Detection of cerebral small vessel disease signs on brain MR images



Vaanathi Sundaresan

Wellcome Centre for Integrative Neuroimaging,
Oxford Centre for Functional MRI of the Brain,
Nuffield Department of Clinical Neurosciences
&
Somerville College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas Term 2019

This thesis is dedicated to
my family, mentors and friends

Coming together is a beginning;
keeping together is progress;
working together is success
- Henry Ford

Acknowledgements

First of all, I would like to express the deepest gratitude to my supervisor, Prof. Mark Jenkinson, for his steadfast support and continued guidance towards enriching my potential throughout the entire DPhil research and for being an inspiring mentor. Many thanks to Dr. Ludovica Griffanti, my co-supervisor, for being such a dear friend and for nurturing in me the skills, confidence and ability to become an independent researcher. This thesis could not have been completed without their timely and thorough reviews despite their busy schedule. I would like to thank my clinical advisor, Dr. Giovanna Zamboni, for her inputs on the clinical aspects of my research.

My endless gratitude goes to Vin Murria, Prem Suki Foundation, Somerville College, University of Oxford and the Oxford India Centre for Sustainable Developement (OICSD) for funding my DPhil research. I would especially like to thank the Principal of Somerville college, Jan Royall, the Development Director at Somerville college, Sara Kalim, the Research Director of OICSD, Radhika Khosla and all my OICSD friends for their kind and incredible support. I am grateful to the EPSRC and MRC for supporting the research work.

My special thanks to Prof. Peter Jezzard, Prof. Michael Chappell, Prof. Thomas E Nichols, Dr. Jacqueline Palace, Prof. Clare E Mackay and Prof. Alison Noble for their valuable inputs and support during my initial training and my DPhil project. I would like to acknowledge our collaborators from the TICH2 trial group at the University of Nottingham, UK and the University of Siena, Italy. I am grateful to the examiners, Dr. Ana Namburete and Dr. Grant Mair for reviewing this thesis, and providing constructive suggestions.

Thank you, my dear friends, Siddharth and Shivani for the wonderful Oxford experience and for always being there for me. Many thanks to Liana, Thanasis, Elisa and Felix for being so supportive. I would like to thank my friends and colleagues: Katie, Hanna, Nicola, Fidel, Matthew, Paul, Rick, Usama, Caitlin, Luca, Silvia, Sana, Clare, Johannes and many others for their progressive ideas and for making every day at work enjoyable.

I would like to thank the IT support team and the academic administration staff of Nuffield Department of Clinical Neurosciences, the college accommodation team and library staff for their timely assistance. I acknowledge the imaging facilities and all the participants for the data that enabled us to perform this research.

Last but not least, I would like to express my heartfelt love and gratitude to my parents and my brother, who have been my pillars of support. I am deeply indebted to my parents for ingraining in me the values of sincerity and perseverance. I thank the Almighty for blessing me with kind, inspiring and lovely people, and wonderful opportunities in life.

Statement of originality

I hereby certify that I am the sole author of this thesis and that neither any part of this thesis nor whole of this thesis has been submitted for a degree to any other University or institution.

I certify that the thesis is the result of own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the thesis. I have summarised my contribution and those of collaborators as follows:

1. For the work regarding modelling white matter hyperintensities (WMH) distribution based on Bayesian inference (chapter 3), I implemented the Bayesian inference method and simulated the data for evaluation purposes, with the assistance of my supervisors. On the simulated data, while I performed the evaluations of the proposed Bayesian method, the results for the Bayesian Spatial Generalised Linear Mixed Model (BSGLMM) method were provided by our collaborators, for the comparative analysis. The real data (OXVASC and UK Biobank) and WMH segmentation masks used in the method were provided by our clinical collaborators.

2. For improvement of WMH segmentation using LOCally Adaptive Threshold Estimation (LOCATE, chapter 4), I implemented the LOCATE algorithm. Our clinical collaborators provided some of the datasets (NDGEN, OXVASC, CADASIL and HC) and their manual segmentations, other than the publicly available MICCAI WMH segmentation challenge training data. I performed all the evaluations, comparison and analyses presented in the chapter.

3. For WMH segmentation using deep learning (chapter 5), I developed the idea of triplanar U-Net and implemented the model. I designed the experiments, evaluated the results and performed subsequent analyses on the datasets, using manual segmentations, both provided by our collaborators (as stated in 2).

4. For detection of cerebral microbleeds (CMBs) in large population (chapter 6), I implemented the population-level CMB candidate subject preselection pipeline and deep learning-based multiscale method for detection of CMBs on individual subjects. For automated rating of CMBs based on the MARS scale, I implemented the method with occasional inputs from my supervisors and clinical collaborators regarding the brain anatomy. As stated before, our collaborators provided the datasets (TICH2, UK Biobank), other than publicly available training data from the Hong Kong research group. I performed the evaluations and analyses specified in the chapter, with an immediate future work involving manual labelling of the preselected data by our collaborators.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyones copyright or violate any proprietary rights and that any ideas, techniques, quotations, or any other material published or otherwise, are fully acknowledged in accordance with the standard referencing practices.

Abstract

Introduction: Accurate and automated detection of brain disease signs could aid in studying their clinical associations in large populations. This research aims to detect on magnetic resonance images two signs of cerebral small vessel disease- white matter hyperintensities (WMHs) and cerebral microbleeds (CMBs)- and to analyse their spatial distribution and characteristics.

Methods: We modelled the distribution of WMHs at the population-level with respect to various parametric factors (e.g. age) using Bayesian inference. For improving the performance of FSL's WMH segmentation tool, BIANCA, we used the above population-level WMH distribution probabilities as an additional feature, tested different classifiers and proposed a locally adaptive thresholding method (LOCATE). Furthermore, we developed a triplanar U-Net ensemble model (TrUE-Net), based on deep learning (DL), spatially-weighted cross-entropy and Dice loss functions to improve WMH segmentation, and explored various domain adaptation techniques. For CMBs, we developed an end-to-end analysis: CMB candidate subject preselection pipeline using machine learning, CMB detection using DL and atlas-based automated CMB rating.

Results and conclusions: Our distribution modelling provided comparable error values, with lower computational load, with respect to existing methods on simulated data. On real data, periventricular WMH probabilities showed significant increases with age, while the probabilities for deep WMHs were higher in hypertensive subjects, in line with literature. Among the proposed improvements for BIANCA, LOCATE gave the highest performance increase on 7 datasets with different lesion characteristics. TrUE-Net performed better than BIANCA on 5 different datasets, especially on low lesion load subjects. The CMB preselection pipeline provided robust performance despite different training modalities. The CMB detection method provided performance comparable with existing methods and higher precision than the state-of-the-art method. Future directions for this work could include improving the DL-based CMB/WMH detection methods and studying the clinical impact of CMBs at the population-level.

Key words: small vessel disease, white matter hyperintensities, cerebral microbleeds, image analysis, machine learning, deep learning, segmentation, structural MRI

Publications based on this work

Journal articles

Sundaresan, V., et al. CMB candidate subject preselection in large datasets. (in preparation)

Sundaresan, V., et al. 3D multiscale patch-based CNN for the CMB detection in big dataset. (in preparation)

Sundaresan, V., et al. Triplanar U-Net ensemble network for WMH segmentation. (in preparation)

Sundaresan, V., Zamboni, G., Le Heron, C., Rothwell, P.M., Husain, M., Battaglini, M., De Stefano, N., Jenkinson, M. and Griffanti, L., 2019. Automated lesion segmentation with BIANCA: impact of population-level features, classification algorithm and locally adaptive thresholding. *NeuroImage*, 202, p.116056.

Sundaresan, V., Griffanti, L., Kindalova, P., Alfaro-Almagro, F., Zamboni, G., Rothwell, P.M., Nichols, T.E. and Jenkinson, M., 2019. Modelling the distribution of white matter hyperintensities due to ageing on MRI images using Bayesian inference. *NeuroImage*, 185, pp.434-445.

Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., **Sundaresan, V.**, Schulz, U.G., Kuker, W., Battaglini, M., Rothwell, P.M. and Jenkinson, M., 2016. BIANCA (Brain Intensity AbNormality Classification Algorithm): a new tool for automated segmentation of white matter hyperintensities. *NeuroImage*, 141, pp.191-205.

Conference proceedings

Sundaresan V., Jenkinson M., Zamboni G. and Griffanti L., 2020. Detection of white matter hyperintensities using Triplanar U-Net ensemble network. 28^{th} Annual Meeting of the International Society for Magnetic Resonance in Medicine (ISMRM) April 18-23 2020, Sydney, Australia (under review).

G. Gentile, M. Battaglini, L. Luchetti, A. Giorgio, L. Griffanti, **V. Sundaresan**, M. Jenkinson, Nicola De Stefano., 2019. BIANCA for an automatic detection of Multiple Sclerosis lesions using machine learning. 35^{th} Congress of the European Committee for Treatment and Research in Multiple Sclerosis (ECTRIMS) September 11-13 2019, Stockolm, Sweden.

Sundaresan V., Arthofer C., Zamboni G., Sotiropoulos S., Rothwell P.M., Dineen R., Auer D., Jenkinson M. and Griffanti L., 2019. Identifying Subjects with Cerebral Microbleeds in Big Datasets: a Pipeline for Candidate Selection. 25^{th} Annual Meeting of the Organization for Human Brain Mapping (OHBM) June 9-13 2019, Auditorium Parco Della Musica, Rome, Italy.

Sundaresan V., Jenkinson M., Zamboni G. and Griffanti L., 2018. Local Intensity Guided Spatially Adaptive Thresholding of White Matter Lesion Probabilities. 24^{th} Annual Meeting of the Organization for Human Brain Mapping (OHBM) June 17-21 2018, Suntec city, Singapore.

Sundaresan V., Griffanti L., Zamboni G. and Jenkinson M., 2017. Modelling white matter hyperintensities distribution within a population using Bayesian Inference. 23^{rd} Annual Meeting of the Organization for Human Brain Mapping (OHBM) June 25-29 2017, Vancouver Convention Centre, Vancouver, Canada.

Contents

1 Background and objective **1**

1.1 Cerebral small vessel disease signs 2

1.1.1 Medical imaging of SVD 4

1.2 White matter hyperintensities (WMHs) 5

1.2.1 Classification of WMHs and clinical impact 5

1.2.2 Exclusion criteria for WMHs 8

1.2.3 WMH segmentation and analysis 8

1.2.4 Brain intensity abnormality classification algorithm (BIANCA) 9

1.2.5 Challenges in WMH segmentation and analysis 10

1.3 Cerebral microbleeds (CMBs) 11

1.3.1 Distribution of CMBs 12

1.3.2 CMB detection and characterisation 14

1.3.3 Challenges in CMB detection 15

1.3.4 CMB mimics 16

1.4 Objectives of this work 19

1.5 Contributions of this work 19

1.6 Thesis outline and overview of the research work 20

2 Literature review **25**

2.1 Clinical impact of WMHs and CMBs 25

2.2 WMH analysis and segmentation 28

2.2.1 Population-level modelling of WMHs 28

2.2.2 Subject-level WMH segmentation 30

2.2.3 WMH segmentation on big datasets 48
2.3 CMB detection 49
2.3.1 Detection of CMBs on individual subjects 49

3 Modelling the distribution of white matter hyperintensities due to ageing on MRI images using Bayesian inference 56

3.1 Introduction 57
3.2 Modelling the distribution of WMHs using Bayesian inference 58
3.2.1 Constructing a representation for the lesion probability distribution using spline basis functions 59
3.2.2 Formulation of the posterior lesion probability 60
3.2.3 Maximisation of the posterior probability by constrained optimisation 63
3.2.4 Optimisation parameters for convergence 65
3.3 Simulations 65
3.3.1 Effect of b-spline knot spacing 68
3.3.2 Comparison with existing alternative methods 68
3.4 Application to real data 69
3.4.1 Evaluation on the UK Biobank dataset 69
3.4.2 Evaluation on the OXVASC dataset 70
3.4.3 Effect of initial values on our algorithm 71
3.5 Results 72
3.5.1 Results on the simulated data 72
3.5.2 Results on the real data 76
3.6 Discussion and conclusion 84
3.7 Summary 89

4 Improving the segmentation of white matter hyperintensities using BIANCA **91**

4.1 Introduction . . . 92

4.2 Dataset details and preprocessing . . . 94

4.2.1 Datasets . . . 94

4.2.2 Image preprocessing . . . 96

4.2.3 BIANCA features and training options . . . 97

4.3 Feature extraction . . . 97

4.3.1 Use of population-level parametric lesion probability map (PPLPM) . . . 97

4.3.2 Use of non-linear registration for estimating spatial features . 98

4.4 Classification: comparison of alternative classifiers within BIANCA . 99

4.5 Thresholding: LOCally Adaptive Threshold Estimation (LOCATE) . 101

4.5.1 Proposed method . . . 101

4.5.2 LOCATE evaluation and validation . . . 104

4.6 Performance evaluation metrics . . . 105

4.7 Results . . . 105

4.7.1 Use of population-level parametric lesion probability map (PPLPM) . . . 105

4.7.2 Use of non-linear registration for estimating spatial features . 107

4.7.3 Comparison of alternative classifiers within BIANCA . . . 107

4.7.4 LOCally Adaptive Threshold Estimation (LOCATE) . . . 112

4.8 Discussion and conclusions . . . 132

4.9 Summary . . . 138

5 Deep learning for segmentation of white matter hyperintensities 139

5.1 Introduction . . . 139

5.2 Dataset preparation . . . 142

5.2.1 Datasets used . . . 142

5.2.2 Data preprocessing . . . 143

5.3 Triplanar U-Net Ensemble Network (TrUE-Net) 143
5.3.1 Network architecture 143
5.3.2 Loss functions 146
5.3.3 Post-processing 148
5.4 Experiment setup 148
5.4.1 Optimisation parameters 148
5.4.2 Data augmentation 148
5.4.3 Implementation details 149
5.5 Experiments 150
5.5.1 Performance evaluation metrics 150
5.5.2 Effect of training parameters on model performance 150
5.5.3 Ablation study: effect of the model dimension and loss function on the segmentation performance 151
5.5.4 Leave-one-out evaluation of WMH segmentation 151
5.5.5 Comparison with BIANCA 152
5.5.6 Harmonisation and domain adaptation: transfer learning and domain adversarial training 155
5.6 Results 160
5.6.1 Effect of training parameters on model performance 160
5.6.2 Ablation study: effect of the model dimension and loss functions on segmentation performance 162
5.6.3 Evaluation of WMH segmentation performance 165
5.6.4 Comparison with BIANCA 172
5.6.5 Domain adaptation: transfer learning and domain adversarial training 179
5.7 Discussion and conclusions 188
5.8 Summary 195

6 Detection of cerebral microbleeds in a large population 196

6.1 Introduction . . . 197

6.2 Dataset details . . . 199

6.2.1 Data preprocessing . . . 201

6.3 Pipeline for preselection of CMB candidate subjects in big datasets . 201

6.3.1 Removal of blood vessels and sulci . . . 201

6.3.2 Voxel-wise detection of initial CMB candidates . . . 204

6.3.3 Filtering of initial candidates using shape-based attributes . . 205

6.3.4 Evaluation of CMB candidate preselection pipeline . . . 206

6.3.5 Performance evaluation metrics for CMB candidate subjects preselection pipeline . . . 208

6.4 Detection of CMBs on individual subjects . . . 208

6.4.1 Initial local maxima identification for ROI extraction . . . 209

6.4.2 Voxel-level CMB initial candidate screening with 3D multiscale encoder-decoder network . . . 210

6.4.3 Object-level CMB candidate discrimination using a 3D classification network . . . 212

6.4.4 Postprocessing for false positive reduction . . . 215

6.4.5 Experiment setup . . . 215

6.4.6 Ablation study: effect of individual scales in the CMB candidate screening step . . . 216

6.4.7 Initial evaluation on the TICH2, HK and OXVASC datasets . 217

6.4.8 Performance evaluation metrics for CMB detection . . . 217

6.5 Automated CMB characterisation and rating . . . 218

6.5.1 CMB size estimation . . . 218

6.5.2 Automated MARS rating of CMBs . . . 218

6.5.3 Characterisation and MARS rating of CMBs on the TICH2 dataset . . . 220

6.5.4 Performance evaluation metrics for automated rating of CMBs 220

6.6 Results . . . 222

6.6.1 Evaluation of CMB candidate subjects preselection pipeline . 222
6.6.2 Evaluation of CMB detection 231
6.6.3 Characterisation of CMBs in the TICH2 dataset 237
6.7 Discussion and conclusions 241
6.7.1 Future work 245
6.8 Summary 247

7 Conclusions and Future Work **249**
7.1 Future work 256

A Dataset details **260**
A.1 The Oxford vascular study (OXVASC) 260
A.2 The UK Biobank study (UKBB) 262
A.3 The neurodegenerative cohort (NDGEN) 264
A.4 The cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) cohort 265
A.5 The healthy control (HC) cohort 266
A.6 The MICCAI WMH segmentation challenge Dataset (MWSC) 266
A.7 The Tranexamic acid for IntraCerebral Haemorrhage 2 (TICH-2) trial dataset 267
A.8 The Hong Kong dataset (HK) 268

B Brief overview of classifiers **270**
B.1 K-nearest neighbour classifier (KNN) 270
B.2 Support vector machine (SVM) 271
B.3 Random forest (RF) 274
B.4 Adaptive boosting of classifiers (AdaBoost) 276
B.5 Neural networks (NNs) 278

C Performance evaluation metrics **280**

C.1 Voxel-wise overlap measures 280

C.2 Cluster-wise measures 281

C.3 Region-level measures 283

C.4 Subject-level measures 285

D The Microbleed Anatomical Rating Scale (MARS): sample form **286**

E Modelling the WMH distribution using Bayesian inference: additional experiments **288**

E.1 Number of iterations in the optimisation step for convergence 288

E.2 Effect of the intermediate b-spline knot spacing values on simulated data 289

F Analysis of LOCATE threshold histograms **292**

F.1 Case 1: LOCATE threshold histogram characteristics within a specific dataset 292

F.2 Case 2: LOCATE threshold histogram characteristics across different datasets 294

Bibliography **297**

List of Figures

1.1 Schematic representation of SVD signs. 4
1.2 Examples of WMHs on various modalities. 6
1.3 Classification criteria of PWMHs and DWMHs. 7
1.4 Schematic representation of the BIANCA algorithm. 10
1.5 Examples of common artefacts confounding WMHs detection. 13
1.6 Difference in the appearance of CMBs with susceptibility effect. . . . 14
1.7 Schematic representation of CMB distribution in the brain. 15
1.8 Examples of CMB mimics. 18
1.9 Workflow for WMH segmentation and analysis. 22
1.10 Workflow for CMB detection and characterisation. 23

2.1 Automated WMH segmentation methods split-up. 38
2.2 A typical CNN architecture for a classification task. 41

3.1 Spline modelling with cubic b-splines. 60
3.2 A sample of data accumulation shown for a single age group. 61
3.3 An instance of initial and final values in the optimisation step. 66
3.4 An instance of simulated lesion data (shown at slice $z = 45$ and age bin number $t = 20$). 67
3.5 Comparison of error maps for various values of knot spacing. 72
3.6 Comparison of simulation analysis results using different methods. . . 74
3.7 Lesion probability values plotted versus age. 75
3.8 Results on the UK Biobank data. 77

3.9 Comparison of the results of the Bayesian spline algorithm with BS-GLMM. 78
3.10 Comparison of the results of the Bayesian spline algorithm with $Smoothed_Average(\sigma = 0.5)$ and $Smoothed_Average(\sigma = 1.5)$ 79
3.11 Results on OXVASC data. 81
3.12 Population-level lesion probability map with respect to MoCA scores on the OXVASC data. 81
3.13 Plot of WMH probability values versus MoCA scores. 82
3.14 Results for different initialisations with various levels of perturbations. 83
3.15 Results of the Bayesian modelling method. 85
3.16 Lesion probability values plotted versus age for various thresholds. . . 86

4.1 Block diagram showing the strategies explored for improving BIANCA performance. 93
4.2 Examples of subjects with different WMH loads. 99
4.3 LOCally Adaptive Threshold Estimation (LOCATE). 102
4.4 Features extracted from each Voronoi polygon plotted against the threshold values. 103
4.5 ROC and SI curves for the NDGEN dataset in the case of use of PPLPM.106
4.6 ROC and SI curves for the cases with linear and non-linear registration.108
4.7 ROC curves and SI curves for alternative classifiers. 109
4.8 ROC curves for alternative classifiers. 110
4.9 SI curves for alternative classifiers for the MWSC dataset. 111
4.10 Results of LOCATE using Voronoi tessellation and SLIC. 112
4.11 Comparison of ROC curves and SI plots for global thresholding. . . . 113
4.12 Example results of LOCATE for different lesion loads. 115
4.13 Example results of BIANCA. 116
4.14 Comparison of ROC curves, SI plots for LOCATE and global thresholding. 117
4.15 ROC curves in the periventricular and the deep regions for LOCATE. 120

4.16 SI curves in the periventricular and the deep regions for LOCATE. . 121
4.17 Comparison of cluster-wise TPR in the periventricular and the deep region for LOCATE and global thresholding. 122
4.18 Example results of LOCATE on the CADASIL data. 128
4.19 Example results of LOCATE on the HC data. 129
4.20 Bland-Altman plots of WMH volumes. 130
4.21 Boxplots of local thresholds obtained from the random forest regression model in LOCATE. 131

5.1 Triplanar U-Net ensemble network (TrUE-Net) architecture. 145
5.2 Example weight maps for weighing voxel-wise cross-entropy loss function. 147
5.3 Triplanar U-Net architecture with attention gate. 154
5.4 TSNE plots showing the combined distribution of FLAIR and T1-weighted intensities. 156
5.5 Sample images from axial and coronal planes. 157
5.6 Transfer learning framework and domain adversarial neural network architectures. 159
5.7 Effect of training parameters on model convergence on the NDGEN dataset. 161
5.8 Effect of the model dimension on the segmentation in the NDGEN dataset. 163
5.9 Effect of loss functions on the segmentation performance. 164
5.10 Regression plot of SI values with respect to lesion load. 165
5.11 Sample results of TrUE-Net segmentation on the MWSC dataset. . . 167
5.12 Sample results of TrUE-Net segmentation on the NDGEN and OX-VASC datasets. 168
5.13 Boxplots of performance measures obtained from LOO evaluation. . . 169
5.14 Boxplots of performance measures obtained in the deep and the periventricular regions. 171

5.15 Sample results for comparison of TrUE-Net segmentation with other methods. 174
5.16 Boxplots of performance measures obtained for TrUE-Net, TrUE-Net-Att and BIANCA. 176
5.17 Regression plot of SI values with respect to lesion loads. 179
5.18 Sample results of domain adaptation experiment test cases (high lesion load). 180
5.19 Sample results of domain adaptation experiment test cases (low lesion load). 181
5.20 Boxplots of performance measures obtained for various test cases. . . 182
5.21 Heatmaps of performance measures for transfer learning. 184
5.22 The effect of domain adaptation on the extracted features. 186

6.1 Histograms of the subject-level CMB counts based on manual segmentations. 202
6.2 Features used for the vessel/sulci removal. 203
6.3 Features extracted for the voxel-wise CMB candidate detection. . . . 205
6.4 An instance of the shape-based filtering of initial candidates. 206
6.5 Overview of the CMB detection method. 209
6.6 CMB candidate screening step. 211
6.7 CMB candidate discrimination step. 214
6.8 MARS rating regions from the structural atlases. 221
6.9 FROC curves for initial CMB candidate detection. 222
6.10 Results of preselection of CMB subjects on the TICH2 dataset. . . . 223
6.11 Examples of CMB subject preselection OXVASC-trained pipeline results. 225
6.12 Results of preselection of CMB subjects on the OXVASC dataset. . . 226
6.13 Examples of CMB subject preselection TICH2-trained pipeline results. 227
6.14 Boxplots of CMB counts in individual subjects (OXVASC-trained model) against three age groups. 228

6.15 Boxplots of CMB counts in individual subjects (TICH2-trained model) against three age groups. 229
6.16 FROC curves for the ablation study on the TICH2 dataset. 232
6.17 FROC curves for the CMB detection on the TICH2, OVXASC and HK datasets. 233
6.18 Examples of CMB detection results on the TICH2, OXVASC and HK datasets. 235
6.19 Results of CMB size estimation. 238
6.20 Confusion matrices for the automated CMB rating using manually segmented CMBs. 239
6.21 Confusion matrices for the automated CMB rating using the detected true positive CMBs. 241
6.22 Barplots showing the number of false negatives and false positives. . . 241

B.1 The effect of k values on KNN classification. 271
B.2 Transformation of the input data using an appropriate kernel mapping function. 272
B.3 Optimisation in SVM. 273
B.4 A simplified version of an RF classifier. 275
B.5 A schematic diagram of an AdaBoost classifier. 276
B.6 Network diagram of a typical NN. 278

D.1 A sample Microbleed Anatomical Rating Scale (MARS) form for CMB rating. 287

E.1 Convergence curves for the optimisation of the Bayesian inference algorithm. 289
E.2 Comparison of error maps for intermediate values of knot spacing. . . 290

F.1 Analysis of LOCATE thresholds within a dataset. 293
F.2 Analysis of LOCATE thresholds for various datasets. 295

List of Tables

1.1 A brief description of MRI artefacts hindering WMH segmentation. . 12
1.2 CMB mimics and their distinguishing characteristics. 17

2.1 Comparison of the existing non-DL WMH segmentation methods. . . 39
2.2 Comparison of the existing DL WMH segmentation methods 44

3.1 Comparison of MSE values and running time on simulation data. . . 75

4.1 Parameters considered for tuning different classifiers. 100
4.2 Best performing set of parameters for different classifiers. 108
4.3 Descriptive statistics and ANOVA results for comparison of TPR values.123
4.4 Descriptive statistics and ANOVA results for comparison of FPR values.124
4.5 Descriptive statistics and ANOVA results for comparison of SI values. 125
4.6 Descriptive statistics and ANOVA results for comparison of cluster-wise TPR values. 126

5.1 Transformations used for data augmentation and their parameter ranges.149
5.2 Leave-one-out evaluation of TrUE-Net. 170
5.3 Comparison of performance measures between PWMHs and DWMHs. 173
5.4 Comparison of TrUE-Net performance with TrUE-Net-Att and BIANCA.177
5.5 Results of two-tailed paired t-tests on the performance measures. . . . 178
5.6 Comparison of performance measures for test cases on the OXVASC dataset. 182
5.7 Results of two-tailed paired t-tests between the individual combinations of test cases. 183

6.1 Transformations used for data augmentation and their parameter ranges. 216
6.2 CMB rating scales BOMBS and MARS, with the atlases. 219
6.3 Demographic/clinical factors and criteria used for shortlisting the potential candidates. 230
6.4 Results of test cases considered in the ablation study. 232
6.5 Performance of the CMB detection method at various steps for the TICH2, OXVASC and HK datasets. 234
6.6 Comparison of the performance of the proposed CMB detection method with the existing methods. 236
6.7 Automated CMB rating of detected CMBs the according to the MARS scale. 240

7.1 Comparison of the existing methods with BIANCA and the methods/improvements proposed in this work. 253

A.1 Demographic and sample details of the datasets used for WMH segmentation. 261
A.2 Demographic and sample details of the datasets used for CMB detection. 262
A.3 MR acquisition parameters for the OXVASC dataset. 263
A.4 MR acquisition parameters for the UK Biobank dataset. 264
A.5 MR acquisition parameters for the NDGEN dataset. 265
A.6 MR acquisition parameters for the MWSC dataset. 267
A.7 MR acquisition parameters for the TICH2 dataset. 268

C.1 Confusion matrix for a multi-class problem. 283

List of Abbreviations

AB	Adaptive Boosting (AdaBoost)
AD	Alzheimer's Disease
AG	Attention Gate
ANOVA	Analysis Of Variance
AVD	Absolute Volume Difference
BET	Brain Extraction Tool
BIANCA	Brain Intensity AbNormality Classification Algorithm
BOMBS	Brain Observer MicroBleed Scale
BP	Blood Pressure
BSGLMM	Bayesian Spatial Generalised Linear Mixed Model
CAA	Cerebral Amyloid Angiopathy
CADASIL	Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy
CE	Cross-entropy
CMBs	Cerebral MicroBleeds
CNNs	Convolutional Neural Networks
CSF	Cerebrospinal Fluid
DA	Domain Adaptation
DANN	Domain Adversarial training of Neural Networks
DL	Deep Learning
FA	Fractional Anisotropy
FCNs	Fully Convolutional Networks
FLAIR	FMRIB Software Library
FLIRT	FMRIB's Linear Image Registration Tool
FN	False Negative
FNIRT	FMRIB's Non-linear Image Registration Tool
FOV	Field-of-view

FP	False Positive
FPCA	Functional Principal Component Analysis
FPR	False Positive Rate
FRST	Fast Radial Symmetry Transform
FSL	FMRIB Software Library
FWHM	Full Width at Half Maximum
GANs	Generative Adversarial Networks
GPU	Graphical Processing Unit
GRE	Gradient Recalled Echo
HC	Healthy Controls
HK	Hong Kong
HT	Hypertensive
IDP	Imaging Derived Phenotypes
INVADE-II	INtervention project on cerebroVAscular disease and Dementia in Ebersberg - II
ISMRM	International Society for Magnetic Resonance in Medicine
KNN	K-Nearest Neighbours
LADIS	Leukoaraiosis And DISability study
LDA	Linear Discriminant Analysis
LGA	Lesion Growth Algorithm
LOCATE	LOCally Adaptive Thresholds Estimation
LoG	Laplacian of Gaussian
LOO	Leave-one-out
LOTS-IM	Limited One-Time Sampling Irregularity Age Map
LPA	Lesion Prediction Algorithm
LST	Lesion Segmentation Tool
MARS	Microbleed Anatomical Rating Scale
MICCAI	Medical Image Computing and Computer Assisted Intervention
MIP	Minimum-Intensity Projections
MD	Mean Diffusivity
ML	Machine Learning
MNI	Montreal Neurological Institute
MoCA	Montreal Cognitive Assessment
MRF	Markov Random Field

MRI	Magnetic Resonance Imaging
MS	Multiple Sclerosis
MSE	Mean-Squared Error
MSD	Mean Square Difference
MWSC	MICCAI WMH Segmentation Challenge
NDGEN	Neurodegenerative data
NIN	Network-in-Network
NNs	Neural Networks
OHBM	Organization for Human Brain Mapping
OPTIMA	Oxford Project to Investigate Memory and Ageing
OXVASC	Oxford Vascular study
PD	Proton Density
QDA	Quadratic Discriminant Analysis
R-CNN	Region-based Convolutional Neural Networks
RFs	Random Forests
ROI	Region Of Interest
RPN	Region Proposal Network
SLIC	Simple Linear Iterative Clustering
SMOTE	Synthetic Minority Over-sampling Technique
SVD	Small Vessel Disease
SVM	Support Vector Machine
SWI	Susceptibility Weighted Imaging
TICH2	Tranexamic acid for IntraCerebral Haemorrhage-2
TL	Transfer Learning
TN	True Negative
TP	True Positive
TPR	True Positive Rate
TrUE-Net	Triplanar U-Net Ensemble Network
UKBB	UK Biobank
VLSM	Voxel-based Lesion-Symptom Mapping
WMHs	White Matter Hyperintensities
YOLO	You Only Look Once

Chapter 1

Background and objective

Cerebral small vessel disease (SVD) is the term used to indicate the clinical damage and neuropathological findings related to perforating arterioles, small arteries, venules and capillaries [1]. Various risk factors have been associated with the occurrence and progression of SVD, including age, sex, smoking and high blood pressure [2] (especially, age and blood pressure [3]). SVD is one of the major causes of stroke (around 20%) [4] and the prevalence of SVD has been shown to be higher in Asians compared to Caucasians [3].

