43. (r =0.11) ['Diagnoses - secondary ICD10 (I420 - I42.0 Dilated cardiomyopathy)']
44. (r =0.11) ['Operation code (1108 - leg artery angioplasty +/- stent)']
45. (r =0.11) ['Non-cancer illness code, self-reported (1408 - alcohol dependency)']
46. (r =0.11) ['Diagnoses - secondary ICD10 (M2551 - M25.51 Pain in joint (Shoulder region))']
47. (r =0.11) ['Non-cancer illness code, self-reported (1400 - peptic ulcer)']
48. (r =0.11) ['Treatment/medication code (1140860806 - ramipril)']
49. (r =0.11) ['Non-cancer illness code, self-reported (1466 - gout)']
50. (r =0.11) ['Non-cancer illness code, self-reported (1121 - pulmonary fibrosis)']

IDP Based Method Men:

IDPS:

T1:

1.(r = -0.55), IDP T1 SIENAX grey normalised volume

2.(r = -0.55), IDP T1 SIENAX grey unnormalised volume

3.(r = -0.52), IDP T1 SIENAX peripheral grey normalised volume

4.(r = -0.51), IDP T1 SIENAX peripheral grey unnormalised volume
5.(r = -0.44), IDP T1 SIENAX brain-normalised volume
6.(r = -0.44), IDP T1 SIENAX brain-unnormalised volume
23.(r = 0.37), IDP T1 SIENAX CSF normalised volume
24.(r = 0.37), IDP T1 SIENAX CSF unnormalised volume
36.(r = -0.34), IDP T1 FAST ROIs L frontal pole
39.(r = -0.34), IDP T1 FIRST left thalamus volume

dMRI:
7.(r = 0.42), IDP dMRI ProbtrackX MD atr l
8.(r = 0.41), IDP dMRI ProbtrackX L1 atr l
9.(r = 0.41), IDP dMRI ProbtrackX L1 str r
10.(r = 0.4), IDP dMRI ProbtrackX MD atr r
11.(r = 0.4), IDP dMRI ProbtrackX L1 atr r
12.(r = 0.39), IDP dMRI ProbtrackX L1 str l
13.(r = 0.39), IDP dMRI ProbtrackX L2 atr l
14.(r = 0.38), IDP dMRI TBSS MD Fornix
15.(r = 0.38), IDP dMRI TBSS L2 Fornix
16.(r = 0.38), IDP dMRI ProbtrackX L3 atr l

T2:
25.(r = 0.36), IDP T2 FLAIR BIANCA WMH volume

rfMRI:
85.(r = -0.29), rfMRI amplitudes (ICA100 node 38)
199.(r = -0.23), rfMRI amplitudes (ICA100 node 19)
201.(r = -0.23), rfMRI amplitudes (ICA100 node 52)
211.(r = -0.23), rfMRI amplitudes (ICA100 node 15)
229.(r = -0.22), rfMRI amplitudes (ICA100 node 17)
233.(r = -0.22), rfMRI amplitudes (ICA100 node 2)
239.(r = -0.22), rfMRI amplitudes (ICA100 node 54)
265.(r = -0.21), rfMRI amplitudes (ICA100 node 18)
270.(r = -0.21), rfMRI amplitudes (ICA100 node 44)
278.(r = -0.21), rfMRI amplitudes (ICA100 node 47)

tfMRI:
635.(r = -0.12), IDP tfMRI 90th-percentile zstat faces
643.(r = -0.11), IDP tfMRI median zstat shapes
652.(r = -0.11), IDP tfMRI 90th-percentile zstat shapes
657.(r = -0.11), IDP tfMRI median zstat faces
687.(r = -0.1), IDP tfMRI median BOLD shapes
710.(r = -0.1), IDP tfMRI median BOLD faces
718.(r = -0.09), IDP tfMRI 90th-percentile BOLD shapes
723.(r = -0.09), IDP tfMRI 90th-percentile BOLD faces
729.(r = -0.09), IDP tfMRI 90th-percentile zstat faces-shapes

IDP Based Method Men:

Non Imaging Variables:

1. (r =0.2) ['Treatment/medication code (1141195224 - formoterol)']
2. (r =-0.19) ['Diagnoses - secondary ICD10 (O701 - O70.1 Second degree perineal laceration during delivery)']
3. (r =0.18) ['Diagnoses - main ICD10 (O16 - O16 Unspecified maternal hypertension)']
4. (r =-0.17) ['Treatment/medication code (1140868460 - progynova 1mg tablet)']

5. (r =-0.16) ['Diagnoses - main ICD10 (R935 - R93.5 Abnormal findings on diagnostic imaging of other abdominal regions, including retroperitoneum)']
6. (r =0.16) ['Treatment/medication code (1140881320 - magnesium carbonate)']
7. (r =0.16) ['Diagnoses - secondary ICD10 (T884 - T88.4 Failed or difficult intubation)']
8. (r =0.16) ['Diagnoses - main ICD10 (Z411 - Z41.1 Other plastic surgery for unacceptable cosmetic appearance)']
9. (r =0.16) ['Diagnoses - secondary ICD10 (M511 - M51.1 Lumbar and other intervertebral disk disorders with radiculopathy)']
10. (r =0.16) ['Diagnoses - secondary ICD10 (M790 - M79.0 Rheumatism, unspecified)']
11. (r =-0.15) ['Operation code (1453 - splenectomy)']
12. (r =0.15) ['Non-cancer illness code, self-reported (1221 - gestational diabetes)']
13. (r =0.15) ['Treatment/medication code (1141157458 - hypromellose product)']
14. (r =0.15) ['Diagnoses - secondary ICD10 (M7909 - M79.09 Rheumatism, unspecified (Site unspecified))']
15. (r =0.15) ['Treatment/medication code (1140872036 - paramax tablet)']
16. (r =-0.14) ['Diagnoses - secondary ICD10 (N701 - N70.1 Chronic salpingitis and oophoritis)']
17. (r =0.14) ['Treatment/medication code (1140866280 - bumetanide)']
18. (r =-0.14) ['Diagnoses - main ICD10 (I843 - I84.3 External thrombosed haemorrhoids)']
19. (r =-0.14) ['Diagnoses - main ICD10 (O682 - O68.2 Labour and delivery complicated by foetal heart rate anomaly with meconium in amniotic fluid)']
20. (r =0.14) ['Operation code (1512 - cone biopsy)']
21. (r =0.14) ['Treatment/medication code (1140871542 - mefenamic acid)']
22. (r =-0.14) ['Diagnoses - main ICD10 (O60 - O60 Preterm delivery)']
23. (r =0.14) ['Non-cancer illness code, self-reported (1522 - grave's disease)"]
24. (r =0.14) ['Treatment/medication code (1140857620 - depo-provera 50mg/1ml injection)']
25. (r =0.13) ['Operation code (1542 - laser treatment cervix)']
26. (r =0.13) ['Diagnoses - main ICD10 (J440 - J44.0 Chronic obstructive pulmonary disease with acute lower respiratory infection)']
27. (r =0.13) ['Treatment/medication code (1140883548 - ipratropium)']
28. (r =0.13) ['Diagnoses - main ICD10 (S430 - S43.0 Dislocation of shoulder joint)']
29. (r =0.13) ['Treatment/medication code (1140921822 - mirena 20mcg/24hrs intrauterine system)']
30. (r =-0.13) ['Diagnoses - main ICD10 (M819 - M81.9 Osteoporosis, unspecified)']
31. (r =0.13) ['Diagnoses - secondary ICD10 (M0690 - M06.90 Rheumatoid arthritis, unspecified (Multiple sites))']
32. (r =0.13) ['Diagnoses - main ICD10 (J441 - J44.1 Chronic obstructive pulmonary disease with acute exacerbation, unspecified)']
33. (r =0.13) ['Diagnoses - secondary ICD10 (N803 - N80.3 Endometriosis of pelvic peritoneum)']
34. (r =0.13) ['Treatment/medication code (1140871732 - buprenorphine)']
35. (r =-0.13) ['Histology of cancer tumour (8520 - Lobular carcinoma, NOS)']
36. (r =-0.13) ['Diagnoses - main ICD10 (O631 - O63.1 Prolonged second stage (of labour))']
37. (r =-0.13) ['Diagnoses - main ICD10 (O049 - O04.9 Complete or unspecified, without complication)']
38. (r =-0.12) ['Non-cancer illness code, self-reported (1606 - incisional hernia)']
39. (r =0.12) ['Diagnoses - secondary ICD10 (M205 - M20.5 Other deformities of toe(s) (acquired))']
40. (r =0.12) ['Treatment/medication code (1140867726 - lofepramine)']
41. (r =0.12) ['Diagnoses - secondary ICD10 (O16 - O16 Unspecified maternal hypertension)']
42. (r =0.12) ['Diagnoses - main ICD10 (N803 - N80.3 Endometriosis of pelvic peritoneum)']
43. (r =0.12) ['Treatment/medication code (1140884600 - metformin)']
44. (r =0.12) ['Diagnoses - main ICD10 (E042 - E04.2 Non-toxic multinodular goitre)']
45. (r =0.12) ['Diagnoses - secondary ICD10 (J069 - J06.9 Acute upper respiratory infection, unspecified)']
46. (r =0.11) ['Non-cancer illness code, self-reported (1556 - menorrhagia (unknown cause))']
47. (r =0.11) ['Diabetes diagnosed by doctor (2.0)']
48. (r =-0.11) ['Diagnoses - main ICD10 (Z400 - Z40.0 Prophylactic surgery for risk-factors related to malignant neoplasms)']
49. (r =0.11) ['Non-cancer illness code, self-reported (1220 - diabetes)']
50. (r =0.11) ['Non-cancer illness code, self-reported (1065 - hypertension)']

C.2 Harmonisation

C.2.1 Comparison to Adding Site as Regressor

Here it is explored whether advantages exist to harmonising the network compared to training a network normally and adding scanner covariates to a linear model that is used for the analysis.

To do this, the situation where there are data from three sites and scanners but only manual segmentation results for a single site is considered, which is a common scenario when working with neuroimaging data as labels are expensive to acquire. The aim was to train a network to complete the segmentation for test subjects from all three sites (e.g. from the ABIDE data) and then use them in a standard GLM to explore the task of interest. In this case, exploring the relationship between age and the normalised grey matter/white matter/CSF volumes was considered.

The model was trained using normal training with just on labelled data from the UM site, and also with unsupervised unlearning, harmonising the data (main task labels for just UM but unlearning scanner information from all three sites). When the model was trained on only the UM data and output predictions for the remaining two sites created, it was found that good predictions were made for the test data from UM but there was a significant performance drop when the method was applied to the other two sites, due to domain shift. It can be seen from the figure below that although the network trained with normal training performs well on UM (average dice = 0.891 ± 0.015) the model suffers significant performance degradation when applied to the unseen sites – MPG (average dice = 0.689 ± 0.020) and Yale (average dice = 0.627 ± 0.039). It can also be seen that this loss in performance leads to significant failures such that image-derived values such as white matter or grey matter volumes are going to be very different

from the expected values. Conversely, it can be seen that the performance is greatly improved using the harmonisation method, even though no manual segmentation labels for MPG or Yale were used during training: average dice = 0.891 ± 0.019 (p=4e-8) and 0.892 ± 0.009 (p=4.5e-10), respectively. This comes at little cost to the performance of the segmentation on UM (average dice = 0.883 ± 0.03, p=0.015). Clearly a covariate approach would be unable to substantially reduce these image-level errors.

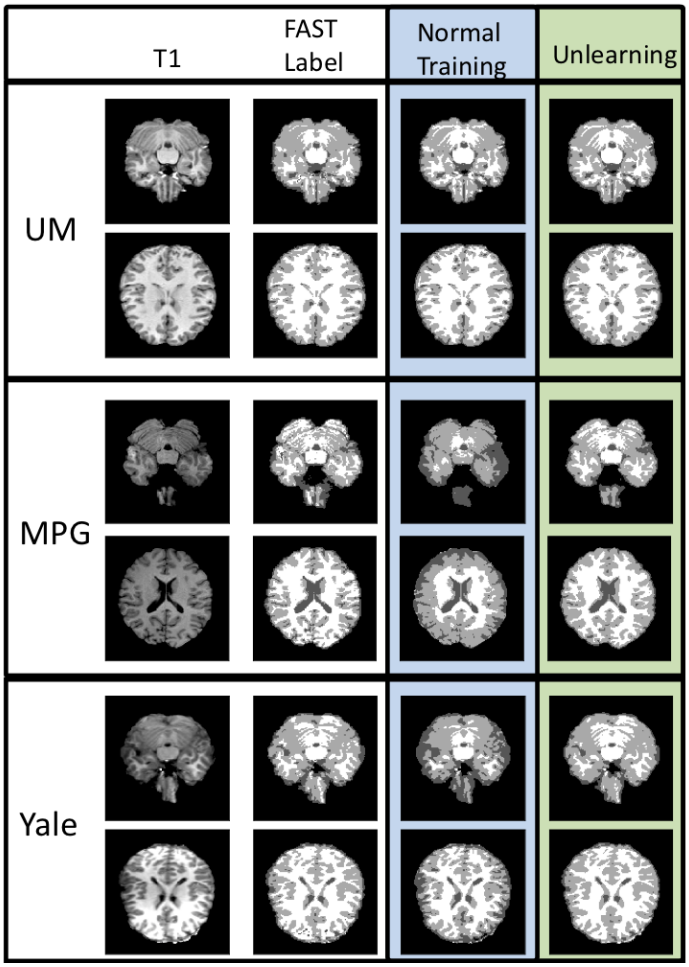


Figure C.9: Segmentation results for the ABIDE data using normal training and unlearning when labels are only available for the UM site. Labels are white matter, grey matter and CSF.

To go further and verify the effect for statistical tests on derived scalar quantities,

the CSF volumes were calculated, normalised by total brain volume. As this is expected to be indicative of age, the subjects are split into a young group and an old group, (n=44) and these are used in a GLM to explore the ability to find group differences. Three scenarios are considered:

1. "normal" input data (no unlearning) but with a design matrix that has separate covariates for the scanners (as well as a simple two group, unpaired, difference regressor – this is x1 below),

2. "unlearning" data with a simple two group (unpaired) difference design matrix,

3. base case of "normal" data and simple design matrix with no scanner-based covariates.

This is computed using the statsmodel python package [397]. The results can be found in the figures below.

It can be seen that, as expected from the image examples, when nothing is done to correct for the scanner differences (scenario 3) there is no significant result (p=0.745). Equally, when scanner is added as a regressor in the model (scenario 1) there is also not a significant result (p=0.128) but when unlearning is used to remove the scanner information, a much stronger statistic is achieved and a significant difference between the groups (p=0.009).

Therefore, there are clear scenarios in which just adding scanner-based covariates in the model leads to reduced statistical differences, in this case preventing significant results from being achieved. Consequently, there is a real need for harmonisation approaches, such as this unlearning method, in situations like this.

Scenario 1:

Generalized Linear Model Regression Results

Dep. Variable:	y	No. Observations:	44
Model:	GLM	Df Residuals:	40
Model Family:	Gaussian	Df Model:	3
Link Function:	identity	Scale:	0.0010448
Method:	IRLS	Log-Likelihood:	90.670
Date:	Wed, 28 Oct 2020	Deviance:	0.041792
Time:	11:29:27	Pearson chi2:	0.0418
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
x1	0.0082	0.005	1.523	0.128	-0.002	0.019
x2	0.1446	0.007	20.013	0.000	0.130	0.159
x3	0.2467	0.010	25.393	0.000	0.228	0.266
x4	0.3048	0.010	30.467	0.000	0.285	0.324

T-values: [1.52273558 20.01274993 25.39284127 30.46712709]

(a) Scenario 1: "normal" data (no unlearning) but design matrix with covariates for the scanners (and group difference)

Scenario 2:

Generalized Linear Model Regression Results

Dep. Variable:	y	No. Observations:	44
Model:	GLM	Df Residuals:	42
Model Family:	Gaussian	Df Model:	1
Link Function:	identity	Scale:	0.00026237
Method:	IRLS	Log-Likelihood:	120.00
Date:	Wed, 28 Oct 2020	Deviance:	0.011019
Time:	11:31:04	Pearson chi2:	0.0110
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
x1	0.0064	0.002	2.612	0.009	0.002	0.011
const	0.1379	0.002	56.414	0.000	0.133	0.143

T-values: [2.61229786 56.41381163]

(b) Scenario 2: "unlearning" data with a simple two group difference design matrix

Scenario 3:

Generalized Linear Model Regression Results

Dep. Variable:	y	No. Observations:	44
Model:	GLM	Df Residuals:	42
Model Family:	Gaussian	Df Model:	1
Link Function:	identity	Scale:	0.0057920
Method:	IRLS	Log-Likelihood:	51.918
Date:	Wed, 28 Oct 2020	Deviance:	0.24326
Time:	11:33:15	Pearson chi2:	0.243
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
x1	-0.0037	0.011	-0.325	0.745	-0.026	0.019
const	0.2156	0.011	18.773	0.000	0.193	0.238

T-values: [-0.32520816 18.77344661]

(c) Scenario 3: base case of "normal" data and simple design matrix

Figure C.10: GLM results for the three explored scenarios.

C.2.2 DANN Results

	Training Data				
	B	O	Biobank MAE	OASIS MAE	Scanner Classification Accuracy %
Normal Training					
1.	✓	×	3.25 ± 2.36	16.50 ± 6.77	-
2.	×	✓	5.61 ± 3.52	4.27 ± 3.79	-
3.	✓	✓	3.30 ± 2.50	4.00 ± 2.78	98 (50)
DANN					
4.	✓	✓	3.35 ± 5.53	4.21 ± 6.02	7.2 (50)

Table C.3: Comparing normal training on the different combinations of datasets to using DANN to remove scanner information.