SVD signs are found commonly in cases of cognitive decline [3, 2], neurodegenerative diseases such as Alzheimer's disease, vascular dementia and mixed (i.e. with both neurodegenerative and vascular components) dementia [5, 6]. However, the real impact on clinical outcome (e.g. cognition) of SVD is not yet fully understood [4]. Studying the clinical signs and imaging biomarkers of SVD is important to further our understanding on the relationship between SVD and its risk factors and clinical outcomes.

This thesis focuses on the detection and analysis of the two SVD signs, white matter hyperintensities (WMHs) and cerebral microbleeds (CMBs), in brain magnetic resonance (MR) images. Accurate segmentation of these SVD signs is the first step towards understanding the clinical associations of these signs in a population.

In this chapter, we provide a brief introduction to SVD signs, explain the need for accurate detection of WMHs and CMBs, and the challenges involved, in sections 1.1,

1.2, 1.3. We also describe our aims and contributions in sections 1.4 and 1.5, followed by an overview of the thesis structure in section 1.6.

1.1 Cerebral small vessel disease signs

Common causes of SVD are blockage of small vessels in the brain leading to ischemia, or leakages resulting in micro haemorrhages in the brain tissue. The main signs of SVD include recent subcortical infarcts, white matter hyperintensities (WMHs), lacunes, enlarged perivascular spaces and cerebral microbleeds (CMBs) [7]. A brief overview of these SVD signs is as follows:

- **Recent subcortical infarcts** are caused by the occlusion of smaller arteries [7]. Even though subcortical infarcts have been shown to cavitate (form fluid filled cavities or lacunes), they could evolve into hyperintensities or lacunar infarcts [7]. Subcortical infarcts cause around 25% of ischemic stroke and are often asymptomatic [8]. However, the progression of subcortical infarcts and their clinical impact are not yet fully explored [9].

- **Lacunes** of presumed vascular origin (commonly known as silent brain infarcts [10]) are fluid-filled cavities formed by healing subcortical infarcts, and occur near perforating subcortical arterioles [7]. Lacunes have been associated with dementia [10], gait impairment [11] and have been shown to increase the risk of stroke [10] and post-stroke depression [12].

- **White matter hyperintensities** of presumed vascular origin (WMHs, also known as white matter lesions) are commonly found in elderly subjects. However, they have been related to various vascular diseases, dementia (including Alzheimer's disease), and are important biomarkers for stroke [13] and gait disturbances [7, 14]. Even though the pathogenesis of WMHs has not yet been fully understood, studies show histopathological correlations between WMHs and partial loss of myelin, hypoxia [15], astrogliosis [16] and enlarged perivascular spaces [17].

- **Cerebral microbleeds** (CMBs) represent focal haemosiderin depositions consisting of macrophages in microhemorraghes [18], and are sometimes surrounded by ischemic areas and gliosis [16]. CMBs could be the early sign of intracerebral haemorrhage (ICH) [16], vascular dementia, Alzheimer's disease [16, 7, 18] and have been associated with cognitive decline [19, 20]. Moreover, CMBs often occur with cerebral amyloid angiopathy (CAA) related vascular damage, mostly in grey-white matter junctions [16].

- **Enlarged perivascular spaces** occur due to dilation of fluid filled space around small arterioles/venules and brain parenchyma tissue [21]. They are associated with other signs of SVD such as WMHs [22] and lacunes [21]. The clinical impact of enlarged perivascular spaces is not yet clear, however they have shown some correlation with age and cognitive decline [23].

In addition to the above, other signs of SVD could include intracerebral haemorrhage and superficial cortical siderosis and brain atrophy [7].

Among the above signs, we focus on WMHs and CMBs in our work. Recent studies indicate a relatively higher burden of WMHs (>65% in [24] and around 36% in [3]) and CMBs (around 10% in [24] and 27% in [3]) among SVD subjects, and show higher correlation with the risk factors of SVD. Given the clinical impact of WMHs and CMBs, their accurate detection is crucial to understand their role in the pathogenesis of various vascular and neurodegenerative diseases occurring in the elderly and to evaluate their usefulness as biomarkers for identifying and staging such diseases. Automated detection of WMHs and CMBs has emerged as an active research area, since they represent a crucial link between vascular and neurodegenerative diseases [25, 16, 26]. Out of the two, WMHs have been better explored, but their segmentation and analysis still have scope for improvement. On the other hand, with only a few methods (and most of them semi-automated) existing for CMB detection, CMBs require more research for their identification and analysis.

1.1.1 Medical imaging of SVD

The manifestation of SVD signs is usually inferred from the findings observed in imaging modalities such as computed tomography (CT) and magnetic resonance (MR) imaging [4, 7]. In particular, structural MR imaging data such as T1-weighted, T2-weighted, diffusion weighted imaging (DWI), fluid attenuation inversion recovery (FLAIR), T2*-weighted (or susceptibility weighted - SWI) sequences are widely used for identifying specific characteristics associated with various SVD signs.

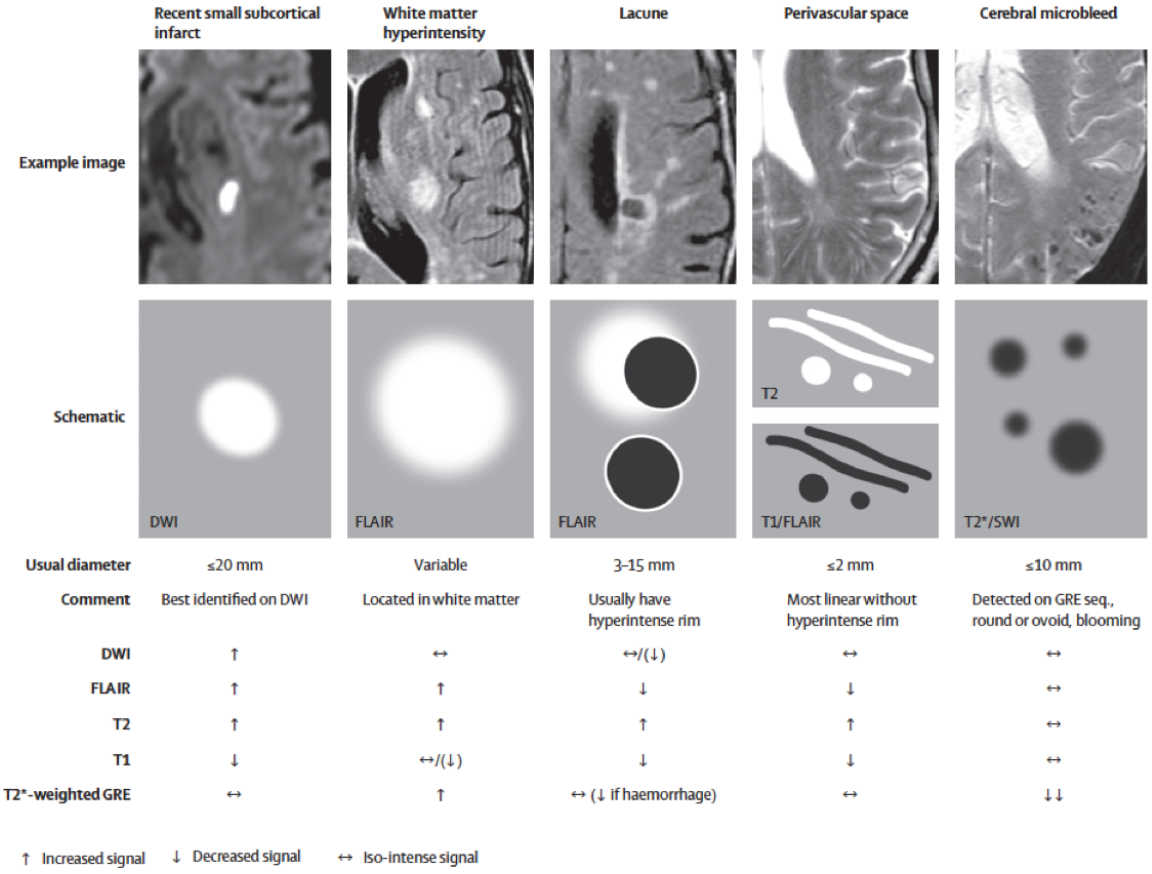


Figure 1.1: Schematic representation of SVD signs. The characteristics of SVD signs (top) along with their schematic representation (bottom). Image courtesy: *Wardlaw et al, Lancet Neurology, 2013.*

Figure 1.1 shows the characteristics of various SVD signs, along with their schematic representation as described in [7]. Briefly, recent subcortical infarcts are focal hyperintense areas of size up to 20 mm on T2-weighted, DWI and FLAIR images. Lacunes have a distinct pattern of ovoid dark cavities (filled with fluid and

having CSF-like signal intensity characteristics) surrounded by hyperintense rims on FLAIR images. Lacunes of presumed vascular origin have size ranging from 3 to 15 mm. Enlarged perivascular spaces look similar to lacunes (except for a less distinct rim), however they are smaller than lacunes with less than 3 mm in size and tubular in shape. WMHs are localised hyperintense regions on T2-weighted, T2*-weighted and FLAIR images, and generally vary in size. CMBs are small focal dark circular lesions on T2*-weighted gradient recalled echo (GRE) sequences ranging from 2 to 10 mm in size. However, CMBs appear prominent on susceptibility weighted images (SWI) due to blooming effect [27, 25].

1.2 White matter hyperintensities (WMHs)

WMHs of presumed vascular origin (also known as white matter lesions or leukoaraiosis [7]) appear as bright localised areas in proton-density, T2-weighted, T2*-weighted and FLAIR images [28]. They can also appear hypointense on T1-weighted images. Figure 1.2 shows the difference in the appearance of WMHs between some of the commonly used image modalities.

1.2.1 Classification of WMHs and clinical impact

WMHs are mainly differentiated into periventricular WMHs (PWMHs) and deep WMHs (DWMHs) [29, 26, 30]. PWMHs occur around ventricles, appearing as caps (especially around the frontal ventricular horns) and thin lining or halo [31]. DWMHs occur in the deep subcortical white matter [31, 26], as punctate, early confluent and confluent lesions [16]. Qualitative assessments by MR studies show that the impact is often distinguished between PWMHs and DWMHs [32, 26, 33]. A detailed account of their clinical impact has been provided in chapter 2. On MR images, WMHs are classified into PWMHs and DWMHs according to various criteria based on distance [33, 30, 34] and continuity from the ventricles [31, 35]. Figure 1.3 shows the schematic representation of different criteria for classification of PWMHs and DWMHs as described in [36]. The continuity from ventricles criterion [31, 35] considers WMHs

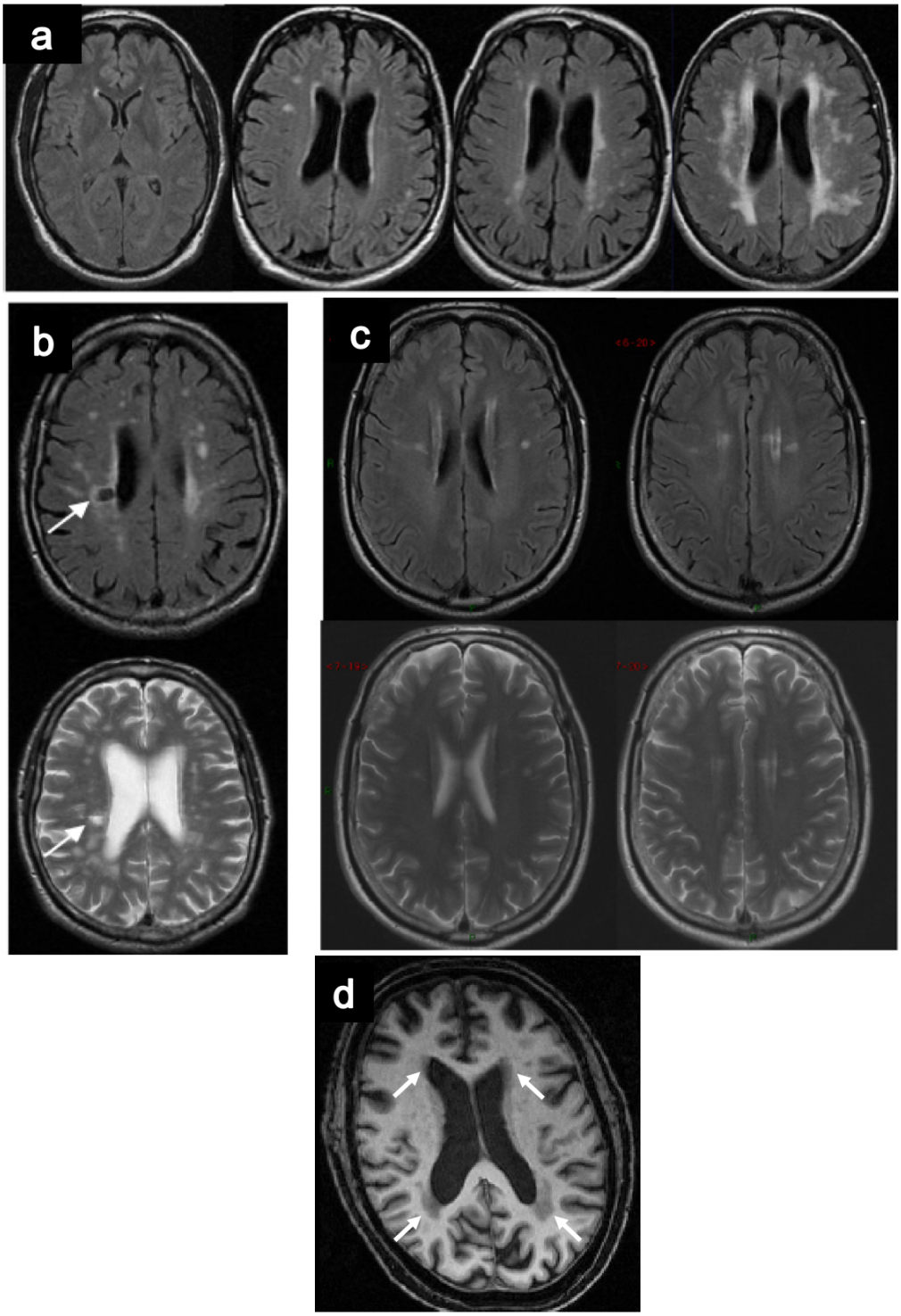


Figure 1.2: Examples of WMHs on various modalities. Examples showing WMHs on (a) FLAIR (from left to right: normal to severe WMHs), (b) FLAIR and T2-weighted images of the same subject showing WMHs and lacune (arrow) (c) FLAIR and T2-weighted images. Note that subtle WMHs are detected more easily on FLAIR. (d) T1-weighted images (WMHs indicated by arrows). (a),(b) and (c) images courtesy: *Wardlaw et al, Journal of the American Heart Association, 2015.*

adjacent to the ventricles as PWMHs. Based on the distance from ventricles, WMHs within 10 mm from ventricles [33] or within 3-13 mm from ventricles [30] are considered as PWMHs, and WMHs lying less than 3 mm are regarded as juxtaventricular WMHs (JWMHs). Wen and Sachdev [34] determined an empirical distance from the ventricles in an elderly population to distinguish PWMHs and DWMHs. Among the above criteria, the 10 mm distance rule has been shown to be more suitable for automatic implementation, is commonly used and agrees with human rater, even in most of the confluent lesion cases [36]. In general, considering the differences in the appearance and clinical impact of PWMHs and DWMHs, it would be useful to understand how their individual characteristics affect automated segmentation methods. For this purpose, we performed comparative analysis on our methods, using various performance measures, between PWMHs and DWMHs in our experiments.

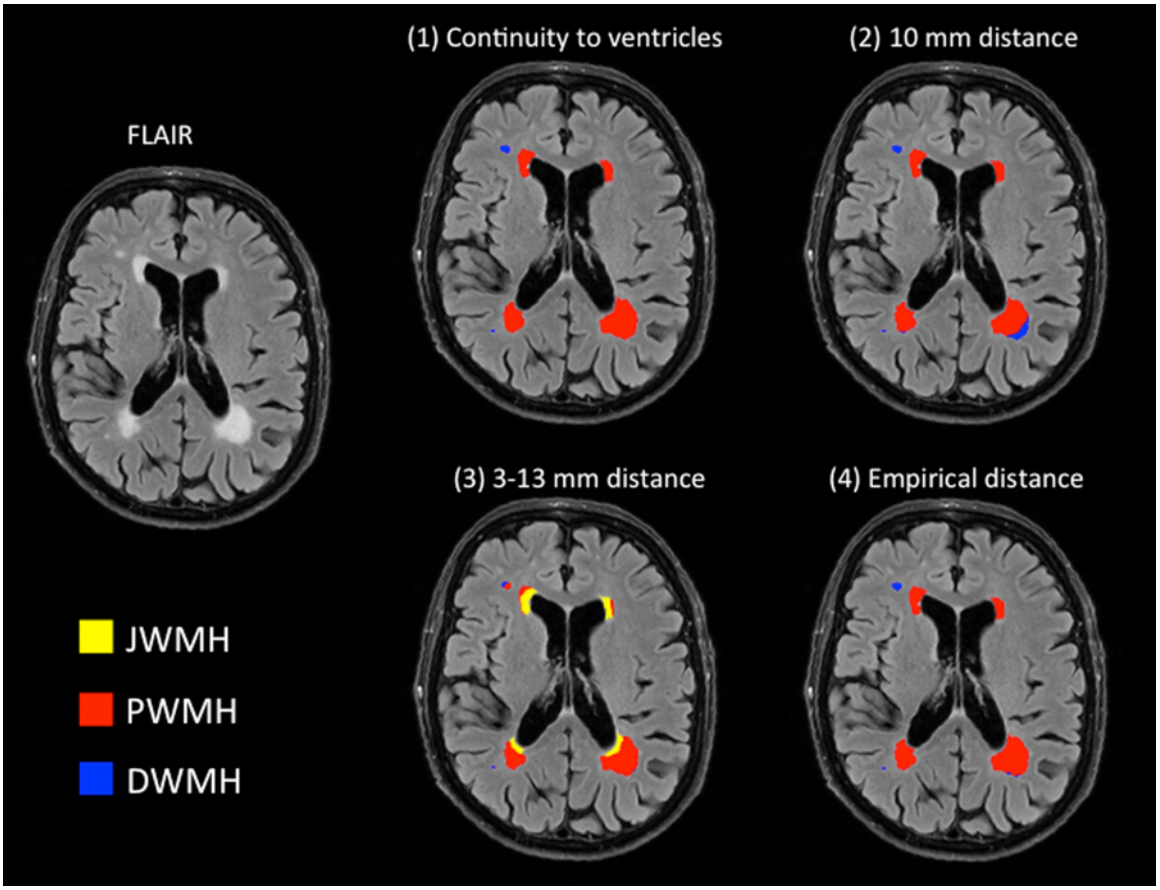


Figure 1.3: Schematic representation of various classification criteria of PWMHs and DWMHs. PWMHs (red) and DWMHs (blue), shown along with juxtaventricular WMHs (JWMHs, yellow). Image courtesy: *Griffanti et al, NeuroImage, 2018.*

1.2.2 Exclusion criteria for WMHs

There have been differences in views on the exclusion criteria for WMHs of presumed vascular origin, and the general consensus [7] is the 'lesions in the subcortical grey matter or brainstem should not be included in the category of white matter hyperintensities of presumed vascular origin unless explicitly stated'. Also, smooth WMHs, particularly in the subcortical deep white matter, have often been confused with and indistinguishable from dirty-white matter [37]. Since dirty-white matter is a biomarker for multiple sclerosis [38] and has a different underlying pathology, we exclude it from WMHs.

A genetic form of SVD is cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) [39], characterised by a genetic mutation with the NOTCH-3 gene. It is the most common heritable cause of stroke and vascular dementia in adults and it is characterised by extensive damage to white matter brain regions [40]. The patients are usually younger than those with sporadic SVD, without the confounding factors of co-existent neurodegenerative pathology [41]. Since they provide a model of pure SVD [42], we include CADASIL lesions for comparative analyses with WMHs.

1.2.3 WMH segmentation and analysis

Various visual rating scales are available for assessing WMHs, the most common ones being the Fazekas scale [31], the Schelten's scale [43] and the age-related white matter changes (ARWMC) scale [44]. However, visual rating is time consuming, causes observer bias leading to inconsistent results among various studies, and only provides qualitative or categorical information about WMHs and limited information about the spatial location of WMHs [45]. Voxel-wise WMHs maps would help us in identifying the relationship between the spatial location/distribution of WMHs and various clinical factors [46]. In addition, getting precise locations for WMHs would be useful to define patterns for discriminating between normal and pathological ageing [47, 48, 49, 50]. Manual segmentation would allow voxel-wise information on

WMHs to be obtained, however, it is time consuming, even more than visual rating, and, similar to visual rating, is prone to inter-rater variability. Therefore, several computer-assisted automated and semi-automated WMH segmentation algorithms have been developed to reliably provide the exact localisation of WMHs at the voxel-level for individual subjects [51]. A review of the existing literature is provided in chapter 2. The FMRIB Software Library (FSL) [52] group has recently developed a machine learning based WMH segmentation tool called Brain Intensity AbNormality Classification Algorithm (BIANCA) [53] (explained below in section 1.2.4). Even though BIANCA provides a reliable segmentation of WMHs (of vascular origin), there is still scope for improvement in the overall segmentation performance, especially in the deep regions, where WMHs are subtler than PWMHs [53].

In addition to the detection of WMHs on individual subjects, determining the voxel-wise distribution of WMHs across a population is important, in order to investigate the relationship between the spatial distribution of lesions and various factors of interest, such as age, presence of vascular risk factors and cognitive function. In fact, although the occurrence of WMHs is common in ageing population [54], clear relationships between volume/distribution of lesions and cognition, demographics, and risk factors have not been established [16, 7].

1.2.4 Brain intensity abnormality classification algorithm (BIANCA)

BIANCA [53] is a supervised algorithm that classifies the image voxels based on intensity and spatial features using the K-Nearest Neighbour (KNN) algorithm. The schematic representation of the BIANCA algorithm is shown in figure 1.4. In KNN, the predicted value for each instance (i.e. voxel) is the average of K nearest neighbouring instances in the feature space (refer to appendix B for a brief overview of KNN). In the output image, the value at each voxel represents the probability of the voxel being WMH. The lesion probability map is then usually thresholded to create a binary lesion map before being used in further analyses. Currently, threshold values of 0.8 - 0.95 have been shown to be the optimal threshold value for most of the

datasets. The false positives can be reduced by applying a mask obtained from the dilated and inverted cortical CSF tissue segmentation, combined with other deep grey exclusion masks, as described in [53] to mask out the voxel values outside the region of interest (ROI) (i.e. the white matter).

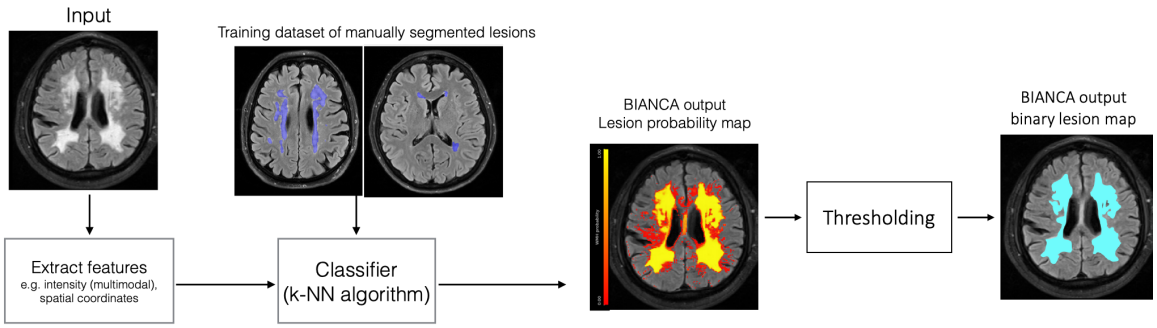


Figure 1.4: Schematic representation of the BIANCA algorithm.

The intensity features of BIANCA are extracted from, but not restricted to, FLAIR, T1 and fractional anisotropy (FA) images, while spatial features are extracted from the corresponding MNI coordinates. Several additional options are available in BIANCA: weights for spatial features, patch (kernel) size for local averaging of intensity features, subjects included in the training set based on their lesion load, number of training points for the WMHs and non-WMHs classes, location of training points for non-WMHs classes (refer to [53] for further details).

BIANCA is quite flexible, computationally lean and provides good reproducibility within scanner. However, robustness of the algorithm requires improvement with respect to the variation in lesion intensities. Also, parameters such as threshold value for obtaining binary lesion maps need tuning for data acquired with different scanners [53].

1.2.5 Challenges in WMH segmentation and analysis

WMH segmentation is challenging due to four main aspects as explained below:

Firstly, segmentation of WMHs is affected by the variability in the intensity characteristics of WMHs in different locations. For instance, PWMHs are brighter and

have more contrast with respect to the background white matter, when compared to DWMHs [55, 56, 57].

The second challenge is the presence of background noise and artefacts in the white matter. Noisy images could lead to spurious detections and could intervene with the correct boundary delineation of WMHs and detection of small lesions, particularly in the deep regions. In addition to noise, other artefacts [28] that occur during acquisition also affect WMH segmentation. Table 1.1 provides a brief description of various MR artefacts confounding WMH segmentation and figure 1.5 shows a few examples of commonly occurring artefacts.

Thirdly, the contrast variations within some subcortical structures near the ventricles (e.g. ends of caudate nucleus) and brain abnormalities (e.g. brain infarcts) [16] can also result in the detection of 'false' WMHs. Even though positional priors could be used to reduce these false detections, these could be counterproductive and can cause the loss of true WMHs.

The fourth challenge is the change in lesion characteristics and distribution with respect to the pathology and demographics of the cohort/dataset. The use of intensity and spatial features alone might not be sufficient for the robust detection of WMHs across various datasets. Hence, there is a need for methods that use contextual information from images in addition to the above features. These include the methods operating on a region (thus using neighbourhood information) or a *receptive field*, rather than working on an individual voxel. For example, convolutional neural networks (CNNs) use convolutional filters with varying sizes of receptive fields, and have been shown to be robust tools for segmentation of WMHs [58, 59] across various datasets.

1.3 Cerebral microbleeds (CMBs)

CMBs appear as small, circular, well-defined hypointense lesions of size ranging from 2 to 10 mm in size on T2*-weighted images. They are more prominent in SWI, since

MRI artefacts	Appearance and characteristics
Motion artefacts	Occurs due to head motion and eye movement. Ghosting effect (objects extended and blurred along the read direction of scans) around the eye; hyperintense concentric rings with ghosting effect and image distortion.
Gibbs ringing	Occurs during image reconstruction due to truncation of Fourier series. Appears as a concentric hyperintense parallel rings immediately adjacent to the high contrast areas (e.g. brain edges).
CSF flow/pulsation artefacts	Occurs due to pulsative flow of CSF. Presence of hyperintense clusters around third, fourth and lateral ventricles. Also occurs in choroid plexuses.
Magnetic susceptibility artefacts	Caused by the presence of ferromagnetic material or at air-tissue interfaces. Sudden changes in the magnetisation result in spin dephasing, causing signal loss and/or image distortion. Occurs at the floor of frontal regions and near air-bone interfaces.
Blood flow artefacts	Small linear hyperintensities near sinuses and major arteries (e.g. carotid artery) due to blood flow.

Table 1.1: A brief description of MRI artefacts hindering WMH segmentation.

haemosiderin (which is paramagnetic) in CMBs causes T2* signal loss and change in phase, making CMBs appear larger visually on images than their actual size. This effect is known as the blooming effect [60, 61, 62]. Figure 1.6 (shown in [25]) indicates the difference in CMB between T2*-weighted GRE image and SWI.

1.3.1 Distribution of CMBs

Based on their histopathological correlates, location and clinical implications, CMBs are broadly classified into deep (infratentorial) CMBs and lobar (cortical-subcortical) CMBs [25, 63], as shown in figure 1.7a. The distribution of CMBs and their significant correlation with clinical factors [64, 63] are shown in figure 1.7b. Deep CMBs are caused by arteriosclerosis and are usually associated with hypertensive vasculopathy [65, 63, 64, 66]. In such cases, CMBs are mostly found in basal ganglia, thalamus,

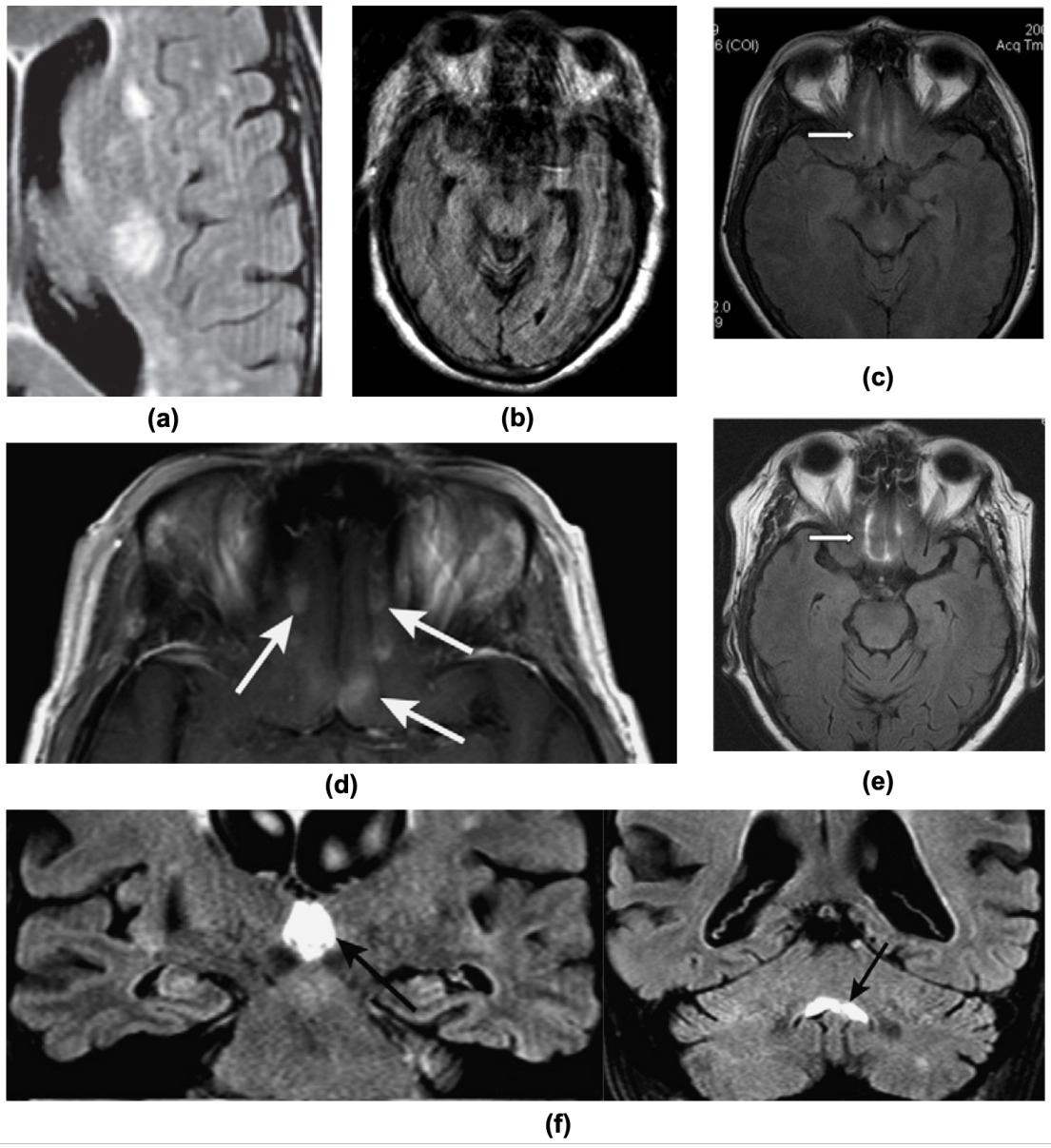


Figure 1.5: Examples of common artefacts confounding WMHs detection (shown on FLAIR images). (a) Gibbs ringing artefact, (b) motion artefact, (c, e) susceptibility artefacts in frontal lobes at air-bone interfaces, (d) susceptibility artefacts in the air-tissue interface in the frontal region, (f) CSF flow artefacts in the third and fourth ventricles (indicated by arrows). Images courtesy: *Wardlaw et al, Lancet Neurology, 2013* (a), *Nyberg et al, American Journal of Neuroradiology, 2012* (b), *Lavdas et al, Clinical Imaging, 2014* (c,e), *Vargas et al, Journal of Neuroradiology, 2009* (d,f).

brainstem and cerebellum [27]. Lobar CMBs are caused by the accumulation of amyloid-β in the arteries, arterioles and capillaries in cerebral cortex and are often associated with CAA, APOE $\epsilon 4$ allele and Alzheimer's disease [67, 68, 64, 69]. Studies suggest that lobar CMBs related to CAA are mainly found in the posterior part of

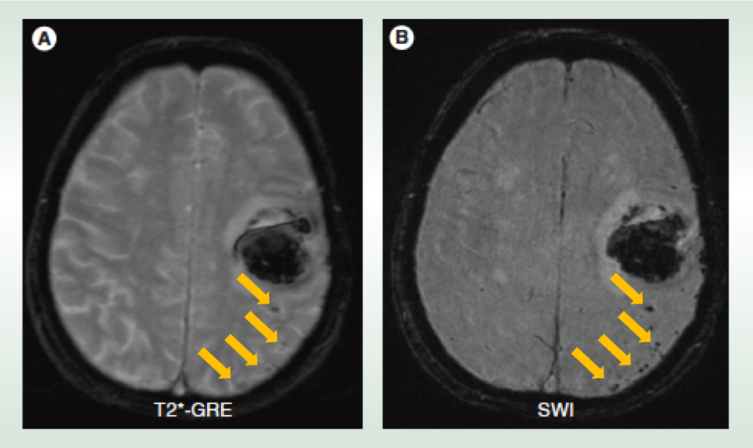


Figure 1.6: Difference in the appearance of CMBs with susceptibility effect. An example showing CMBs in an intracerebral haemorrhage subject on (A) T2*-weighted GRE and (B) SWI. Note 4 definite CMBs in the bottom right (indicated by arrows). Image courtesy: *Charidimou et al, Future Neurology, 2011.*

temporal and parietal lobes [70, 71], and in occipital lobes [72]. A detailed overview of the clinical impact of CMBs, as provided in various studies has been reported in chapter 2.

1.3.2 CMB detection and characterisation

Various visual rating scales such as the microbleed anatomical rating scale (MARS) [73] and the brain observer microbleed scale (BOMBS) [74] are available for rating CMBs in different regions of the brain. The BOMBS scale provides the CMB ratings in the lobar and deep regions of the brain, subdivided by CMB size (<5 mm, 5-10 mm) [74], while the MARS scale, in addition, provides CMB ratings in the individual lobes. Similar to WMH visual rating scales, they are time consuming and prone to inter-rater variability. Voxel-level detection of CMBs would allow their exact localisation, determination of their count, and the possibility of their characterisation. For instance, obtaining a voxel-wise CMB map would help in determining the size and count of CMBs. An automated CMB segmentation tool with reliable accuracy and precision would be useful to derive the voxel-wise CMB maps for individual subjects efficiently in large populations.

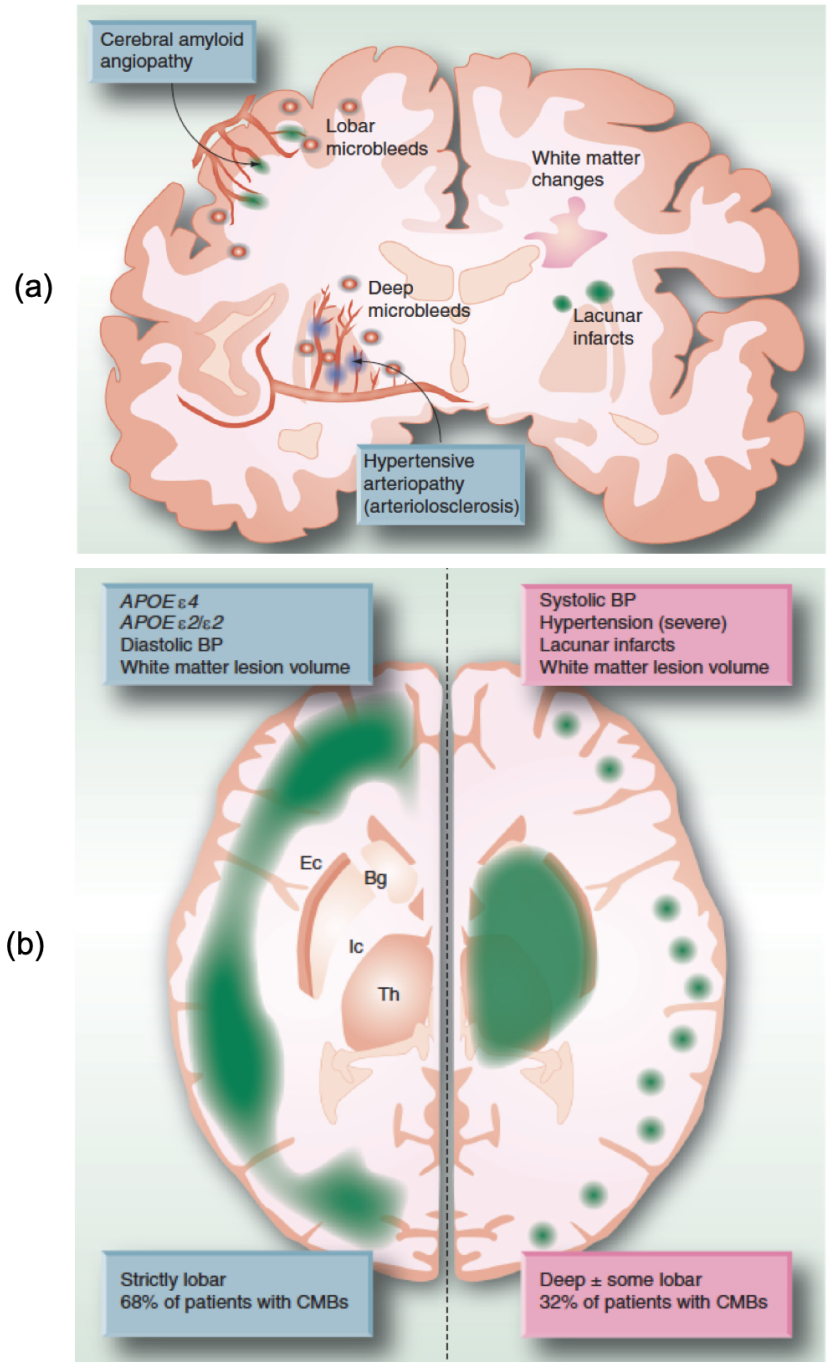


Figure 1.7: Schematic representation of CMB distribution in the brain. (a) Histopathology of deep and lobar CMBs. (b) Distribution of CMBs and their clinical implications as reported in the Rotterdam Scan Study. Bg - basal ganglia, BP - blood pressure, Ec - external capsule, Ic - internal capsule, Th - thalamus. Images courtesy: *Charidimou et al, Future Neurology, 2011.*

1.3.3 Challenges in CMB detection

The main challenges in CMB detection are that CMBs are small in size, sparsely distributed and are not very common. This makes their detection and analysis highly

difficult both at population- and subject-level. It would be quite useful to observe the prevalence and clinical/demographic associations of CMBs in larger populations. But these population-level analyses are limited by the fact that CMBs occur sparsely, are difficult to detect due to their size, and are often accompanied by other signs (e.g. haemorrhages).

At the subject-level, automated voxel-wise detection of CMBs is still highly challenging due to their small size, contrast variations and the other mimics. The use of SWI generally improves the CMB contrast when compared to T2*-weighted GRE sequences. However, SWI also enhances haemorrhagic CMB mimics (explained in section 1.3.4) such as cavernous malformation, micrometastases and small haemorrhages, making it difficult to identify true CMBs. Distinguishing CMBs from their mimics can benefit from using multiple imaging modalities, prior clinical information (especially in case of flow voids and partial volume effects) and good quality (high resolution) images. Semi-automated methods have been proposed as a possible solution to overcome the spurious detection of CMB mimics, although with a trade-off of post-detection manual intervention.

1.3.4 CMB mimics

The SWI has been shown to aid in identifying more CMBs (at least >67% [75]) compared to T2*-weighted GRE images. However, the presence of other paramagnetic substances, apart from haemosiderin, causes the enhanced appearance of dark structures that resemble CMBs, known as CMB 'mimics' [27]. The description and features of CMB mimics [27, 25] are listed in table 1.2. Figure 1.8 shows a few examples of commonly observed CMB mimics [27]. The major haemorrhagic/paramagnetic CMB mimics include cavernous malformations, haemorrhagic micrometastases, diffusion axonal injury, small haemorrhages nearer to the infarcts and ICH areas. In addition to the above, there are various non-haemorrhagic mimics such as flow voids, calcifications and partial volume artefacts at air-bone interfaces [27, 25]. However, the use of T2*-weighted GRE and SWI images together could help in the removal of

non-haemorrhagic mimics, since they are non-paramagnetic and hence do not exhibit the blooming effect on SWI.

CMB mimics	**Description and characteristics**
Cavernous malformations	Visible on T2 (subtypes I - IV). Leakage of blood from the vessels at various stages of degradation [27], with subtype II lesions appearing with popcorn-like structure (looks similar to ringing artefacts found on T2*-weighted images). Subtype - IV lesions are punctate and are more difficult to differentiate from CMBs [25].
Haemorrhagic micrometastases	Metastatic melanoma often can bleed and appear hypointense on T2*-weighted images since melanin is paramagnetic, but has concomitant hyperintensities on T1-weighted images surrounded by oedema. Note that other than metastatic melanoma, other haemorrhagic metastases are possible. However, small non-oedemous lesions are difficult to distinguish from CMBs.
Diffuse axonal injury	Concomitant presence of abnormalities (e.g. skull fracture and contusions) [25, 76]. The clinical history of subjects is recommended for distinguishing from CMBs.
Small haemorrhages near infarction and ICH areas	Visible on T2, T2*-weighted, FLAIR, DWI sequences. Sometimes, they differ from CMBs in size. Would be easier to use infarctions as exclusion biomarkers.
Flow voids	Do not show blooming effect. They have linear/ curvilinear tubular structures extending through contiguous slices, usually appear in the cortical regions and are visible on T2 images [25]
Calcifications	They are usually found in the basal ganglia. However, calcifications could also occur in choroid plexuses and the pineal gland. They appear as hyperintense blobs on CT images [77]
Partial volume artefacts (air-bone interfaces)	Do not show blooming effect. Particularly seen in frontal and temporal lobes (orbit and mastoid bones) and at the edges of cerebellum. Distinguishable from CMBs using location priors and observation of contiguous slices.

Table 1.2: CMB mimics and their distinguishing characteristics.

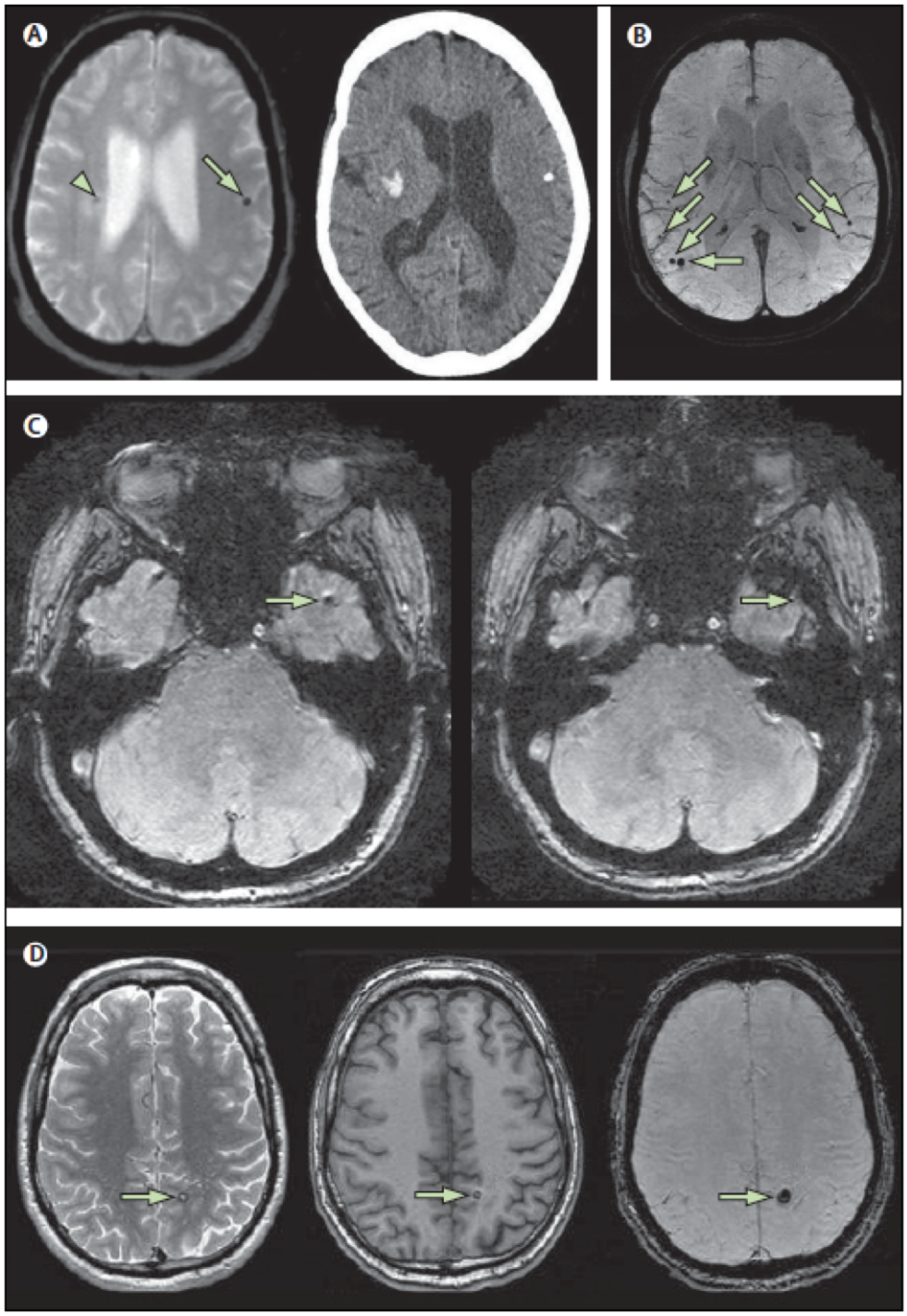


Figure 1.8: Examples of CMB mimics. (a) Calcification in the left hemisphere (surrounded by hyperintense signal). In the CT image on right, the calcification is shown as an hyperintense area. (b) Minimum intensity projection of axial slices shows flow voids nearer to the brain surface. (c) Two contiguous axial slices of T2*-weighted GRE images showing partial volume artefacts. The dark spot would have been misidentified as CMB by observing left slice alone, but the slice below shows that it is actually a partial volume artefact at the air-bone interface (in this example, sphenoid bone). (d) cavernous malformations shown on proton density-weighted (left), T1-weighted (middle) and T2*-weighted (right) images. Image courtesy: *Greenberg et al, Lancet Neurology, 2009.*

1.4 Objectives of this work

Our **primary objectives** are:

- To study the distribution of WMHs at the population-level with respect to various clinical factors.
- To identify and segment WMHs on structural MR images with equally high accuracy in all regions of the brain white matter.
- To detect CMBs automatically in large datasets.

Given our primary research objectives, our **secondary objectives** are as follows:

- To obtain population-level WMH distribution maps with respect to various demographic factors, and determine their statistical significance.
- To experiment with various strategies at different steps of BIANCA for improving WMH segmentation.
- To explore the possibility of using machine learning techniques, especially deep learning models for WMH segmentation as a potential and better alternative to BIANCA.
- To automatically preselect candidate subjects with CMBs in a large dataset.
- To detect CMBs at the voxel-level for individual subjects.
- To characterise CMBs lesion load based on their number and size and provide an automated rating of CMBs in different regions of the brain.

1.5 Contributions of this work

The research presented in this thesis has contributed to the development of the following:

- **Bayesian inference-based modelling algorithm:** A robust modelling approach for obtaining voxel-wise probability map at the population-level has been proposed.

- **Locally adaptive thresholding of lesion probability maps:** A novel thresholding method, adaptive to local probabilities, has been presented as an alternative for the global thresholding. Our proposed method is robust across lesion loads and pathological conditions.

- **Ensemble of U-Nets in triplanar architecture with spatially weighted loss function for WMH segmentation:** An ensemble model has been proposed for WMH segmentation using anatomical information regarding WMHs, and has been extensively evaluated on various datasets.

- **Pipeline for CMB candidate subject preselection:** A learning based pipeline has been described for the preselection of CMB candidates subjects in large datasets and evaluated on different image modalities.

- **A multiscale deep learning-based CMB detection on individual subjects:** The proposed method involves initial candidate detection, candidate discrimination and false positives removal for CMB detection, with efficient removal of CMB mimics.

- **An atlas-based method for automated rating of CMBs:** Structural atlases from the MNI space have been used for determining CMB ratings based on the MARS scale.

1.6 Thesis outline and overview of the research work

This section outlines the overall organisation of the thesis. Figures 1.9 and 1.10 show the overview of the proposed workflow for WMHs and CMBs respectively, catalogued by the corresponding chapter numbers.

Chapter 2 reports the relevant literature for our research work on WMHs and CMBs. The chapter provides an account of the literature available for WMH segmentation and CMB detection to the best of our knowledge. It also describes the population-level studies that analyse the relationship between the distribution of WMHs, CMBs and clinical factors.

Chapter 3 describes our proposed method for modelling the distribution of WMHs using Bayesian inference within a population, with respect to a clinical factor. The individual subject-level lesion probability maps obtained from BIANCA provided a good starting point for the development of our algorithm for obtaining the population-level map. The chapter explains (i) the proposed Bayesian modelling algorithm (shown in the top panel of figure 1.9), (ii) results of our algorithm on a simulated dataset, followed by the evaluation on the real datasets and (iii) comparative analysis of the performance of our algorithm with existing methods. The preliminary results of the algorithm were presented at the 2017 OHBM conference, Vancouver (Sundaresan et al., OHBM 2017) and the final results of this work, including the results of the comparative analysis, are reported in Sundaresan et al., 2019 [78].

In chapter 4, we improve the existing WMH segmentation with BIANCA. The results obtained from chapter 3 were used for part of the experiments done in this work. The chapter explains the strategies for experiments performed on different steps of BIANCA: (i) feature extraction, (ii) classification and (iii) thresholding of probability maps. The middle panel of figure 1.9 shows the schematic representation of the work explained in the chapter. The main contribution of this work is our proposed method for locally adaptive thresholding of subject-level lesion probability maps. The preliminary results of the adaptive thresholding algorithm were presented at OHBM conference 2018, Singapore (Sundaresan et al., OHBM 2018). The final results of this work, including the other two strategies for the improvement of BIANCA, are reported in Sundaresan et al., 2019 [79].

Chapter 5 addresses our work on the development of a hybrid deep learning model for WMH segmentation. The chapter outlines the proposed model architecture and

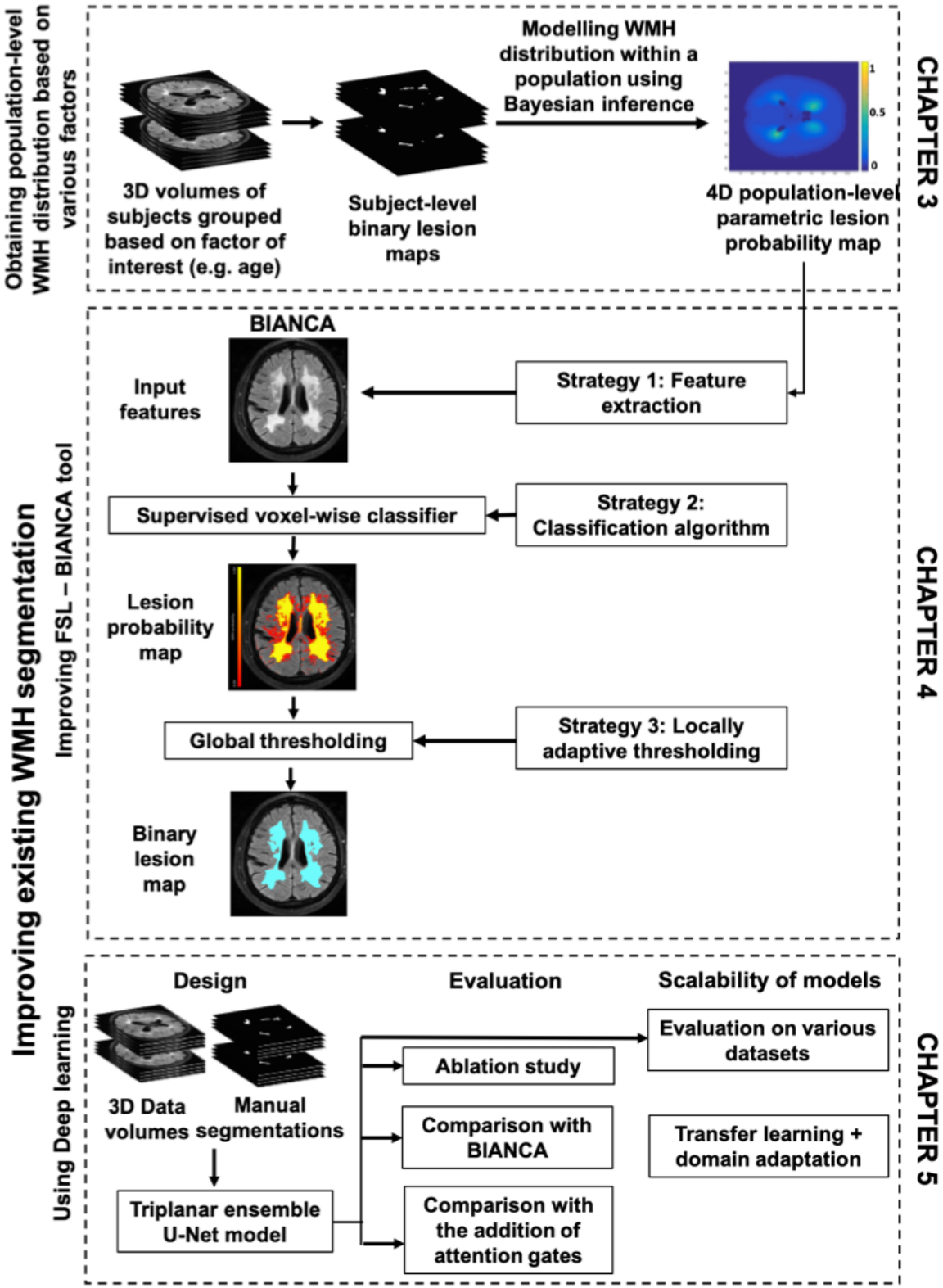


Figure 1.9: Flow chart depicting the overall workflow for WMH segmentation and analysis.

discusses various data considerations such as limited number of training subjects, pre-processing and augmentation. The chapter reports (i) evaluation results on various datasets, (ii) results of comparison of our proposed model with BIANCA. Finally, the generalisability of the model on various datasets has been analysed. The chapter explains about the domain adaptation of our model on various datasets with different characteristics by using transfer learning and domain adversarial training. The preliminary evaluation results of the proposed model have been submitted to ISMRM conference 2020, Sydney (Sundaresan et al., ISMRM 2020) and the corresponding publication is currently in preparation. The bottom panel of figure 1.9 depicts the overview of the work explained in the chapter.

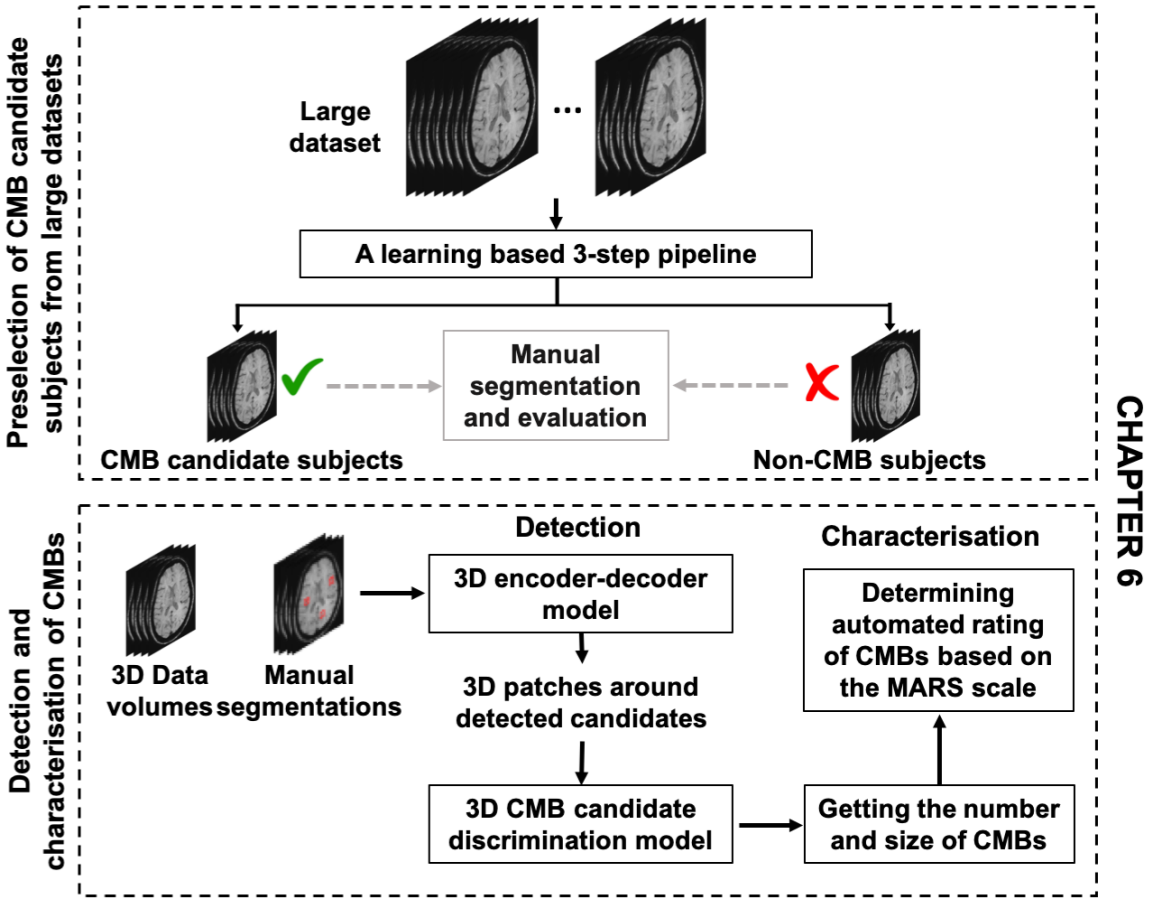


Figure 1.10: Flow chart depicting the overall workflow for CMB detection and characterisation.

Chapter 6 describes our work on the end-to-end automated analysis of CMBs in a large dataset. The flowchart in figure 1.10 shows the overall workflow of CMB detec-

tion. We considered the detection of CMBs in a large dataset as a two-level problem and hence handled it in two phases: (i) preselecting the CMB candidate subjects for further analysis and obtaining manual segmentations (top panel in figure 1.10) and (ii) detection and characterisation of CMBs in individual subjects (the bottom panel in figure 1.10). The chapter explains the work done at both levels and focuses on the challenges at each level. The work described in this chapter was done in collaboration with the TICH2 trial group [80, 81] at University of Nottingham, UK. Our collaborators provided the TICH2 trial data and manual segmentations for the development of both the preselection pipeline and the deep learning model for CMB detection. In the first phase, we developed the preselection pipeline and identified the potential CMB subjects in a large dataset (UK Biobank). Our collaborators will provide the manual segmentations for these selected subjects, which would enable us to proceed with the large-scale validation in future. During the second phase, we developed a deep learning model for the detection of CMBs and evaluated the model on various datasets. The chapter provides a detailed explanation of work done in both phases and it reports (i) the evaluation results of the preselection pipeline and (ii) the results of our proposed CMB detection model on various datasets. Finally, the chapter describes our work on obtaining clinically more useful information on CMBs such as size, number and rating for further analyses. The preliminary results of the work done in the first phase were presented at 2019 OHBM conference, Rome (Sundaresan et al., OHBM 2019).

Finally, chapter 7 provides the conclusions and possible future directions of our research. The details of datasets used for various experiments in our research are described in appendix A. A brief overview of various supervised classification algorithms used in chapter 4 are provided in appendix B. The performance measures used for the evaluation of our experiments are described in appendix C and finally, a sample MARS rating form has been provided in appendix D.

Chapter 2

Literature review

This chapter presents a review of the existing literature for WMH and CMB detection and analysis at both the population- and subject-level.

Section 2.1 provides an overview of various clinical correlates of WMHs and CMBs that have been established in large population studies in the literature. Section 2.2 provides a review of the existing work done in WMH segmentation and analysis. For WMHs, section 2.2.1 reports work done towards modelling the population-level voxel-wise distribution of WMHs, and studying the relationship between WMH voxel-wise distribution and various factors of interest. At subject-level, section 2.2.2 provides an account of WMH segmentation methods proposed in the literature, including the recent development in deep learning methods. For CMBs, section 2.3.1 provides a detailed review of various types of existing CMB detection methods. Although this is not an exhaustive systematic review, for clinical impact, we performed a search through keywords (e.g. white matter hyperintensities, cerebral microbleeds, age, hypertension etc.). For methods, we used a combination of search through keywords and going through the references of important work.