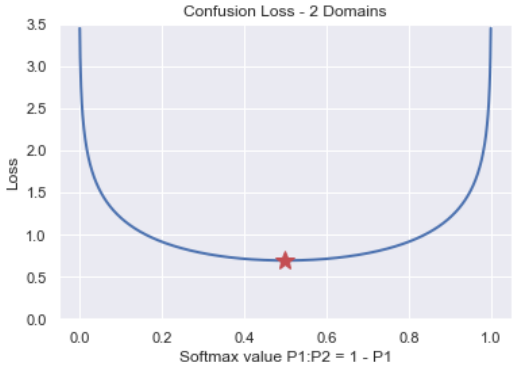
A DANN-style domain adaptation approach was also tried due to its simpler implementation. As confirmed across the literature, it was unstable to train, limiting its performance. The results from training can be see in Table C.3 where it can be seen that the scanner classification accuracy was much lower than random chance, indicating that scanner information remained in the feature representation. It can also be seen that the performance on the main task was reduced, with the domain adaptation leading to lower performance compared to normal training.

C.2.3 Visualisation of Confusion Loss

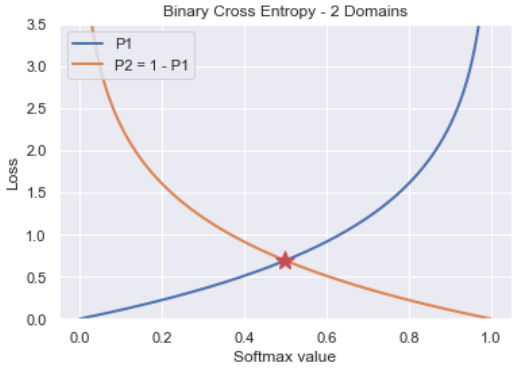
The iterative unlearning approach was explored because of the increased stability provided when creating a classifier which is equally uninformative across all of the classes. The equation for the confusion loss is given by:

$$L_{conf}(\boldsymbol{X}_u, \boldsymbol{d}_u, \boldsymbol{\Theta}_d; \boldsymbol{\Theta}_{repr}) = -\frac{1}{S_u}\sum_{s=1}^{S_u}\sum_{k=1}^{N}\frac{1}{N}log(p_{s,k}) \tag{C.6}$$

and is minimised when the softmax values are equal, as shown by the star. It can be seen that the region around the minimum is much flatter than the equivalent area for binary cross entropy, indicating that this minimum is more stable than that found using the DANN approach.



(a) Confusion loss



(b) Binary Cross Entropy

Figure C.11: Comparison of the confusion loss to binary cross entropy.

C.2.4 p-Values

Paired t-tests have been completed between experiments and the uncorrected p-values are reported separately for each dataset. For each table in the results section, the rows compared are indicated in the first column of the table.

Age prediction - basic fully-supervised learning

Rows Compared	Biobank	OASIS	Whitehall
4) - 9)	0.038	0.567	0.621
5) - 10)	0.094	1.63×10^{-14}	0.001
6) - 11)	1.69×10^{-14}	0.001	1.21×10^{-10}
7) - 12)	0.002	0.042	4.61×10^{-7}
4) - 13)	0.223	0.923	0.032
5) - 14)	0.099	9.42×10^{-15}	8.62×10^{-20}
6) - 15)	4.21×10^{-23}	0.012	0.008
7) - 16)	0.145	3.25×10^{-12}	1.96×10^{-22}
9) - 13)	0.416	0.004	0.070
10) - 14)	0.893	0.034	0.041
11) - 15)	3.96×10^{-24}	1.30×10^{-4}	3.61×10^{-8}
12) - 16)	0.011	0.666	0.098

1) - 16) Biobank Only: 0.103
2) - 16) OASIS Only: 5.62×10^{-31}
3) - 16) Whitehall Only: 9.71×10^{-71}

Age prediction - balanced datasets

Rows Compared	Biobank	OASIS	Whitehall
1) - 3)	4.65×10^{-11}	2.34×10^{-18}	2.98×10^{-10}
2) - 4)	5.67×10^{-4}	1.98×10^{-3}	7.45×10^{-6}

Age prediction - biased datasets

Rows Compared	Biobank	OASIS
1) - 2)	5.68×10^{-89}	8.62×10^{-89}
1) - 3)	2.98×10^{-92}	5.78×10^{-93}
2) - 3)	5.63×10^{-4}	0.098
4) - 5)	3.69×10^{-65}	7.14×10^{-75}
4) - 6)	3.48×10^{-72}	9.01×10^{-83}
5) - 6)	9.72×10^{-4}	0.032
7) - 8)	6.91×10^{-42}	4.78×10^{-61}
7) - 9)	9.01×10^{-51}	1.00×10^{-63}
8) - 9)	0.045	0.029

Age prediction - sex correlated with scanner

Rows Compared	Biobank	OASIS
1) - 2)	0.456	0.892
1) - 3)	0.416	0.493

Age prediction - sex correlated with age

Rows Compared	Biobank	OASIS
1) - 2)	0.001	0.009
1) - 3)	0.752	0.042

Segmentation - location of domain classifier

Rows Compared	Biobank	OASIS

1) - 2)	9.72×10^{-4}	3.00×10^{-3}
1) - 3)	0.003	0.091

Segmentation - method comparison

Rows Compared	Biobank	OASIS
3) - 4)	0.645	0.986

1) - 4) Biobank Only: 0.833
2) - 4) OASIS Only: 0.654

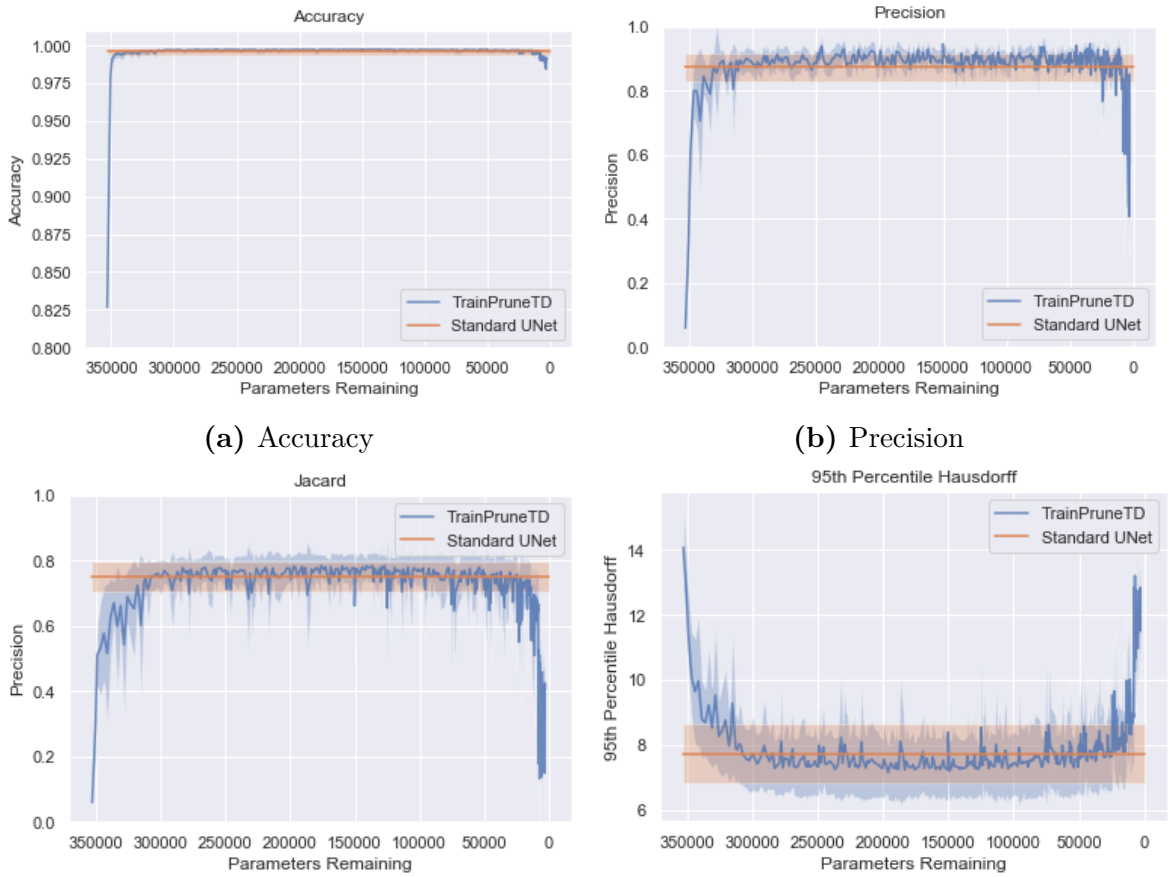
C.3 Pruning

C.3.1 Evaluation Metric - HarP Data

Dice scores were reported throughout this work. Here the results for comparing the *Standard UNet* and *TrainPruneTD* for the HarP data is evaluated using different metrics for segmentation performance, and the results shown in Fig. C.12.

C.3.2 ABIDE Dataset Distribution

Three sites from the ABIDE dataset were used during these experiments. Data was used from UM, Yale and MPG. UM was used as the training dataset, due to having the largest number of subjects available for training. The two other sites were then chosen to complement the distribution of the data points in UM: Yale has a very similar distribution in terms of age, and MPG has a very different distribution with some subjects much older than the subjects in UM. The distributions can be seen in Fig C.13. Representative slices from each site can be seen in Fig. C.14 where it can be seen that the subjects from the Yale scanner are lower quality than the other two sites, representing a challenge for generalisability to this site, and explaining why the results for the Yale scanner are lower than the other two sites.



(a) Accuracy

(b) Precision

(c) Jacard

(d) 95th Percentile Hausdorff

Figure C.12: Results with different evaluation metrics, comparing *Standard UNet* and *TrainPruneTD*.

C.3.3 Choice of Evaluation Metric - ABIDE Data

The results throughout are quoted using the Dice metric to evaluate performance. Here several evaluation metrics are assessed and it can be seen that the performance is consistent across the choice of evaluation metric.

C.3.4 Pruning Percentage for PruneFinetune

In the main body, the result for *PruneFinetune* with varying percentages are only shown for MPG. Here the results for all three sites can be found for $1, 5, 10$ and

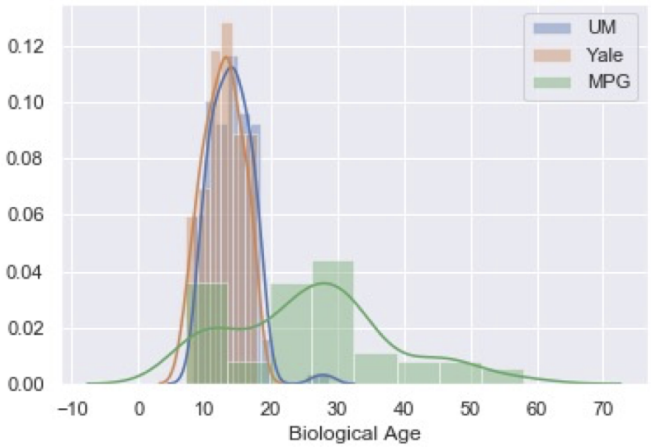


Figure C.13: Distribution of the data from the three sites used: UM, Yale and MPG with age.

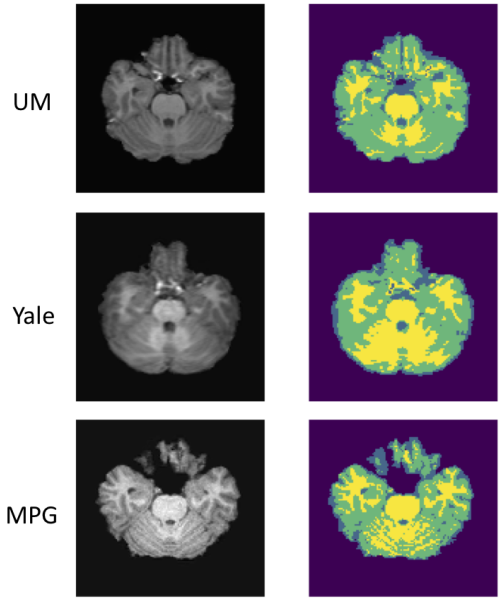


Figure C.14: Example slices from the three sites being used.

20 percent pruning.

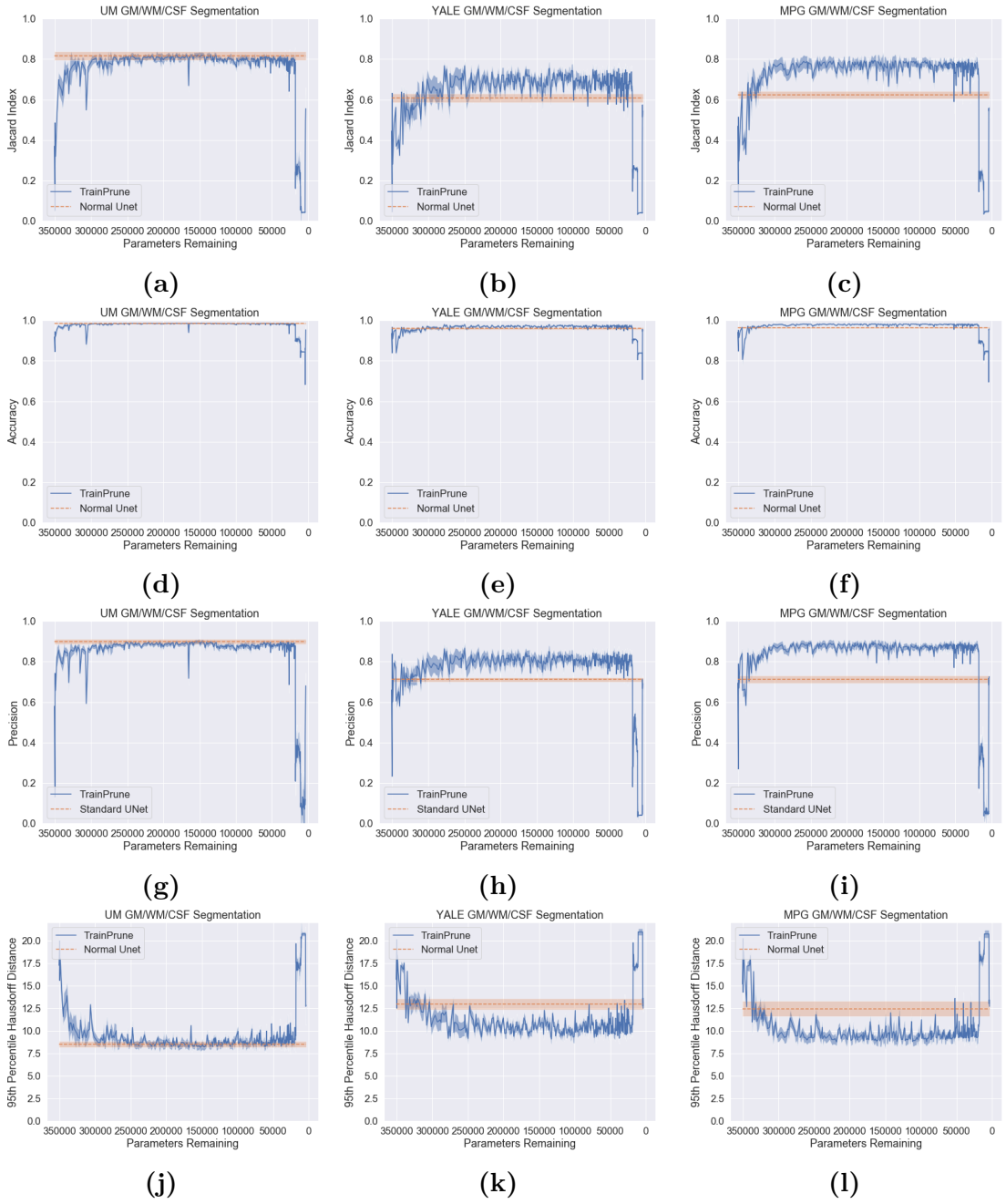
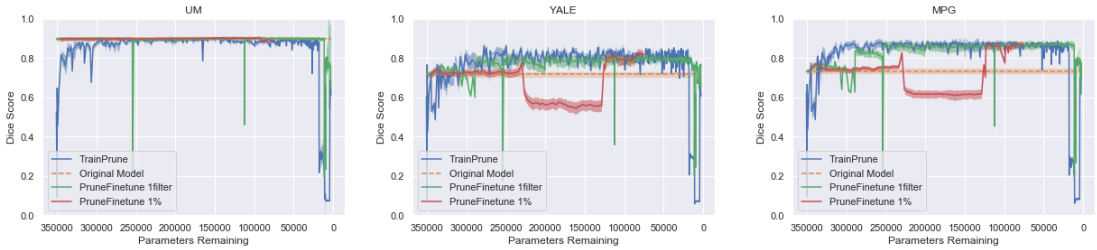
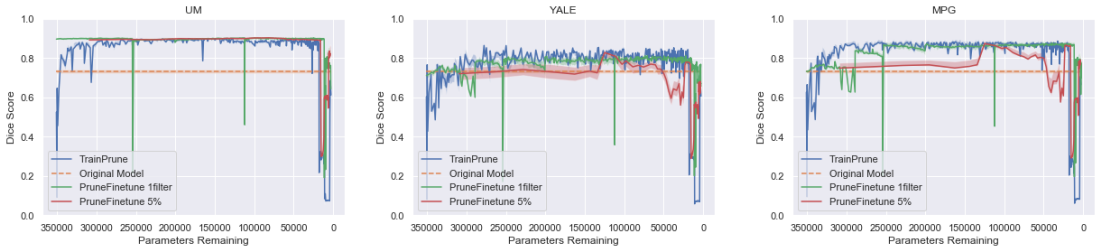


Figure C.15: TrainPrune results for the ABIDE data with targeted dropout. It can be seen that the result is consistent across the choice of evaluation metric. (a-c) Jaccard Index, (d-f) Accuracy, (g-i) Precision, (j-l) 95th percentile symmetric Hausdorff Distance.