2.1 Clinical impact of WMHs and CMBs

Clinical impact of WMHs: This section focuses on some of the most important large scale studies that were done on thousands of subjects for observing the clinical correlates of WMHs.

The crucial large population studies include the Rotterdam study [82], the Leukoaraiosis and Disability (LADIS) study [83], the Lothian birth cohort study [84], AGES-Reykjavik Study [85], the Oxford vascular study (OXVASC) [86], the UK Biobank study [87] and the INtervention project on cerebroVAscular disease and Dementia in Ebersberg (INVADE II) study [88]. In the earlier studies [82, 83], semi-quantitative ratings, involving visual rating [89], were used and significant correlations of WMHs with cognitive decline [82, 90], vascular risk factors [90], gait imbalance [91] and stroke [32] were reported. The development of automated WMH segmentation methods aided in accurate quantification and in studying the relationship between the clinical correlates and spatial distribution of lesions. For instance, deep WMHs (DWMHs) have been associated with brain atrophy [84] and WMH volume was found to be significantly higher in apathy subjects [85]. Significant correlations have also been established between WMH load and cognition [88, 92, 86]. However, a clear relationship between the spatial distribution of WMHs and apathy or cognition has not been established. In the UK Biobank study, the largest MR population ever (currently acquired for $\approx$ 48,000 subjects, will reach 100,000), WMHs have been found to be positively correlated with age, blood pressure [87] and APOE $\epsilon 4$ genotypes [93].

Regarding the individual clinical impact of PWMHs and DWMHs, PWMHs have been shown to have higher correlation with age and cognitive decline [54, 94], while DWMHs have been associated with depression [94, 26]. In the existing literature, both PWMHs [36] and DWMHs [50] have been associated with high blood pressure (BP). Therefore, while WMHs are associated with high BP in general, a clearer localisation of the effect is lacking. Studies show differences in the functional and histopathological characteristics of PWMHs and DWMHs [30, 95], indicating slightly more ischemic damage in DWMHs when compared to PWMHs [31, 96].

From the above studies, WMHs have been associated with age, BP, gait imbalance, cognitive decline and brain atrophy. They have also been commonly found in subjects with vascular risk factors and depression. However, the results are still contradictory for these associations (e.g. positive correlation of BP with DWMHs in some literature,

while with PWMHs in others), and there is a lack of clear understanding regarding the origin and the impact of WMHs. Hence, further work is needed in terms of detection and studying the clinical associations of WMHs.

Clinical impact of CMBs: While several studies exist for CMBs [27, 97, 25], this section provides a review of the studies done at larger scale (200 subjects or more) and highlights the commonly established clinical correlates and risk factors of CMBs.

In the large population studies, usually CMBs were detected and rated manually based on the MARS scale [73]. Some of the large population cross-sectional studies include the Austrian Stroke Prevention Study [98], the Japanese study [99, 100], the Japanese 'brain docking' study [101], the Framingham study [102], the AGES-Reykjavik Study [68, 103], the Rotterdam scan study [63, 64], the Japanese Kishiwada City Hospital study [104] and the Japanese rural community study [105]. Strong associations have been established between CMBs and risk factors such as age [98, 101, 102, 63, 64, 105], hypertension [98, 99, 101, 63, 64], smoking habits [99, 63, 64], sex (found more commonly in male participants) [102, 105], white matter damage [98] and several vascular risk factors [104, 97]. Moreover, CMBs have often been found in subjects that have experienced transient ischemic attack (TIA) [106, 54, 97, 107, 108, 109] and vascular dementia [110]. CMBs have also been correlated with microangiopathy, hemorrhagic stroke [100], APOE $\epsilon 4$ [68] and diabetes [68]. Cordonnier and colleagues [111] studied 772 subjects and found that CMBs are associated with age, WMHs and lacunar infarcts, which was later confirmed in the Rotterdam scan study [63, 64]. CMBs have been shown to affect neurological functions causing gait disturbances [112] and are found in CADASIL subjects [113, 114, 115]. Strong correlations have been found between CMBs and cognitive decline [103, 105, 116, 117, 20, 118, 119, 110] and cerebral amyloid angiopathy (CAA) in elder subjects [120]. However, recent studies on their predictive value for long-term cognitive outcome have shown inconsistent results, therefore the specific role of CMBs in cognitive impairment remains unclear. A few longitudinal studies have also been performed to observe the associations between the increase in CMBs and various

factors. A longitudinal study performed by Gregoire and colleagues [107] using the MARS rating scale showed that 50% of subjects with CMBs had more risk of developing new CMBs over a mean interval of around 5 years, indicating the progressive nature of CMBs. Another study [121] performed on a European cohort of 487 subjects, with a median follow-up time of 2.2 years, observed that patients with CMBs have higher risk of developing new ischemic strokes.

From the above studies, the common risk factors of CMBs are age, hypertension, smoking and APOE $\epsilon 4$ alleles. Also, CMBs have been commonly associated with white matter damage, cognitive decline, stroke and dementia. In chapter 6, we propose an automated pipeline to preselect the candidate CMB subjects in big datasets. We apply the above factors as preliminary conditions to shortlist the candidates for the application of the preselection pipeline. Note that this shortlisting might affect the possibility of considering the presence of CMBs as an independent biomarker. However, shortlisting the subjects before preselection would considerably increase the accuracy/specificity of the preselection in an efficient way in very large populations.

2.2 WMH analysis and segmentation

2.2.1 Population-level modelling of WMHs

As mentioned earlier, WMHs have been associated with various clinical factors, the most common ones being age, BP, cardiovascular risk factors, cognitive decline, stroke and atrophy. However, in addition to the overall correlations, obtaining voxel-based population-level WMH maps would help in characterising the relationship between WMH spatial distribution and various clinical and demographic factors.

The simplest method used in the literature for obtaining voxel-wise lesion distribution maps with various parametric factors (such as age, cognitive function and so on) consists of averaging subject-level lesion probability maps that are grouped according to those parametric factors, as proposed by Rostrup and colleagues [50]. A limitation of this approach is that it does not necessarily provide continuity or smoothness in

space or across the parametric dimension for the lesion probabilities. This limitation is quite prominent in the population with small number of subjects or with lower lesion load (sparse lesions), where there are not enough lesion samples available at the voxel-level. Alternatively, several *mass-univariate* methods have been proposed. For instance, Bates and colleagues [122] proposed voxel-based lesion-symptom mapping (VLSM), while Charil and colleagues [123, 124] proposed voxel-wise linear regression of lesion probabilities against various factors. In their later work [124], they studied the correlation between the cortical thickness and various factors including lesion distribution in multiple sclerosis (MS) using voxel-wise linear regression. However, these methods are based on models that work independently in each voxel and do not capture the relationship between neighbouring voxels. Hence, they are not ideal for modelling the lesion distribution, since lesions are clustered regions rather than isolated voxels. Moreover, the linear models used in these methods are not optimal for binary data [124, 122].

Spatially varying coefficient processes establish a local spatial relationship between the coefficients of regression models [125, 126, 127]. For example, Ge and colleagues [127] proposed Bayesian Spatial Generalised Linear Mixed Model (BSGLMM) based on spatially varying coefficients to determine the relationship between the spatial distribution of lesions and subject specific covariates such as MS subtype, age, gender, disease duration and disease severity measures. The spatially varying coefficients are modelled jointly using a multivariate pairwise difference prior model, a particular instance of the Multivariate Conditional Autoregressive model. However, the main implementation problem with spatially varying coefficients is that the dimension of the correlation matrices involved is very high and their inversion becomes computationally infeasible for very large imaging datasets [127].

Several solutions have been proposed to overcome this problem. Some of the most effective methods are Gaussian predictive processes [128], functional principal component analysis (FPCA) approach [129, 130] and covariance tapering [131]. Crainiceanu and colleagues [132] used a truncated Karhunen-Love expansion to estimate spatially

varying coefficients also as an alternative solution. However, methods such as Gaussian predictive processes require fine-tuning of several parameters to precise values for good results [128]. In FPCA, the choice of optimal number of components becomes a crucial parameter to be tuned for the trade-off between bias and variance. More components, despite reducing the bias, result in higher variance due to the addition of random error [133]. Also, FPCA could provide a spatially less continuous model for sparse data.

In chapter 3, we have proposed our own novel method to model the distribution of WMHs within a population, in a spatially continuous manner, with respect to various factors of interest.

2.2.2 Subject-level WMH segmentation

Obtaining a generalisable WMH distribution model and studying the population-level clinical correlates of WMHs relies on their accurate segmentation. As explained in section 1.2.3 of chapter 1, correct automated segmentation of WMHs should be able to identify subtle DWMHs as well as properly delineate PWMHs. More generally, getting the precise location of WMHs at the single subject level would be useful to define patterns for discriminating between normal and pathological ageing.

Several methods are available [51, 134], including the FSL tool, BIANCA [53], for WMH segmentation. This section provides a critical review of various semi-automated and automated methods for WMH segmentation. This review also includes some of the relevant work done on segmentation of other pathological signs (e.g. MS lesions), that share some of the characteristics of WMHs on MR images, due to their potential application to WMHs.

2.2.2.1 Semi-automated methods for WMH segmentation

Semi-automated methods require manual intervention as a part of the algorithm, usually to initialise the segmentation through seed placement [135, 136, 137, 138] and/or at the final step for the exclusion of false positives [136, 139, 140] or brain/region of

interest (ROI) extraction and refinement [141, 140, 142]. Itti and colleagues [135] segmented WMHs using recursive region growing from manually provided seed points. Alternatively, Payne and colleagues [136] used a non-parametric statistical method [137] involving manual seeding of voxels for various tissues types. The identified areas were further checked manually to distinguish WMHs from other tissues. Burton and colleagues [141] used morphological operations and a final step of manual editing to remove non-brain voxels. Sheline and colleagues [143, 144] used bispectral fuzzy class means for segmenting tissue types and WMHs. The peak search on the 2D histogram had to be done semi-automatically, and had to be further constrained by a manually selected ROI. Kawata and colleagues [139] proposed a method that required manual intervention for both seed initialisation in the region growing step and removal of false positives (FPs). Another method, named as Lesion Explorer was proposed by Ramirez and colleagues [140] for both tissue segmentation and brain extraction. The brain was subdivided into small 3D regions and thresholds were estimated for each of them using their intensity histograms. However, the final step of manual intervention for WMH segmentation took 15-20 min per subject. In general, the manual intervention required for semi-automated methods takes a long time and could introduce inter and intra-operator variability, especially in big datasets. Hence, a fully automated algorithm is preferable as it avoids manual intervention and leads to efficient and more objective WMH segmentation.

2.2.2.2 Automated segmentation of WMH based on features

Automated methods provide an end-to-end single- or multi-step pipeline, without manual intervention on every image. Existing methods for automated WMH segmentation have utilised various characteristics of WMHs such as intensity, spatial distribution, contrast and texture. In addition, anatomical information regarding brain white matter and subcortical structures have also been utilised for improving the segmentation accuracy. This section provides a review of automated methods for WMH segmentation based on the characteristics used: (i) intensity (thresholding,

intensity and histogram-based), (ii) intensity and anatomy (probabilistic and atlas-based), and (iii) intensity, anatomy and appearance (texture, shape, contrast and so on) of WMHs (morphological operations and co-occurrence matrices). Figure 2.1a shows percentages of the WMH segmentation approaches based on various features discussed in this review. Table 2.1 provides the comparison of various exiting automated WMH segmentation methods based on the type of features, methods used and their performance.

Intensity: The most crucial and discriminating characteristic of WMHs is their intensity. WMHs are brighter than normal white matter on FLAIR images (and are usually hypointense, to a lesser degree, in T1-weighted images). Hence, most of the straightforward methods in the existing literature involve thresholding [145, 146] and histogram-based modelling [147, 148, 149]. The thresholding methods, though simpler to implement, usually require an additional post-processing step for FP reduction [145, 146]. Also, determining an optimal threshold is quite challenging and varies according to the data. For instance, Jack and colleagues [145] determined the threshold using regression analysis, while Yoo and colleagues [146] determined the relationship between thresholds and lesion load. Simões and colleagues [147] fitted Gaussian mixture model (GMM) on histograms, with a modified expectation-maximisation step to consider the probabilities of neighbouring voxels. In general, thresholding and histogram based methods do not work well on pathological images or in the presence of atrophy. For instance, Wei and colleagues [148] had to develop two individual statistical models for lesion detection based on the subject groups (MS and normal older subjects). Hence, using intensity alone typically results in spurious detections and is not robust across datasets.

Intensity and anatomy: In order to overcome the drawback of intensity-based approaches, a few approaches combined intensity and the spatial location of lesions for more accurate detection. These include outlier detection in histograms constrained spatially [150, 151, 149], Bayesian framework with intensity, location and anatomical priors [152, 153, 154, 155, 156], Markov random field (MRF) models [157, 158, 159]

and atlas-based segmentation [160, 161]. Methods that detect WMHs as outliers initially perform tissue segmentation on histogram by GMM modelling [154, 155, 149], and then detect WMHs as outliers due to their extreme intensity values. They add spatial constraints to the detected outliers using anatomical priors [154, 155, 156, 149]. Ong and colleagues [150], in their work on MS lesions, used modified trimmed mean to truncate the histogram ends, so that the extreme outliers could become probable indicators of lesions. Another interesting approach by Yang and colleagues [162] involved detection of WMHs as outliers spatially on a probability map. They proposed a novel method by calculating the joint probability maps (χ^2-maps) of WMH occurrence in T1- and T2-weighted images. WMHs were detected as outliers since they showed very low joint probabilities when compared to healthy brain tissue voxels due to the difference in their intensity characteristics. In general, the intensities/probabilities of subtle lesions might not be extreme and hence get ignored during the estimation of overall subject-level probability threshold. Hence the methods provided relatively low performance in subtle pathological cases [150]. In those cases, the use of a combination of intensity and location priors at the voxel-level in Bayesian approaches [152, 153] provided more accurate detection. Wu and colleagues [160] used multi-atlases as likelihood functions and as a spatial constraint for the estimation of posterior probabilities of WMHs. In addition to having voxel-wise anatomical and intensity priors, the effect of neighbouring voxels using MRF model have been shown to improve WMH detection [157, 158]. Generally, while probabilistic and atlas-based methods perform well on most cases, the presence of image artefacts and other pathological signs such as infarcts (due to their deviation from normal prior/atlas) could lead to the misclassification of voxels. Therefore it is useful to consider, along with voxel-wise features, lesion-level characteristics such as shape and contrast for the detection of lesions in order to obtain accurate segmentation.

Intensity, anatomy and appearance: A wide range of approaches have been proposed in the literature based on texture, contrast and appearance, the most common among them based on morphological operations [163, 164] and region growing [165].

Morphological methods have been used not only for WMH detection [163, 165], but also for post-processing steps such as FP reduction [166] in combination with other methods. Shi and colleagues [163] adopted a coarse-to-fine approach using 4 steps of morphological operations: binary dilation, grayscale closing, binary reconstruction and grayscale reconstruction. However, the main drawback of these methods is that they require parameter tuning (e.g. size of the morphological kernels, minimum contrast threshold for region growing). The optimal values for these parameters vary with datasets and affect their generalisability. Another texture-based method using multisort co-occurrence matrices was proposed by Kruggel and colleagues [167]. Co-occurrence matrices contain information about pairs of voxels, and could be used to derive the level of similarity between neighbouring voxels. An interesting method was proposed by Khademi and colleagues [168] with an aim of detecting WMH to sub-voxel precision, accounting for partial volume artefacts. Since these artefacts occur along the edges, they model the mixture of proportion of various tissues at the edge voxels by considering the change in their values from the 'pure' tissue types. While the method provides reliable edge detection of WMHs, local intensities of the voxels on the borders are prone to be affected by the presence of noise. Another recent method proposed by Rachmadi and colleagues [169] was based on 'limited one-time sampling irregularity age map (LOTS-IAM)' algorithm. LOTS-IAM worked at a patch-level at multiple scales and was based on the assumption that any WMH patch should have a higher average of maximum and mean intensity values with respect to a normal patch. Generally, morphological methods are affected by the presence of image artefacts, especially noise [166, 168].

2.2.2.3 Automated segmentation of WMH based on the classification method

Automated machine learning (ML) based methods could be either unsupervised (does not require manually labelled data and works based on inherent characteristics of the data) or supervised (requires manual labelling and *learns* to make decisions from the input data). This section provides a review of methods based on their approaches

such as (i) unsupervised clustering methods and (ii) supervised classification and regression algorithms. Figure 2.1a shows percentages of the unsupervised clustering methods and supervised classifiers for WMH segmentation that are discussed in this review.

Unsupervised clustering methods: Out of the various clustering algorithms, fuzzy c-means clustering algorithms have been most widely used [170, 171, 172, 173, 174, 175]. In fuzzy c-means clustering, datapoints are clustered based on their degree of membership, determined by fuzzy sets (e.g. intensity, contrast), towards the individual clusters with a mean value c (e.g. mean intensity). Some of the commonly used fuzzy sets or features were intensity [170, 171, 174] and location [171, 173]. Anitha and colleagues [175] proposed geostatistical fuzzy clustering by adding a spatial penalty to the objective function of the clustering algorithm. Often, the clustering step had to be combined with additional steps such as initial seed selection for clusters [172] or postprocessing thresholding/consensus step [174]. In general, while clustering methods are good at detecting lesions, underestimation of WMH occurs at the edges of the lesions (even with spatial context [175]) or with low lesion loads and hence requires some additional steps of processing [174].

Supervised classification algorithms: Supervised machine learning (ML) based classification algorithms have been widely used for WMH segmentation. These algorithms, although requiring manual segmentation for training, constitute the most popular class of methods for segmentation. Various supervised classifiers like K-nearest neighbour (KNN), support vector machines (SVM), random forest (RF) and adaptive boosting (AdaBoost) algorithms have been used either individually or in combination for WMH segmentation. More details on these algorithms are provided in Appendix B. Figure 2.1b shows the bar plot of various classifiers for lesion segmentation reported in this review.

One very commonly used algorithm is KNN [176, 55, 177, 56, 178, 179, 180], including BIANCA [53]. With respect to other algorithms, the KNN algorithm is simpler to implement, with fewer parameters to fine-tune and has been shown to provide

good results [55, 177]. Other classifiers that were often used in the literature include artificial neural networks [181, 57], AdaBoost [182], SVM [183, 184] and RF [185]. In [184], Damangir and colleagues used a segmentation framework CASCADE based on a cascade of SVMs and achieved accurate segmentation results. With individual classifiers, additional information such as anatomical priors [180], preprocessing steps involving morphological operations [182] or postprocessing steps [185] are required at times for improving the segmentation accuracy.

Another approach used to improve the segmentation and to avoid excessive additional processing is to implement combinations/ensemble of classifiers, with SVM and RF being the preferred options. SVM classifiers have been used in combination with the AdaBoost classifier [34, **?**], RF [187] and KNN [188, 189]. RF has been used along with the AdaBoost classifier [190, 191] and MRF model [192]. Stone and colleagues [193] used RF classifiers built upon Advanced Normalization Tools (ANTs). Manjón and colleagues [194] used an ensemble learning strategy to minimise the classification error, exploring both bagging and boosting techniques. Hence the RF classifier has often been used in combination with other classifiers or with additional steps for FP reduction. Lesion prediction algorithm (LPA) in the Lesion segmentation Tool (LST) [195], a freely available toolbox from SPM [196], is based on a supervised logistic regression model with adaptation to fit large-scale regression models. LPA has been validated and shown to provide good results [197, 198], comparable with those from BIANCA [53].

Comparative analyses: Other than individual supervised algorithms, a few comparative analyses [134, 199, 200, 169] have also been performed. Kloppel and colleagues [134] used T1 and FLAIR from 20 subjects with MCI and dementia to compare Otsu thresholding, SVM and KNN. The features included intensities extracted from spherical neighbourhood, anatomical position and Gabor filter outputs. SVM provided the best segmentation, further improved with intensity normalisation. Sweeney and colleagues [199] used T1, T2 and FLAIR from 98 patients with MS and compared logistic regression, NN, SVM, quadratic discriminant analysis (QDA), linear

discriminant analysis (LDA), GMM, KNN, RF and superlearner (ensemble of RFs). They used features based on unnormalised, normalised, smoothed and moments of intensities. They found that with more processed features (smoothed and moments of intensities), linear classifiers provided better performance. Similar classifiers such as Naive Bayes, LDA, QDA, logistic regression, decision trees, RF, KNN , SVM, Bagging, AdaBoost were compared by Dadar and colleagues [200] on 201 subjects with Alzheimers disease and MCI. The features were extracted from T1w, T2w, PD, and FLAIR based on their intensities, probability density functions of intensities from WMH and non-WMH voxels and their ratios. RF and LDA gave the best performance amongst nonlinear and linear methods respectively. Rachmadi and colleagues [169] compared their LOTS-IAM method with LST, SVM, RF and patch-based 2D CNN methods such as DeepMedic with global spatial information [201], UResNet [202] and U-Net [58]. Features were extracted from T2 FLAIR from 50 images with AD and MS, and found that LOTS-IAM outperformed LST [195] and supervised ML methods, while the DL methods [201, 202, 58] provided the best performance for WMH segmentation. In general, the comparative methods used a consistent preprocessing pipeline: registration, skull stripping, bias correction and intensity normalisation/scaling. Data imbalance between the classes (WMH/non-WMH) was handled in [199] by making features more discriminative, while [134] sampled equal, fixed amount (500 voxels) of data from both classes, which could at times lead to undersampling. Also, [199] and [200] determined the effect of feature subsets on the classifier performance, and found that FLAIR features were the most useful, and extracting neighbourhood information (intensity and location) improved the performance. The main advantage of [200] and [169] is that they performed the validation on data from different scanners/sites for improving model generalisability. Additionally, the effect of lesion size on classifier performance was analysed in [200]. AdaBoost, KNN, SVM and RF were more accurate on medium/large WMHs, but slightly oversegmented small WMHs.

From our comparison of the performance of exiting methods in table 2.1, it can be observed that the use of anatomy and appearance based features in addition to inten-

sity improved the performance [159, 163, 171, 162, 195]. From the data perspective, a few methods used only small amount of data for evaluation [150, 174, 55, 56]. Some methods used data with different pathological conditions or acquired with different scanners [173, 164, 195, 169, 147, 184, 53], to improve method generalisability, while others might require extensive validation by retraining or fine-tuning parameters. Regarding method availability, only a few methods are available as open source tools [195, 184, 53], out of which [195] was developed and trained on MS subjects. Out of supervised methods, most of the methods were based on KNN [180, 56, 55], SVM [184], RF [190], AB [182, 190] and NN [57]. Especially, RF and SVM performed well in the comparative analyses [134, 200]. Therefore, we have explored the potential of RF, SVM, AB and NN in improving BIANCA performance in chapter 4.

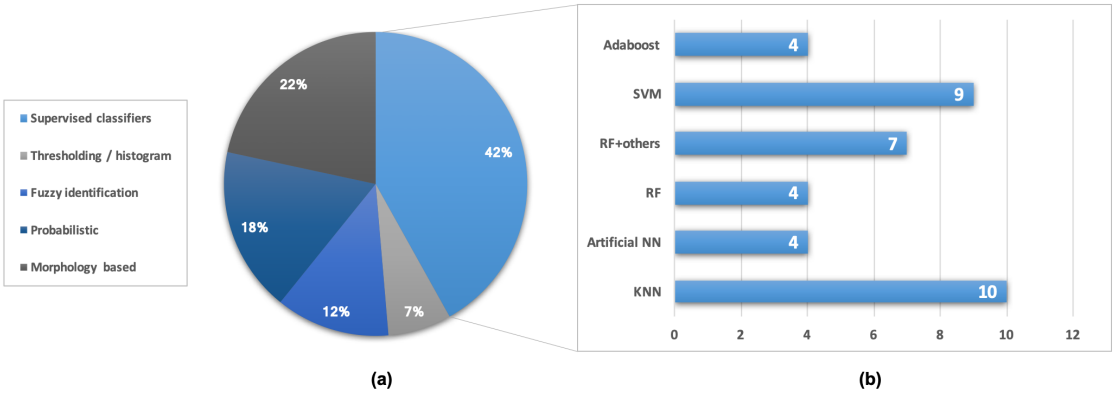


Figure 2.1: Automated WMH segmentation methods split-up. (a) Pie-chart of various WMH segmentation approaches discussed in the review (shown in percentages), (b) Bar plot of the reported supervised classifiers used for WMH segmentation (including comparative analyses).

2.2.2.4 Deep learning for segmentation of WMHs

The conventional ML methods are limited in their ability to process data in their original form, and often require the estimation of hand-crafted features. Deep learning (DL), on the other hand, allows computational models of multiple layers to learn data representations at different layers of abstraction [203]. However, compared to ML, they require a large amount of training data, which are often difficult to obtain in medical images, and are computationally intensive. The increase in computational resources such as GPUs and constant updates in DL concepts by an active computer

Method	Type[1]	Alg.[2]	Population - study (subjects)	Image modalities	Dice[3]	Sens[4]	Spec[5]
Wang et al., [149]	I,U	GMM	Singapore ageing cohort (272)	T1, T2, FLAIR	0.77	0.81	0.97
Ong et al., [150]	IA,U	Ad. mean	HC (19)	T1, FLAIR	0.47	0.67	0.40
Sudre et al., [154]	IA,U	GMM	BrainWeb (simulated)	T1, T2, PD	0.68^{A}	74.1^{A}	-
Jeon et al., [159]	IA,U	MRF	SVD - AMPETIS (45)	PD, T2, FLAIR	0.90	-	-
Yang et al., [162]	IA,U	χ^2 RF	Dementia - LEILA (30)	FLAIR	0.81	ROC	ROC
Shi et al., [163]	IAAp,U	Thr	Acute infarction (91)	DWI, T1, FLAIR	0.84	0.80	-
Samaille et al., [164]	IAAp,U	WS	MCI, CADASIL (67)	T1, FLAIR	0.72	-	-
Schmidt et al., [195]	IA,U	MRF	HC, MS (70)	T1, FLAIR	0.75	0.80	0.99
Khademi et al., [168]	IAAp,U	PVA	Subjects with lesions (24)	FLAIR	0.83	0.82	0.99
Rachmadi et al., [169]	IAAp,U	LOTS-IAM	MCI, Alzheimer's - ADNI (50)	T2, FLAIR	0.59	-	-
Admiraal-Behloul et al., [171]	IA,U	FZ	Vasc. disease - PROSPER (100)	PD, T2, FLAIR	0.75	-	-
Seghier et al., [173]	A,U	FZ	HC, stroke (64)	T1	0.64	ROC	ROC
Gibson et al., [174]	I,U	FZ	WM disease (18)	FLAIR	0.81	-	-
Herskovits et al., [152]	IA,S	BY	Diabetes - ACCORD-MIND (42)	T1, T2, FLAIR	0.60	ROC	ROC
Yoo et al., [146]	I,S	Var thr	Longitudinal/ dementia - Korean study (32)	FLAIR	0.76	-	-
Simões et al., [147]	I,S	GMM	HC, MCI (40)	FLAIR	0.68	-	-
Anbeek et al., [55]	IA,S	KNN	Arterial vascular disease (20)	T1, IR, PD, T2, FLAIR	0.80	0.97	0.97
de Boer et al., [56]	I,S	KNN	HC - Rotterdam scan study(6)	T1, PD, FLAIR	0.72	0.79	-
Steenwijk et al., [180]	I,S	KNN	Hypertension (20)	T1, FLAIR	0.84	-	-
Dyrby et al., [57]	IA,S	NN	Elderly subjects - LADIS (639)	T1, T2, FLAIR	0.56	-	-
Beare et al., [182]	IAAp,S	AB	HC - TASCOG (232)	T1, T2, FLAIR	0.58	-	-
Damangir et al., [184]	IA,S	SVM	HC, Alzheimer's, dementia - DemWest (102)	T1, FlAIR	-	0.90	0.99
Ghafoorian et al., [190]	IA,S	RF+AB	SVD (50)	PD, T2, FLAIR	-	0.73	-
Griffanti et al., [53]	IAAp,S	KNN	Neurodegenerative - OPTIMA (85)	Multimodal	0.75	-	-
	IAAp,S	KNN	Vascular - OXVASC (474)	Multimodal	0.52	-	-

[1] Type: I - intensity, IA - intensity + anatomy, IAAp - intensity + anatomy + appearance, S - supervised, U - unsupervised.
[2] Algorithm used: GMM - Gaussian mixture model, Var thr - variable thresholding, Ad. mean - adaptive trimmed mean, BY - Bayesian classifier, MRF - Markov random field, χ^2 RF - χ^2 random field [162], Thr - thresholding, WS - watershed segmentation, PVA - partial volume averaging, LOTS-IAM - limited one-time sampling irregularity age map, FZ - fuzzy clustering/inference, KNN - K-nearest neighbour classifier, NN - neural networks, AB - AdaBoost, SVM - suppport vector machine, RF - random forest.
[3,4,5] Dice, sensitivity, specificity: ^{A} - average of mild, moderate and severe cases, ROC - only ROC curves shown without individual Sens and Spec values.

Table 2.1: Comparison of the existing automated WMH segmentation methods (non-DL approaches) in the literature.

vision community have made DL a quickly emerging tool over the last few decades. Techniques such as data augmentation (inflating the amount of training data) [204] have been used to overcome the shortage of training data, reduce the possibility of overfitting and improve the accuracy of the DL models. Based on representation learning, DL architectures are multilayered stacks of non-linear modules that compute input-output mappings during training. One of the most popular feed-forward DL architectures, the convolutional neural network (CNN), has been widely used for several tasks in computer vision [205, 206], and has been shown to outperform conventional ML methods [201]. A typical CNN architecture is shown in figure 2.2. CNNs usually consist of a series of alternating convolutional and pooling layers, followed by fully connected layers and finally a softmax layer for prediction. The convolutional layer, the main block of CNNs, applies the filters systematically by sliding them patch-wise along the images to get feature maps. During each application of the filter, a correlation operation (using dot product) is performed between the filter and a patch in the input image to obtain the values in the output feature map. The pooling layer downsamples these output feature maps, while still retaining the useful information. For instance, maximum pooling retains the maximum values (i.e. the values indicating the best correlation with the convolutional filter) within a patch of the output feature map. Pooling the feature maps also reduces number of parameters, speeds up training and resists overfitting. Activation functions are gating functions that feeds the output feature maps as inputs to next layers. Activation functions are generally non-linear and monotonically increasing, that helps in training the CNN to model complex mappings between the feature maps in different layers. In the fully connected layers, the features maps are flattened and non-linearly combined to get meaningful features that are used for classification in the final softmax layer. The CNN is trained using an optimisation algorithm in an iterative manner. During the forward-pass, the CNN makes predictions to match the distribution of target variable. A loss function is an objective function that determines the error (or loss) between the predicted and target distributions. The optimisation aims to minimise the loss value by estimating the loss gradient and updating the weights and bias values of

the layers, based on the gradient. This process of updating the weights is called back-propagation. For binary classification tasks, the cross-entropy loss function is most commonly used, while for segmentation tasks, the Dice loss function (using Dice similarity index) is used to achieve good overall segmentation. This section provides a review of various CNN models used for WMH segmentation and other similar pathologies. This review also includes the generic concepts that are relevant to our work, adopted from other segmentation tasks.

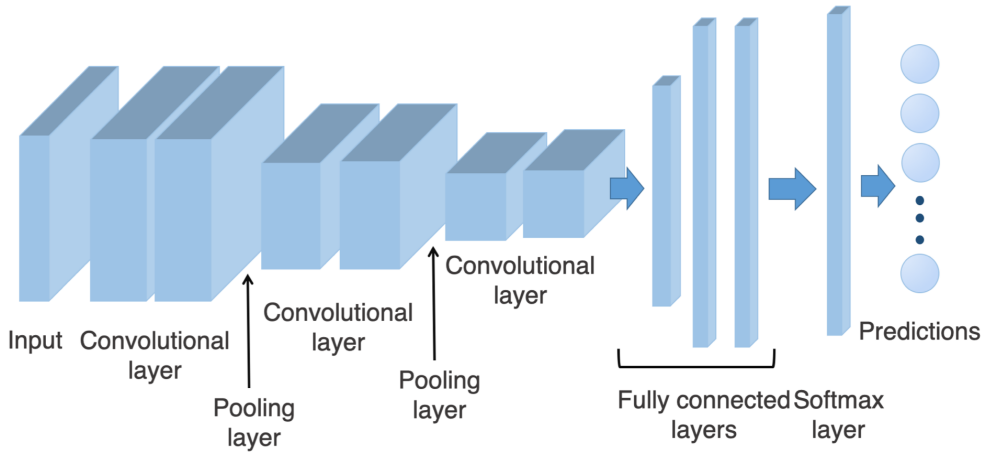


Figure 2.2: A typical CNN architecture for a classification task.

Deep CNN architectures: Ever since AlexNet [207] was successful in the ImageNet challenge, various DL architectures have been increasingly used for different tasks, leveraging their own advantages. For instance, VGGNets [208] enabled building deeper models using smaller convolutional kernels. GoogLeNet [209] established sparse connections between the activations to increase the computational efficiency. Residual networks (ResNets, [210]) introduced *shortcut connections* using identity mapping and gated layers (units that are trained to regulate information flow) as in Highway networks [211]. They overcame the vanishing gradient problem [212], where gradients of the loss function approach zero as the network gets deeper, making the model difficult to train. Hence, ResNets allowed deeper architecture with relatively less number of parameters. DenseNets [213] connected each layer to every other layer, thus reusing features and strengthening their propagation. In order to pool the performances of multiple models, ensemble of models were created for various tasks. The

first ensemble architecture, called Network-in-network (NIN) [214] replaced convolutional block with a multi-layer micro network. The ensemble networks, in general, perform better than their individual counterparts and avoid over-fitting, a common problem with complex models. Inception network [209], based on NIN, allowed parallel stacking of CNNs with far fewer parameters than AlexNets. The later version, Inception-V4 [215] combined original Inception architectures with shortcut connections as in ResNets.

Semantic segmentation of medical images: DL models are being increasingly used for automated semantic segmentation of images (i.e. associating each voxel of an image with a class label). This is demonstrated by the results of various challenges, including the MICCAI WMH segmentation challenge 2017 (MWSC 2017): out of 20 methods competing in the initial call, 14 of them were based on deep learning [216]. In general, for semantic segmentation, the end-to-end trainable networks such as encoder-decoder networks [217, 218] (e.g. DeepNAT [219], QuickNAT [220], LedNet [221]) and fully convolutional neural networks (FCNs) [222] are the most popular categories used. In encoder-decoder framework, the encoder takes an image as input and outputs a feature map. This feature map holds rich information representative of input images. The decoder takes the feature map and provides the output image close to the target image. In FCNs, the fully connected layers of CNNs are replaced by convolutional layers for upsampling/deconvolving the high-level feature maps to obtain the final segmentation maps. Encoder-decoder models provide better segmentation results compared to FCNs [222]. This is because in FCNs the high-level features are directly projected into image dimensions to get segmentation outputs. On the other hand, in encoder-decoder model, the decoder is an independent network that is trained to provide accurate output maps by taking the high-level features (from the encoder) as input. Even though many models have been implemented in the encoder-decoder framework [223, 221], perhaps one the most successful ones for medical image segmentation is U-Net [224] and its modified 3D architectures [225, 226].

The DL methods used for segmentation of WMHs include various ensemble models [58, 227, 216], encoder-decoder models, especially U-Nets [58, 216, 202, 228, 229] and multiscale patch-based models [230, 59, 231]. Table 2.2 compares the performance of various DL methods proposed in the literature for WMH segmentation.

U-Nets have the encoder-decoder structure with skip connections between the encoder and decoder layers at the same levels of abstraction. The encoder arm in the U-Net captures the fine contextual information through successive convolutional and pooling operations. The decoder arm uses a series of upsampling/deconvolution layers to increase the resolution of features. To aid in precise localisation, the high resolution features from the encoders are combined with the upsampled features (through skip connections) and go through another convolutional layer in each level of abstraction. The combination of features and the increased number of feature channels in the decoder arm allows the propagation of contextual information in the decoder, resulting in accurate segmentation. Ensemble models and combination of various networks have been shown to improve segmentation accuracy. In fact, Li and colleagues [58] used ensemble of U-Nets for WMH segmentation, and won the MICCAI WMH segmentation challenge 2017 (MWSC 2017) [216]. Guerrero and colleagues [202] used a combination of U-Net and ResNet architectures for segmentation of WMHs and stroke lesions.

Utilisation of information at different scales: While 3D models have been predominantly used for segmentation of larger lesions such as brain tumours [232, 233, 234], 2D models have been often used for WMH segmentation [58, 216]. Alternatively, 2D/3D patch-based models have also been used successfully [235, 230, 59]. An important point to keep in mind is that, in patch-based methods, smaller patches lack enough contextual information, while larger patches increase computational load and reduces localisation accuracy [224]. To overcome this dependency on patch size, some existing methods have used multiscale approaches [231, 236, 232], adding location-based information [237] and context-based models (e.g. conditional random

fields [232]). Ghafoorian and colleagues [230] proposed a novel method involving non-uniformly sampled patches (sparser sampling towards edges), outperforming multi-scale approaches. Later, they [59] compared various patch techniques and found that adding location-based information provided substantial improvement in WMH segmentation. The patch-based methods have achieved the second best performances in MWSC 2017: 3D multi-dimensional gated recurrent units [238] and FCN trained on multiscale patches, along with location based features [216].

Methods	**Datasets**[1]	**Network**[2]	**Loss func.**[3]	**Dim.**	**Dice**	**cl. TPR**	**F1**	**Notes**[4]
Li et al., [58]	MWSC TS	U-Net	DSC	2D	0.80	0.84	0.76	EN, DA, Po
Andermatt et al., [238]	MWSC TS	MDGRU	multinom. log.	2D	0.78	0.83	0.70	DA, PB
Berseth et al., [227]	MWSC TS	CNN	CE loss	2D	0.77	0.73	0.78	EN
Valverde et al., [235]	MWSC TS	CNN	CE loss	3D	0.77	0.75	0.71	DA, Po, Npre
Mehrtash et al., [216]	MWSC TS	U-Net	DSC	2D	0.77	0.59	0.70	EN, DA
Luna et al., [216]	MWSC TS	CNN	Mean Sq. err.	3D	0.72	0.63	0.68	-
Xu et al., [239]	MWSC TS	VGG-16	Multinom. log.	2D	0.73	0.63	0.67	DA
Jin et al., [216]	MWSC TS	U-Net	CE loss	2D	0.68	0.59	0.54	DA, Po
Zeng et al., [240]	MWSC TS	U-Net	CE loss	2D	0.69	0.44	0.57	DA
Mckinley et al., [241]	MWSC TS	DenseNet	CE loss	2D	0.63	0.55	0.51	-
Guerrero et al., [202]	Mild Isc. stroke (167)	ResNets in Unet	BC CE loss	2D	0.69	-	-	-
Moeskops et al., [231]	MRBrainS13 (20)	CNN	CE loss	2D	0.67	-	-	PB, MS
Ghafoorian et al., [59]	Elder SVD (50)	LF CNN IW	CE loss	2D	0.78	-	-	MS, PB, LF

[1] Datasets: MWSC TS - MICCAI WMH segmentation Challenge test dataset [216], MRBrainS13 - MRBrainS challenge test dataset [242]

[2] Network: MDGRU - multi-dimensional gated recurrent units, LF CNN IW - late fusion of parallel CNNs with independent weights [59]

[3] Loss functions: DSC - Dice loss function, CE loss - cross-entropy loss function, BC CE - bootstrapped categorical cross-entropy loss function

[4] Notes: DA - Data augmentation performed, EN - ensemble network, MS - multiscale network, LF - location-based features used, PB - patch-based training/testing, Po - postprocessing done, NPre - No preprocessing

Table 2.2: Comparison of the existing DL methods for WMH segmentation in the literature.

Leveraging salient features for improved segmentation: Since the development of CNNs, several techniques have been proposed for understanding and using

salient (most useful) regions in the features at different layers for improved segmentation. Peng and colleagues [243] proposed global convolution by larger kernels for localising objects, using contextual information from a wider neighbourhood. Other common approaches include parallel *'local'* (using local features) and *'global'* (using global contextual features with a wider support/neighbourhood) pathways [233]. Similarly, non-local networks [244] establishing *long-range* dependencies between consecutive frames/slices, therefore obtaining the overall contextual information. Jetley and colleagues [245] initially proposed end-to-end attention gates for bootstrapping the performance of CNNs. Attention gates are additional trainable modules that can be added to a network and propose soft region proposals on the fly. In attention gates, the low-level features (containing local information) at the initial layers are weighed using the attention weights determined from *coarser* or *global* features. In order to achieve this, the attention gate module takes the global/coarser feature map, and combines it (in a multiplicative or additive manner) with the low-level feature map to provide score/weight map that would enhance the salient regions of features, and suppress the irrelevant and unhelpful regions. Therefore, as the network trains, the local features that lead to a better global context are selectively enhanced, and would eventually highlight salient features corresponding to the object to be segmented. Kuen and colleagues [246] trained recurrent attention gates for saliency detection in images. Attention gates have also been used for pancreas segmentation [247] and scan plane localisation in ultrasound images [248].

Data harmonisation and domain adaptation: Another desired characteristic of CNN is generalisability across various datasets from different *'domains'* (e.g. different scanners), irrespective of the variations in their intensity, resolutions, acquisition and pathological characteristics. Domain adaptation (DA) addresses the problem where data are from various domains, with the distributional properties of the data differing between the domains. Let $\mathcal{D}_s$ be the *source* domain with the learning task $\mathcal{T}_s$ and $\mathcal{D}_t$ be the *target* domain with the task $\mathcal{T}_t$. Domain adaptation methods become necessary in the case where the domains $\mathcal{D}_s \neq \mathcal{D}_t$ or tasks $\mathcal{T}_s \neq \mathcal{T}_t$ differ [249]. In our case, we

adapt the task $\mathcal{T}_s = \mathcal{T}_t$ (WMH segmentation) for $\mathcal{D}_s \neq \mathcal{D}_t$. The simplest step towards DA is data augmentation [204], wherein the amount of training data is increased by adding variations through geometric transformations [250], neural style transfer [251] and generative adversarial neural networks (GANs) [252, 253, 254, 255]. *One-shot* [256, 257] and *zero-shot learning* [258, 259] have also been proposed to create descriptive attributes with extremely limited training data. In MWSC 2017, Mehrtash and Ghafoorian [216] used a two-level approach, by initially identifying the scanner type, and then training scanner-specific models. Some of the successful methods proposed for DA includes minimisation of divergence or domain discrepancy measures [260, 261], using features that could form intermediate representation between the domains [262, 263], reconstruction methods [264] (e.g. CycleGANs [265]) and transfer learning [266].

Transductive transfer learning paradigm [266] is the most commonly used DA technique and transfers the knowledge between source and target domains for a given task. In transfer learning (TL), the model trained on the source domain data is fine-tuned using target domain data. Fine-tuning is the process of updating the weights of selected layers (usually, the ones with high-level features) of a pretrained network using the data from target domain, while fixing or freezing the remaining layers. Usually a smaller learning rate is used for fine-tuning, and the number of layers to be fine-tuned usually depends on the size of the target dataset and the similarity in their distribution to that of the source dataset [262]. Fine-tuning has been shown to improve the performance [267] and performs better with increasing similarity between features [262, 268]. TL has been widely used for medical image segmentation [269, 270, 267], including lesion segmentation [271]. While the main aim of TL is to maximise the model performance in the target domain, the training on different domains occurs separately and does not enable us to mix data from different domains and relies on having a reasonable amount of target training data [267].

Domain adversarial training [272, 273], on the other hand, simultaneously trains the source and target domain data, and has been shown to perform well even with

limited target training data. A specific instance, *domain adversarial training of neural networks* (DANNs) [272] learns the domain invariant features using two parallel predictors: (1) a label predictor that predicts class labels and (2) a domain predictor that discriminates between domains, simultaneously optimising features and weights. The adversarial training at the domain predictor using a *gradient-reversal layer* enables the model to predict the class labels successfully, while minimising the shift between the domains. Originally, the label predictor in DANNs does not require manual segmentations from the target dataset (e.g. haemorrhage segmentation task in [274]). Later, semi-supervised extensions of DANNs were used, where the model used a small fraction of labelled target training data for label prediction. Various Siamese architectures [275, 276, 277] have also been developed for effective domain transfer. A supervised Siamese architecture using classification and contrastive semantic alignment (CCSA) loss [277] was proposed for effective domain transfer. The semantic alignment loss minimised the distance between the same class samples from different domains, while the separation loss maximised the distance between different class samples between domains. Another update of adversarial training involves optimisation for domain invariance with respect to both data and tasks [273], and could be trained both in either unsupervised or semi-supervised manner. The model uses a *label distribution matching loss* to transfer information between the tasks, and a *domain confusion loss* for adaptation between different domain datasets.

From table 2.2, most of the existing DL methods for WMH segmentation are based on U-Nets [58, 216, 240, 202] and use cross-entropy (CE) loss [227, 235, 216, 240, 241, 202, 231, 59], while the best performing method uses Dice loss (DSC) [58]. Except a few methods [235, 216], almost all methods use 2D models, some of them in ensemble architecture [278, 227, 216] and in multiscale patch-based networks [238, 231, 59]. Hence, it could be inferred from the literature that 2D ensemble, multiscale models based on U-Nets using CE and DSC loss have provided good results for WMH segmentation.

In chapter 5, we have proposed a DL model for WMH segmentation using an ensemble U-Net architecture and have explored various domain adaptations techniques such as TL and DANNs for different datasets.

2.2.2.5 Manual segmentation in automated methods

Different methods with varying degrees of accuracy have been used for manual segmentation. Other than being hand drawn [55, 171, 33, 183], manual segmentations have also been generated using various semi-automated methods including region growing [279, 165] and random walker algorithms [187] using manually selected seed points. Also, there could be uncertainties in manual segmentations due to the lack of consensus regarding the pathological criteria (e.g. ventricle lining as a part of WMHs), artefacts (e.g. noise, motion artefacts), limitation of methods (e.g. over-segmentation in low contrast areas) and so on. Hence, it is crucial to understand that, even though manual segmentations are used as a 'gold standard', they are not perfect. Therefore, any result with respect to manual segmentation has to be handled and interpreted carefully. In order to overcome this limitation, some methods have also used visual rating scales for performance evaluation [34, 153]. Although they are still subject to intra- and inter-rater variability, they do not rely on voxel-level accuracy. Another way of evaluating segmentation results is to use external standards, i.e. variables that are known to be related to the lesion load such as age [53, 280, 281].

2.2.3 WMH segmentation on big datasets

Quantifying WMHs in big datasets enables clinical and demographic correlates of WMHs to be obtained for large populations. In earlier studies such as the Rotterdam study and the LADIS study, semi-quantitative ratings involving visual rating [89] and lesion volumes calculated from the manual segmentation were used for population-level analysis. Some of the automated methods specified in section 2.2.2.3 were applied in the prominent large population studies as follows: multispectral image processing tool, MXMxxxVI [282] in the Lothian birth cohort study [84], ANN [283] in the AGES-Reykjavik study [85] and BIANCA in the OXVASC study [86, 92] and

the UK Biobank study [87, 284]. Alternatively semi-automated methods were also used in large datasets: software package AMIRA 3.1.1 (Mercury Computer System Inc.) in the INVADE II study [88]. Hence, the use of automated and semi-automated methods have led to the more efficient and accurate identification and characterisation of WMHs in large datasets. Therefore, after our initial evaluations on multiple datasets, we will finally apply our methods on large datasets such as the UK Biobank and OXVASC.

2.3 CMB detection

2.3.1 Detection of CMBs on individual subjects

Compared to WMHs, work done on CMB detection is much sparser. In most of the population-level studies, CMBs have been detected and rated manually for studying their clinical associations [107]. This section provides a review of the existing semi-automated and automated methods for CMB detection, including deep learning methods, to the best of our knowledge. As specified in chapter 1, a detailed account of the criteria for CMB detection has been provided by Greenberg and colleagues [27]. While SWI is the most commonly used modality for CMB detection [285], among T2*-weighted image modalities, accelerated 3D T2*-weighted GRE images depict more CMBs than conventional 2D T2*-weighted GRE images [286].

For CMB detection, in general, the methods follow a common pattern with two steps: CMB candidate detection and post-processing for removal of FPs. The first step generally achieves high sensitivity, while the second step is more challenging and leads to improvement in the precision.

2.3.1.1 Semi-automated methods for CMB detection

In the semi-automated methods for CMB detection, manual intervention has often been used in the cleaning-up step for reducing FPs [287, 288, 289, 290]. Occasionally, candidate detection [291] and ground truth verification [292, 293] in the rating procedure also involves manual intervention.

Manual CMB detection: De Bresser and colleagues [291] identified minimum intensity projection (MIP) processing conditions to detect more CMBs with a limited number of slices on 7T images. On the MIP processed images, CMBs were then manually detected and rated based on the MARS scale [73].

Manual FP reduction: Seghier and colleagues [287] developed a semi-automated tool 'Microbleed detection using automated segmentation' (MIDAS). Their tool spatially normalises GRE T2*-weighted images, and identifies CMBs as an extra tissue class for abnormal voxels within a unified segmentation-normalisation model. In the final FP reduction step, image artefacts were excluded manually. After manual reduction of FPs, the method showed good agreement when validated with the MARS visual rating system. Supervised classification algorithms such as SVM [288] and RF [289] based on intensity and shape based features have also been used for candidate detection, and were followed by manually removal of FPs. In [289], a user-guided interface was added to review and revise classified candidates, with a reported additional time of 13 min for 80 subjects for manual revising step alone. Exploiting the circular shape of CMBs, fast radial symmetry transform (FRST) have been used for candidate detection [292, 290], while the FPs are further reduced by manual inspection [292] or in a rule-based manner (depending on the CMB size) [290]. In [290], additional processing steps such as application of vessel exclusion mask and 3D region growing were performed before manual inspection. They managed to reduce manual intervention time to almost half of the time reported in [289] while achieving comparable sensitivity.

Manual revision for ground truth improvement: Kuijf and colleagues [293] extended their earlier work [292] by using 3D FRST to adapt to the spherical nature of CMBs. Both these methods involved a manual intervention step, where the raters (who were not involved in the original visual rating) revised the visual rating and number of CMBs which were missing in the ground truth. Hence, their semi-automated method rather optimised and assisted human rating using their computer-

aided pipeline. The required human rater time for visual revision of the results was about 1-2 min per subject on average.

Even though the manual revising step removes FP candidates effectively, it typically takes at least a few minutes per subject [289, 293, 290] and it is not an efficient solution for CMB detection in very large datasets. Hence, end-to-end automated methods are more useful for big data applications.

2.3.1.2 Automated CMB detection

In automated methods, most of the candidate detection steps use hand-crafted shape [294, 295], intensity [296] and geometric features [295] with supervised classifier framework [295, 296, 297, 298].

Shape: Bian and colleagues [294] applied a 2D radial symmetry transform on the MIP of SWI images from 10 subjects with gliomas and used shape-based features for reducing FPs. They achieved a sensitivity of 86% (263 detected out of 304 CMBs) with 45 FPs per subject. Fazlollahi and colleagues [295] used shape descriptors to train a cascade of RFs. They reported a sensitivity of 92% with 17 FPs per subject on SWI from 40 subjects (7 with Alzheimer's disease, 9 with cognitive impairment and 25 HC). In their later work [296], they used a multiscale Laplacian of Gaussian (LoG) filter for candidate detection. In the second stage, they extracted a set of 3D Radon transform and Hessian-based shape descriptors for training a 3-layer cascade of RF models. The consecutive independent RF classifiers in the cascade with low to high posterior probability thresholds progressively ruled out obvious FP candidates, and handled any imbalance in the CMB and non-CMB classes. Manual labelling consisted of 'definite' and 'possible' CMBs, and the method achieved a sensitivity of 93% and 10 FPs per subject in the definite set. Even though this method detects quite small CMBs (0.93 mm), the presence of small vessels resulted in the detection of FPs.

Intensity and geometry: Ghafaryasl and colleagues [298] used more comprehensive features that integrated the geometry, intensity, scale and local image structures for

CMB candidate detection. For the following FP removal stage, in addition to above features, they used proton density (PD) values, since CMBs are not visible on PD-weighted images, as opposed to the blood vessels. On comparing various classifiers for both stages, they obtained the best performance at the first stage with a Parzen classifier [299], and QDA classifier at the second stage. They achieved an overall sensitivity of 91% with a relatively low FP count of 4.1 per subject.

Hierarchical feature extraction: Dou and colleagues [297] developed a 3-step approach to reduce FPs: candidate screening based on intensity values, compact 3D hierarchical feature extraction and FP removal with an SVM classifier. The feature extraction step used a stacked convolutional independent subspace analysis network, which is an unsupervised learning algorithm that extracts invariant features for object detection by exploring the underlying subspace structure of the data. They reported a low FP count of 8 per subject, for a sensitivity of 90%.

Hence, using structural (e.g. intensity and shape) and anatomical information in combination with the local characteristics (e.g. local contrast) could aid in reduction of FPs and more accurate detection of CMBs.

Deep learning for CMB detection

Conventional ML methods require the extraction of meaningful features capable of distinguishing CMBs from the background and mimics. However, due to the small size, variation in shape and intensities of CMBs on SWI images, designing robust, descriptive and cost effective features is highly challenging. The use of CNNs could overcome this challenge and provide more efficient CMB detection. This section provides an overview of DL models used for CMB detection and other generic DL models/concepts used for detecting small objects with similar characteristics to CMBs.

Similar to [297], Chen and colleagues [300] proposed a novel framework consisting of three steps: CMB candidate localisation, feature representation and classification. Firstly, they thresholded the images to get the initial candidates with high sensitivity.

Secondly, they used a CNN for extracting hierarchical 3D feature representation. In the final step, they trained an SVM classifier. Their method achieved a sensitivity of 89% with 6 FPs per subject.

Dou and colleagues [301] developed a 2-step cascaded framework for CMB candidate detection and later their discrimination from mimics. They initially trained a 3D CNN with positive CMB samples (patches containing CMBs) and randomly selected negative samples, and applied the initial model on the training set. The FP patches obtained from the above step were then provided as additional negative samples for fine-tuning the discrimination capability of the network, by enlarging the training dataset. During the testing phase, the fully connected layers of the model were converted into convolutional layers, and the model was applied on the whole volumes. On an SWI dataset of 320 subjects, their method achieved a sensitivity of 93.2% with 2.7 FPs per subject. Later, Chen and colleagues [302] proposed a model based on deep ResNets on SWI images and reported a sensitivity of 95% with 11 FPs per subject. In a recent work, Liu and colleagues [278] proposed a 2-stage detection method, that used 3D FRST in the CMB candidate detection stage. For the FP reduction stage, the method used deep ResNets on both SWI and high-pass filtered phase images. They also additionally used morphological operations as such dilation and erosion as post-processing steps to achieve a sensitivity of 96% with a very low FP count of 1.8 per subject.

For the identification of small objects and fine-grained localisation, networks such region-based CNN models (R-CNN) [303, 304, 305, 306, 307], dilated convolutional networks [308, 309] and probabilitistic learning [310] have been widely used. Most of these networks operate on multiscale patches [311, 310, 217, 312] or derive features at different scales [313, 303, 314] to avoid loss of finer details. An earlier multiscale approach [313] used a CNN model to extract features from multiscale regions centred on each voxel. Motivated by this approach, R-CNN models [303] used a *region based recognition* [315, 316] framework by applying bottom-up region proposals for

localising and segmenting the objects. The object detection happened in 3 steps: region proposal, feature extraction and classification. Initially, Girshick and colleagues [303] used a selective search algorithm to propose the candidate regions at multiple scales, that were then fed into a CNN for feature extraction. The regions were finally classified based on the extracted features using an SVM classifier. In addition to detecting the objects, the SVM classifier also predicts the four values corresponding to the bounding box coordinates of the proposed candidate regions. The framework provided accurate object localisation, but only a specific number of regions could be proposed, making the selective search algorithm for region proposal inefficient. In their later work [304], they overcame this problem by using a CNN on the whole image to generate a feature map. They then selected the regions from the feature map, followed by an ROI Pooling layer that would enable feeding the proposed regions into fully connected layers for prediction. The faster version of R-CNN, faster R-CNN [305] completely eliminated selective search and uses CNN for both region proposal and classification. Mask R-CNN [307] extended faster R-CNN by adding a parallel network for detecting a more precise object segmentation. A regression model, *you only look once* (YOLO) [317], separated images into bounding boxes and predicted the probabilities within each bounding box. They later combined the predictions at the image-level using another network, forming an end-to-end trainable framework. In general, for classification problems, CNNs reduce the feature dimensions using pooling layers until they get compact at the fully connected level. However, this results in the loss of spatial acuity and might not be useful for detecting details in the image. Hence, dilated convolutions [308] (where the filters are dilated) were used to increase the resolution of the feature maps for the prediction of smaller details. Dilated convolutions increase the size of filter kernels by introducing spaces (specified by dilation rate) between the kernel elements. Hence the dilated convolution enables us to increase/widen the receptive field without adding additional parameters. Dilated convolutions were combined at multiple scales for better detection [309]. He and colleagues [310] proposed a multiscale approach wherein the features were extracted at multiple scales and combined using a probabilistic framework. Another multiscale

approach [311] used an attention model to softly weight features at different scales and predicted the final probability map as the weighted sum of all the scales. Several networks have used features at different scales: starting from lower scale and 'zooming-out' to capture their representation at the image-level [314], forming *hypercolumns* at each voxel as the vector of activations of all CNN units related to that voxel [312].

To summarise, CNN models with location-based information [301] and anatomical information (information regarding small vessels, shape of CMBs) [278] with a good post-processing [278, 298] have provided better segmentation results. In general, multiscale patches/features based networks [305, 311, 313, 308, 314] have provided accurate detection of finer objects, and hence could be adopted for CMB detection due to the similarity of their characteristics.

In chapter 6, we have proposed a 3D multiscale patch-based object-level CMB detection with 3-steps that are widely used in the literature: CMB candidate screening, candidate discrimination and postprocessing for FP reduction.

Chapter 3

Modelling the distribution of white matter hyperintensities due to ageing on MRI images using Bayesian inference

In this chapter, we propose a probabilistic approach to model the population-level parametric lesion probability map in order to investigate the relationship between the distribution of WMHs and various parametric factors.

In section 3.2, we explain our proposed method for modelling the distribution of WMHs within a population, and report the optimisation parameters. We initially evaluated our model on a simulated dataset and compared it to existing methods. The simulation details and analyses are specified in section 3.3 and the results are reported in section 3.5.1. We then validated our model on two real datasets as explained in section 3.4. The results of our experiments on real data are reported in section 3.5.2. We have provided the discussion of our results on simulations and real data, our main conclusion and potential future avenues for this work in section 3.6.

The preliminary results of this work were presented at the 2017 OHBM conference (Sundaresan et al., OHBM 2017) and the final results of this work are published in Sundaresan et al., 2019 [78].

3.1 Introduction

As mentioned in chapter 1, WMHs are strongly associated with cerebrovascular disease and vascular risk factors [318], frequently found in neurodegenerative diseases such as vascular dementia [319, 95, 320, 82] and Alzheimer's disease [95, 319]. While the occurrence of WMHs is common in the ageing population, clear relationships between the amount and distribution of WMHs and cognition, demographics and vascular risk factors are not yet well understood.

Even though several visual rating scales for WMHs are available and commonly used [44, 89], determining the voxel-wise distribution of WMHs at the population-level is important in order to investigate the relationship between the spatial distribution of lesions and various parametric factors such as age, vascular risk factors and cognitive function (e.g. measured with the Montreal Cognitive Assessment (MoCA) score). In a population-level parametric lesion probability map, the value at each voxel reflects the probability of finding a lesion in a specific population for a parametric factor of interest. Therefore, modelling population-level parametric lesion probability maps would aid in studying differences in lesion distribution between normal and pathological ageing [47, 46] and be useful in the clinical context of the studied population.

We have provided an account of existing literature on population-level voxel-wise lesion probability maps in chapter 2. Several methods have been proposed, ranging from the simplest method involving averaging subject-level lesion probability maps [50] to the more complex Bayesian Spatial Generalised Linear Mixed Model (BSGLMM) proposed by [127]. The averaging method does not provide spatially continuous population-level lesion maps since it does not consider the neighbouring voxels. BSGLMM, based on spatially varying coefficients, improves the local spatial relationship and provides a more spatially continuous model. However, it is computationally intensive and requires the use of graphical processing units (GPUs).

Our proposed method is computationally lean, and uses a spline representation to ensure continuity and smoothness between neighbouring voxels in both spatial and

parametric dimensions. In our model, the lesion distribution at each voxel is obtained by the linear combination of all the splines at the neighbouring voxels. Moreover, our model adopts a Bayesian framework in order to overcome the discretely sampled nature of the input binary maps obtained from lesion segmentation. Our approach for modelling the spatial distribution of lesions is suitable for very large imaging datasets such as UK Biobank [284] and allows the flexibility of observing the effect of various parametric factors on the lesion distribution. In fact, we used a subsample of the UK Biobank data for the initial evaluation of our method. In this evaluation, we tested the relationship between WMHs and age comparing subjects with and without hypertension. We chose hypertension as our grouping variable, since it has been found to be one of the important risk factors of WMHs [321, 322] in addition to age. Also in this case we compared our results against the existing methods. Our next validation was on the second dataset, a clinical population (vascular population of subjects who had a transient ischemic attack or minor stroke) from the Oxford Vascular Study (OXVASC) [323], and analysed the results with respect to age. We chose age as our factor of interest since a strong relationship between WMHs and age has already been established [54]. Moreover, in order to validate the fact that our model could work with any factor, we modelled the voxel-wise distribution of lesions with respect to cognitive function measured by MoCA scores. Finally, we evaluated the robustness of our method with respect to the initialisation for optimisation and the initial thresholds used for obtaining binary maps.

3.2 Modelling the distribution of WMHs using Bayesian inference

Our proposed Bayesian spline model takes as input the binary lesion map and the age (or other parametric factor of interest) for each individual subject. We modelled the distribution of lesions in three steps: 1) constructing a representation for the lesion probability distribution using by spline basis functions, 2) formulation of the posterior

lesion probability; and 3) maximisation of the posterior probability by constrained optimisation.

Let D_s be the 3D binary lesion map of individual subject S. Our aim was to form a 4D spatio-temporal parametric lesion probability map with the 4^{th} dimension indicating the parametric factor of interest.