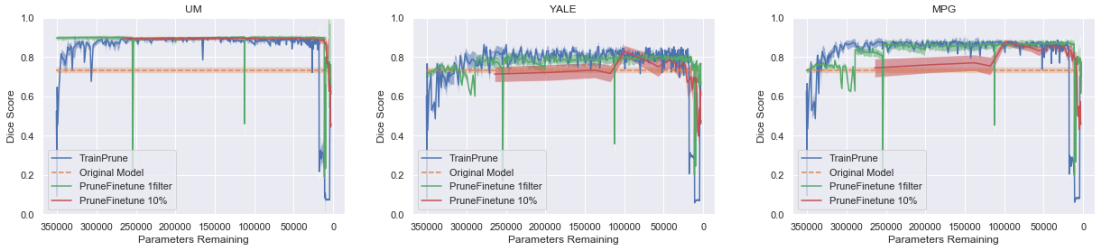
(a) **(b)** **(c)**

Figure C.16: *PruneFinetune* results with 1% pruning for each site separately.



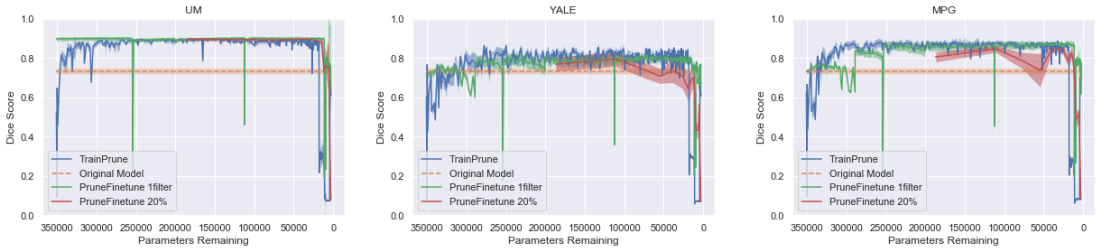
(a) **(b)** **(c)**

Figure C.17: *PruneFinetune* results with 5% pruning for each site separately.



(a) **(b)** **(c)**

Figure C.18: *PruneFinetune* results with 10% pruning for each site separately.



(a) **(b)** **(c)**

Figure C.19: *PruneFinetune* results with 20% pruning for each site separately.

Bibliography

[1] 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. 12:e1001779, 03 2015.

[2] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, AnnetteKopp-Schneider, Bennett Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald Summers, Bram Ginneken, Michel Bilello, Patrick Bilic, Patrick Christ, Richard Do, Marc Gollub, Stephan Heckers, Henkjan Huisman, William Jarnagin, and Manuel Jorge Cardoso. The medical segmentation decathlon. *ArXiv*, 06 2021.

[3] Mark Jenkinson, Christian Beckmann, Timothy Behrens, Mark Woolrich, and Stephen Smith. Fsl. *NeuroImage*, 62:782–90, 09 2011. doi: 10.1016/j.neuroimage.2011.09.015.

[4] Daniel Marcus, Tracy Wang, Jamie Parker, John Csernansky, John Morris, and Randy Buckner. Open access series of imaging studies (OASIS): Cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19:1498–507, 10 • 2007. doi: 10.1162/jocn.2007.19.9.1498.

[5] Russell T. Shinohara, Elizabeth M. Sweeney, Jeff Goldsmith, Navid Shiee, Farrah J. Mateen, Peter A. Calabresi, Samson Jarso, Dzung L. Pham, Daniel S. Reich, and Ciprian M. Crainiceanu. Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical*, 6:9–19, 2014. doi: https://doi.org/10.1016/j.nicl.2014.08.008.

[6] Stephen Smith. Fast robust automated brain extraction. *Human brain mapping*, 17:143–55, 11 2002. doi: 10.1002/hbm.10062.

[7] Bruce Fischl. FreeSurfer. *NeuroImage*, 62(2):774–781, 2012. doi: 10.1016/j.neuroimage.2012.01.021.

[8] Mark Jenkinson and Stephen Smith. Global optimisation method for robust affine registration of brain images. *Medical image analysis*, 5:143–56, 07 2001. doi: 10.1016/S1361-8415(01)00036-6.

[9] Jesper LR Andersson, Mark Jenkinson, and Stephen Smith. Non-linear registration, aka spatial normalisation. *FMRIB technical report TR07JA2*, 2010.

[10] Günther Grabner, Andrew L. Janke, Marc M. Budge, David Smith, Jens Pruessner, and D. Louis Collins. Symmetric atlasing and model based segmentation: An application to the hippocampus in older adults. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006*, pages 58–66, 2006. doi: 10.1007/11866763-8.

[11] YY Zhang, Michael Brady, and S.A. Smith. Segmentation of brain MR images through a hidden markov random field model and the expectation maximization algorithm. *IEEE Transactions on Medical Imaging*, 20:45–57, 02 2001. doi: 10.1109/42.906424.

[12] Brian Patenaude, Stephen Smith, David Kennedy, and Mark Jenkinson. A Bayesian model of shape and appearance for subcortical brain segmentation. *NeuroImage*, 56:907–22, 02 2011. doi: 10.1016/j.neuroimage.2011.02.046.

[13] Stephen Smith, Nicola De Stefano, Mark Jenkinson, and Paul Matthews. Normalized accurate measurement of longitudinal brain change. *Journal of computer assisted tomography*, 25:466–75, 05 2000. doi: 10.1016/S1053-8119(00)91589-1.

[14] Stephen Smith, Yongyue Zhang, Mark Jenkinson, Jacqueline Chen, Paul Matthews, Antonio Federico, and Nicola De Stefano. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *NeuroImage*, 17:479–89, 10 2002. doi: 10.1006/nimg.2002.1040.

[15] John Ashburner and Karl Friston. Voxel-based morphometry—the methods. *NeuroImage*, 11:805–821, 07 2000. doi: 10.1006/nimg.2000.0582.

[16] Ludovica Griffanti, Giovanna Zamboni, Aamira Khan, Linxin Li, Guendalina Bonifacio, Vaanathi Sundaresan, Ursula Schulz, Wilhelm Kuker, Marco Battaglini, Peter Rothwell, and Mark Jenkinson. BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities. *NeuroImage*, 141, 07 2016. doi: 10.1016/j.neuroimage.2016.07.018.

[17] Peter Basser and Derek Jones. Diffusion-tensor MRI: Theory, experimental design and data analysis - a technical review. *NMR in biomedicine*, 15:456–67, 11 2002. doi: 10.1002/nbm.783.

[18] T.E.J. Behrens, H. Johansen Berg, S. Jbabdi, M.F.S. Rushworth, and M.W. Woolrich. Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? *NeuroImage*, 34(1):144–155, 2007. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2006.09.018.

[19] Ahmad Hariri, Alessandro Tessitore, Venkata Mattay, Francesco Fera, and Daniel Weinberger. The amygdala response to emotional stimuli: A comparison of faces and scenes. *NeuroImage*, 17:317–23, 10 2002. doi: 10.1006/nimg.2002.1179.

[20] Yann LeCun, Y. Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–44, 05 2015. doi: 10.1038/nature14539.

[21] Yann. LeCun, B.ernhard Boser, John S. Denker, Donnie. Henderson, R. E. Howard, Wayne. Hubbard, and Lawrence D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. doi: 10.1162/neco.1989.1.4.541.

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[23] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, 11–13 Apr 2011.

[24] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *IEEE International Conference on Computer Vision (ICCV 2015)*, 1502, 02 2015. doi: 10.1109/ICCV.2015.123.

[26] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456. PMLR, 07–09 Jul 2015. URL `http://proceedings.mlr.press/v37/ioffe15.html`.

[27] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *ArXiv*, 07 2016.

[28] Jimmy Ba, Jamie Kiros, and Geoffrey Hinton. Layer normalization. *ArXiv*, 07 2016.

[29] Leslie N. Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472, 2017. doi: 10.1109/WACV.2017.58.

[30] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6980`.

[31] Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P. Bradley, Aaron Carass, Carolin Feldmann, Alejandro F. Frangi, Peter M. Full, Bram van Ginneken, Allan Hanbury, Katrin Honauer, Michal Kozubek, Bennett A. Landman, Keno März, Oskar Maier, Klaus Maier-Hein, Bjoern H. Menze, Henning Müller, Peter F. Neher, Wiro Niessen, Nasir Rajpoot, Gregory C. Sharp, Korsuk Sirinukunwattana, Stefanie Speidel, Christian Stock, Danail Stoyanov, Abdel Aziz Taha, Fons van der Sommen, Ching-Wei Wang, Marc-André Weber, Guoyan Zheng, Pierre Jannin, and Annette Kopp-Schneider. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications*, 9(1):5217, 2018. doi: 10.1038/s41467-018-07619-7.

[32] Nigam H. Shah, Arnold Milstein, and Steven C. Bagley, PhD. Making Machine Learning Models Clinically Useful. *JAMA*, 322(14):1351–1352, 10 2019. doi: 10.1001/jama.2019.10306.

[33] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10:e0118432, 03 2015. doi: 10.1371/journal.pone.0118432.

[34] Pearse Keane and Eric Topol. With an eye to AI and autonomous diagnosis. *npj Digital Medicine*, 1, 12 2018. doi: 10.1038/s41746-018-0048-y.

[35] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, and Christian Wachinger. QuickNAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy. *NeuroImage*, 186, 11 2018. doi: 10.1016/j.neuroimage.2018.11.042.

[36] Ali Hatamizadeh, Demetri Terzopoulos, and Andriy Myronenko. *End-to-End Boundary Aware Networks for Medical Image Segmentation*, pages 187–194. 10 2019. doi: 10.1007/978-3-030-32692-0_22.

[37] Yu Zhang, Bo Liu, Yinuo Wang, Zhengzhou Gao, Xiangzhi Bai, and Fugen Zhou. BDB-Net: Boundary-enhanced dual branch network for whole brain segmentation.

In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 188–197, Cham, 2020. doi: 10.1007/978-3-030-59728-319.

[38] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, R. Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2, 1990.

[39] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

[40] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. ImageNet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 01 2012. doi: 10.1145/3065386.

[41] O. Ronneberger, P.Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015.

[42] Ozgun Cicek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016*, pages 424 – 432, 2016. doi: 10.1007/978.3.319.46723.8.49.

[43] F. Milletari, N. Navab, and S. Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, 2016. doi: 10.1109/3DV.2016.79.

[44] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Neural Information Processing Systems*, 2018.

[45] Michal Drozdzal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal. The importance of skip connections in biomedical image segmentation. In *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, International Workshop on Deep Learning in Medical Image Analysis*, 08 2016. doi: 10.1007/978-3-319-46976-8_19.

[46] Simon Jegou, Michal Drozdzal, David Vázquez, Adriana Romero, and Y. Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1175–1183, 07 2017. doi: 10.1109/CVPRW.2017.156.

[47] Nahian Siddique, Sidike Paheding, Colin P. Elkin, and Vijay Devabhaktuni. U-Net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, 9:82031–82057, 2021. doi: 10.1109/ACCESS.2021.3086020.

[48] Xindi Wang, Xin-Hui Li, Jae Wook Cho, Brian E. Russ, Nanditha Rajamani, Alisa Omelchenko, Lei Ai, Annachiara Korchmaros, Stephen Sawiak, R. Austin Benn, Pamela Garcia-Saldivar, Zheng Wang, Ned H. Kalin, Charles E. Schroeder, R. Cameron Craddock, Andrew S. Fox, Alan C. Evans, Adam Messinger, Michael P. Milham, and Ting Xu. U-net model for brain extraction: Trained on humans for transfer to non-human primates. *NeuroImage*, 235:118001, 2021. doi: 10.1016/j.neuroimage.2021.118001.

[49] Martin Kolarik, Radim Burget, Vaclav Uher, and Malay Kishore Dutta. 3D Dense-U-Net for MRI brain tissue segmentation. In *2018 41st International Conference on Telecommunications and Signal Processing (TSP)*, pages 1–4, 2018. doi: 10.1109/TSP.2018.8441508.

[50] Fabian Isensee, Paul Jaeger, Simon Kohl, Jens Petersen, and Klaus Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18:1–9, 02 2021. doi: 10.1038/s41592-020-01008-z.

[51] Diederik Kingma and Max Welling. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014*, 12 2014.

[52] Pradeep Kumar Mallick, Seuc Ho Ryu, Sandeep Kumar Satapathy, Shruti Mishra, Gia Nhu Nguyen, and Prayag Tiwari. Brain MRI image classification for cancer detection using deep wavelet autoencoder-based deep neural network. *IEEE Access*, 7:46278–46287, 2019. doi: 10.1109/ACCESS.2019.2902252.

[53] Kees M van Hespen, Jaco JM Zwanenburg, Jan W Dankbaar, Mirjam I Geerling, Jeroen Hendrikse, and Hugo Kuijf. An anomaly detection approach to identify chronic brain infarcts on MRI. *Scientific Reports*, 11(7714), 2021. doi: 10.1038/s41598-021-87013-4.

[54] Hans E. Atlason, Askell Love, Sigurdur Sigurdsson, Vilmundur Gudnason, and Lotta M. Ellingsen. SegAE: Unsupervised white matter lesion segmentation from brain MRIs using a CNN autoencoder. *NeuroImage: Clinical*, 24:102085, 2019. ISSN 2213-1582. doi: 10.1016/j.nicl.2019.102085.

[55] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Y. Bengio. Generative adversarial networks. *Advances in Neural Information Processing Systems*, 3, 06 2014. doi: 10.1145/3422622.

[56] M. Sun, K. Liang, W. Zhang, Q. Chang, and X. Zhou. Non-local attention and densely-connected convolutional neural networks for malignancy suspiciousness classification of gastric ulcer. *IEEE Access*, 8:15812–15822, 2020. doi: 10.1109/ACCESS.2020.2967350.

[57] Jue Jiang, Yu-Chi Hu, Neelam Tyagi, Pengpeng Zhang, Andreas Rimner, Gig S. Mageras, Joseph O. Deasy, and Harini Veeraraghavan. Tumor-aware, adversarial domain adaptation from CT to MRI for lung cancer segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 777–785, 2018. doi: 10.1007/978-3-030-00934-2-86.

[58] Yuan Xue, Tao Xu, Han Zhang, Rodney Long, and Xiaolei Huang. SegAN: Adversarial network with multi-scale l_1 loss for medical image segmentation. *Neuroinformatics*, 16, 10 2018. doi: 10.1007/s12021-018-9377-x.

[59] Clifford Jack, Matt Bernstein, Nick Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula Britson, Jennifer Whitwell, Chadwick Ward, Anders Dale, Joel Felmlee, Jeffrey Gunter, Derek Hill, Ron Killiany, Norbert Schuff, Sabrina Fox-Bosetti, Chen Lin, Colin Studholme, and Michael Weiner. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of magnetic resonance imaging : JMRI*, 27:685–91, 05 2008. doi: 10.1002/jmri.21049.

[60] Mohammad Amin Morid, Alireza Borjali, and Guilherme Del Fiol. A scoping review of transfer learning research on medical image analysis using ImageNet. *Computers in Biology and Medicine*, 128:104115, 2021. doi: https://doi.org/10.1016/j.compbiomed.2020.104115.

[61] Bjoern Menze, András Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahaniy, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboomy, Roland Wiest, Levente Lancziy, Elizabeth Gerstnery, Marc-Andr´e Webery, Tal Arbel, Brian Avants, Nicholas Ayache, Patricia Buendia, Louis Collins, Nicolas Cordier, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 99, 12 2014. doi: 10.1109/TMI.2014.2377694.

[62] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[63] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (Canadian institute for advanced research). *Technical Report*, 2009. URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

[64] Junghwan Cho, Kyewook Lee, Ellie Shin, G. Choy, and Synho Do. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy. *arXiv: Learning*, 2015. doi: arXiv:1511.06348.

[65] Tong He, Ru Kong, Avram J. Holmes, Minh Nguyen, Mert R. Sabuncu, Simon B. Eickhoff, Danilo Bzdok, Jiashi Feng, and B.T. Thomas Yeo. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *NeuroImage*, 206:116276, 2020. ISSN 1053-8119. doi: doi.org/10.1016/j.neuroimage.2019.116276.

[66] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning with applications to medical imaging. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 02 2019.

[67] Muhammed Talo, Ulas Baloglu, Ozal yıldırım, and U Rajendra Acharya. Application of deep transfer learning for automated brain abnormality classification using MRimages. *Cognitive Systems Research*, 12 2018. doi: 10.1016/j.cogsys.2018.12.007.

[68] Atif Mehmood, Shuyuan Yang, Zhixi Feng, Min Wang, AL Smadi Ahmad, Rizwan Khan, Muazzam Maqsood, and Muhammad Yaqub. A transfer learning approach for early diagnosis of Alzheimer's disease on MRI images. *Neuroscience*, 460: 43–52, 2021. doi: https://doi.org/10.1016/j.neuroscience.2021.01.002.

[69] Veronika Cheplygina, Marleen de Bruijne, and Josien P.W. Pluim. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54:280–296, 2019. doi: https://doi.org/10.1016/j.media.2019.03.009.

[70] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 2018. doi: 10.23915/distill.00010.

[71] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[72] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B. Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3D medical image analysis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 384–393, 2019.

[73] Bin Lu, Hui-Xian Li, Zhi-Kai Chang, Le Li, Ning-Xuan Chen, Zhi-Chen Zhu, Hui-Xia Zhou, Xue-Ying Li, Yu-Wei Wang, Shi-Xian Cui, Zhao-Yu Deng, Zhen Fan, Hong Yang, Xiao Chen, Paul M. Thompson, Francisco Xavier Castellanos, and Chao-Gan Yan. A practical Alzheimer disease classifier via brain imaging-based deep learning on 85,721 samples. *ArXiv*, 2021. doi: 10.1101/2020.08.18.256594.