3.2.1 Constructing a representation for the lesion probability distribution using spline basis functions

Since the binary lesion maps contain localised areas, the lesion voxels are sparse and have discrete values (0 or 1). Therefore, to ensure spatial and temporal continuity in the probability distribution of lesions, we approximated the lesion probabilities using cubic b-splines.

Consider cubic b-splines with spline coefficients C_i and basis functions B_{ij}, where i denotes the indices of knot points of the basis functions and $j = (x, y, z, t)$ indicate the spatial coordinates in the first 3 dimensions and the parametric factor in the 4^{th} dimension. The lesion probability θ_j at each voxel is related to C_i and B_{ij} by,

$$\theta_j = \sum_i C_i B_{ij} = C^T B \tag{3.1}$$

The probability value at each voxel is calculated as the linear combination of spline basis functions. We formulated the splines as a separable outer product of individual 1D splines in 4 dimensions (3 in space and one in the parametric dimension). To implement the above calculation, we generated 1-dimensional cubic b-splines and convolved the binary map with scaled 1D b-splines independently in all 4 dimensions to get spatially continuous lesion probabilities θ as shown in figure 3.1. This operation is more computationally efficient than off-the-shelf spline fitting toolboxes and is well suited to large datasets.

It is worth noting that after each convolution edge artefacts occurred due to convolution with the tails of the splines and zero padding. In order to get the corrected convolution output f_{norm}, we divided the convolution output f_{in} with a normalising

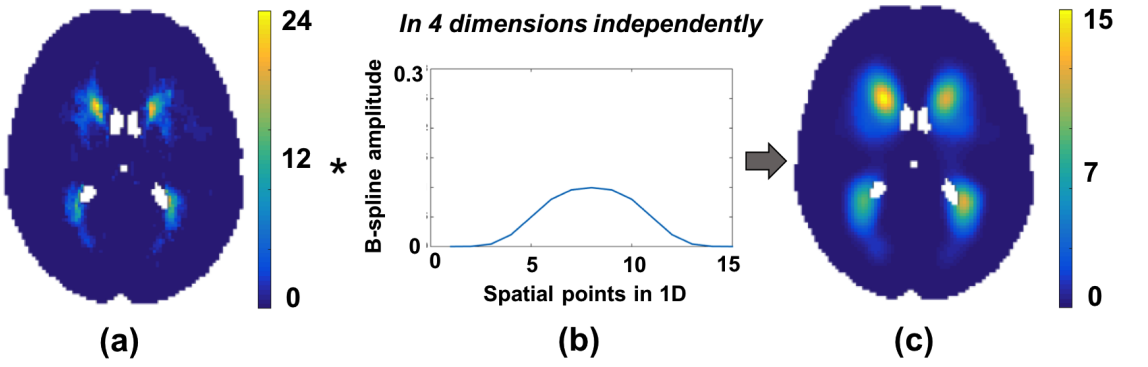


Figure 3.1: Spline modelling with cubic b-splines. (a) Average of binary lesion maps that are grouped according to a parametric factor, (b) Cubic b-spline used for modelling, (c) Spatially continuous lesion probability map obtained as a result of spline fitting. The convolution of (b) with (a) is done in all 4 dimensions independently to get (c).

.

image (formed by convolving the same b-splines B with an unity image $u(x)$, where $u(x) = 1$ inside the valid FOV - brain voxels - and $u(x) = 0$ outside, thus forming a similar pattern at the edges) as shown in eqn. 3.2.

$$f_{norm}(x_0) = \frac{\sum_x f_{in}(x).B(x_0 - x)}{\sum_x u(x).B(x_0 - x)} \tag{3.2}$$

3.2.2 Formulation of the posterior lesion probability

The most common lesion probability estimation method is averaging lesion maps across subjects. However, the accuracy of this approach is sensitive to the amount of data, especially considering the fact that there might not be any subject representing specific age groups or a few subjects might not have any WMHs. Bayesian methods are well suited to this problem since they allow for the uncertainty associated with the limited amount of data, while the spline model ensures continuity between the neighbouring voxels.

As we considered age as our parametric factor, the 4^{th} dimension, t, of our data indicates the bin number corresponding to the age groups. We formed these bins t by grouping the subjects into age groups age_t (e.g., $age_1 = (21, 23]$, $age_2 = (23, 25]$, etc.). The age groups can be defined to have a duration as short as we require (even in months) and hence binning the ages does not necessarily restrict the values to be

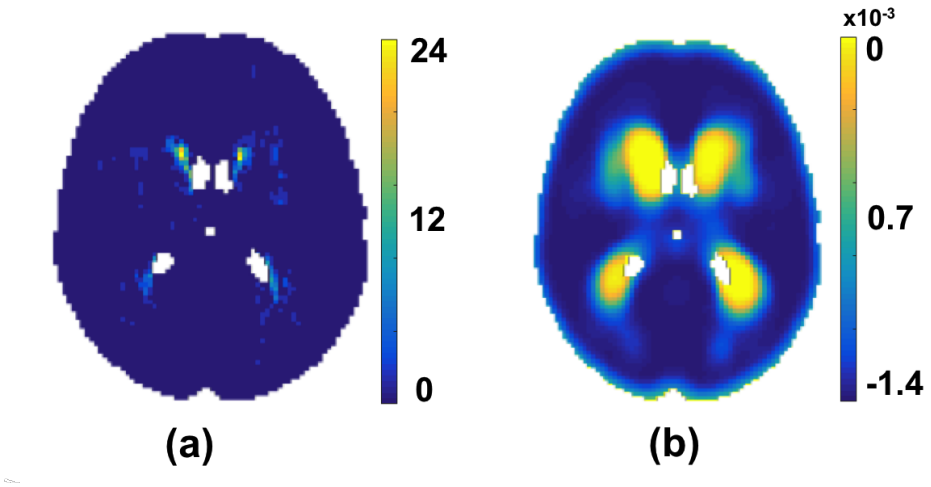


Figure 3.2: A sample of data accumulation shown for a single age group. (a) The map indicating the number of subjects having lesions (R_j) for a specific age group $age_5 = (29, 31]$. (b) Gradient magnitude of the log-probability distribution obtained by Eqn. 3.16

overly discretised in the parametric dimension. Let R (shown in figure. 3.2(a)) be the 4D spatio-temporal volume formed by summing the binary lesion maps D_{Sj} of the subjects S in each age group age_t. The value at each voxel $R_j = R(x, y, z, t)$ denotes the number of subjects having lesions at that voxel in the age group age_t. Let N be a 4D volume in which the value at each voxel $N_j = N(x, y, z, t)$ denotes the number of subjects in each age group age_t. R and N are formed by,

$$R_j = \sum_{S \in T} D_{Sj}, \quad \text{where } T = \{S \mid \text{age of } S \in age_t\} \tag{3.3}$$

$$N_j = \#\text{ subjects in } age_t \tag{3.4}$$

For an age group age_t, R_j indicates number of observations of binary data (lesion occurrence) in the subjects belonging to this age group age_t, with the total number of subjects in age_t given by N_j. Thus the probability of observing R_j number of binary outputs given N_j and probability of lesion occurrence θ_j for an age group age_t is given by the binomial likelihood distribution,

$$p(R_j \mid N_j, \theta_j) = {}^{N_j}C_{R_j}\, \theta_j^{R_j}\, (1 - \theta_j)^{(N_j - R_j)} \tag{3.5}$$

where ${}^{N_j}C_{R_j} = \frac{N_j!}{R_j!(N_j-R_j)!}$. We assume independence between age groups and voxel locations for the further simplification of our model. Hence, the full likelihood distribution for all age groups and at all voxel locations is given by the product of individual likelihoods,

$$p(R \mid N, \theta) = \prod_j p(R_j \mid N_j, \theta_j) \tag{3.6}$$

$$= \prod_j ({}^{N_j}C_{R_j}\, \theta_j^{R_j}\, (1-\theta_j)^{(N_j-R_j)}) \tag{3.7}$$

Using Bayes theorem, the posterior lesion probability distribution $p(\theta \mid N, R)$ is given by,

$$p(\theta \mid N, R) \propto p(R \mid N, \theta)\, p(\theta \mid N) \tag{3.8}$$

$$= \alpha\, p(R \mid N, \theta)\, p(\theta \mid N) \tag{3.9}$$

In order to avoid any spatial bias, we assumed that the prior probability of lesion occurrence is the same for any point in space and time. Hence with uniform prior $p(\theta) = 1$,

$$p(\theta \mid N, R) = \alpha\, p(R \mid N, \theta) \tag{3.10}$$

$$= \alpha \prod_j ({}^{N_j}C_{R_j}\, \theta_j^{R_j}\, (1-\theta_j)^{(N_j-R_j)}) \tag{3.11}$$

Our aim was to maximise $p(\theta \mid N, R)$ to obtain a parametric model of the lesion probability distribution over a population. To make the calculation of derivatives simpler, we determined the log-posterior function since the logarithm is a monotonically increasing function. Log-posterior $L(\theta \mid N, R)$ is given by,

$$L(\theta \mid N, R) = \log(p(\theta \mid N, R)) \tag{3.12}$$

$$= \log(\alpha \prod_j ({}^{N_j}C_{R_j}\, \theta_j^{R_j}\, (1-\theta_j)^{(N_j-R_j)})) \tag{3.13}$$

$$= \log(\alpha) + \sum_j \log({}^{N_j}C_{R_j}) + \sum_j R_j\, \log(\theta_j) + \sum_j (N_j - R_j)\, \log(1-\theta_j) \tag{3.14}$$

Since the lesion probability values θ have been approximated with spline basis functions as explained in the section above, we substituted the values of θ from the eqn. 3.1 in the above log-posterior equation,

$$\begin{aligned} L(\theta \mid N, R) = \log(\alpha) + \sum_j \log({}^{N_j}C_{R_j}) + \sum_j R_j \log(\sum_i C_i B_{ij}) \\ + \sum_j (N_j - R_j) \log(1 - \sum_i C_i B_{ij}) \end{aligned} \tag{3.15}$$

The derivative of $L(\theta|N, R)$ with respect to the spline coefficients C_k is given by,

$$\frac{\partial(L(\theta \mid N, R))}{\partial C_k} = \sum_j \left(\frac{R_j B_{kj}}{\sum_i C_i B_{ij}} \right) + \sum_j \left(\frac{(N_j - R_j) B_{kj}}{\sum_i C_i B_{ij} - 1} \right) \tag{3.16}$$

Note that the derivative function requires more spline modelling steps (basically, each summation requires a 4D spline approximation), however, performing 1D convolutions independently in 4 dimensions considerably reduces the running time.

3.2.3 Maximisation of the posterior probability by constrained optimisation

In order to estimate the final lesion probability θ_i, we need to determine the spline coefficient values that would maximise the log-posterior. Also, by maintaining the value of C_i within the range [0,1] the final θ_i values will be constrained to the range [0,1]. Hence, to determine the value of C_i corresponding to the maximum posterior estimate, we formulated it as a constrained local optimisation problem as specified below:

$$\begin{aligned} \underset{C}{\text{argmin}} \quad & -L(\theta \mid N, R) \\ \text{Subject to} \quad & 0 \leq C_i \leq 1 \\ & \sum_i C_i = 1 \end{aligned} \tag{3.17}$$

We used the steepest gradient descent algorithm, which is a first order optimisation algorithm that uses an iterative polynomial line search method. The step size γ for each iteration is determined by a polynomial line search method satisfying Armijo - Wolfe conditions [324]. In our case, the function to be minimised is $-L(\theta \mid N, R)$.

For $L : \mathbb{R}^n \to \mathbb{R}$, the step size for each iteration follows the following condition:

$$\min_{C} \quad -L(\theta_k + \gamma q_k) \tag{3.18}$$

where, θ_k is the current guess, q_k indicates the search direction and γ is the step size. Moreover, in order to speed up the convergence, we redefined C_i using a normalising function (with parameter $\lambda_i \in (-\infty, +\infty)$) in order to maintain the range of C_i within [0,1] and relax the optimisation constraints as given by,

$$\begin{aligned} \underset{\lambda}{\text{argmin}} \quad & -L(\theta \mid N, R) \\ \text{Subject to} \quad & -\infty < \lambda_i < \infty \\ & \sum_i C_i = 1 \end{aligned} \tag{3.19}$$

where $C_i = (1 + \tanh(\lambda_i))/2$ so that $C_i \in [0, 1]$ when $\lambda_i \in (-\infty, +\infty)$. Using the chain rule, the derivative of $L(\theta|N, R)$ (eqn(3.16)) can be modified as

$$\frac{\partial(L(\theta \mid N, R))}{\partial \lambda} = \frac{\partial(L(\theta \mid N, R))}{\partial C_k} \frac{\partial C_k}{\partial \lambda_k} \tag{3.20}$$

$$= \left[\left(\frac{\sum_j R_j}{\sum_i C_i B_{ij}} \right) B_{kj} + \left(\frac{\sum_j (N_j - R_j)}{\sum_i C_i B_{ij} - 1} \right) B_{kj} \right] \frac{\text{sech}^2(\lambda_i)}{2} \tag{3.21}$$

We performed a step of spline modelling (given by eqn. 3.1) on the optimal C_i that maximises the log-posterior $-L(\theta \mid N, R)$ to get the final estimated parametric lesion probability map $\hat{\theta}$. We determined the average of lesion binary maps of subjects within the age group age_t to get the age group-wise average 4D lesion map $Average = \sum_j (R_j/N_j)$. We provided smoothed $Average$ as the initial estimate for C_i during optimisation.

Figure 3.3 illustrates the initial and final values of C_i and λ_i involved in the optimisation step for two different age groups (taken from different age bins of the 4D spatio-temporal map). The initial θ values (fig. 3.3(d)) and the final $\hat{\theta}$ values (fig. 3.3(e)) are shown with their difference (fig. 3.3(f)). The results show little difference between the initial θ values and final $\hat{\theta}$ values for the younger age group ($\leq$ 0.005 in fig. 3.3(f, bottom row)) and a higher difference ($\approx$ 0.01 in fig. 3.3(f, top row))

in the elder age group. This could be attributed to the difference in the amount of lesion voxels between the age groups.

3.2.4 Optimisation parameters for convergence

Implementation of steepest gradient descent optimisation algorithm was done in MATLAB [325]. We found the following parameters optimal for convergence by trial-and-error method during preliminary testing. We specified the initial step-size γ to be 1×10^{-3}, while step sizes for the subsequent iterations were determined by the polynomial line search algorithm, for which lower and upper bounds were set to $0.1 \times$ initial γ and $0.5 \times$ initial γ respectively. We specified the maximum number of iterations as 50 empirically, setting more number of iterations than needed, allowing for convergence of difficult cases (convergence curves for a few sample cases shown in appendix E.1). The main parameter for determining γ for each iteration is the 2-norm $\|\mathscr{G}\|$ of the 4D gradient matrix $\mathscr{G}$ (calculated from eqn. 3.20). We set the final convergence tolerance value for the decrease in $\|\mathscr{G}\|$ as 1×10^{-4} (since gradient of the function must be zero at the function minimum).

3.3 Simulations

In order to evaluate the accuracy of our model, we initially tested it on a simulated dataset. For this experiment, we compared the ground truth distribution (θ_{true}) with the probability distribution ($\hat{\theta}$) estimated by our algorithm and compared it with other methods.

We simulated volumes of size $91 \times 109 \times 91$ arranged in 6 months age bins to form 60 age groups, thus creating a simulated set of subject images with dimensions $91 \times 109 \times 91 \times 60$. For this simulation, we assigned a random number of subjects with a maximum number of subjects N_S in each age group $N_{sim} \in [1, N_S]$. We chose the value of N_S as 50 for our experiments, since this yielded the total number of subjects comparable with those used for real data experiments. For each subject, we generated a lesion map by randomly sampling voxels within a specified ROI (in

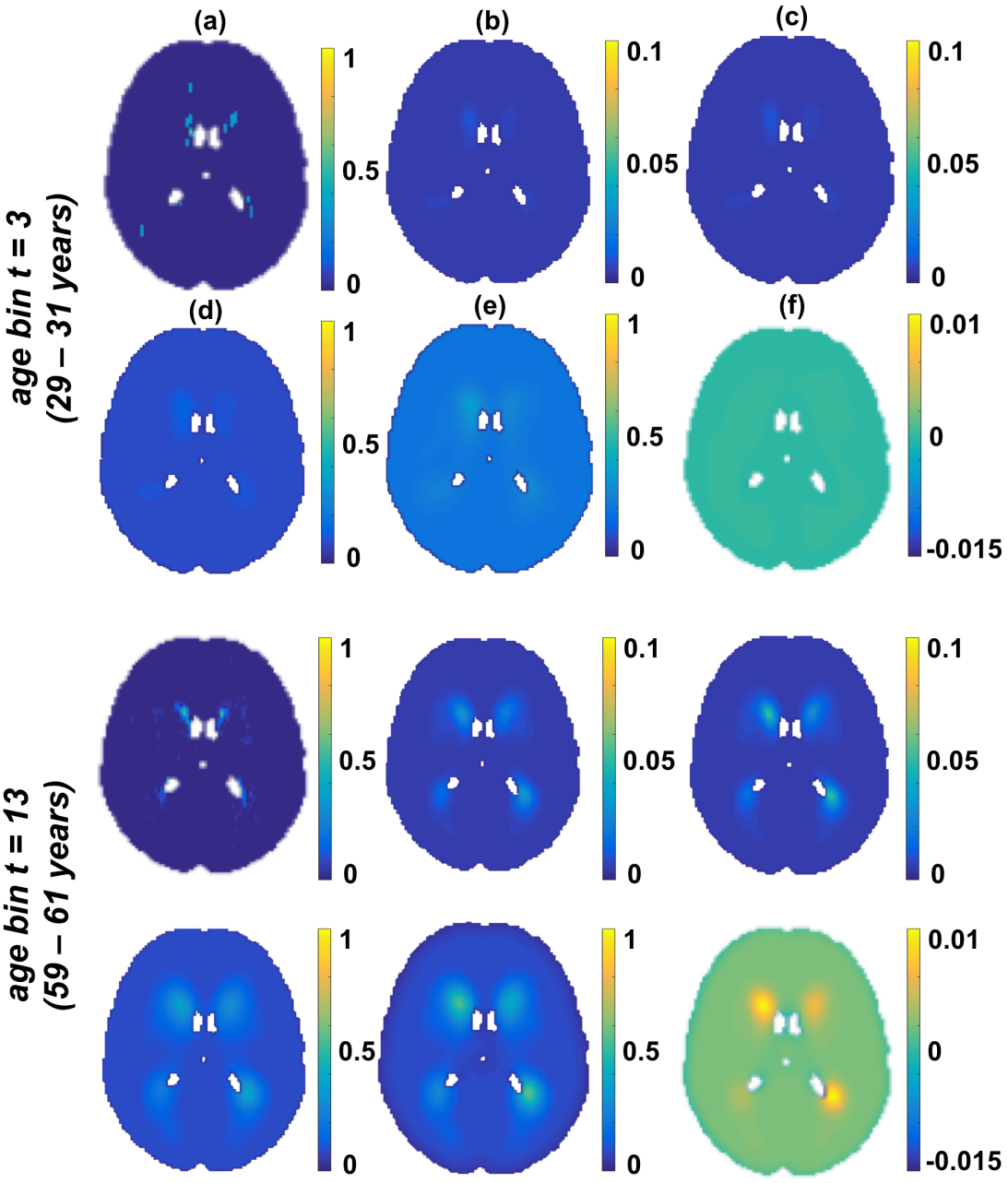


Figure 3.3: An instance of initial and final values in the optimisation step. The instances relate to two different age groups (younger age group in top rows). All the images are shown at slice $z = 45$. (a) Average of lesions across the age group (R_j/N_j). (b) Initial C values obtained by smoothing R_j/N_j, (c) Final C values after convergence, (d) θ values from initial C values, (e) $\hat{\theta}$ values estimated from final C values after convergence. (f) Difference between initial θ values and final $\hat{\theta}$ values. Note that in the younger age group there is not much difference (second row from top) in θ before (d) and after (e) convergence, since the neighbouring voxels are mostly non-lesions in the young population, whereas $\hat{\theta}$ values are larger and differ more in the bottom row between (d) and (e) as shown in (f) in older population.

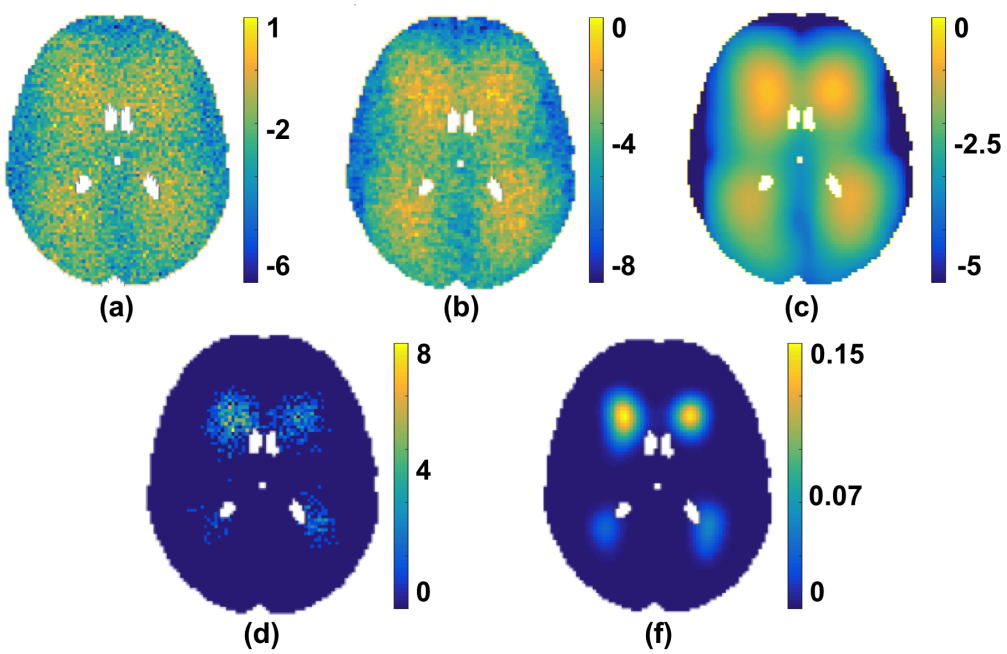


Figure 3.4: An instance of simulated lesion data (shown at slice $z = 45$ and age bin number $t = 20$). (a) Addition of random noise to μ indicating randomly sampled voxels, (b) lesion voxels after smoothing (a) with K, (c) $K * \mu$, (d) Sum of binary lesion maps (R_{sim}) in a specific age group, (e) Simulated θ_{true} image

our case, the MNI brain mask). For sampling, we used the smoothed average of lesion probability maps from a previous study (vascular cohort in [53]) as the ground truth probabilities to obtain a realistic WMH spatial distribution. The probability of sampling each lesion voxel in the data is initially represented by a random normal variable $N(\mu(x), 1)$. In order to avoid isolated lesion voxels and to establish spatial dependencies between the neighbouring voxels (so that sampled voxels would look like plausible lesions), we smoothed the sampled lesion voxels using a Gaussian kernel K with a standard deviation of 0.8. Since voxels of $\mu(x)$ could be negative when the ground truth probabilities were low, we obtained binary lesion maps by thresholding the smoothed lesion map $K * N(\mu(x), 1)$ above 0. As a result of convolution with K, the ground truth distribution at voxel x is now represented by $N(K * \mu, K^2 * M)$, where M is a binary brain mask. We obtained the true lesion probability for the simulation, θ_{true}, by calculating the cumulative distribution function of the normal variate $z = K * \mu / \sqrt{K^2 * M}$ or more specifically, $p(z > 0)$.

Applying our algorithm, for each age group, we accumulated the sum of binary

volumes as R_{sim} and applied our algorithm to R_{sim} and N_{sim} to model the distribution of lesions over the simulated population, to get $\hat{\theta}$ (algorithm output). An instance of data obtained from the simulation is shown in figure 3.4.

3.3.1 Effect of b-spline knot spacing

As a part of our simulation experiments, we studied the effect of the knot spacing on the final lesion probabilities. We analysed how it relates to the smoothness of kernel K used for the generation of the simulation data, by varying them independently and evaluated their combined effect on the algorithm result $\hat{\theta}$.

3.3.2 Comparison with existing alternative methods

We applied our algorithm to the simulated data and compared its performance with respect to some alternative methods with different levels of complexity. The simplest method for estimating the lesion probability with respect to age groups is to determine the group-level average of lesion probability maps as done in [50]. We performed the similar average on binary lesion maps and calculated $Average = R_t/N_t$ for each age group age_t. However, the resulting lesion probability map $Average$ is sensitive to the number of subjects in each age bin age_t and does not provide a continuous/generalised estimate of lesion distribution. Hence another option is to smooth the average lesion distribution map $Average$, using a Gaussian kernel with different standard deviations σ (0.5, 1.5 and 3.0) to get a more spatio-temporally continuous lesion probability map $Smoothed_Average(\sigma)$. We also applied the Bayesian Spatial Generalised Linear Mixed Model (BSGLMM) based on spatially varying coefficients proposed by [127]. We evaluated the performance of the above methods by determining the error values (difference between the ground truth probabilities and the outputs of the above methods) and the corresponding mean-squared error (MSE) values.

3.4 Application to real data

We applied our algorithm to a subset of subjects from the UK Biobank data and compared the lesion distribution between subjects with and without hypertension using the methods that provided better MSE values than those reported in the literature, as determined from tests on the simulated dataset. We further validated our observations by applying our method on a clinical dataset (OXVASC) consisting of a vascular population of subjects who had experienced a transient ischemic attack or minor stroke.

3.4.1 Evaluation on the UK Biobank dataset

Modelling the distribution of WMHs with respect to age and hypertension

The dataset used for initial validation is a subset of UK Biobank data [284]. In the UK Biobank data, about 10% of the sample has been identified as being diagnosed with hypertension, corresponding to around 7,975 subjects. Our goal was to model the lesion distribution with respect to age, comparing the subjects diagnosed with hypertension (HT group) to subjects without hypertension, where the latter act as a control group (non-HT group). In this case, to reduce the amount of computation, we selected a balanced subset of randomly sampled 1000 subjects, 500 subjects from the HT group (age range 45.5 - 78.3, mean age 66.3 $\pm$ 6.1 years, with female to male ratio, F:M = 202:298), while the remaining were from the non-HT group (age range 45.5 - 78.4, mean age 62.0 $\pm$ 7.6 years, with female to male ratio, F:M = 257:243). For more details on UK Biobank dataset and MRI parameters, refer to appendix A.

For these subjects, binary masks were generated using BIANCA [53] with information from FLAIR and T1-weighted images [87], to be used as input for our algorithm. The subject-level lesion probability maps obtained were binarised after thresholding at 0.8. We then registered the single subject binary lesion maps to the MNI space using linear and non-linear registrations, re-thresholding them at 0.5 and binarised them to compensate for interpolation. We then applied our algorithm to the two groups separately and compared the results with those from other methods. When

applying our method, we arranged the subjects in each group into one year age bins to form a 4D image of dimension $91 \times 109 \times 91 \times 33$. We compared the two groups using an unpaired t-test using permutation testing (randomise, [326]) to analyse the significance of the difference in the lesion probability values between HT and non-HT groups, and calculated the corresponding z-scores. Results were considered significant for $p_{corr} < 0.05$, corrected for multiple comparisons by using the null distribution of the max (across the image) voxelwise test statistic. For BSGLMM, the t-values were obtained by dividing the posterior mean by posterior standard deviation [127].

3.4.2 Evaluation on the OXVASC dataset

3.4.2.1 Modelling the distribution of WMHs with respect to age

We validated our algorithm on a clinical dataset of subjects at risk of vascular cognitive impairment. The dataset consists of MRI data from 474 consecutive eligible participants from OXVASC, who had recently experienced a minor non-disabling stroke or transient ischemic attack (TIA) (age range 20 - 102 years, mean age 67.4 $\pm$ 14.3 years, with female to male ratio, F:M = 240:234). For more details on the OXVASC dataset and MRI parameters, refer to appendix A. Similar to the UK Biobank dataset, we obtained the lesion binary masks using BIANCA and registered the single-subject masks to the MNI space. We grouped the data in the age bins of 3 years to form a 4D image of dimensions $91 \times 109 \times 91 \times 27$ and applied our model. We also performed permutation-based statistical analysis (corrected for multiple comparisons) to test the statistical significance of the effect of age on the lesion probabilities. For this analysis, in each iteration, we permuted the data with respect to the age groups and applied our model on the permuted data to generate results for that permutation, that were then used to assess statistical significance, with correction for multiple comparisons (using the maximum statistic across space).

3.4.2.2 Modelling the distribution of WMHs with respect to MoCA score

In addition to age, we modelled the distribution of WMHs on the OXVASC dataset considering MoCA score as the parametric factor of interest. In fact, apart from age,

cognitive impairment is another factor that has been widely associated with WMHs. In particular, it has been shown that MoCA score is correlated with the presence of WMHs in frontal periventricular areas [92]. Out of the initial 474 subjects, we considered the 411 subjects for which the MoCA score was available (MoCA scores range 12 - 30, mean MoCA score = 26.17 $\pm$ 3.54). We grouped the subjects having MoCA scores lower than 20 (N = 22) in a single bin and grouped the rest of the subjects in MoCA score bins of 1 to form a 4D image of dimensions 91 $\times$ 109 $\times$ 91 $\times$ 12 and applied our model.

3.4.3 Effect of initial values on our algorithm

3.4.3.1 Effect of initialisation on algorithm convergence

In order to test the robustness of our method, we evaluated the convergence of our algorithm by providing different starting points (initialisations) on the OXVASC dataset. Since the scope of our Bayesian modelling method is its application to WMHs of presumed vascular origin, we used the smoothed R_i/N_i map as our initialisation. Hence, as sensible options for different initialisations, we introduced perturbations in the R_i, smooth the ratio R_i/N_i and provide them as initialisation.

For providing different initialisations, we modified each voxel of the R_i map by sampling a value randomly within a specific percentage of the existing voxel value. For example, for a perturbation of 10% in a voxel which has an initial R_i value of 20, a value was chosen randomly in the interval [18, 22]. For this experiment, we considered three percentages of perturbation: 10%, 30% and 50%.

3.4.3.2 Effect of initial subject-level thresholds on our algorithm results

We also evaluated the sensitivity of our Bayesian modelling method with respect to the thresholds applied on the subject-level lesion probability maps obtained from BIANCA. To this aim, we used the 474 subjects from the OXVASC dataset. The subject-level lesion probability maps were obtained using BIANCA with the modalities and parameters specified in [53]. As shown in the supplementary figure S4 (F) of [53], threshold values of 0.85 to 0.95 gave reasonably good similarity indices (Dice

index) during BIANCA optimisation. The threshold value of 0.9 gave the maximum Dice index and hence was used for obtaining the binary lesion maps in [53] and in our work. We have now used the values 0.85 and 0.95 for the test.