[74] Bo Cheng, Mingxia Liu, Daoqiang Zhang, Brent C. Munsell, and Dinggang Shen. Domain transfer learning for MCI conversion prediction. *IEEE Transactions on Biomedical Engineering*, 62(7):1805–1817, 2015. doi: 10.1109/TBME.2015.2404809.

[75] Benedikt Jonsson, Gyda Bjornsdottir, Thorgeir Thorgeirsson, Lotta Ellingsen, Gudmundur Walters, Daniel Guobjartsson, Hreinn Stefansson, K. Stefansson, and Magnus Ulfarsson. Brain age prediction using deep learning uncovers associated sequence variants. *Nature Communications*, 10, 11 2019. doi: 10.1038/s41467-019-13163-9.

[76] Vishnu Bashyam, Guray Erus, Jimit Doshi, Mohamad Habes, Ilya Nasralah, Monica Truelove-Hill, Dhivya Srinivasan, Liz Mamourian, Raymond Pomponio, Yong Fan, Lenore Launer, Colin Masters, Paul Maruff, Chuan Zhuo, Henry Volzke, Sterling Johnson, Jurgen Fripp, Nikolaos Koutsouleris, Theodore Satterthwaite, and Christos Davatzikos. MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. *Brain : a journal of neurology*, 143, 06 2020. doi: 10.1093/brain/awaa160.

[77] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, 2020.

[78] Raia Hadsell, Sumit Chopra, and Yann Lecun. Dimensionality reduction by learning an invariant mapping. pages 1735 – 1742, 02 2006. ISBN 0-7695-2597-0. doi: 10.1109/CVPR.2006.100.

[79] 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, 2021. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2020.117401.

[80] C. Wachinger, M. Reuter, and T. Klein. DeepNAT: Deep convolutional neural network for segmenting neuroanatomy. *NeuroImage*, 170:434–445, 2018.

[81] Bumshik Lee, Nagaraj Yamanakkanavar, and Jae Young Choi. Automatic segmentation of brain MRI using a novel patch-wise U-Net deep architecture. *PLOS ONE*, 15(8):1–20, 08 2020. doi: 10.1371/journal.pone.0236493.

[82] Leonie Henschel, Sailesh Conjetia, Santiago Estrada, Kersten Diers, Bruce Fischl, and Martin Reuter. FastSurfer - a fast and accurate deep learning based neuroimaging pipeline. *Neuroimage*, 219, 10 2020. doi: 10.1016/j.neuroimage.2020.117012.

[83] Ashish Sinha and Jose Dolz. Multi-scale self-guided attention for medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, PP:1–1, 04 2020. doi: 10.1109/JBHI.2020.2986926.

[84] C.W. Zitnick, N. Jojic, and Sing Bing Kang. Consistent segmentation for optical flow estimation. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1308–1315 Vol. 2, 2005. doi: 10.1109/ICCV.2005.61.

[85] Konstantinos Kamnitsas, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36, 03 2016. doi: 10.1016/j.media.2016.10.004.

[86] Bijen Khagi, Chung Ghiu Lee, and Goo-Rak Kwon. Alzheimer's disease classification from brain MRI based on transfer learning from CNN. In *2018 11th Biomedical Engineering International Conference (BMEiCON)*, pages 1–4, 2018. doi: 10.1109/BMEiCON.2018.8609974.

[87] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL `http://jmlr.org/papers/v15/srivastava14a.html`.

[88] Hongyi Zhang, Moustapha Cisse, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. 10 2017.

[89] C. Chen, C. Qin, H. Qiu, C. Ouyang, S. Wang, L. Chen, G. Tarroni, W. Bai, and D. Rueckert. Realistic adversarial data augmentation for MR image segmentation. In *23rd International Conference on Medical Image Computing and Computer Assisted Intervention*, June 2020.

[90] Egor Krivov, Maxim Pisov, and Mikhail Belyaev. *MRI Augmentation via Elastic Registration for Brain Lesions Segmentation*, pages 369–380. 12 2018. doi: 10.1007/978-3-319-75238-9_32.

[91] Nicholas Tustison, Brian Avants, Zixuan Lin, Xue Feng, Nicholas Cullen, Jaime Mata, Lucia Flors, James Gee, Talissa Altes, John Mugler, and Kun Qing. Convolutional neural networks with template-based data augmentation for functional lung image quantification. *Academic Radiology*, 26, 09 2018. doi: 10.1016/j.acra.2018.08.003.

[92] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *ArXiv*, 11 2014. doi: arXiv:1411.1784.

[93] Veit Sandfort, Ke Yan, Perry Pickhardt, and Ronald Summers. Data augmentation using generative adversarial networks (cycleGAN) to improve generalizability in CT segmentation tasks. *Scientific Reports*, 9, 11 2019. doi: 10.1038/s41598-019-52737-x.

[94] Yuya Onishi, Atsushi Teramoto, Masakazu Tsujimoto, Tetsuya Tsukamoto, Kuniaki Saito, Hiroshi Toyama, Kazuyoshi Imaizumi, and Hiroshi Fujita. Investigation of pulmonary nodule classification using multi-scale residual network enhanced with 3D GAN-synthesized volumes. *Radiological Physics and Technology*, 13, 05 2020. doi: 10.1007/s12194-020-00564-5.

[95] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017. doi: 10.1109/ICCV.2017.244.

[96] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 07 2017. doi: 10.1109/CVPR.2017.632.

[97] Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, and Annette Haworth. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 06 2021. doi: 10.1111/1754-9485.13261.

[98] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco Castellanos, Kaat Alaerts, Jeffrey Anderson, Michal Assaf, Susan Bookheimer, Mirella Dapretto, Ben Deen, Sonja Delmonte, Ilan Dinstein, Ertl-Wagner Birgit, Damien Fair, Louise Gallagher, Daniel Kennedy, Christopher Keown, Christian Keysers, and Michael Milham. The Autism Brain Imaging Data Exchange: Towards large-scale evaluation of the intrinsic brain architecture in Autism. *Molecular psychiatry*, 19, 06 2013. doi: 10.1038/mp.2013.78.

[99] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Association for Computing Machinery 2017*, pages 506–519, 04 2017. doi: 10.1145/3052973.3053009.

[100] Meichen Yu, Kristin Linn, Philip Cook, Mary Phillips, Melvin McInnis, Maurizio Fava, Madhukar Trivedi, Myrna Weissman, Russell Shinohara, and Yvette Sheline. Statistical harmonization corrects site effects in functional connectivity

measurements from multisite fMRI data. *Human Brain Mapping*, 39, 07 2018. doi: 10.1002/hbm.24241.

[101] Xiao Han, Jorge Jovicich, David Salat, Andre Kouwe, Brian Quinn, Silvester Czanner, Evelina Busa, Jenni Pacheco, Marilyn Albert, Ronald Killiany, Ralph Maguire, Herm Rosas, Nikos Makris, Anders Dale, Bradford Dickerson, and Bruce Fischl. Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage*, 32:180–94, 09 2006. doi: 10.1016/j.neuroimage.2006.02.051.

[102] Hidemasa Takao, Naoto Hayashi, and Kuni Ohtomo. Effects of study design in multi-scanner voxel-based morphometry studies. *NeuroImage*, 84, 08 2013. doi: 10.1016/j.neuroimage.2013.08.046.

[103] Hidemasa Takao, Naoto Hayashi, and Kuni Ohtomo. Effect of scanner in longitudinal studies of brain volume changes. *Journal of magnetic resonance imaging : JMRI*, 34:438–44, 08 2011. doi: 10.1002/jmri.22636.

[104] Jorge Jovicich, Silvester Czanner, Douglas Greve, Elizabeth Haley, Andre Kouwe, Randy Gollub, David Kennedy, Franz Schmitt, Gregory Brown, James Macfall, Bruce Fischl, and Anders Dale. Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data. *NeuroImage*, 30:436–43, 05 2006. doi: 10.1016/j.neuroimage.2005.09.046.

[105] Suheyla Cetin Karayumak, Sylvain Bouix, Lipeng Ning, Anthony James, Tim Crow, Martha Shenton, Marek Kubicki, and Yogesh Rathi. Retrospective harmonization of multi-site diffusion MRI data acquired with different acquisition parameters. *NeuroImage*, 184:180–200, January 2019. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2018.08.073.

[106] Blake Dewey, Can Zhao, Jacob Reinhold, Aaron Carass, Kathryn Fitzgerald, Elias Sotirchos, Shiv Saidha, Jiwon Oh, Dzung Pham, Peter Calabresi, Peter Zijl, and Jerry Prince. DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. *Magnetic Resonance Imaging*, 64, 07 2019.

[107] Daniel Moyer, Greg Ver Steeg, Chantal Tax, and Paul Thompson. Scanner invariant representations for diffusion MRI harmonization. *Magnetic Resonance in Medicine*, 04 2020. doi: 10.1002/mrm.28243.

[108] Nicola K. Dinsdale, Mark Jenkinson, and Ana I.L. Namburete. Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. *NeuroImage*, 228:117689, 2021. doi: 10.1016/j.neuroimage.2020.117689.

[109] Hao Henry Zhou, Yilin Zhang, Vamsi K. Ithapu, Sterling C. Johnson, Grace Wahba, and Vikas Singh. When can multi-site datasets be pooled for regression?

Hypothesis tests, ℓ_2-consistency and neuroscience applications. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 4170–4179, 06–11 Aug 2017.

[110] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

[111] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 05 2015.

[112] Sergi Valverde, Mostafa Salem, Mariano Cabezas, Deborah Pareto, Joan C. Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Joaquim Salvi, Arnau Oliver, and Xavier Lladó. One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. *NeuroImage: Clinical*, 21: 101638, 2019. doi: https://doi.org/10.1016/j.nicl.2018.101638.

[113] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *Information Processing in Medical Imaging*, pages 597–609, 2017. doi: 10.1007/978-3-319-59050-947.

[114] Christian S. Perone, Pedro Ballester, Rodrigo C. Barros, and Julien Cohen-Adad. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 194:1–11, 2019. doi: https://doi.org/10.1016/j.neuroimage.2019.03.026.

[115] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. *International Conference on Computer Vision*, pages 4068–4076, 12 2015.

[116] Vaanathi Sundaresan, Giovanna Zamboni, Nicola K. Dinsdale, Peter M. Rothwell, Ludovica Griffanti, and Mark Jenkinson. Comparison of domain adaptation techniques for white matter hyperintensity segmentation in brain MR images. *bioRxiv*, 2021. doi: 10.1101/2021.03.12.435171.

[117] Hao Guan, Yunbi Liu, Erkun Yang, Pew-Thian Yap, Dinggang Shen, and Mingxia Liu. Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification. *Medical Image Analysis*, 71:102076, 2021. doi: 10.1016/j.media.2021.102076.

[118] Benjamin Billot, Douglas Greve, Koen Van Leemput, Bruce Fischl, Juan Iglesias, and Adrian Dalca. A learning strategy for contrast-agnostic MRI segmentation. In *Proceedings of Machine Imaging with Deep Learning (MIDL) 2020*, 03 2020. doi: arXiv:2003.01995v3.

[119] Siddhesh Thakur, Jimit Doshi, Sarthak Pati, Saima Rathore, Chiharu Sako, Michel Bilello, Sung Ha, Gaurav Shukla, Adam Flanders, Aikaterini Kotrotsou, Mikhail Milchenko, Spencer Liem, Gregory Alexander, Joseph Lombardo, Joshua Palmer, Pamela LaMontagne, Arash Nazeri, Sanjay Talbar, Uday Kulkarni, and Spyridon Bakas. Brain extraction on MRI scans in presence of diffuse glioma: Multi-institutional performance evaluation of deep learning methods and robust modality-agnostic training. *NeuroImage*, 220:117081, 06 2020. doi: 10.1016/j.neuroimage.2020.117081.

[120] Benjamin Billot, Eleanor Robinson, Adrian V Dalca, and Juan Eugenio Iglesias. Partial volume segmentation of brain MRI scans of any resolution and contrast. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 177–187, 2020. doi: 10.1007/978-3-030-59728-318.

[121] Juan Iglesias, Benjamin Billot, Yaël Balbastre, Azadeh Tabari, John Conklin, Daniel Alexander, Polina Golland, Brian Edlow, and Bruce Fischl. Joint super-resolution and synthesis of 1 mm isotropic MP-RAGE volumes from clinical MRI exams with scans of different orientation, resolution and contrast. *ArXiv*, 12 2020. doi: arXiv:2012.13340.

[122] Clare Clement, Lucy Selman, Patrick Kehoe, Beth Howden, Athene Lane, and Jeremy Horwood. Challenges to and facilitators of recruitment to an Alzheimer's disease clinical trial: A qualitative interview study. *Journal of Alzheimer's Disease*, 69:1–9, 05 2019. doi: 10.3233/JAD-190146.

[123] Jason Karlawish, Mark S. Cary, Jonathan Rubright, and Tom TenHave. How redesigning AD clinical trials might increase study partners' willingness to participate. *Neurology*, 71(23):1883–1888, 2008. doi: 10.1212/01.wnl.0000336652.05779.ea.

[124] Christopher Hostage, Kingshuk Choudhury, Pudugramam Doraiswamy, and Jeffrey Petrella. Dissecting the gene dose-effects of the APOE η4 and η2 alleles on hippocampal volumes in aging and Alzheimer's disease. *PloS one*, 8:e54483, 02 2013. doi: 10.1371/journal.pone.0054483.

[125] Jessica Paulus and David Kent. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *npj Digital Medicine*, 3:99, 07 2020. doi: 10.1038/s41746-020-0304-9.

[126] David R. Williams and Ronald Wyatt. Racial Bias in Health Care and Health: Challenges and Opportunities. *JAMA*, 314(6):555–556, 08 2015. ISSN 0098-7484. doi: 10.1001/jama.2015.9260.

[127] Lipeng Ning, Elisenda Bonet-Carne, Francesco Grussu, Farshid Sepehrband, Enrico Kaden, Jelle Veraart, Stefano B. Blumberg, Can Son Khoo, Marco Palombo, Iasonas Kokkinos, Daniel C. Alexander, Jaume Coll-Font, Benoit Scherrer, Simon K. Warfield, Suheyla Cetin Karayumak, Yogesh Rathi, Simon Koppers, Leon Weninger, Julia Ebert, Dorit Merhof, Daniel Moyer, Maximilian Pietsch, Daan Christiaens, Rui Azeredo Gomes Teixeira, Jacques-Donald Tournier, Kurt G. Schilling, Yuankai Huo, Vishwesh Nath, Colin Hansen, Justin Blaber, Bennett A. Landman, Andrey Zhylka, Josien P.W. Pluim, Greg Parker, Umesh Rudrapatna, John Evans, Cyril Charron, Derek K. Jones, and Chantal M.W. Tax. Cross-scanner and cross-protocol multi-shell diffusion MRI data harmonization: Algorithms and results. *NeuroImage*, 221:117128, 2020. doi: https://doi.org/10.1016/j.neuroimage.2020.117128.

[128] Winston Wong, Thomas Laveist, and Joshua M. Sharfstein. Achieving health equity by design. *JAMA*, 313, 03 2015. doi: 10.1001/jama.2015.2434.

[129] Abdur R Feyjie, Reza Azad, Marco Pedersoli, Claude Kauffman, Ismail Ben Ayed, and Jose Dolz. Semi-supervised few-shot learning for medical image segmentation. *arXiv*, 2020. doi: arXiv:2003.08462.

[130] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N. Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, 2020. doi: https://doi.org/10.1016/j.media.2020.101693.

[131] Lu Wang, Dong Guo, Guotai Wang, and Shaoting Zhang. Annotation-efficient learning for medical image segmentation based on noisy pseudo labels and adversarial learning. *IEEE Transactions on Medical Imaging*, page 1–1, 2020. doi: 10.1109/tmi.2020.3047807.

[132] Hancan Zhu, Zhenyu Tang, Hewei Cheng, Yihong Wu, and Yong Fan. Multi-atlas label fusion with random local binary pattern features: Application to hippocampus segmentation. *Scientific Reports*, 9:16839, 11 2019. doi: 10.1038/s41598-019-53387-9.

[133] Xinyang Feng, Jie Yang, Andrew Laine, and Elsa Angelini. Discriminative localization in CNNs for weakly-supervised segmentation of pulmonary nodules. volume 10435, pages 568–576, 09 2017. doi: 10.1007/978-3-319-66179-765.

[134] Martin Rajchl, Matthew C. H. Lee, Ozan Oktay, Konstantinos Kamnitsas, Jonathan Passerat-Palmbach, Wenjia Bai, Mellisa Damodaram, Mary A. Rutherford, Joseph V. Hajnal, Bernhard Kainz, and Daniel Rueckert. DeepCut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE Transactions on Medical Imaging*, 36(2):674–683, 2017. doi: 10.1109/TMI.2016.2621185.

[135] Reuben Dorent, Samuel Joutard, Jonathan Shapey, Sotirios Bisdas, Neil Kitchen, Robert Bradford, Shakeel Saeed, Marc Modat, Sébastien Ourselin, and Tom Vercauteren. Scribble-based domain adaptation via co-segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 479–489, 2020. doi: 10.1007/978-3-030-59710-8-47.

[136] Xiangde Luo, Guotai Wang, Tao Song, Jingyang Zhang, Michael Aertsen, Jan Deprest, Sebastien Ourselin, Tom Vercauteren, and Shaoting Zhang. MIDeepSeg: Minimally interactive segmentation of unseen objects from medical images using deep learning. *Medical Image Analysis*, page 102102, 2021. doi: https://doi.org/10.1016/j.media.2021.102102.

[137] R. McEver and B. Manjunath. PCAMs: Weakly supervised semantic segmentation using point supervision. *ArXiv*, 07 2020.

[138] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z. Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, pages 399–407, 2017.

[139] Vishwesh Nath, Dong Yang, Bennett A. Landman, Daguang Xu, and Holger R. Roth. Diminishing uncertainty within the training pool: Active learning for medical image segmentation. *ArXiv*, 2021. doi: abs/2101.02323.

[140] Ilija Radosavovic, Piotr Dollar, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4119–4128, 06 2018. doi: 10.1109/CVPR.2018.00433.

[141] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192, 06–11 Aug 2017.

[142] Lorenzo Venturini, Aris T. Papageorghiou, J. Alison Noble, and Ana I. L. Namburete. Uncertainty estimates as data selection criteria to boost omni-supervised learning. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 689–698, 2020. doi: 10.1007/978-3-030-59710-867.

[143] Federico Cabitza, Davide Ciucci, and Raffaele Rasoini. A giant with feet of clay: On the validity of the data that feed machine learning in medicine. In Federico Cabitza, Carlo Batini, and Massimo Magni, editors, *Organizing for the Digital World*, pages 121–136, 2019. doi: 10.1007/978-3-319-90503-710.

[144] Or Shwartzman, Harel Gazit, Ilan Shelef, and Tammy Riklin-Raviv. The worrisome impact of an inter-rater bias on neural network training. *ArXiv*, 2020. doi: arXiv:1906.11872.

[145] Mark L Graber. The incidence of diagnostic error in medicine. *BMJ Quality & Safety*, 22(Suppl 2):ii21–ii27, 2013. doi: 10.1136/bmjqs-2012-001615.

[146] Federico Cabitza, Andrea Campagner, Domenico Albano, Alberto Aliprandi, Alberto Bruno, Vito Chianca, Angelo Corazza, Francesco Di Pietto, Angelo Gambino, Salvatore Gitto, Carmelo Messina, Davide Orlandi, Luigi Pedone, Marcello Zappia, and Luca Maria Sconfienza. The elephant in the machine: Proposing a new metric of data reliability and its application to a medical case to assess classification reliability. *Applied Sciences*, 10(11), 2020. doi: 10.3390/app10114014.

[147] Carl Svensson, Ron Hübler, and Marc Figge. Automated classification of circulating tumor cells and the impact of interobsever variability on classifier training and performance. *Journal of Immunology Research*, 2015:1–9, 10 2015. doi: 10.1155/2015/573165.

[148] Martin Visser, Domenique Müller, R.J.M. Duijn, M. Smits, Niels Verburg, Eef Hendriks, R.J.A. Nabuurs, J.C.J. Bot, Roelant Eijgelaar, M. Witte, M.B. Herk, F. Barkhof, Philip De Witt Hamer, and Jan De Munck. Inter-rater agreement in glioma segmentations on longitudinal MRI. *NeuroImage: Clinical*, 22:101727, 02 2019. doi: 10.1016/j.nicl.2019.101727.

[149] Julia K. Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, and Holger A. Haenssle. Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatology*, 155(10):1135–1141, 10 2019. doi: 10.1001/jamadermatol.2019.1735.

[150] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

[151] R. Selvaraju Ramprasaath, Michael Cogswell, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks

via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 10 2017. doi: 10.1109/ICCV.2017.74.

[152] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, 03 2017. doi: 10.5555/3305890.3306024.

[153] Alexander Binder, Sebastian Lapuschkin, Gregoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for deep neural network architectures. In *Information Science and Applications (ICISA) 2016*, pages 913–922, 01 2016. doi: 10.1007/978-981-10-0557-2_87.

[154] 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*, 11:194, 2019. doi: 10.3389/fnagi.2019.00194.

[155] Somosmita Mitra, Subhashis Banerjee, and Yoichi Hayashi. Volumetric brain tumour detection from MRI using visual saliency. *PLOS ONE*, 12(11):1–14, 11 2017. doi: 10.1371/journal.pone.0187209.

[156] Petra Takacs and Andrea Manno-Kovacs. MRI brain tumor segmentation combining saliency and convolutional network features. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, 09 2018. doi: 10.1109/CBMI.2018.8516544.

[157] Matthew Zeiler and Rob Fergus. Visualizing and understanding convolutional neural networks. In *2014 IEEE European Conference on Computer Vision (ECCV)*, volume 8689, 11 2014. ISBN 978-3-319-10589-5. doi: 10.1007/978-3-319-10590-1_53.

[158] Anees Abrol, Manish Bhattarai, Alex Fedorov, Yuhui Du, Sergey Plis, and Vince Calhoun. Deep residual learning for neuroimaging: An application to predict progression to Alzheimer's disease. *Journal of Neuroscience Methods*, 339:108701, 2020.

[159] C Bass, MD Silva, CH Sudre, PD Tudosiu, SM Smith, and EC Robinson. ICAM: Interpretable classification via disentangled representations and feature attribution mapping. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020. doi: arxiv.org/abs/2006.08287.

[160] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 9525–9536, 2018. doi: 10.5555/3327546.3327621.

[161] C. F. Baumgartner, L. M. Koch, K. C. Tezcan, J. X. Ang, and E. Konukoglu. Visual feature attribution using wasserstein GANs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8309–8319, 2018. doi: 10.1109/CVPR.2018.00867.

[162] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Bianchi-Berthouze. Evaluating saliency map explanations for convolutional neural networks: A user study. In *IUI '20: 25th International Conference on Intelligent User Interfaces*, pages 275–285, 03 2020. doi: 10.1145/3377325.3377519.

[163] JongChan Park, Sanghyun Woo, Joon-Young Lee, and Inso Kweon. BAM: Bottleneck attention module. In *Proceedings of the British Machine Vision Conference (BMVC) 2018*, 07 2018. doi: arXiv:1807.06514.

[164] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. doi: 10.1109/CVPR.2018.00813.

[165] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. doi: 10.1109/CVPR.2018.00745.

[166] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, 53:197–207, 2019. doi: https://doi.org/10.1016/j.media.2019.01.012.

[167] Manhua Liu, Fan Li, Hao Yan, Kundong Wang, Yixin Ma, Li Shen, and Mingqing Xu. A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease. *NeuroImage*, 208:116459, 2020. doi: https://doi.org/10.1016/j.neuroimage.2019.116459.

[168] Giovanni Frisoni and Clifford Jack. HarP: The EADC-ADNI harmonized protocol for manual hippocampal segmentation. a standard of reference from a global working group. *Alzheimer's and Dementia*, 11, 02 2015. doi: 10.1016/j.jalz.2014.05.1761.

[169] Nicola K. Dinsdale, Mark Jenkinson, and Ana I. L. Namburete. Spatial warping network for 3D segmentation of the hippocampus in MR images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 284–291, 2019. doi: 10.1007/978-3-030-32248-9.32.

[170] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1050–1059, 20–22 Jun 2016. doi: 10.5555/3045390.3045502.

[171] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. doi: 10.5555/2627435.2670313.

[172] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, and Christian Wachinger. Inherent brain segmentation quality control from fully ConvNet Monte Carlo sampling. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 664–672, 2018. doi: 10.1007/978-3-030-00928-1-75.

[173] Yongchan Kwon, Joong Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics and Data Analysis*, 142, February 2020. doi: 10.1016/j.csda.2019.106816.

[174] Jeppe Thagaard, Søren Hauberg, Bert van der Vegt, Thomas Ebstrup, Johan D. Hansen, and Anders B. Dahl. Can you trust predictive uncertainty under real dataset shifts in digital pathology? In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 824–833, Cham, 2020. doi: 10.1007/978-3-030-59710-8-80.

[175] Max-Heinrich Laves, Sontje Ihler, Jacob F. Fast, Lüder A. Kahrs, and Tobias Ortmaier. Well-calibrated regression uncertainty in medical imaging with deep learning. In *Medical Imaging with Deep Learning*, 2020.

[176] Meenakshi Khosla, Keith Jamison, Amy Kuceyeski, and Mert R. Sabuncu. Ensemble learning with 3D convolutional neural networks for functional connectome-based prediction. *NeuroImage*, 199:651–662, 2019. doi: https://doi.org/10.1016/j.neuroimage.2019.06.012.

[177] Fabian Isensee, Paul Jaeger, Simon Kohl, Jens Petersen, and Klaus Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18:1–9, 02 2021. doi: 10.1038/s41592-020-01008-z.

[178] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS 2014 Deep Learning Workshop*, 03 2014. doi: arXiv.

[179] Mohsen Ghafoorian, Jonas Teuwen, Rashindra Manniesing, Frank-Erik Leeuw, Bram Ginneken, Nico Karssemeijer, and Bram Platel. Student beats the teacher: Deep neural networks for lateral ventricles segmentation in brain MR. In *Proceedings Volume 10574, Medical Imaging 2018: Image Processing; 105742U (2018)*, 01 2018. doi: doi.org/10.1117/12.2293569.

[180] Balamurali Murugesan, Sricharan Vijayarangan, Kaushik Sarveswaran, Keerthi Ram, and Mohanasankar Sivaprakasam. KD-MRI: A knowledge distillation framework for image reconstruction and image restoration in MRI workflow. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, volume 121 of *Proceedings of Machine Learning Research*, pages 515–526. PMLR, 06–08 Jul 2020.

[181] S. Vaze, W. Xie, and A. I. L. Namburete. Low-memory CNNs enabling real-time ultrasound segmentation towards mobile deployment. *IEEE Journal of Biomedical and Health Informatics*, 24(4):1059–1069, 2020. doi: 10.1109/JBHI.2019.2961264.

[182] Yann Le Cun, John S. Denker, and Sara A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems 2*, page 598–605, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc. ISBN 1558601007.

[183] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient transfer learning. In *International Conference on Learning Representations*, 11 2016.

[184] Nick Pawlowski and Ben Glocker. Is texture predictive for age and sex in brain MRI? In *International Conference on Medical Imaging with Deep Learning – Extended Abstract Track*, 08–10 Jul 2019. doi: openreview.net/forum?id=BJxgXfab94.

[185] Latanya Sweeney. K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, October 2002. doi: 10.1142/S0218488502001648.

[186] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020. doi: 10.1109/MSP.2020.2975749.

[187] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.*, 10(2), January 2019. doi: 10.1145/3298981.

[188] Nicola Rieke, Jonny Hancox, Fausto Milletari, Holger Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu Galtier, Bennett Landman, Klaus Maier-Hein, Sebastien Ourselin, Micah Sheller, Ronald Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and Manuel Jorge Cardoso. The future of digital health with federated learning. *npj Digital Medicine volume*, 03 2020. doi: 10.1038/s41746-020-00323-1.

[189] Li Huang, Andrew L. Shea, Huining Qian, Aditya Masurkar, Hao Deng, and Dianbo Liu. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical

records. *Journal of Biomedical Informatics*, 99:103291, 2019. doi: 10.1016/j.jbi.2019.103291.

[190] Micah J. Sheller, G. Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 92–104, 2019. doi: 10.1007/978-3-030-11723-8-9.

[191] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014. doi: 10.1561/0400000042.

[192] Alexander Ziller, Dmitrii Usynin, Rickmer Braren, Marcus Makowski, Daniel Reuckert, and George Kaissis. Medical imaging deep learning with differential privacy. *Nature Scientific Reports*, 2021. doi: 10.1038/s41598-021-93030-0.

[193] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 909–910, 2015. doi: 10.1109/ALLERTON.2015.7447103.

[194] Kelly Rootes-Murdy, Harshvardhan Gazula, Eric Verner, Ross Kelly, Thomas DeRamus, Sergey Plis, Anand Sarwate, and Vince Calhoun. Federated analysis of neuroimaging data: A review of the field. *Neuroinformatics*, 11 2021. doi: 10.1007/s12021-021-09550-7.

[195] UN. World population ageing 2015. *United Nations Department of Economic and Social Affairs*, 2015.

[196] James H Cole and Katja Franke. Predicting age using neuroimaging: innovative brain ageing biomarkers. *Trends Neuroscience*, 40(12), 2017.

[197] A Pfefferbaum, Daniel Mathalon, Edith Sullivan, J Rawles, Robert Zipursky, and Kar Onn Lim. A quantitative magnetic resonance imaging study of changes in brain morphology from infancy to late adulthood. *Archives of neurology*, 51: 874–87, 10 1994. doi: 10.1001/archneur.1994.00540210046012.

[198] Christina E. Hugenschmidt, Ann M. Peiffer, Robert A. Kraft, Ramon Casanova, Andrew R. Deibler, Jonathan H. Burdette, Joseph A. Maldjian, and Paul J. Laurienti. Relating imaging indices of white matter integrity and volume in healthy older adults. *Cerebral Cortex*, 18(2):433–442, 2007.

[199] Naftali Raz, Paolo Ghisletta, Karen M. Rodrigue, Kristen M. Kennedy, and Ulman Lindenberger. Trajectories of brain aging in middle-aged and older adults: Regional and individual differences. *NeuroImage*, 51(2):501 – 511, 2010.

[200] Rachael I. Scahill, Chris Frost, Rhian Jenkins, Jennifer L. Whitwell, Martin N. Rossor, and Nick C. Fox. A Longitudinal Study of Brain Volume Changes in Normal Aging Using Serial Registered Magnetic Resonance Imaging. *JAMA Neurology*, 60(7):989–994, 07 2003.

[201] Kristine Walhovd, Lars Westlye, Inge Amlien, Thomas Espeseth, Ivar Reinvang, Naftali Raz, Ingrid Agartz, David Salat, Doug Greve, Bruce Fischl, Anders Dale, and Anders Fjell. Consistent neuroanatomical age-related volume differences across multiple samples. *Neurobiology of aging*, 32:916–32, 07 2009. doi: 10.1016/j.neurobiolaging.2009.05.013.

[202] A. Fjell, L. Westlye, H. Grydeland, I. Amlien, T. Espeseth, I. Reinvang, N. Raz, A. Dale, and K. Walhovd. Accelerating cortical thinning: unique to dementia or universal in aging? *Cerebral cortex*, 24 4:919–34, 2014. doi: 10.1093/cercor/bhs379.

[203] Andreas B. Storsve, Anders M. Fjell, Christian K. Tamnes, Lars T. Westlye, Knut Overbye, Hilde W. Aasland, and Kristine B. Walhovd. Differential longitudinal changes in cortical thickness, surface area and volume across the adult life span: Regions of accelerating and decelerating change. *Journal of Neuroscience*, 34 (25):8488–8498, 2014.

[204] Anders Fjell, Lars Westlye, Hakon Grydeland, Inge Amlien, Thomas Espeseth, Ivar Reinvang, Naftali Raz, Dominic Holland, Anders Dale, and Kristine Walhovd. Critical ages in the life course of the adult brain: Nonlinear subcortical aging. *Neurobiology of aging*, 34, 05 2013. doi: 10.1016/j.neurobiolaging.2013.04.006.

[205] Mohamad Habes, Guray Erus, Jon Toledo, Tianhao Zhang, Nick Bryan, Lenore Launer, Yves Rosseel, Deborah Janowitz, Jimit Doshi, Sandra Van der Auwera, Bettina von Sarnowski, Katrin Hegenscheid, Norbert Hosten, Georg Homuth, Henry Völzke, Ulf Schminke, Wolfgang Hoffmann, Hans Grabe, and Christos Davatzikos. White matter hyperintensities and imaging patterns of brain ageing in the general population. *Brain*, 139:aww008, 02 2016. doi: 10.1093/brain/aww008.

[206] Joanna Wardlaw, Eric Smith, Geert Biessels, Charlotte Cordonnier, Franz Fazekas, Richard Frayne, Richard Lindley, John O'Brien, Frederik Barkhof, Oscar Benavente, Sandra Black, Carol Brayne, Monique Breteler, Hugues Chabriat, Charles DeCarli, Frank-Erik Leeuw, Fergus Doubal, Marco Duering, Nick Fox, and STandards v. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *The Lancet Neurology*, 12: 822–838, 08 2013. doi: 10.1016/S1474-4422(13)70124-8.

[207] Shu-Wei Sun, Sheng-Kwei Song, Michael P. Harms, Shiow-Jiuan Lin, David M. Holtzman, Kalpana M. Merchant, and John J. Kotyk. Detection of age-dependent

brain injury in a mouse model of brain amyloidosis associated with Alzheimer's disease using magnetic resonance diffusion tensor imaging. *Experimental Neurology*, 191(1):77–85, 2005. doi: 10.1016/j.expneurol.2004.09.006.

[208] Sheng-Kwei Song, Jun Yoshino, Tuan le, Shiow-Jiuan Lin, SW Sun, Anne Cross, and Regina Armstrong. Demylination increases radial diffusivity in corpus callosum of mouse brain. *NeuroImage*, 26:132–40, 06 2005. doi: 10.1016/j.neuroimage.2005.01.028.

[209] S. Cox, S. Ritchie, E. Tucker-Drob, D. Liewald, S. Hagenaars, G. Davies, J. Wardlaw, C. Gale, M. Bastin, and I. Deary. Ageing and brain white matter structure in 3,513 UK Biobank participants. *Nature Communications*, 7, 2016.

[210] Lars Marstaller, Mark Williams, Anina Rich, Greg Savage, and Hana Burianová. Aging and large-scale functional networks: White matter integrity, gray matter volume, and functional connectivity in the resting state. *Neuroscience*, 290, 01 2015. doi: 10.1016/j.neuroscience.2015.01.049.

[211] Roser Sala-Llonch, David Bartrés-Faz, and Carme Junqué. Reorganization of brain networks in aging: A review of functional connectivity studies. *Frontiers in psychology*, 6:663, 05 2015. doi: 10.3389/fpsyg.2015.00663.

[212] A. Cherubini, M. E. Caligiuri, P. Péran, U. Sabatini, C. Cosentino, and F. Amato. Importance of multimodal MRI in characterizing brain tissue and its potential application for individual age prediction. *IEEE Journal of Biomedical and Health Informatics*, 20(5):1232–1239, 2016.