3.5 Results

3.5.1 Results on the simulated data

3.5.1.1 Effect of b-spline knot spacing

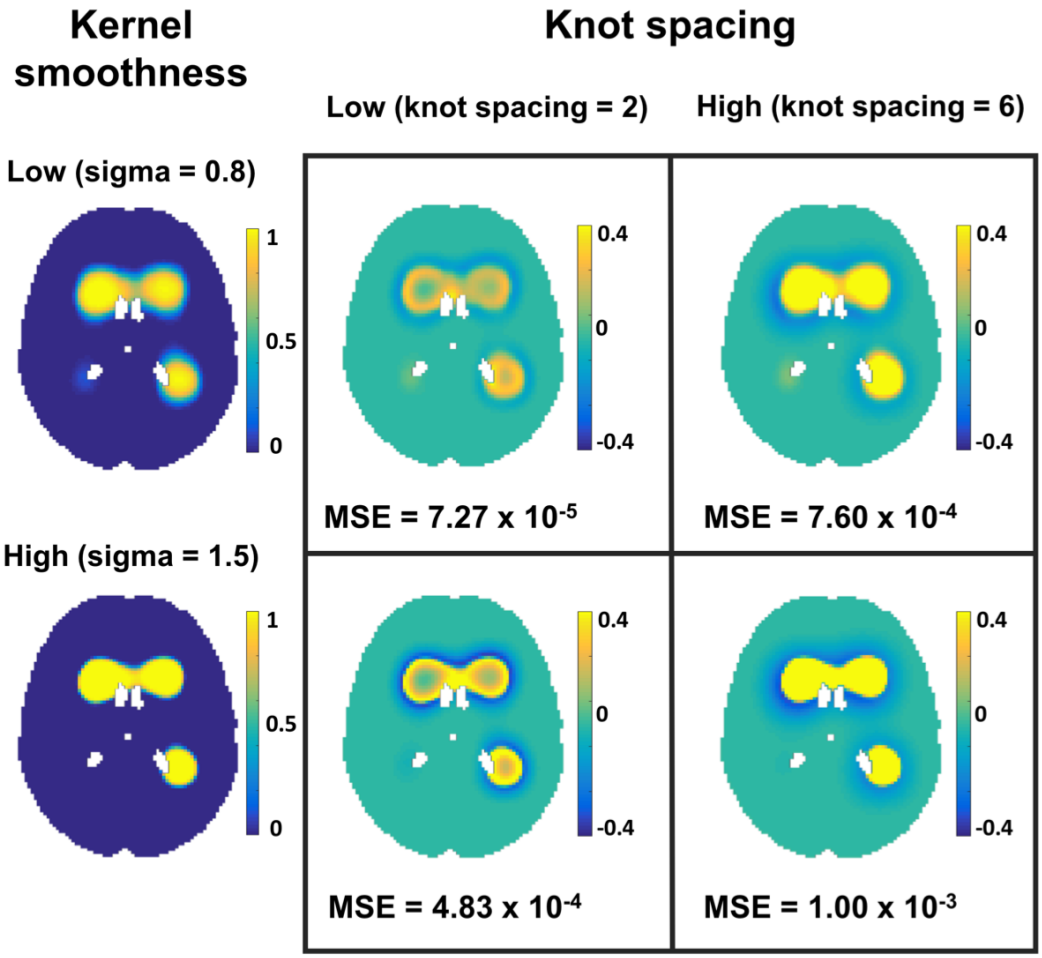


Figure 3.5: Comparison of error maps for various values of knot spacing in our Bayesian spline method $\hat{\theta}$ and standard deviation (sigma) values for smoothing kernel in simulation ground truth θ_{true}. Within the table, the top row shows error maps and mean-squared error (MSE) values, when θ_{true}, generated using a smaller symmetric kernel (sigma = 0.8) is compared against $\hat{\theta}$ modelled with smaller and larger b-spline knot spacing of 2 and 6 voxels respectively. Similarly, the bottom row shows error maps and MSE values for comparison of smoothing kernel K with larger symmetric kernel (sigma = 1.5) in θ_{true} against $\hat{\theta}$ with two different b-spline knot spacings.

Figure 3.5 shows the effect of change in spline knot spacing and in the standard deviation of kernel K used in the ground truth estimation independently on the error values on the simulated data. For this experiment we used knot spacings of 2 and 6 voxels in our algorithm, while in ground truth generation we used a Gaussian smoothing kernel K with standard deviation values of 0.8 and 1.5. The additional results for intermediate values (3, 4 and 5 voxels) are shown in appendix E.2. As we can observe from figure 3.5, the error image corresponding to the smaller knot spacing has lower MSE values of 7.27×10^{-5} and 4.83×10^{-4} for both the standard deviation values used in the smoothing kernel, and the MSE values increases for increase in the knot spacing. With the small knot spacing the splines are still able to model a smooth function with lower computational complexity, and gave lower error value for both the cases. On the other hand, the splines with larger knot spacing performed much worse with respect to smoother ground truth distribution (generated using standard deviation of 1.5) with the highest MSE of 1.00×10^{-3}. Also, with a large knot spacing there is a limitation to how well it can model less smooth functions. Hence, the MSE values are higher for the larger knot spacing for any given smoothness of ground truth distribution. Overall, we observed that the cubic b-splines with the smaller knot spacing of 2 voxels provided the best result on the simulated data and hence we used the same specifications henceforth on the simulated and real data experiments.

3.5.1.2 Comparison with the existing alternative methods

Figure 3.6 shows the error values of the outputs of the different methods with respect to ground truth, while table 3.1 reports the corresponding MSE values and running time. For the $Smoothed_Average(\sigma)$ method, we tried various values of standard deviations σ for the Gaussian kernel and have shown the results for the three σ values that provided minimum MSE values for this simulated dataset.

On a simulated dataset consisting of 698 subjects, we obtained an MSE value of 7.27×10^{-5} for our Bayesian spline method, while BSGLMM method gave a lower MSE value of 5.22×10^{-5} with ≈ 15 times slower running time. *Average* is the fastest (with < 1 second running time) but gives twice as much as error (18.790×10^{-5}) as

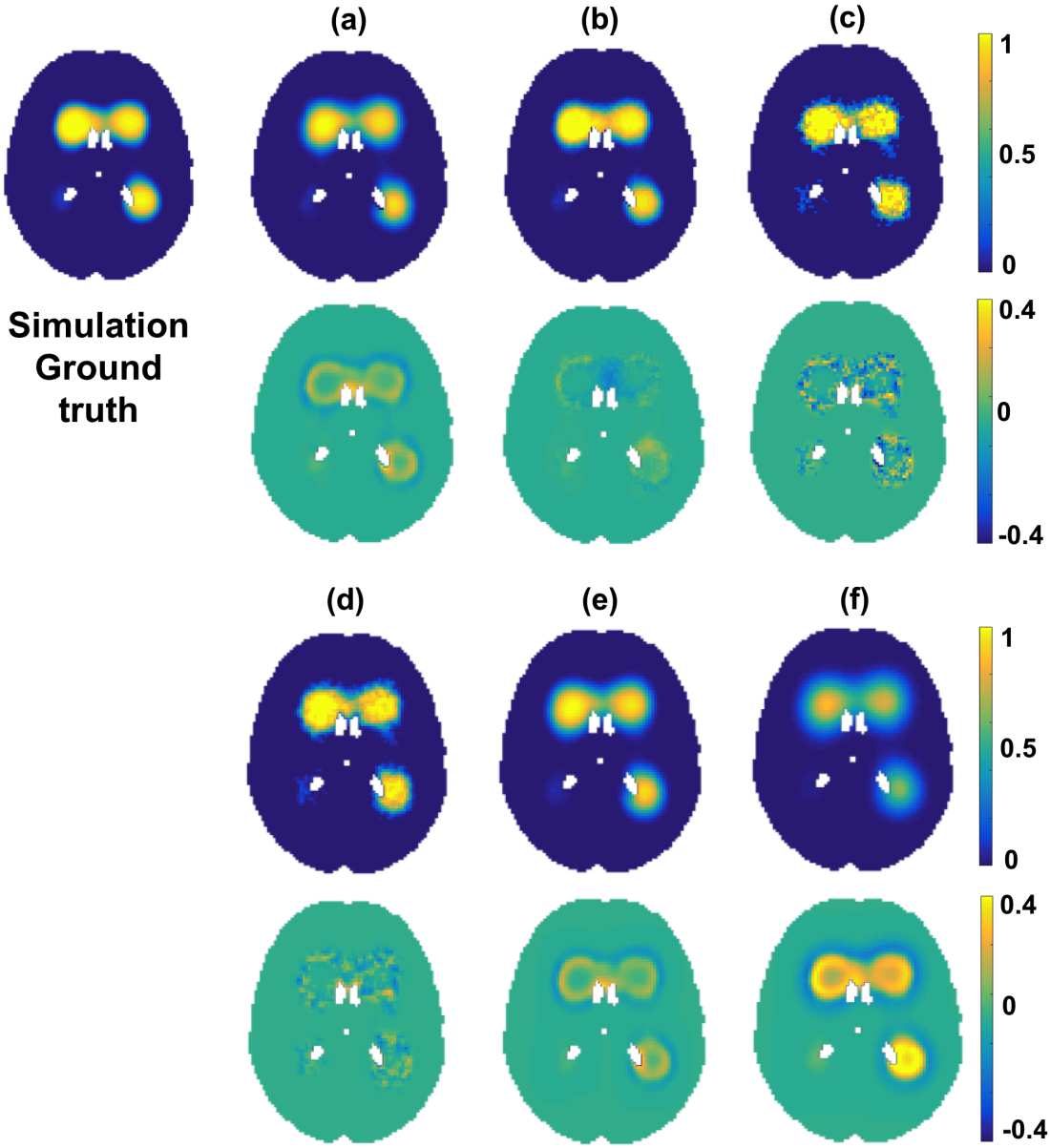


Figure 3.6: Comparison of simulation analysis results using different methods. The results are shown at slice $z = 45$ and age bin number $t = 33$. Ground truth distribution θ_{true} shown along with the outputs of various methods $\hat{\theta}$ (a - f) and the corresponding error maps immediately underneath the probability maps. The outputs are shown for our Bayesian spline method (a), BS-GLMM (b), *Average* (c), *Smoothed_Average*($\sigma = 0.5$) (d), *Smoothed_Average*($\sigma = 1.5$) (e) and *Smoothed_Average*($\sigma = 3$) (f).

our method. [127] reported an MSE value of 12×10^{-5} on simulated data consisting of 100 randomly sampled subjects. From table 3.1, we can observe that only *Average* and *Smoothed_Average*($\sigma = 3$) give MSE values greater than 12×10^{-5} (reported in [127]). Moreover, both MSE values and running times increases as the standard

Method	MSE values	Running time / Processor specs
Our Bayesian spline method	7.27×10^{-5}	$<$ 20 mins (84 secs per iteration) / iMac CPU with 2.9 GHz Intel Core i5 processor
BSGLMM [127]	5.22×10^{-5}	$\approx$ 293 mins / NVIDIA Tesla K80 GPU with 12GB
Average	18.79×10^{-5}	$<$ 1 sec / iMac CPU with 2.9 GHz Intel Core i5 processor
$Smoothed_Average(\sigma = 0.5)$	3.86×10^{-5}	$\approx$ 8 secs / iMac CPU with 2.9 GHz Intel Core i5 processor
$Smoothed_Average(\sigma = 1.5)$	8.56×10^{-5}	$\approx$ 11 secs / iMac CPU with 2.9 GHz Intel Core i5 processor
$Smoothed_Average(\sigma = 3)$	33.71×10^{-5}	$\approx$ 17 secs / iMac CPU with 2.9 GHz Intel Core i5 processor

Table 3.1: Comparison of MSE values and running time of different methods on simulation data.

deviation σ of the smoothing kernel increases due to increase in the kernel size.

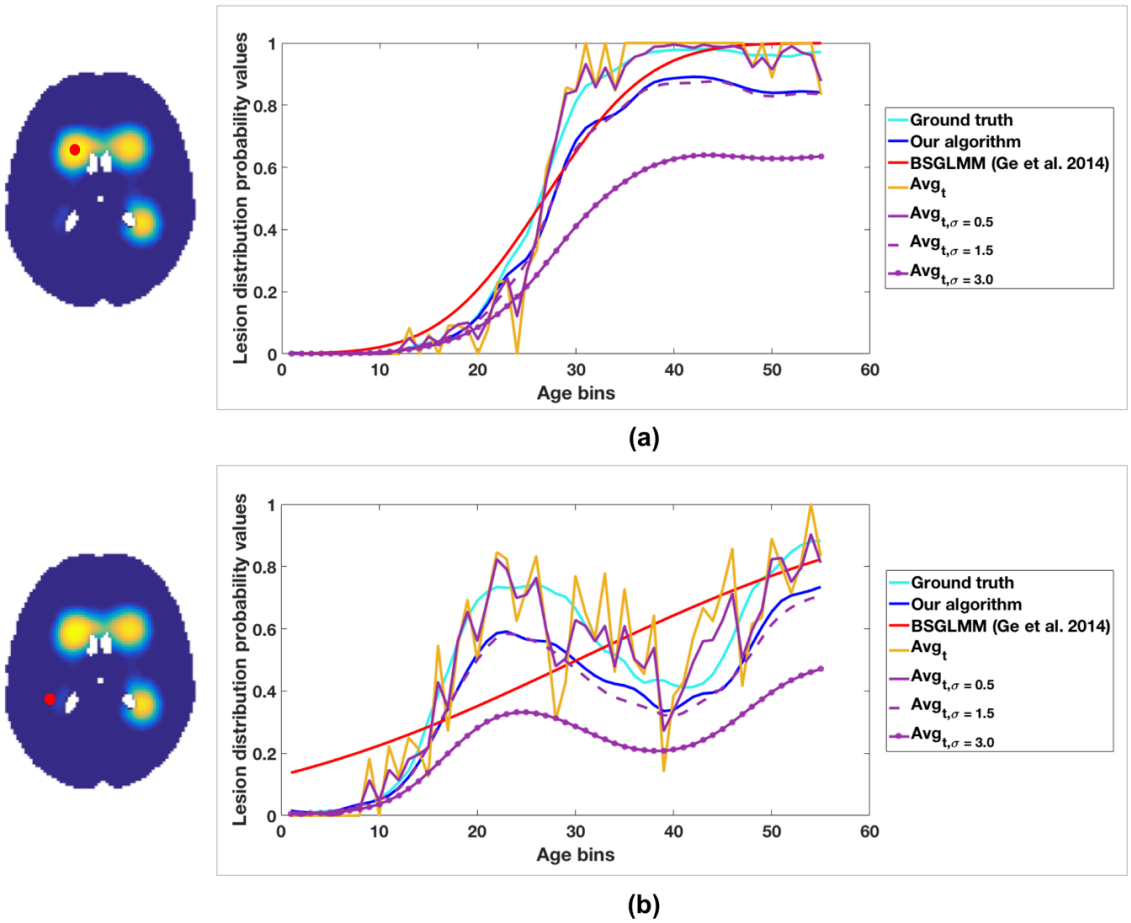


Figure 3.7: Lesion probability values plotted versus age. The plot of lesion probability values shown for (a) anterior and (b) posterior voxels. The red dot on the ground truth images indicate the voxel locations corresponding to the plots.

Figure 3.7 shows the lesion probability with respect to age at two representative voxels (one more anterior and one more posterior) for the tested methods. The curves obtained from *Average* and *Smoothed_Average*($\sigma = 0.5$) are very close to the ground truth, but they show spikes due to their dependence on the number and WMH characteristics of the specific subjects in each group. *Smoothed_Average*($\sigma = 1.5$) and our method give curves that capture the true trend quite well. BSGLMM posterior probabilities tend to increase monotonically with age throughout the white matter irrespective of the location of the lesions.

3.5.2 Results on the real data

3.5.2.1 Evaluation on the UK Biobank dataset

We show the results of our algorithm on the UK Biobank dataset to estimate the lesion distribution probabilities with respect to age within the HT and non-HT subjects in figure 3.8.

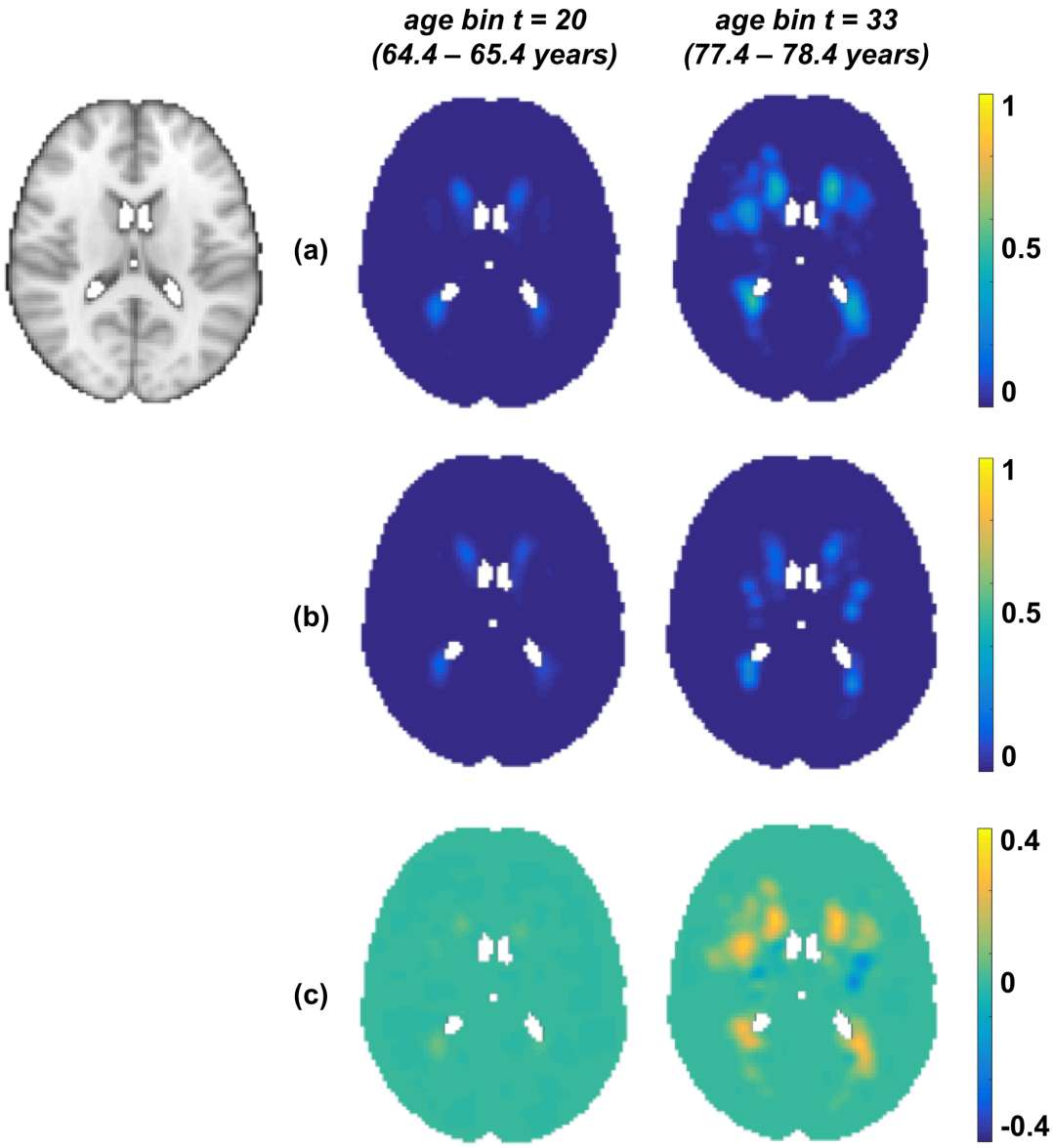


Figure 3.8: Results on the UK Biobank data. Output of our Bayesian spline method on HT group (a) and non-HT group (b) shown along with their difference maps (c). First column shows the age group t = 20 (age_t = 64.4 - 65.4 years) and second column shows t = 33 (age_t = 77.4 - 78.4 years). All the results have been shown at the slice z = 45 (MNI template on the left)

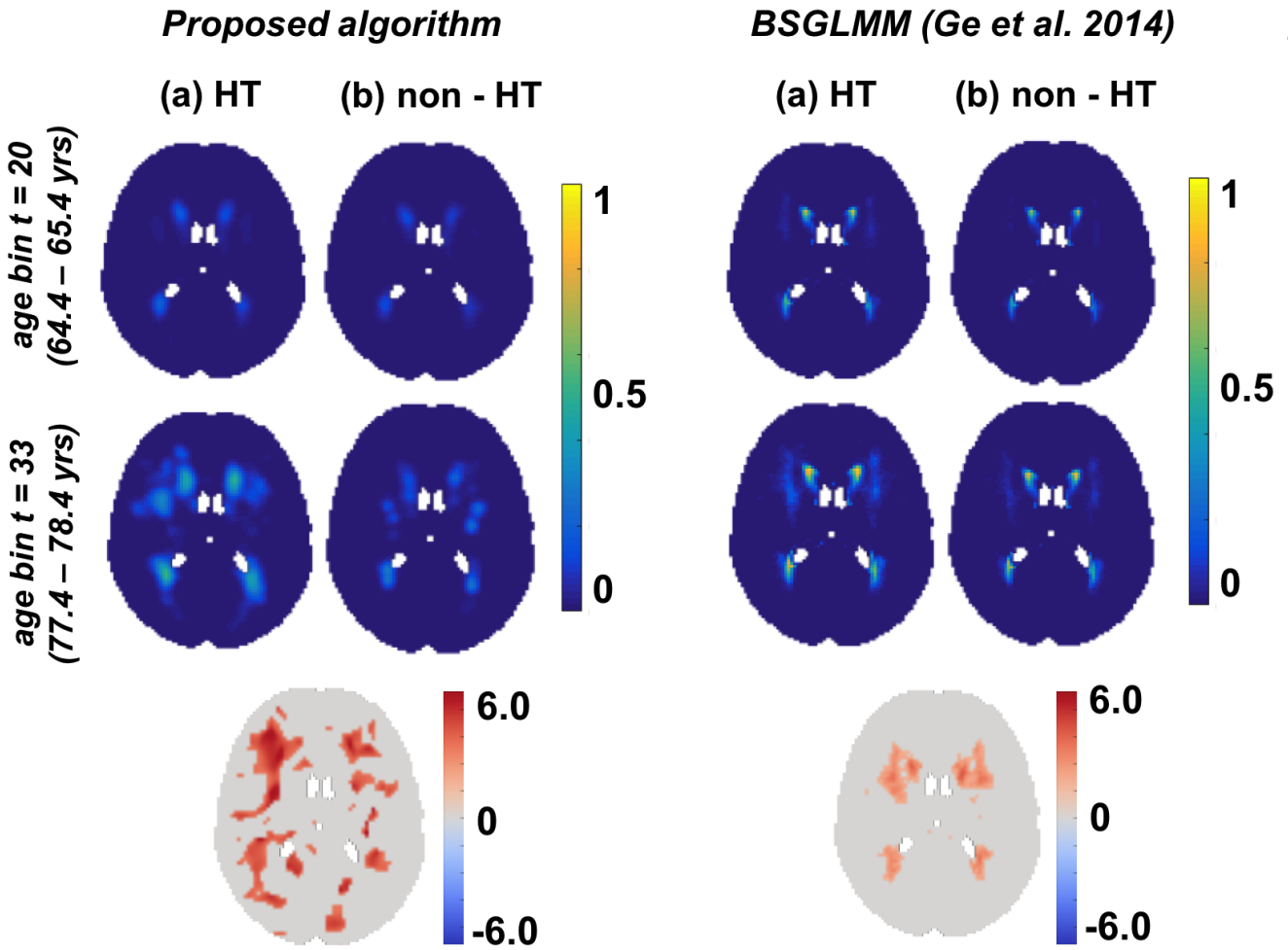


Figure 3.9: Comparison of the results of the Bayesian spline algorithm with BSGLMM on the UK Biobank data. Top two rows show the results of the methods for HT and non-HT groups for two age groups. The bottom row shows the z-scores in the significant regions ($p_{corr} < 0.05$), obtained from the significance test on the differences of the lesion probabilities between HT and non-HT groups obtained by the corresponding methods.

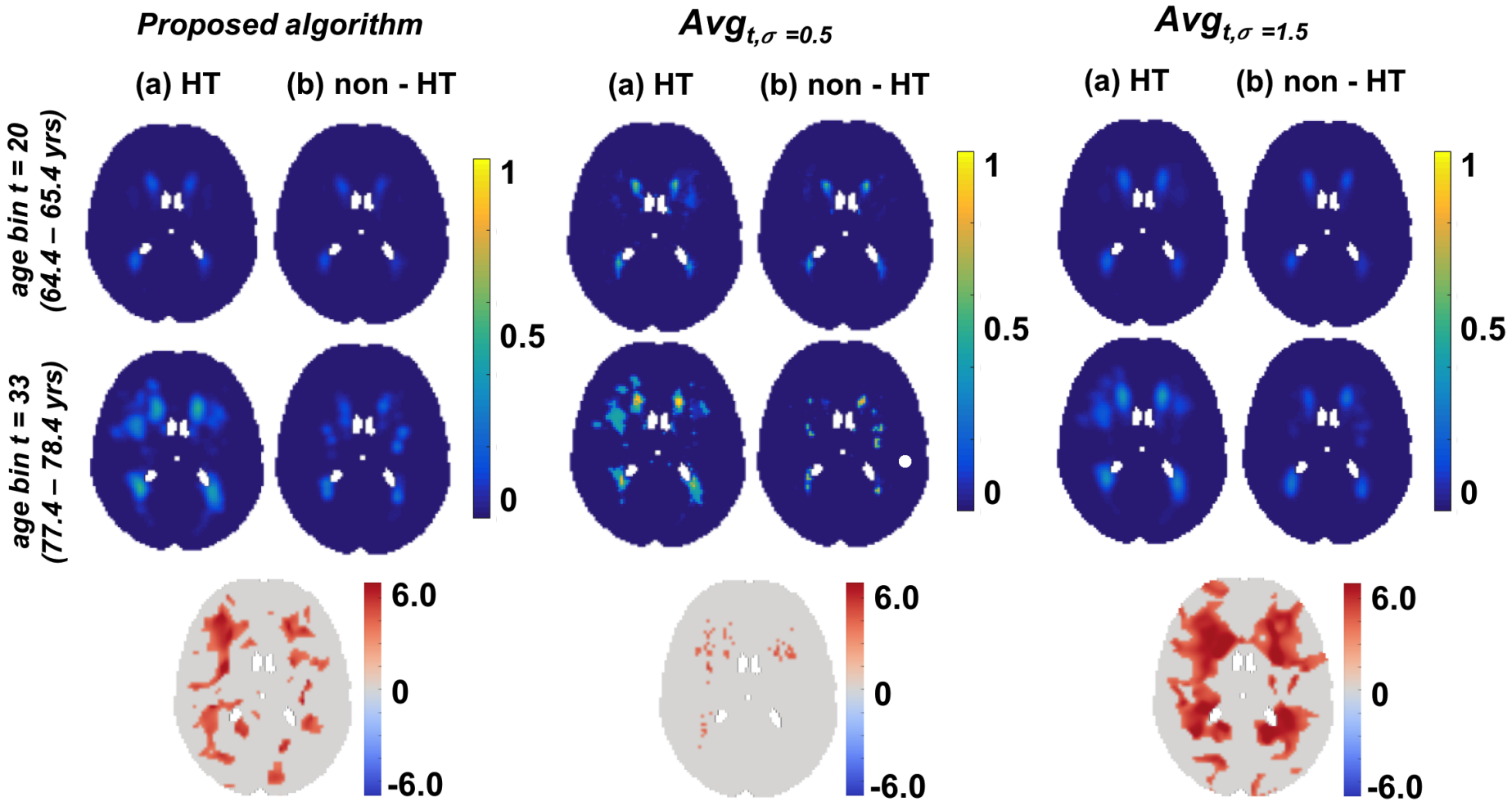


Figure 3.10: Comparison of the results of the Bayesian spline algorithm with $Smoothed_Average(\sigma = 0.5)$ and $Smoothed_Average(\sigma = 1.5)$ on the UK Biobank data. Top two rows show the results of the methods for HT and non-HT groups for two age groups. The bottom row shows the z-scores in the significant regions ($p_{corr} < 0.05$), obtained from the significance test on the differences of the lesion probabilities between HT and non-HT groups obtained by the corresponding methods.

The lesion probabilities increase with age for both groups and are higher in the HT group (≈ 0.4 as shown in fig. 3.8(c)) than the non-HT group. We can observe that this difference is higher in the older age groups (≈ 0.3 between the columns in row (c)) compared with the younger group. So not only is there a strong relationship between lesion probabilities and hypertension, but also this relationship is more pronounced as the population gets older.

In figures 3.9 and 3.10, we show the results of our algorithm compared with the posterior probabilities obtained from the BSGLMM method and the outputs of smoothed average maps ($Smoothed_Average(\sigma = 0.5)$, $Smoothed_Average(\sigma = 1.5)$) respectively. We chose to compare the above methods with our Bayesian spline algorithm since these methods gave MSE values less than 12×10^{-5} [127] on the simulated data (table 3.1). We observed that the lesion probabilities in the periventricular regions increase with age for both HT and non-HT groups, while the probabilities in the deep regions increase only in the HT group (between the first two rows in (a) in fig. 3.9 and 3.10). The bottom rows of figures 3.9 and 3.10 show the z-scores for significant regions ($p_{corr} < 0.05$), obtained from a statistical test to determine the significance of differences between the HT and non-HT groups. We observed increased lesion probabilities in the HT group compared with non-HT group, especially in the deep regions. This effect is particularly strong in our method and $Smoothed_Average(\sigma = 1.5)$ when compared with BSGLMM and $Smoothed_Average(\sigma = 0.5)$ (fig. 3.9 and 3.10 (bottom row)).

3.5.2.2 Evaluation on the OXVASC dataset

Figure 3.11 (top row) shows the estimated lesion probabilities from our Bayesian spline method on the OXVASC dataset for two representative age groups. Similar to our results on the UK Biobank data, it can be observed that the overall lesion probability values are higher in the older age group compared to the younger age group (≈ 0.2, between the second and third columns). On performing the permutation test, we observed that periventricular voxels change significantly with age ($p < 0.05$) as shown in figure 3.11 (bottom row).