[213] R Schmidt, H Schmidt, J Haybaeck, M Loitfelder, S Weis, M Cavalieri, S Seiler, C Enzinger, S Ropele, T Erkinjuntti, L Pantoni, P Scheltens, F Fazekas, and K Jellinger. Heterogeneity in age-related white matter changes. *acta neuropathologica*, 122(2):171–185, 2011.

[214] Marco Lorenzi, Xavier Pennec, Giovanni B. Frisoni, and Nicholas Ayache. Disentangling normal aging from alzheimer's disease in structural magnetic resonance images. *Neurobiology of Aging*, 36:S42 – S52, 2015. ISSN 0197-4580. Novel Imaging Biomarkers for Alzheimer's Disease and Related Disorders (NIBAD).

[215] Nhat Trung Doan, Andreas Engvig, Krystal Zaske, Karin Persson, Martina Jonette Lund, Tobias Kaufmann, Aldo Cordova-Palomera, Dag Alnæs, Torgeir Moberget, Anne Brækhus, Maria Lage Barca, Jan Egil Nordvik, Knut Engedal, Ingrid Agartz, Geir Selbæk, Ole A. Andreassen, and Lars T. Westlye. Distinguishing early and late brain aging from the Alzheimer's disease spectrum: consistent morphological patterns across independent samples. *NeuroImage*, 158: 282 – 295, 2017. ISSN 1053-8119.

[216] Jennifer H. Barnett, Lily Lewis, Andrew D. Blackwell, and Matthew Taylor. Early intervention in Alzheimer's disease: a health economic study of the effects of diagnostic timing. *BMC Neurology*, 14(1), 2014.

[217] James Cole, Rudra Poudel, Dimosthenis Tsagkrasoulis, Matthan Caan, Claire Steves, Tim Spector, and Giovanni Montana. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, 163, 12 2016. doi: 10.1016/j.neuroimage.2017.07.059.

[218] Katja Franke, Georg Hagemann, Ekkehard Schleussner, and Christian Gaser. Changes of individual BrainAGE during the course of the menstrual cycle. *NeuroImage*, 115:1 – 6, 2015.

[219] Stephen M Smith, Lloyd T Elliott, Fidel Alfaro-Almagro, Paul McCarthy, Thomas E Nichols, Gwenaëlle Douaud, and Karla L Miller. Brain aging comprises many modes of structural and functional change with distinct genetic and biophysical associations. *eLife*, 9, mar 2020. doi: 10.7554/eLife.52677.

[220] 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, 2021. ISSN 1361-8415. doi: 10.1016/j.media.2020.101871.

[221] Seyed Abolfazl Valizadeh, Jürgen Hänggi, Susan Mérillat, and Lutz Jäncke. Age prediction on the basis of brain anatomical measures. *Human Brain Mapping*, 38, 11 2016. doi: 10.1002/hbm.23434.

[222] James Cole, S Ritchie, M Bastin, M Hernández, Susana Muñoz Maniega, N Royle, Janie Corley, Alison Pattie, Sarah Harris, Qian Zhang, Naomi Wray, P Redmond, R Marioni, J Starr, Simon Cox, Joanna Wardlaw, D Sharp, and I Deary. Brain age predicts mortality. *Molecular Psychiatry*, 23, 04 2017.

[223] Jason Steffener, Christian Habeck, Deirdre O'Shea, Qolamreza Razlighi, Louis Bherer, and Yaakov Stern. Differences between chronological and brain age are related to education and self-reported physical activity. *Neurobiology of Aging*, 40:138 – 144, 2016.

[224] John Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38:95–113, 11 2007. doi: 10.1016/j.neuroimage.2007.07.007.

[225] Katja Franke, Gabriel Ziegler, Stefan Klöppel, and Christian Gaser. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. *NeuroImage*, 50:883–92, 04 2010. doi: 10.1016/j.neuroimage.2010.01.005.

[226] Nikolaos Koutsouleris, Christos Davatzikos, Stefan Borgwardt, Christian Gaser, Ronald Bottlender, Thomas Frodl, Peter Falkai, Anita Riecher-Rössler, Hans-Jürgen Möller, Maximilian Reiser, Christos Pantelis, and Eva Meisenzahl. Accelerated Brain Aging in Schizophrenia and Beyond: A Neuroanatomical Marker of Psychiatric Disorders. *Schizophrenia Bulletin*, 40(5):1140–1153, 10 2013. doi: 10.1093/schbul/sbt142.

[227] Longfei Su, Lubin Wang, and Dewen Hu. Predicting the age of healthy adults from structural MRI by sparse representation. In *Intelligent Science and Intelligent Data Engineering*, pages 271–279, Berlin, Heidelberg, 2013. doi: 10.1007/978-3-642-36669-7-34.

[228] Jenessa Lancaster, Romy Lorenz, Rob Leech, and James H. Cole. Bayesian optimization for neuroimaging pre-processing in brain age classification and prediction. *Frontiers in Aging Neuroscience*, 10:28, 2018. doi: 10.3389/fnagi.2018.00028.

[229] Harini Eavani, Mohamad Habes, Theodore Satterthwaite, Yang An, Meng-Kang Hsieh, Nicolas Honnorat, Guray Erus, Jimit Doshi, Luigi Ferrucci, Lori Beason-Held, Susan Resnick, and Christos Davatzikos. Heterogeneity of structural and functional imaging patterns of advanced brain aging revealed via machine learning methods. *Neurobiology of Aging*, 71, 06 2018. doi: 10.1016/j.neurobiolaging.2018.06.013.

[230] Deepthi P. Varikuti, Sarah Genon, Aristeidis Sotiras, Holger Schwender, Felix Hoffstaedter, Kaustubh R. Patil, Christiane Jockwitz, Svenja Caspers, Susanne Moebus, Katrin Amunts, Christos Davatzikos, and Simon B. Eickhoff. Evaluation of non-negative matrix factorization of grey matter in age prediction. *NeuroImage*, 173:394–410, 2018. doi: 10.1016/j.neuroimage.2018.03.007.

[231] Bing Wang and Tuan D. Pham. MRI-based age prediction using hidden Markov models. *Journal of Neuroscience Methods*, 199(1):140–145, 2011. doi: https://doi.org/10.1016/j.jneumeth.2011.04.022.

[232] J. Cole, R. Leech, and D. Sharp. Prediction of brain age suggests accelerated atrophy after traumatic brain injury. *Annals of Neurology*, 77:571 – 581, 2015.

[233] C. Kondo, K. Ito, K. Wu, K. Sato, Y. Taki, H. Fukuda, and T. Aoki. An age estimation method using brain local features for T1-weighted images. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 666–669, 2015.

[234] R. Fujimoto, K. Ito, K. Wu, K. Sato, Y. Taki, H. Fukuda, and T. Aoki. Brain age estimation from T1-weighted images using effective local features. In *2017*

39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 3028–3031, 2017.

[235] Gemma Monté-Rubio, Carles Falcón, Edith Pomarol-Clotet, and John Ashburner. A comparison of various MRI feature types for characterizing whole brain anatomical differences using linear pattern recognition methods. *NeuroImage*, 178, 06 2018. doi: 10.1016/j.neuroimage.2018.05.065.

[236] Timothy T. Brown, Joshua M. Kuperman, Yoonho Chung, Matthew Erhart, Connor McCabe, Donald J. Hagler, Vijay K. Venkatraman, Natacha Akshoomoff, David G. Amaral, Cinnamon S. Bloss, B.J. Casey, Linda Chang, Thomas M. Ernst, Jean A. Frazier, Jeffrey R. Gruen, Walter E. Kaufmann, Tal Kenet, David N. Kennedy, Sarah S. Murray, Elizabeth R. Sowell, Terry L. Jernigan, and Anders M. Dale. Neuroanatomical assessment of biological maturity. *Current Biology*, 22(18):1693–1698, 2012. doi: 10.1016/j.cub.2012.07.002.

[237] Jieqiong Wang, Wenjing Li, Wyn Mew, Dai Dai, Jing Hua, and Huiguang He. Age estimation using cortical surface pattern combining thickness with curvatures. *Medical & biological engineering & computing*, 52, 01 2014. doi: 10.1007/s11517-013-1131-9.

[238] Budhachandra Khundrakpam, Jussi Tohka, and Alan Evans. Prediction of brain maturity based on cortical thickness at different spatial resolutions. *NeuroImage*, 02 2015. doi: 10.1016/j.neuroimage.2015.02.046.

[239] Benjamín Gutierrez Becker, Tassilo Klein, and Christian Wachinger. Gaussian process uncertainty in age estimation as a measure of brain abnormality. *NeuroImage*, 175, 04 2018. doi: 10.1016/j.neuroimage.2018.03.075.

[240] Heath Pardoe and Ruben Kuzniecky. NAPR: a cloud-based framework for neuroanatomical age prediction. *Neuroinformatics*, 16(1):43–49, 01 2018. doi: 10.1101/099309.

[241] Christopher Madan and Elizabeth Kensinger. Predicting age from cortical structure across the lifespan. *Eur J Neurosci.*, 47(5):399–416, 01 2018. doi: 10.1101/248518.

[242] Habtamu M. Aycheh, Joon-Kyung Seong, Jeong-Hyeon Shin, Duk L. Na, Byungkon Kang, Sang W. Seo, and Kyung-Ah Sohn. Biological brain age prediction using cortical thickness data: A large scale cohort study. *Frontiers in Aging Neuroscience*, 10:252, 2018. doi: 10.3389/fnagi.2018.00252.

[243] Matthias Guggenmos, Katharina Schmack, Maria Sekutowicz, Maria Garbusow, Miriam Sebold, Christian Sommer, Michael Smolka, Hans-Ulrich Wittchen, Ulrich Zimmermann, Andreas Heinz, and Philipp Sterzer. Quantitative neurobiological

evidence for accelerated brain aging in alcohol dependence. *Translational Psychiatry*, 7, 12 2017. doi: 10.1038/s41398-017-0037-y.

[244] Epifanio Bagarinao, Hirohisa Watanabe, Satoshi Maesawa, Daisuke Mori, Kazuhiro Hara, Kazuya Kawabata, Noritaka Yoneyama, Reiko Ohdake, Kazunori Imai, Michihito Masuda, Takamasa Yokoi, Aya Ogura, Toshihiko Wakabayashi, Masafumi Kuzuya, Norio Ozaki, Minoru Hoshiyama, Haruo Isoda, Shinji Naganawa, and Gen Sobue. An unbiased data-driven age-related structural brain parcellation for the identification of intrinsic brain volume changes over the adult lifespan. *NeuroImage*, 169, 12 2017. doi: 10.1016/j.neuroimage.2017.12.014.

[245] Nico Dosenbach, Binyam Nardos, Alexander Cohen, Damien Fair, Jonathan Power, Jessica Church, Steven Nelson, Gagan Wig, Alecia Vogel, Christina Lessov-Schlaggar, Kelly Barnes, Joseph Dubis, Eric Feczko, Rebecca Coalson, John Pruett, Deanna Barch, Steve Petersen, and Bradley Schlaggar. Prediction of individual brain maturity using fMRI. *Science (New York, N.Y.)*, 329:1358–61, 09 2010. doi: 10.1126/science.1194144.

[246] Kamen Tsvetanov, Richard Henson, Lorraine Tyler, Adeel Razi, Linda Geerligs, Timothy Ham, and James Rowe. Extrinsic and intrinsic brain network connectivity maintains cognition across the lifespan despite accelerated decay of regional brain activation. *Journal of Neuroscience*, 36:3115–3126, 03 2016. doi: 10.1523/JNEUROSCI.2733-15.2016.

[247] Lubin Wang, Longfei Su, Hui Shen, and Dewen Hu. Decoding lifespan changes of the human brain using resting-state functional connectivity MRI. *PLOS ONE*, 7(8):1–12, 08 2012. doi: 10.1371/journal.pone.0044530.

[248] Hongming Li, Theodore D. Satterthwaite, and Yong Fan. Brain age prediction based on resting-state functional connectivity patterns using convolutional neural networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 101–104, 2018. doi: 10.1109/ISBI.2018.8363532.

[249] 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, 12 2015.

[250] Cheol E. Han, Luis R. Peraza, John-Paul Taylor, and Marcus Kaiser. Predicting age across human lifespan based on structural connectivity from diffusion tensor imaging. In *2014 IEEE Biomedical Circuits and Systems Conference (BioCAS) Proceedings*, pages 137–140, 2014. doi: 10.1109/BioCAS.2014.6981664.

[251] Benson Mwangi, Khader M. Hasan, and Jair C. Soares. Prediction of individual subject's age across the human lifespan using diffusion tensor imaging: A machine

learning approach. *NeuroImage*, 75:58–67, 2013. doi: https://doi.org/10.1016/j.neuroimage.2013.02.055.

[252] Adrian R. Groves, Stephen M. Smith, Anders M. Fjell, Christian K. Tamnes, Kristine B. Walhovd, Gwenaëlle Douaud, Mark W. Woolrich, and Lars T. Westlye. Benefits of multi-modal fusion analysis on a large-scale dataset: Life-span patterns of inter-subject variability in cortical morphometry and white matter microstructure. *NeuroImage*, 63(1):365–380, 2012. doi: 10.1016/j.neuroimage.2012.06.038.

[253] Andrei Irimia, Carinna Torgerson, Matthew Goh, and John Van Horn. Statistical estimation of physiological brain age as a descriptor of senescence rate during adulthood. *Brain imaging and behavior*, 9, 11 2014. doi: 10.1007/s11682-014-9321-0.

[254] Melis Anaturk, Tobias Kaufmann, James Cole, Sana Suri, Ludovica Griffanti, Enikő Zsoldos, Nicola Filippini, Archana Singh-Manoux, Mika Kivimäki, Lars Westlye, Klaus Ebmeier, and Ann-Marie G de Lange. Prediction of brain age and cognitive age: quantifying brain and cognitive maintenance in aging. *Human Brain Mapping*, 06 2020. doi: 10.31234/osf.io/gwqnt.

[255] James H. Cole. Multimodality neuroimaging brain-age in UK Biobank: relationship to biomedical, lifestyle, and cognitive factors. *Neurobiology of Aging*, 92: 34–42, 2020. doi: 10.1016/j.neurobiolaging.2020.03.014.

[256] James H. Cole, Katja Franke, and Nicolas Cherbuin. *Quantification of the Biological Age of the Brain Using Neuroimaging*, pages 293–328. 2019. doi: 10.1007/978-3-030-24970-0_19.

[257] 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. doi: 10.1016/j.neuroimage.2012.08.001.

[258] Benedikt Jonsson, Gyda Bjornsdottir, Thorgeir Thorgeirsson, Lotta Ellingsen, Gudmundur Walters, Daniel Guobjartsson, Hreinn Stefansson, K. Stefansson, and Magnus Ulfarsson. Brain age prediction using deep learning uncovers associated sequence variants. *Nature Communications*, 10, 11 2019. doi: 10.1038/s41467-019-13163-9.

[259] Jin Hong, Zhangzhi Feng, Shui-Hua Wang, Andrew Peet, Yu-Dong Zhang, Yu Sun, and Ming Yang. Brain age prediction of children using routine brain [mr.

[260] Huiting Jiang, Na Lu, Kewei Chen, Li Yao, Ke Li, Jiacai Zhang, and Xiaojuan Guo. Predicting brain age of healthy adults based on structural MRI parcellation using convolutional neural networks. *Frontiers in Neurology*, 10:1346, 2020. doi: 10.3389/fneur.2019.01346.

[261] Guangxiang Rao, Ang Li, Yong Liu, and Bing Liu. A high-powered brain age prediction model based on convolutional neural network. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1915–1919, 2020. doi: 10.1109/ISBI45749.2020.9098376.

[262] Ana Namburete, Weidi Xie, and Julia Noble. Robust regression of brain maturation from 3D fetal neurosonography using CRNs. pages 73–80, 09 2017.

[263] Deepthi P. Varikuti, Sarah Genon, Aristeidis Sotiras, Holger Schwender, Felix Hoffstaedter, Kaustubh R. Patil, Christiane Jockwitz, Svenja Caspers, Susanne Moebus, Katrin Amunts, Christos Davatzikos, and Simon B. Eickhoff. Evaluation of non-negative matrix factorization of grey matter in age prediction. *NeuroImage*, 173:394 – 410, 2018.

[264] K. Franke, M. Ristow, and C. Gaser. Gender-specific effects of health and lifestyle markers on individual BrainAGE. In *2013 International Workshop on Pattern Recognition in Neuroimaging*, pages 94–97, 2013.

[265] Osamu Abe, Hidenori Yamasue, Haruyasu Yamada, Yoshitaka Masutani, Hiroyuki Kabasawa, Hiroki Sasaki, Kunio Takei, Motomu Suga, Kiyoto Kasai, Shigeki Aoki, and Kuni Ohtomo. Sex dimorphism in gray/white matter volume and diffusion tensor during normal aging. *NMR in biomedicine*, 23:446–58, 06 2010.

[266] Stephen Smith, Diego Vidaurre, Fidel Alfaro-Almagro, Thomas Nichols, and Karla Miller. Estimation of brain age delta from brain imaging. 02 2019.

[267] Stephen Smith, Lloyd Elliott, Fidel Alfaro-Almagro, Paul McCarthy, Thomas Nichols, Gwenaëlle Douaud, and Karla Miller. Brain aging comprises many modes of structural and functional change with distinct genetic and biophysical associations. *eLife*, 9, 03 2020. doi: 10.7554/eLife.52677.

[268] Karla Miller, Fidel Alfaro-Almagro, Neal Bangerter, David Thomas, Essa Yacoub, Junqian Xu, Andreas Bartsch, Saad Jbabdi, Stamatios Sotiropoulos, Jesper Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Thomas Okell, Peter Weale, Iulius Dragonu, Steve Garratt, Sarah Hudson, Rory Collins, Mark Jenkinson, and Stephen Smith. Multimodal population brain imaging in the UK biobank prospective epidemiological study. *Nature neuroscience*, 19, 09 2016.

[269] Fidel Alfaro-Almagro, Mark Jenkinson, Neal Bangerter, Jesper Andersson, Ludovica Griffanti, Gwenaelle Douaud, Stamatios Sotiropoulos, Saad Jbabdi, Moises Hernandez Fernandez, Emmanuel Vallee, Diego Vidaurre, Matthew Webster, Paul McCarthy, Christopher Rorden, Alessandro Daducci, Daniel Alexander, Hui Zhang, Iulius Dragonu, Paul Matthews, and Stephen Smith. Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage*, 166:130385, 04 2017. doi: 10.1016/j.neuroimage.2017.10.034.

[270] Jessica Bernard and Rachael Seidler. Moving forward: Age effects on the cerebellum underlie cognitive and motor declines. *Neuroscience and Biobehavioral Reviews*, 02 2014.

[271] François Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015.

[272] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, and et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `http://tensorflow.org/`. Software available from tensorflow.org.

[273] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67:301 – 320, 04 2005.

[274] Usama Pervaiz, Diego Vidaurre, Mark Woolrich, and Stephen Smith. Optimising network modelling methods for fMRI. *NeuroImage*, 211:116604, 02 2020. doi: 10.1016/j.neuroimage.2020.116604.

[275] Heath Pardoe and Ruben Kuzniecky. NAPR: a cloud-based framework for neuroanatomical age prediction. *Neuroinformatics*, 16, 10 2017.

[276] Feng Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Yen Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6458, 2017.

[277] Kegang Hua, Jiangyang Zhang, Setsu Wakana, Hangyi Jiang, Xin Li, Daniel Reich, Peter Calabresi, James Pekar, Peter van zijl, and Susumu Mori. Tract probability maps in stereotaxic spaces: Analyses of white matter anatomy and tract-specific quantification. *NeuroImage*, 39:336–47, 02 2008.

[278] Ellen O'Flaherty. Modeling normal aging bone loss, with consideration of bone loss in osteoporosis. *Toxicological sciences : an official journal of the Society of Toxicology*, 55:171–88, 06 2000. doi: 10.1093/toxsci/55.1.171.

[279] Christopher Gallagher and Sri Harsha Tella. Prevention and treatment of postmenopausal osteoporosis. *The Journal of steroid biochemistry and molecular biology*, pages 155–170, July 2014.

[280] Franziskus Liem, Gael Varoquaux, Jana Kynast, Frauke Beyer, Shahrzad Kharabian Masouleh, Julia Huntenburg, Leonie Lampe, Mehdi Rahim, Alexandre Abraham, Cameron Craddock, Steffi Riedel-Heller, Tobias Luck, Markus Loeffler, Matthias Schroeter, Veronica Witte, and Daniel Margulies. Predicting brain-age from multimodal imaging data captures cognitive impairment. *NeuroImage*, 148, 11 2016.

[281] Stéphanie Debette and H S Markus. The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. *BMJ*, 341, 2010. doi: 10.1136/bmj.c3666. URL `https://www.bmj.com/content/341/bmj.c3666`.

[282] John Van Horn and Arthur Toga. Multi-site neuroimaging trials. *Current opinion in neurology*, 22:370–8, 07 2009. doi: 10.1097/WCO.0b013e32832d92de.

[283] Stephen Smith and Thomas Nichols. Statistical challenges in "big data" human neuroimaging. *Neuron*, 97:263–268, 01 2018. doi: 10.1016/j.neuron.2017.12.018.

[284] Colin Hawco, Joseph Viviano, Sofia Chavez, Erin Dickie, Navona Calarco, Peter Kochunov, Miklos Argyelan, Jessica Turner, Anil Malhotra, Robert Buchanan, and Aristotle Voineskos. A longitudinal human phantom reliability study of multi-center T1-weighted, DTI, and resting state fMRI data. *Psychiatry Research: Neuroimaging*, 282, 06 2018. doi: 10.1016/j.pscychresns.2018.06.004.

[285] Elisabetta Pagani, Jochen G. Hirsch, Petra J.W. Pouwels, Mark A. Horsfield, Elisabetta Perego, Achim Gass, Stefan D. Roosendaal, Frederik Barkhof, Federica Agosta, Marco Rovaris, Domenico Caputo, Antonio Giorgio, Jacqueline Palace, Silvia Marino, Nicola De Stefano, Stefan Ropele, Franz Fazekas, and Massimo Filippi. Intercenter differences in diffusion tensor MRI acquisition. *Journal of Magnetic Resonance Imaging*, 31(6):1458–1468, 2010. doi: https://doi.org/10.1002/jmri.22186.

[286] R.T. Shinohara, J Oh, Govind Nair, P.A. Calabresi, C Davatzikos, Jimit Doshi, R.G. Henry, G Kim, Kristin Linn, N Papinutto, D Pelletier, D.L. Pham, Daniel Reich, W Rooney, Snehashis Roy, William Stern, Subhash Tummala, F Yousuf, and A Zhu. Volumetric analysis from a harmonized multisite brain MRI study of a single subject with multiple sclerosis. *AJNR. American journal of neuroradiology*, 38, 06 2017. doi: 10.3174/ajnr.A5254.

[287] Tong Zhu, Rui Hu, Xing Qiu, Michael Taylor, Yuen So, Constantin Yiannoutsos, Bradford Navia, Susumu Mori, Sven Ekholm, Giovanni Schifitto, and Jianhui Zhong. Quantification of accuracy and precision of multi-center DTI measurements: A diffusion phantom and human brain study. *NeuroImage*, 56:1398–411, 02 2011. doi: 10.1016/j.neuroimage.2011.02.010.

[288] Jiayu Chen, Jingyu Liu, Vince Calhoun, Alejandro Arias-Vásquez, Marcel Zwiers, Cota Gupta, Barbara Franke, and Jessica Turner. Exploration of scanning effects in multi-site structural MRI studies. *Journal of neuroscience methods*, 230, 04 2014. doi: 10.1016/j.jneumeth.2014.04.023.

[289] Vincent Magnotta, Joy Matsui, Dawei Liu, Hans Johnson, Jeffrey Long, Bradley Bolster Jr, Bryon Mueller, Susumu Mori, Karl Helmer, Jessica Turner, Mark

Lowe, Elizabeth Aylward, Greg Bonett, and Jane Paulsen. Multicenter reliability of diffusion tensor imaging. *Brain connectivity*, 2, 10 2012. doi: 10.1089/brain.2012.0112.

[290] Hidemasa Takao, Naoto Hayashi, and Kuni Ohtomo. Effects of the use of multiple scanners and of scanner upgrade in longitudinal voxel-based morphometry studies. *Journal of magnetic resonance imaging : JMRI*, 38, 11 2013. doi: 10.1002/jmri.24038.

[291] Christian Vollmar, Jonathan O'Muircheartaigh, Gareth Barker, Mark Symms, Pamela Thompson, Veena Kumari, John Duncan, Mark Richardson, and Matthias Koepp. Identical, but not the same: Intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0 T scanners. *NeuroImage*, 51:1384–94, 03 2010. doi: 10.1016/j.neuroimage.2010.03.046.

[292] Nico Dario Papinutto, Francesca Maule, and Jorge Jovicich. Reproducibility and biases in high field brain diffusion MRI: An evaluation of acquisition and analysis variables. *Magnetic Resonance Imaging*, 31(6):827–839, 2013. doi: https://doi.org/10.1016/j.mri.2013.03.004.

[293] Antonio Torralba and Alexei Efros. Unbiased look at dataset bias. pages 1521 – 1528, 07 2011. doi: 10.1109/CVPR.2011.5995347.

[294] Ben Glocker, Rob Robinson, Daniel Coelho de Castro, Qi Dou, and Ender Konukoglu. Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. 10 2019.

[295] Christian Wachinger, Anna Rieckmann, and Sebastian Polsterl. Detect and correct bias in multi-site neuroimaging datasets. *Medical Image Analysis*, 02 2021. doi: 10.1016/j.media.2020.101879.

[296] William Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8: 118–27, 01 2007.

[297] Jean-Philippe Fortin, Nicholas Cullen, Yvette Sheline, Warren Taylor, Irem Aselcioglu, Philip Cook, Phil Adams, Crystal Cooper, Maurizio Fava, Patrick McGrath, Melvin Mcinnis, Mary Phillips, Madhukar Trivedi, Myrna Weissman, and Russell Shinohara. Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, 167, 11 2017. doi: 10.1016/j.neuroimage.2017.11.024.

[298] Raymond Pomponio, Guray Erus, Mohamad Habes, Jimit Doshi, Dhivya Srinivasan, Elizabeth Mamourian, Vishnu Bashyam, Ilya Nasrallah, Theodore Satterthwaite, Yong Fan, Lenore Launer, Colin Masters, Paul Maruff, Chuanjun

Zhuo, Henry Valzke, Sterling Johnson, Jurgen Fripp, Nikolaos Koutsouleris, Daniel Wolf, and Christos Davatzikos. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, 208: 116450, 12 2019. doi: 10.1016/j.neuroimage.2019.116450.

[299] Jean-Philippe Fortin, Drew Parker, Birkan Tunc, Takanori Watanabe, Mark Elliott, Kosha Ruparel, David Roalf, Theodore Satterthwaite, Ruben Gur, Raquel Gur, Robert Schultz, Ragini Verma, and Russell Shinohara. Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, 161, 08 2017. doi: 10.1016/j.neuroimage.2017.08.047.

[300] Joaquim Radua, Eduard Vieta, Russell Shinohara, Peter Kochunov, Yann Quidé, Melissa J. Green, Cynthia S. Weickert, Thomas Weickert, Jason Bruggemann, Tilo Kircher, Igor Nenadić, Murray J. Cairns, Marc Seal, Ulrich Schall, Frans Henskens, Janice M. Fullerton, and et al. Increased power by harmonizing structural mri site differences with the combat batch adjustment method in enigma. *NeuroImage*, 218:116956, 2020. doi: https://doi.org/10.1016/j.neuroimage.2020.116956.

[301] Andrew A. Chen, Joanne C. Beer, Nicholas J. Tustison, Philip A. Cook, Russell T. Shinohara, and Haochang Shou. Removal of scanner effects in covariance improves multivariate pattern analysis in neuroimaging data. *bioRxiv*, 2020. doi: 10.1101/858415.

[302] Joanne C. Beer, Nicholas J. Tustison, Philip A. Cook, Christos Davatzikos, Yvette I. Sheline, Russell T. Shinohara, and Kristin A. Linn. Longitudinal combat: A method for harmonizing longitudinal multi-scanner imaging data. *bioRxiv*, 2020. doi: 10.1101/868810.

[303] Suheyla Cetin Karayumak, Katharina Stegmayer, Sebastian Walther, Philip Szeszko, Tim Crow, Anthony James, Matcheri Keshavan, Marek Kubicki, and Yogesh Rathi. Exploring the limits of ComBat method for multi-site diffusion MRI harmonization. *ArXiv*, 11 2020. doi: 10.1101/2020.11.20.390120.

[304] Rafael Garcia-Dias, Cristina Scarpazza, Lea Baecker, Sandra Vieira, Walter Pinaya, Aiden Corvin, Alberto Redolfi, Barnaby Nelson, Benedicto Crespo-Facorro, Colm Mcdonald, Diana Tordesillas-Gutiérrez, Dara Cannon, David Mothersill, Dennis Hernaus, Daisy Morris, Esther Setién-Suero, Gary Donohoe, Giovanni Frisoni, Giulia Tronchin, and Andrea Mechelli. Neuroharmony: A new tool for harmonizing volumetric MRI data from unseen scanners. *NeuroImage*, page 117127, 07 2020. doi: 10.1016/j.neuroimage.2020.117127.

[305] Julia Wrobel, Melissa Martin, Russell Shinohara, and Jeff Goldsmith. Intensity warping for multisite MRI harmonization. *Neuroimage*, 223:117242, 12 2020. doi: 10.1016/j.neuroimage.2020.117242.

[306] Suheyla Cetin Karayumak, Marek Kubicki, and Yogesh Rathi. Harmonizing diffusion MRI data across magnetic field strengths. *Medical Image Computing and Computer Assisted Intervention - Conference Proceedings*, pages 116–124, 09 2018.

[307] Stefano Blumberg, Marco Palombo, Can Khoo, Chantal Tax, Ryutaro Tanno, and Daniel Alexander. Multi-stage prediction networks for data harmonization. *Medical Image Computing and Computer Assisted Intervention - Conference Proceedings*, pages 411–419, 10 2019.

[308] Khoi Minh Huynh, Geng Chen, Ye Wu, Dinggang Shen, and Pew-Thian Yap. Multi-site harmonization of diffusion MRI data via method of moments. *IEEE transactions on medical imaging*, 38(7):1599 – 1609, July 2019. ISSN 0278-0062. doi: 10.1109/TMI.2019.2895020.

[309] Hengameh Mirzaalian, Amicie Pierrefeu, Peter Savadjiev, Ofer Pasternak, Sylvain Bouix, Marek Kubicki, Carl-Fredrik Westin, Martha Shenton, and Yogesh Rathi. Harmonizing diffusion MRI data across multiple sites and scanners. *Medical Image Computing and Computer Assisted Intervention - Conference Proceedings*, 9349:12–19, 10 2015.

[310] H. Mirzaalian, Lipeng Ning, P. Savadjiev, Ofer Pasternak, Sylvain Bouix, O. Michailovich, M. Kubicki, Carl-Fredrik Westin, Martha Shenton, and Yogesh Rathi. Inter-site and inter-scanner diffusion MRI data harmonization. *NeuroImage*, 135, 04 2016. doi: 10.1016/j.neuroimage.2016.04.041.

[311] Hengameh Mirzaalian, Lipeng Ning, Peter Savadjiev, Ofer Pasternak, Sylvain Bouix, Oleg Michailovich, Sarina Karmacharya, Gerald Grant, Christine Marx, Rajendra Morey, Laura Flashman, Mark George, Thomas Mcallister, Norberto Andaluz, Lori Shutter, Raul Coimbra, Ross Zafonte, Mike Coleman, Marek Kubicki, and Yogesh Rathi. Multi-site harmonization of diffusion MRI data in a registration framework. *Brain Imaging and Behavior*, 12:284–295, 02 2018. doi: 10.1007/s11682-016-9670-y.

[312] Samuel St-Jean, Max Viergever, and Alexander Leemans. Harmonization of diffusion MRI datasets with adaptive dictionary learning. *Human Brain Mapping*, 41(16):4478–4499, 10 2020. doi: 10.1002/hbm.25117.

[313] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017.

[314] Fenqiang Zhao, Zhengwang Wu, Li Wang, Weili Lin, Shunren Xia, and Gang Li. Harmonization of infant cortical thickness using surface-to-surface cycle-

consistent adversarial networks. *Medical Image Computing and Computer Assisted Intervention - Conference Proceedings*, pages 475–483, 10 2019.

[315] Blake Dewey, Lianrui Zuo, Aaron Carass, Yufan He, Yihao Liu, Ellen Mowry, Scott Newsome, Jiwon Oh, Peter Calabresi, and Jerry Prince. A disentangled latent space for cross-site MRI harmonization. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 720–729, 09 2020. doi: 10.1007/978-3-030-59728-3_70.

[316] Rob Robinson, Q. Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, M. Groot, R. Summers, D. Rueckert, and Ben Glocker. Image-level harmonization of multi-site data using image-and-spatial transformer networks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 710–719, 06 2020. doi: 10.1007/978-3-030-59728-3_69.

[317] Christian Wachinger and Martin Reuter. Domain adaptation for Alzheimer's disease diagnostics. *NeuroImage*, 139, 06 2016. doi: 10.1016/j.neuroimage.2016.05.053.

[318] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. *27th International Conference on Artificial Neural Networks*, 2018.

[319] Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. Beyond sharing weights for deep domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP, 03 2016. doi: 10.1109/TPAMI.2018.2814042.

[320] Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. *European Conference on Computer Vision*, 07 2016.

[321] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *International Conference on Machine Learning*, 09 2014.

[322] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. *Workshop on Bias Estimation in Face Analytics, ECCV*, 09 2018.

[323] Katja Franke and Christian Gaser. Ten years of BrainAGE as a neuroimaging biomarker of brain aging: What insights have we gained? *Frontiers in Neurology*, 10, 08 2019. doi: 10.3389/fneur.2019.00789.

[324] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and

Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems 32 (NIPS 2019)*, 2019.

[325] Nicola Filippini, Enikő Zsoldos, Rita Haapakoski, Claire Sexton, Abda Mahmood, Charlotte Allan, Anya Topiwala, Vyara Valkanova, Eric Brunner, Martin Shipley, Edward Auerbach, Steen Moeller, Kâmil Urbil, Junqian Xu, Essa Yacoub, Jesper Andersson, Janine Bijsterbosch, Stuart Clare, Ludovica Griffanti, and Klaus Ebmeier. Study protocol: The Whitehall II imaging sub-study. *BMC psychiatry*, 14:159, 05 2014. doi: 10.1186/1471-244X-14-159.

[326] Laurens van der Maaten and Geoffrey Hinton. Viualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.

[327] José Lezama. Overcoming the disentanglement vs reconstruction trade-off via Jacobian supervision. *International Conference on Learning Representations*, 2019.

[328] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein GANs. 2017.

[329] Daniel Moyer and Polina Golland. Harmonization and the worst scanner syndrome. *ArXiv*, 01 2021.

[330] Robin Flores, Renaud La Joie, and Gaël Chételat. Structural imaging of hippocampal subfields in healthy aging and Alzheimer's disease. *Neuroscience*, 309, 08 2015. doi: 10.1016/j.neuroscience.2015.08.033.

[331] G. Balakrishnan, Amy Zhao, M. Sabuncu, J. Guttag, and Adrian V. Dalca. VoxelMorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38:1788–1800, 2019.

[332] X. Li, H. Chen, X. Qi, Q. Dou, C. W. Fu, and P. A. Heng. H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Transactions on Medical Imaging*, 37(12):2663–2674, 2018. doi: 10.1109/TMI.2018.2845918.

[333] Yilong Chen, Kai Wang, Xiangyun Liao, Yinling Qian, Qiong Wang, Zhiyong Yuan, and Pheng-Ann Heng. Channel-Unet: A spatial channel-wise convolutional neural network for liver and tumors segmentation. *Frontiers in Genetics*, 10: 1110, 2019. doi: 10.3389/fgene.2019.01110.

[334] Grzegorz Chlebus, Andrea Schenk, Jan Moltz, Bram Ginneken, Hans Meine, and Horst Hahn. Automatic liver tumor segmentation in CT with fully convolutional neural networks and object-based postprocessing. *Scientific Reports*, 8, 10 2018. doi: 10.1038/s41598-018-33860-7.

[335] M. Yahyatabar, P. Jouvet, and F. Cheriet. Dense-Unet: a light model for lung fields segmentation in chest x-ray images. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 1242–1245, 2020. doi: 10.1109/EMBC44109.2020.9176033.

[336] Madeleine K. Wyburd, Mark Jenkinson, and Ana I. L. Namburete. Cortical plate segmentation using CNNs in 3D fetal ultrasound. In *Medical Image Understanding and Analysis*, pages 56–68, 2020. doi: 10.1007/978-3-030-52791-4.5.

[337] Davis M. Vigneault, Weidi Xie, Carolyn Y. Ho, David A. Bluemke, and J. Alison Noble. Omega-Net: Fully automatic, multi-view cardiac MR detection, orientation, and segmentation with deep neural networks. *Medical Image Analysis*, 48:95–106, 2018. doi: 10.1016/j.media.2018.05.008.

[338] Babak Hassibi and David Stork. Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in Neural Information Processing Systems*, volume 5, 1993. doi: 10.5555/645753.668069.

[339] Xin Dong, Shangyu Chen, and Sinno Jialin Pan. Learning to prune deep neural networks via layer-wise optimal brain surgeon. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 4860–4874, 2017. doi: 10.5555/3295222.3295239.

[340] Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, page 1135–1143, 2015. doi: 10.5555/2969239.2969366.

[341] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 11 2019.

[342] Hengyuan Hu, Rui Peng, Yu-Wing Tai, and Chi-Keung Tang. Network trimming: A data-driven neuron pruning approach towards efficient deep architectures. *ArXiv*, 07 2016.

[343] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and H.P. Graf. Pruning filters for efficient convnets. In *International Conference on Learning Representations*, 08 2017.

[344] Yiwen Guo, Anbang Yao, and Yurong Chen. Dynamic network surgery for efficient DNNs. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 1387–1395, 2016. doi: 10.5555/3157096.3157251.

[345] Suraj Srinivas and R. Venkatesh Babu. Data-free parameter pruning for Deep Neural Networks. *Proceedings of BMVC*, July 2015. doi: dx.doi.org/10.5244/C.29.31.

[346] Y. He, X. Zhang, and J. Sun. Channel pruning for accelerating very deep neural networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1398–1406, 2017. doi: 10.1109/ICCV.2017.155.

[347] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *International Conference on Learning Representations*, 10 2019.

[348] Guillaume Bellec, David Kappel, Wolfgang Maass, and Robert Legenstein. Deep rewiring: Training very sparse deep networks. 11 2017.

[349] Shibo Shen, Rongpeng Li, Zhifeng Zhao, Honggang Zhang, and Yugeng Zhou. Learning to prune in training via dynamic channel propagation. *ArXiv*, 2020. doi: arXiv:2007.01486.

[350] Joseph Cohen, Henry Lo, and Wei Ding. Randomout: Using a convolutional gradient norm to win the filter lottery. In *ICLR Workshop*, 02 2016.

[351] Karttikeya Mangalam and Mathieu Salzamann. On compressing Unet using knowledge distillation. *ArXiv*, 12 2018. doi: arXiv:1812.00249v1.

[352] Franck Mamalet and Christophe Garcia. Simplifying convnets for fast learning. In *Proc. of Int. Conf on Artificial Neural Networks (ICANN'12)*, 09 2012. doi: 10.1007/978-3-642-33266-18.

[353] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2017. doi: 10.1109/CVPR.2017.195.

[354] Andrew Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv*, 04 2017. doi: arXiv:1704.04861.

[355] Haifeng Wang, Qianqian Zhang, Hongya Lu, Daehan Won, and Sang Won Yoon. 3d medical image classification with depthwise separable networks. *Procedia Manufacturing*, 39:349–356, 2019. doi: https://doi.org/10.1016/j.promfg.2020.01.369.

[356] Zia Khan, Norashikin Yahya, Khaled Alsaih, and Fabrice Meriaudeau. Segmentation of prostate in MRI images using depth separable convolution operations. In *Intelligent Human Computer Interaction*, pages 132–141, 2021. doi: 10.1007/978-3-030-68449-514.

[357] Zhuang Li, Lihong Xu, and Shuwei Zhu. Pruning of network filters for small dataset. *IEEE Access*, PP:1–1, 12 2019. doi: 10.1109/ACCESS.2019.2963080.

[358] Aidan Gomez, Ivan Zhang, Kevin Swersky, Yarin Gal, and Geoffery Hinton. Targeted dropout. *Advances in Neural Information Processing Systems 32 (NIPS 208)*, 2018.

[359] Saihui Hou and Zilei Wang. Weighted channel dropout for regularization of deep convolutional neural network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:8425–8432, 07 2019. doi: 10.1609/aaai.v33i01.33018425.

[360] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256. JMLR Workshop and Conference Proceedings, 13–15 May 2010. URL `http://proceedings.mlr.press/v9/glorot10a.html`.

[361] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 648–656, 2015. doi: 10.1109/CVPR.2015.7298664.

[362] C. Tobon-Gomez, A. J. Geers, J. Peters, J. Weese, K. Pinto, R. Karim, M. Ammar, A. Daoudi, J. Margeta, Z. Sandoval, B. Stender, Y. Zheng, M. A. Zuluaga, J. Betancur, N. Ayache, M. A. Chikh, J. L. Dillenseger, B. M. Kelm, S. Mahmoudi, S. Ourselin, A. Schlaefer, T. Schaeffter, R. Razavi, and K. S. Rhode. Benchmark for algorithms segmenting the left atrium from 3D CT and MRI datasets. *IEEE Transactions on Medical Imaging*, 34(7):1460–1473, 2015. doi: 10.1109/TMI.2015.2398818.

[363] Amber Simpson, Julie Leal, Amudhan Pugalenthi, Peter Allen, Ronald DeMatteo, Y. Fong, Mithat Gönen, William Jarnagin, T Kingham, Michael Miga, Jinru Shia, Martin Weiser, and Michael D'Angelica. Chemotherapy-induced splenic volume increase is independently associated with major complications after hepatic resection for metastatic colorectal cancer. *Journal of the American College of Surgeons*, 220, 12 2014. doi: 10.1016/j.jamcollsurg.2014.12.008.

[364] Geert Litjens, Oscar Debats, Wendy van de Ven, Nico Karssemeijer, and Henkjan Huisman. A pattern recognition approach to zonal segmentation of the prostate on MRI. volume 15, pages 413–20, 10 2012. ISBN 978-3-642-33417-7. doi: 10.1007/978-3-642-33418-4_51.

[365] A.T. Papageorghiou, Eric Ohuma, Douglas Altman, Tullia Todros, Leila Ismail, Ann Lambert, Yasmin Jaffer, Enrico Bertino, Michael Gravett, Manorama Purwar, Julia Noble, Ruyan Pang, Cesar Victora, Fernando Colugnati, Maria

Carvalho, Laurent Salomon, Zulfiqar Bhutta, Stephen Kennedy, and José Villar. International standards for fetal growth based on serial ultrasound measurements: The fetal growth longitudinal study of the INTERGROWTH-21st project. *The Lancet*, 384:869–879, 09 2014. doi: 10.1016/S0140-6736(14)61490-2.

[366] Ali Gholipour, Caitlin Rollins, Clemente Velasco-Annis, Abdelhakim Ouaalam, Alireza Akhondi-Asl, Onur Afacan, Cynthia Ortinau, Sean Clancy, Catherine Limperopoulos, Edward Yang, Judy Estroff, and Simon Warfield. A normative spatiotemporal MRI atlas of the fetal brain for automatic segmentation and analysis of early brain growth. *Scientific Reports*, 7, 12 2017. doi: 10.1038/s41598-017-00525-w.

[367] Wenjun Yan, Yuanyuan Wang, Shengjia Gu, Lu Huang, Fuhua Yan, Liming Xia, and Qian Tao. The domain shift problem of medical image segmentation and vendor-adaptation by Unet-GAN. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 623–631, 2019. ISBN 978-3-030-32245-8. doi: 10.1007/978-3-030-32245-8.69.

[368] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *ArXiv*, abs/1207.0580, 2012. doi: http://arxiv.org/abs/1207.0580.

[369] Connor Shorten and Khoshgoftaar Taghi M. A survey on image data augmentation for deep learning. *Journal of Big Data volume*, 6, 2019. doi: 10.1186/s40537-019-0197-0.

[370] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

[371] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. doi: 10.2307/1932409.

[372] Fernando Pérez-García, Rachel Sparks, and Sebastien Ourselin. TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *arXiv:2003.04696 [cs, eess, stat]*, March 2020. URL `http://arxiv.org/abs/2003.04696`. arXiv: 2003.04696.

[373] T. Huang, H. Chen, R. Fujimoto, K. Ito, K. Wu, K. Sato, Y. Taki, H. Fukuda, and T. Aoki. Age estimation from brain MRI images using deep learning. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 849–852, 2017.

[374] Lianrui Zuo, Blake E. Dewey, Aaron Carass, Yihao Liu, Yufan He, Peter A. Calabresi, and Jerry L. Prince. Information-based disentangled representation

learning for unsupervised MR harmonization. In *Information Processing in Medical Imaging*, pages 346–359, 2021. doi: 10.1007/978-3-030-78191-0_27.

[375] Mengting Liu, Piyush Maiti, Sophia Thomopoulos, Alyssa Zhu, Yaqiong Chai, Hosung Kim, and Neda Jahanshad. Style transfer using generative adversarial networks for multi-site MRI harmonization. *BioArXiv*, 03 2021. doi: 10.1101/2021.03.17.435892.

[376] Amber Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Ginneken, Annette Kopp-Schneider, Bennett Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald Summers, Patrick Bilic, Patrick Christ, Richard Do, Marc Gollub, Jennifer Golia-Pernicka, Stephan Heckers, William Jarnagin, and Manuel Jorge Cardoso. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *ArXiv*, 02 2019. doi: arXiv:1902.09063v1.

[377] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. doi: 10.1109/CVPR.2015.7298594.

[378] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[379] Jihoon Oh, Baek-Lok Oh, Kyong-Uk Lee, Jeong-Ho Chae, and Kyongsik Yun. Identifying schizophrenia using structural MRI with a deep learning algorithm. *Frontiers in Psychiatry*, 11, 02 2020. doi: 10.3389/fpsyt.2020.00016.

[380] Pradyumna Lanka, Rangaprakash D., Michael Dretsch, Jeffrey Katz, Thomas Denney, and Gopikrishna Deshpande. Supervised machine learning for diagnostic classification from large-scale neuroimaging datasets. *Brain Imaging and Behavior*, 14, 12 2020. doi: 10.1007/s11682-019-00191-8.

[381] CK Fisher, AM Smith, JR Walsh, AJ Simone, C Edgar, Clifford Jack, D Holtzman, D Russell, Derek Hill, D Grosset, F Wood, H Vanderstichele, J Morris, K Blennown, K Marek, LM Shaw, Mercy Albert, M Weiner, N Fox, and W Kubick. Machine learning for comprehensive forecasting of Alzheimer's disease progression. *Scientific Reports*, 9, 01 2019.

[382] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, jun 2015. doi: 10.1109/CVPR.2015.7298965.

[383] Salome Kazeminia, Christoph Baur, Arjan Kuijper, Bram van Ginneken, Nassir Navab, Shadi Albarqouni, and Anirban Mukhopadhyay. GANs for medical image analysis. *Artificial Intelligence in Medicine*, 109:101938, 2020. doi: 10.1016/j.artmed.2020.101938.

[384] Ender Konukoglu, Ben Glocker, Darko Zikic, and Antonio Criminisi. Neighbourhood approximation using randomized forests. *Medical Image Analysis*, 17(7): 790–804, 2013. doi: 10.1016/j.media.2013.04.013.

[385] 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*, 25(6): 1676–1684, 01 2014. doi: 10.1093/cercor/bht425.

[386] Obada Al Zoubi, Chung Ki Wong, Rayus T. Kuplicki, Hung-wen Yeh, Ahmad Mayeli, Hazem Refai, Martin Paulus, and Jerzy Bodurka. Predicting age from brain EEG signals—a machine learning approach. *Frontiers in Aging Neuroscience*, 10:184, 2018. doi: 10.3389/fnagi.2018.00184.

[387] Johnny Wang, Maria J. Knol, Aleksei Tiulpin, Florian Dubost, Marleen de Bruijne, Meike W. Vernooij, Hieab H. H. Adams, M. Arfan Ikram, Wiro J. Niessen, and Gennady V. Roshchupkin. Gray matter age prediction as a biomarker for risk of dementia. *Proceedings of the National Academy of Sciences*, 116(42): 21213–21218, 2019. doi: 10.1073/pnas.1902376116.

[388] Lea Baecker, Jessica Dafflon, Pedro Ferreira da Costa, Rafael Garcia-Dias, Sandra Vieira, Cristina Scarpazza, Vince Calhoun, João Sato, Andrea Mechelli, and Walter Pinaya. Brain age prediction: A comparison between machine learning models using region- and voxel-based morphometric data. *Human Brain Mapping*, 42, 03 2021. doi: 10.1002/hbm.25368.

[389] Loredana Bellantuono, Luca Marzano, Marianna La Rocca, Dominique Duncan, Angela Lombardi, Tommaso Maggipinto, Alfonso Monaco, Sabina Tangaro, Nicola Amoroso, and Roberto Bellotti. Predicting brain age with complex networks: From adolescence to adulthood. *NeuroImage*, 225:117458, 2021. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2020.117458. URL `https://www.sciencedirect.com/science/article/pii/S1053811920309435`.

[390] Inpyeong Hwang, Eung Koo Yeon, Ji Ye Lee, Roh-Eul Yoo, Koung Mi Kang, Tae Jin Yun, Seung Hong Choi, Chul-Ho Sohn, Hyeonjin Kim, and Ji hoon Kim. Prediction of brain age from routine T2-weighted spin-echo brain magnetic resonance images with a deep convolutional neural network. *Neurobiology of Aging*, 105:78–85, 2021. doi: https://doi.org/10.1016/j.neurobiolaging.2021.04.015.

[391] Simon Koppers, Luke Bloy, Jeffrey Berman, Chantal Tax, J. Christopher Edgar, and Dorit Merhof. Spherical harmonic residual network for diffusion signal harmonization. In *MICCAI 2019: Computational Diffusion MRI*, pages 173–182, 08 2019. doi: 10.1007/978-3-030-05831-9_14.

[392] Maria Ines Meyer, Ezequiel de la Rosa, Koen Van Leemput, and Diana M. Sima. Relevance vector machines for harmonization of MRI brain volumes using image descriptors. In *OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging*, pages 77–85, 2019. doi: 10.1007/978-3-030-32695-1-9.

[393] Vishwesh Nath, Prasanna Parvathaneni, Colin B. Hansen, Allison E. Hainline, Camilo Bermudez, Samuel Remedios, Justin A. Blaber, Kurt G. Schilling, Ilwoo Lyu, Vaibhav Janve, Yurui Gao, Iwona Stepniewska, Baxter P. Rogers, Allen T. Newton, L. Taylor Davis, Jeff Luci, Adam W. Anderson, and Bennett A. Landman. Inter-scanner harmonization of high angular resolution DW-MRI using null space deep learning. In *Computational Diffusion MRI*, pages 193–201, 2019. doi: 10.1007/978-3-030-05831-9_16.

[394] Ayumu Yamashita, Noriaki Yahata, Takashi Itahashi, Giuseppe Lisi, Takashi Yamada, Naho Ichikawa, Masahiro Takamura, Yujiro Yoshihara, Akira Kunimatsu, Naohiro Okada, Hirotaka Yamagata, Koji Matsuo, Ryuichiro Hashimoto, Go Okada, Yuki Sakai, Jun Morimoto, Jin Narumoto, Yasuhiro Shimada, Kiyoto Kasai, Nobumasa Kato, Hidehiko Takahashi, Yasumasa Okamoto, Saori C. Tanaka, Mitsuo Kawato, Okito Yamashita, and Hiroshi Imamizu. Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. *PLOS Biology*, 17(4):1–34, 04 2019. doi: 10.1371/journal.pbio.3000042.

[395] Gourav Modanwal, Adithya Vellal, Mateusz Buda, and Maciej A. Mazurowski. MRI image harmonization using cycle-consistent generative adversarial network. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, pages 259 – 264. SPIE, 2020. doi: 10.1117/12.2551301.

[396] Jie Zhong, Ying Wang, Jie Li, Xuetong Xue, Simin Liu, Miaomiao Wang, Xinbo Gao, Quan Wang, Jian Yang, and Xianjun Li. Inter-site harmonization based on dual generative adversarial networks for diffusion tensor imaging: Application to neonatal white matter development. *BioMedical Engineering OnLine*, 19, 01 2020. doi: 10.1186/s12938-020-0748-9.

[397] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.