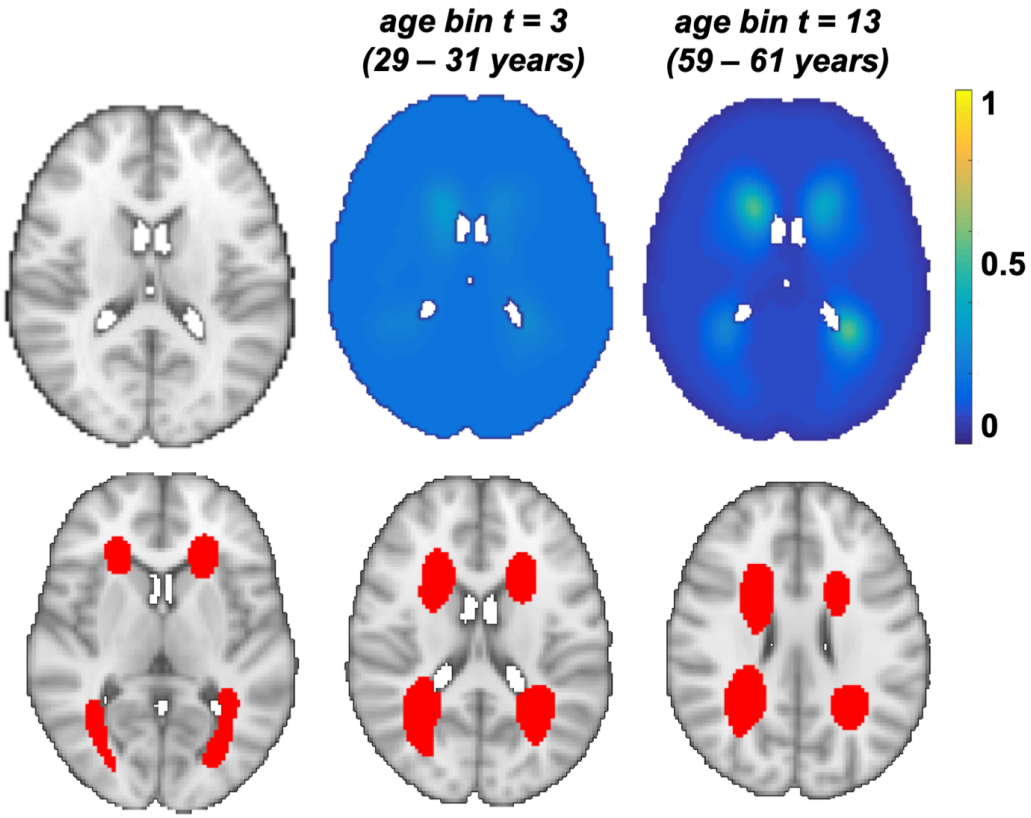


Figure 3.11: Results on the OXVASC data. (Top row) From left to right: MNI brain, lesion probability distribution at bin numbers $t = 3$ (age_t = 29 - 31 years) and $t = 13$ (age_t = 59 - 61 years). The results are shown at slice $z = 45$. (Bottom row) Results of permutation analysis. From left to right: voxels that change significantly ($p < 0.05$) with age (red) shown on slices $z = 39, 45$ and 49.

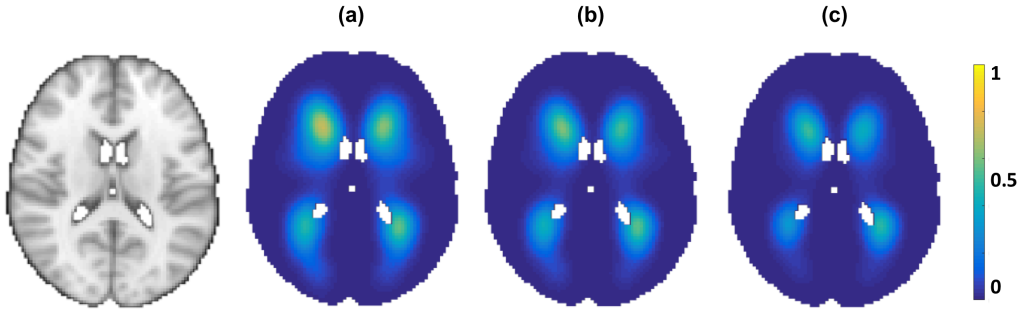


Figure 3.12: Population-level lesion probability map with respect to MoCA scores on the OXVASC data. The result of our Bayesian modelling method is shown for three representative cases: (a) MoCA score < 20 (severe cognitive impairment), (b) MoCA score $= 25$ (mild cognitive impairment) and (c) MoCA score $= 30$ (cognitively healthy). The MNI template is shown on the left and all the results have been shown at the slice z $= 45$.

Figure 3.12 shows the resulting population-level parametric lesion probability map obtained from our Bayesian modelling algorithm on the OXVASC data for three

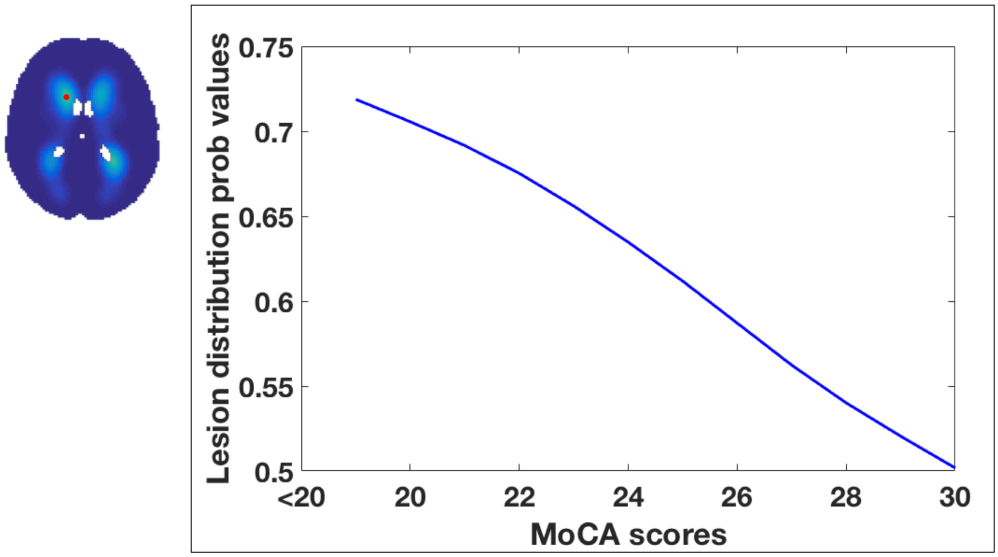


Figure 3.13: Plot of WMH probability values versus MoCA scores. The plot shows lesion probability values for a frontal periventricular voxel (indicated by the red dot in the image on the left).

MoCA scores: MoCA score < 20 (severe cognitive impairment), MoCA score $= 25$ (mild cognitive impairment) and MoCA score $= 30$ (cognitively healthy subject). We observed that the population-level WMH probability values decrease with increase in the MoCA scores. This trend can also be observed in the plot of lesion probability values shown for a frontal periventricular voxel in figure 3.13. This lower probability of finding WMHs in cognitively healthy subjects compared with the subjects having cognitive impairment is in line with the literature [92].

3.5.2.3 Evaluation of effect of initial values on our algorithm

Figure 3.14 shows the results of our Bayesian modelling method with different initialisations based on different percentages of perturbations. The figure shows the difference maps obtained by comparing the algorithm results obtained for different initialisations with the result for the original/unperturbed initialisation. From the figure, we can observe that for initialisations obtained by perturbing the R_i by 10% and 30%, the results of our Bayesian modelling method show very small differences with respect to our results obtained from the unperturbed R_i. The corresponding difference maps show negligible mean square difference (MSD) values of 4.43×10^{-6} and

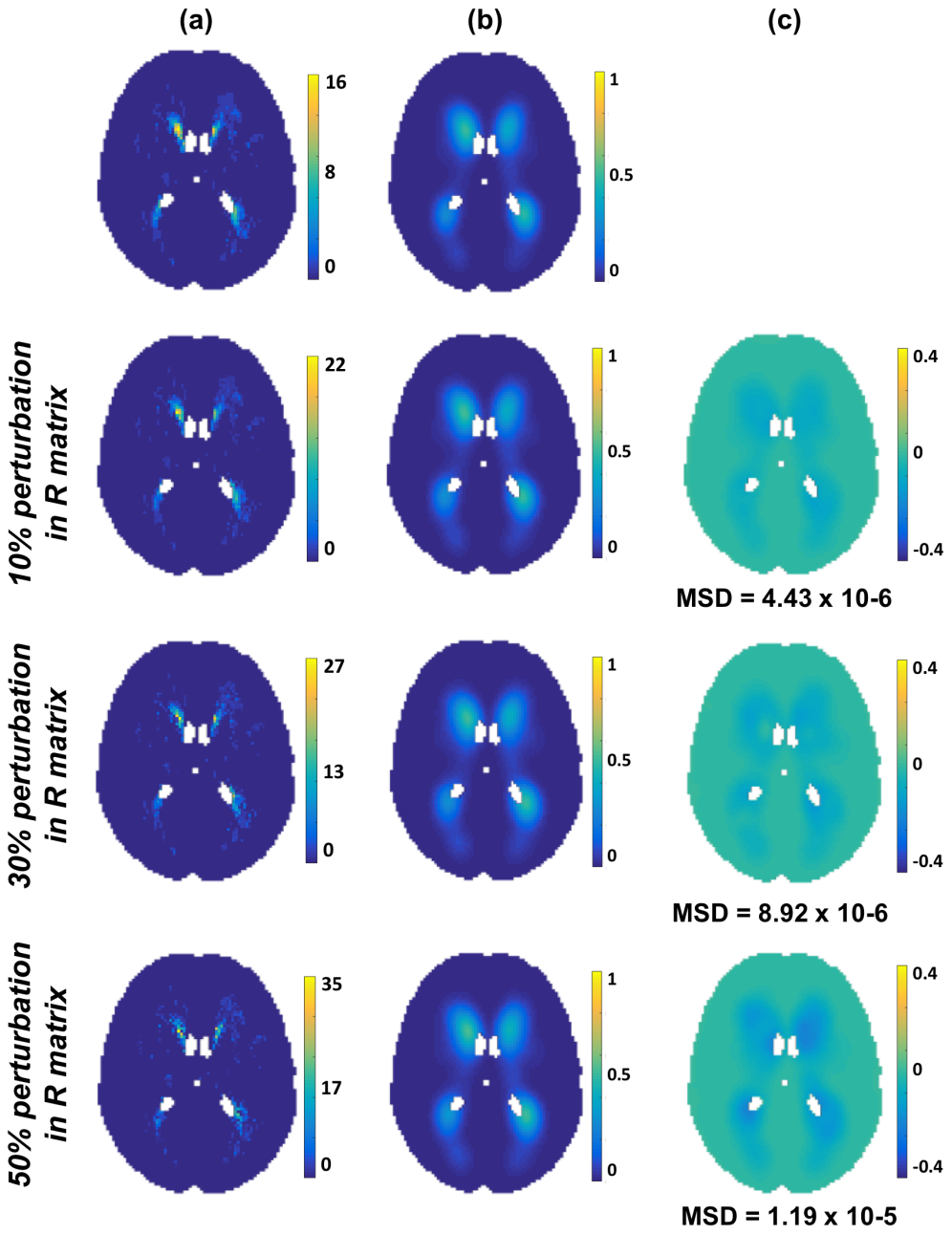


Figure 3.14: Results for different initialisations with various levels of perturbations. The second to fourth rows correspond to the results for different initialisations obtained by introducing 10%, 30% and 50% perturbation in the group-wise average 4D map R_i (top row, column a). (a) group-wise average 4D map R_i, (b) result of our Bayesian modelling method and (c) difference maps with respect to the result of initialisation from unperturbed R_i (top row, column b).

8.92×10^{-6} (with respect to the unperturbed case) for the perturbations of 10% and 30% respectively. Here, we use the term difference rather than error since we compare two different cases of algorithm results rather than with respect to the ground

truth. For comparison, these values are lower than the MSE value of 7.7×10^{-5} that we obtained in our simulation analysis for our method with respect to the ground truth distribution. The difference maps corresponding to the case of 50% perturbed initialisation gives an MSD value of 1.19×10^{-5} (which is quite close to the MSE values obtained in our simulation analysis).

These additional results show that our method is quite stable and converging on very similar results as shown numerically and visually (in figure 3.14). For robust results, up to 30% perturbations can be tolerated by the method, since the magnitude of the difference (from the unperturbed case) is smaller than the errors obtained in our simulation data.

Figure 3.15 shows the results of the Bayesian modelling method when our subject-level lesion probability maps were thresholded at 0.85 and 0.95, compared with those obtained at the threshold of 0.9. From the figure, we can observe that both the MSD values of 1.3×10^{-4} and 1.5×10^{-4} are higher than the MSE value of 7.7×10^{-5} obtained in our simulation analysis. This is due to the fact that changing the threshold value from 0.9 to 0.85 or from 0.9 to 0.95 results in the increase or decrease in lesion volumes (by $\approx$ 300 and 400 mm^3 per subject respectively) from the volumes obtained at the optimal threshold (0.9) on the binary lesion maps provided as input to our algorithm.

However, the overall relationship of the lesion distribution probabilities with respect to age remains same (i.e. lesion distribution probability values increase with age) for all the thresholds as shown in figure 3.16.

3.6 Discussion and conclusion

In this work we compared several methods including our proposed Bayesian spline method to model the distribution of WMHs with respect to a parametric factor within different populations. For the experiments in this chapter, we considered age as our parameter of interest, but the framework is general and can work with any continuous parameter. As an example, we modelled the distribution of WMHs with respect to MoCA scores on the OXVASC data.

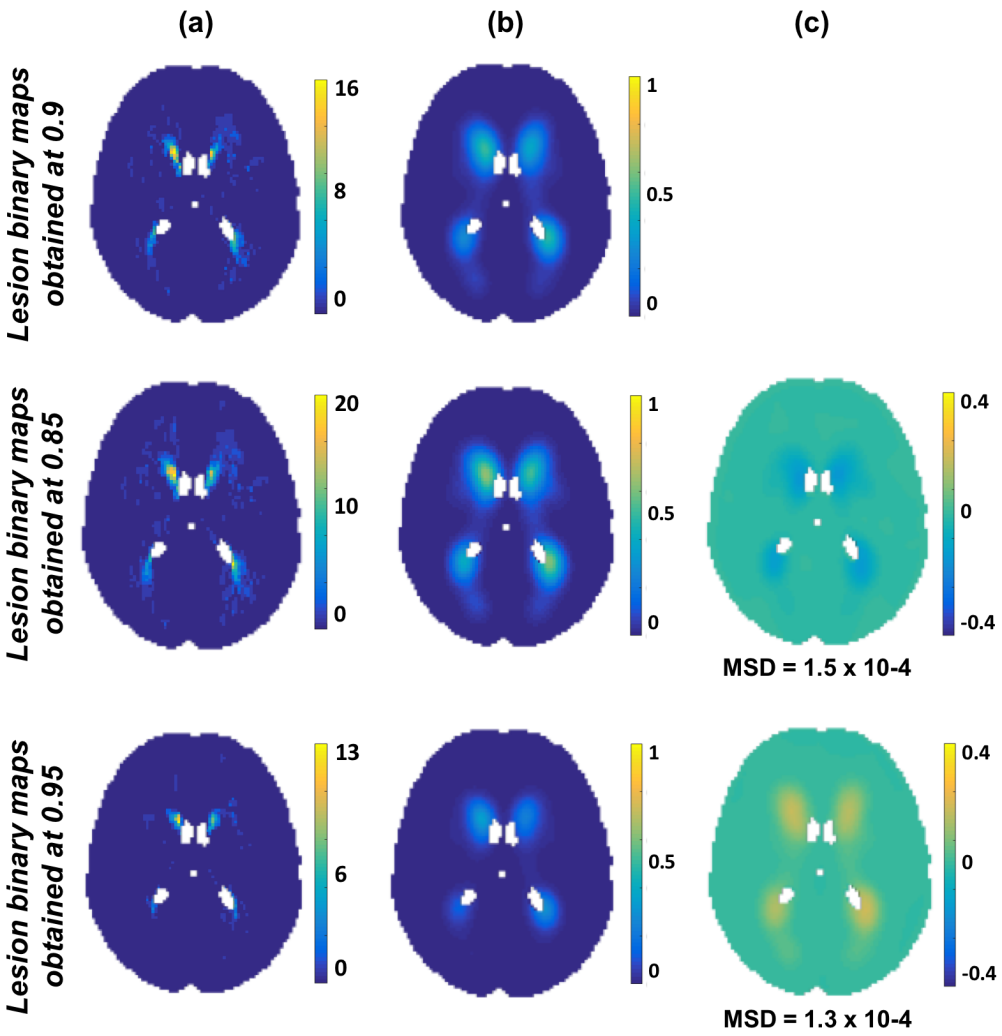


Figure 3.15: Results of the Bayesian modelling method when subject-level lesion binary maps are obtained using different thresholds. The rows correspond to the results for individual binary maps obtained using thresholds 0.9 (top row), 0.85 and 0.95 (middle and bottom row, respectively). (a) group-wise average 4D map R_i, (b) result of our Bayesian modelling method and (c) difference maps with respect to the result of binary maps obtained using the threshold of 0.9 (top row, column b).

During the development, we analysed the effect of various parameters of our algorithm on the resulting lesion probabilities. One of the most important parameters in the algorithm is the knot spacing in the spline model. We studied its effect and how it relates to the size of smoothing kernel K used for the generation of the simulation data by varying them independently and evaluated their combined effect on the algorithm result $\hat{\theta}$. The best results were obtained for the knot spacing of 2 voxels with MSE of 7.270×10^{-5}. We also evaluated the effect of the initialisation provided for

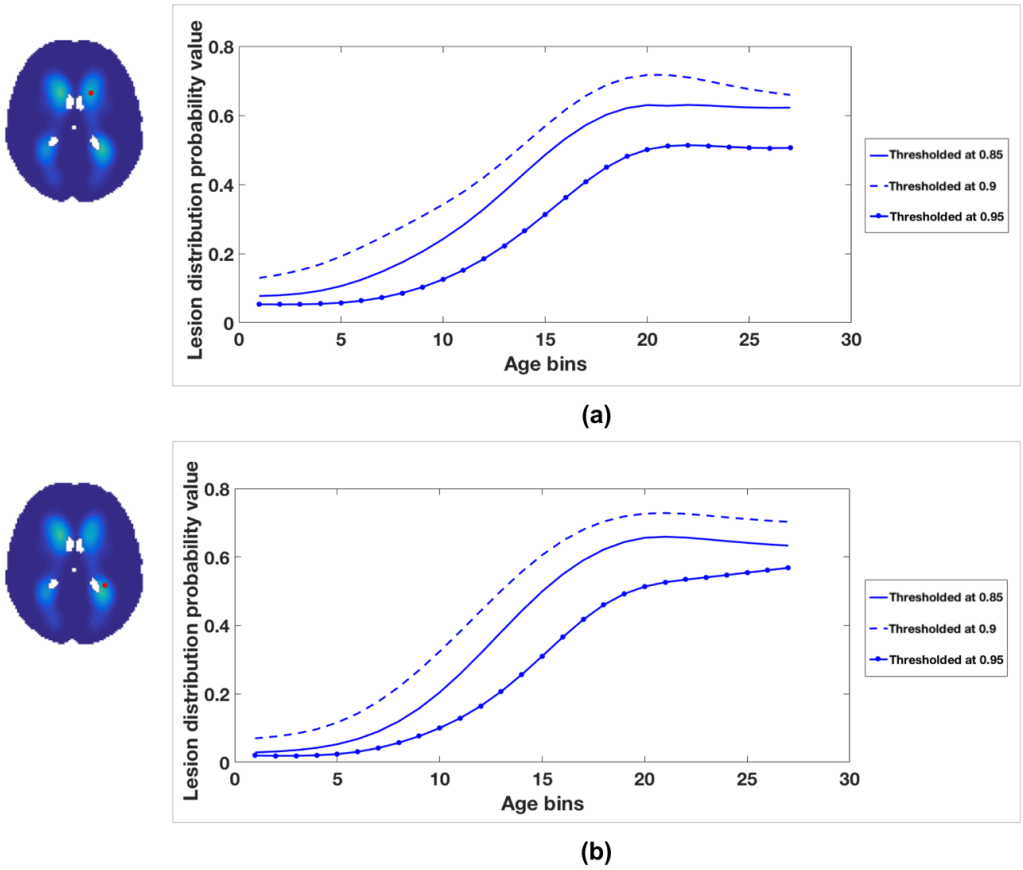


Figure 3.16: Lesion probability values plotted versus age for various thresholds. The plot of lesion probability values shown for (a) an anterior and (b) a posterior voxel. The red dots on the population-level lesion probability map indicate the voxel locations corresponding to the plots.

the optimisation function and observed that the algorithm converges quicker when we provide the smoothed average lesion map as the initial value for λ_i, whereas it fails to converge when a zero image or unity image is provided as initialisation for λ_i. Also, our algorithm converges to a similar output when the average lesion maps are smoothed with kernels of similar standard deviation. We additionally verified on the OXVASC data that our algorithm is robust with respect to changes in the initialisations due to various levels of random perturbations (up to 30%) in the group-wise average 4D map R_i.

When comparing the results of different methods on the simulated dataset, we found that the MSE values are lower than that reported in [127] for most of the methods, except *Average* and $Smoothed_Average(\sigma = 3.0)$. We obtained the high-

est MSE value and highest speed for the simplest method *Average*. However, this also gave a resulting image that was not as spatially continuous as the other methods. Moreover, the accuracy of *Average* (as well as of its Gaussian smoothed results $Smoothed_Average(\sigma)$) depends on the number of subjects in the age groups. While they provide reliable estimates in larger populations, their estimates are not so accurate when there are fewer subjects or when the lesion load is lower, since the data becomes sparse. Also, in the case of smoothed $Smoothed_Average(\sigma)$ method, determining the optimum value of standard deviation σ for the Gaussian kernel is dependent on the dataset and the lesion characteristics. However, the smoothed $Smoothed_Average(\sigma)$ outputs provide good initial estimates for our method and lead to better convergence, as we discussed above. We observed that, while results of both the Bayesian spline method and BSGLMM show relatively similar MSE values, the spline modelling in our algorithm is spatially more robust when compared to BSGLMM method. For instance, from figure 3.7, while lesion probabilities obtained from BSGLMM method fit well in periventricular areas, they do not fit the ground truth distribution as well in deep areas (figure 3.7). Another important advantage of our method is that it is much faster than BSGLMM (table 3.1) and can easily scale up to very large numbers of subjects. Our modelling algorithm takes less than 20 minutes (84 seconds per iteration), when run on an iMac with 2.9 GHz Intel Core i5 processor, irrespective of the number of subjects in the dataset. The main reason for the faster run time is that from the point R and N images are formed, number of subjects in the dataset becomes irrelevant. From then on, the amount of data depends on the spatial resolution and number of parameter bins. Since the time taken for forming R and N is relatively trivial, the run time is mostly taken by the optimisation steps. Another reason is that we independently perform the convolution in all 4 dimensions with a 1D spline, which reduces the computational load. Also, by calculating the analytical gradient (Eqn. 3.16) we avoid the numerical computation of gradient for 4D data in the optimisation step. Even though the maximum number of iterations allowed for the optimisation step is 50, our algorithm typically converges earlier (10 iterations on simulated data and < 20 on real data). On the other hand,

the BSGLMM algorithm is computationally demanding since it has to estimate the posterior distribution via Markov Chain Monte Carlo model using Gibbs sampler at each voxel [127].

Our results on the real data (the UK Biobank and OXVASC datasets) show that the lesion distribution probabilities are higher in older ages, which is in line with the existing literature [54, 87]). Moreover, the fact that we observed an increase in lesion probabilities in the deep regions in the HT group with respect to the non-HT group for all the methods (fig. 3.9 and 3.10), suggests that deep WMHs (DWMHs) might be associated with hypertension. The same phenomenon has also been reported in [50]. We further verified it with statistical tests and observed higher value of z-scores with greater significance ($p < 0.05$), particularly in the deep regions between the HT and non-HT groups. Though this is generally observable in all the methods, it is more evident for our Bayesian spline method and smoothed average map $Smoothed_Average(\sigma = 1.5)$ i.e. the deep regions are affected significantly in the HT group compared with the non-HT group for these two methods.

The results on the UK Biobank data are similar to those on the simulated data when comparing the Bayesian spline method with BSGLMM and $Smoothed_Average$. $Smoothed_Average(\sigma = 0.5)$, did not yield a smooth lesion distribution and had isolated voxels in the lesion probability maps. Since this method gave lower MSE on simulated data with respect to our method, and in the absence of ground truth, we cannot exclude the possibility that spline interpolation in our method induces spurious smoothness that is not inherent to the data. However, the simulation results also suggest that $Smoothed_Average(\sigma = 0.5)$ method produces less smooth data because it is affected by the sparsity of the observed data and hence is dependent on data/lesion characteristics. Also, results of the statistical test using $Smoothed_Average(\sigma = 0.5)$ show only a few isolated voxels that are significantly different between the HT group and the non-HT group, while the relationship between WMHs and blood pressure is well known [321, 50] and detected quite strongly by the other methods (including ours), that produce a smoother lesion distributions.

The significant increase in periventricular WMHs (PWMHs) is consistent between the two datasets, showing that our model is robust across datasets. We also verified that this trend of change in lesion probabilities with age remains the same irrespective of the initial thresholds used for obtaining subject-level lesion binary maps. As future work, the Bayesian spline method could be used to model the voxel-wise relationship between the distribution of WMHs and other population-level parametric factors that can be useful in the clinical context of the studied population. Another interesting future direction could be to extend the scope of our method in modelling the distribution of other types of lesions (e.g. Multiple sclerosis) and test the robustness of the spline method in modelling lesions that have a different spatial pattern from WMHs. In addition, the lesion probability distribution map obtained from our model can be used either as an additional feature or as a spatial prior to improve lesion segmentation algorithms, as described in chapter 4 and [327].

3.7 Summary

1. We propose a Bayesian inference based approach to model the voxel-wise distribution of WMHs within a population with respect to a parametric factor (e.g. age, cognitive function).

2. On a simulated data, our method provides an MSE value of 7.27×10^{-5}. This is comparable with existing methods, with the advantage of being computationally lean and providing spatially continuous population-level parametric lesion probability maps.

3. On evaluating our model on real datasets (the UK Biobank and OXVASC datasets), we observed that WMH distribution probabilities increases with age, especially in the periventricular region. Additionally, in the UK Biobank dataset, we observed the increase in DWMH probabilities with hypertension.

4. Our method is robust to the changes in initialisations (up to 30% of random perturbations). Also, the lesion probabilities remain consistent with initial thresholds used for obtaining the subject-level binary lesion maps.

Chapter 4

Improving the segmentation of white matter hyperintensities using BIANCA

The work described in this chapter aims to improve BIANCA [53] to overcome population- and subject-level variabilities in the lesion characteristics.

Section 4.2 describes the datasets used for our experiments, preprocessing pipeline and BIANCA options used for each dataset. Sections 4.3, 4.4 and 4.5 illustrate our experiments on three steps of BIANCA: feature extraction, classification and thresholding respectively. In particular, for the thresholding step, we propose the estimation of locally adaptive thresholds to automatically determine thresholds for different regions of the brain. Sections 4.7.1 and 4.7.2 report the results at the feature extraction step. Section 4.7.3 provides the results of the experiments in the classification step, while section 4.7.4 describes the results at the thresholding step using LOCATE. Section 4.8 contains the discussion on our results, conclusions and possible future directions for this work.

The preliminary results of this work were presented at the 2018 OHBM conference (Sundaresan et al., OHBM 2018) and the final results of this work are published in Sundaresan et al., 2019 [79].

4.1 Introduction

Accurate automated detection of WMHs is very challenging due to the high variability of their characteristics both between- and within- subjects. For example, at the population-level, the amount and distribution of WMHs have been associated with various factors such as cognition, vascular risk factors and neurodegenerative diseases [318, 95]. WMHs also occur more commonly in older age [54, 328]. At the subject-level, generally WMHs exhibit spatial heterogeneity and do not appear with the same contrast in all regions of the brain on structural MR images [329]. For example, typically, PWMHs are brighter on T2-weighted, FLAIR or proton density images and bigger compared to DWMHs [30], and hence can be detected more easily.

The variability in lesion characteristics can affect the performance of supervised lesion segmentation methods at different stages. The estimation of the probability for each voxel to be a lesion (lesion probability map) is affected by population-level factors (such as age, cognition, vascular risk factors and pathological conditions) [50, 92] and lesion load [57, 55]. Once an individual subject's lesion probability map is generated, the choice of the threshold used to obtain a binary map is affected by lesion load and the variation in local intensities. In fact, heterogeneity in signal intensity affects lesion probability values determined by segmentation algorithms [56, 55, 57], since brighter lesions (e.g. PWMHs) are typically assigned higher probabilities than those with lower contrast (e.g. DWMHs). This causes difficulty in determining the optimal threshold for obtaining the final binary lesion map. In fact, applying a global threshold (i.e. the same threshold to all the voxels of the lesion probability map) typically results either in the exclusion of deep lesions or overestimation of periventricular lesions. This often results in a trade-off between sensitivity and specificity of lesion detection [55]. An alternative proposed by [330] consists of determining thresholds based on Fazekas score or by visual inspection of individual subjects. However, these thresholds would also be global and determining them would require manual intervention.

We aim to improve our existing lesion segmentation with BIANCA by systematically exploring ways of improving various steps of the lesion probability map estimation, and also by developing a method for thresholding that takes into account the spatial variability of the lesion probability map. Figure 4.1 shows the block diagram describing the strategies explored for improving BIANCA performance.

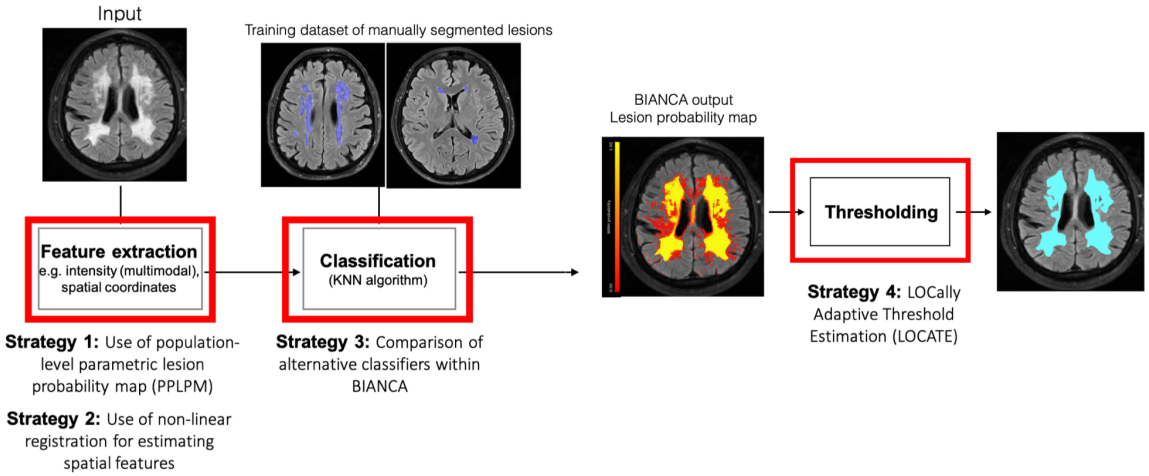


Figure 4.1: Block diagram showing the strategies explored for improving BIANCA performance. The red highlighted blocks indicate the steps we focused on in this work.

The first strategy consists of the inclusion of population-level WMH probabilities with respect to a demographic factor (in our case age) at the feature extraction step. In chapter 3, we modelled the spatial distribution of WMHs within a population with respect to age to obtain a population-level parametric lesion probability map (PPLPM). Here we used it as either a spatial prior or an additional feature in the WMH segmentation to explore the improvement in the results. The second strategy we tested is the use of non-linear registration for spatial feature extraction instead of the linear registration currently used in BIANCA.

At the classification step, we explored the performance of alternative machine learning methods as the third strategy. We have provided a detailed account of other supervised classifiers used in the existing literature in chapter 2. In particular, neural networks [57], support vector machines and adaboost classifiers [34] have been already

used for WMH segmentation, while random forests have been used for the segmentation of various structures and pathological signs that have similar characteristics to WMHs on structural MR images [331, 332, 333]. Therefore, in the classification step of BIANCA, we explored these above four supervised classifiers and compared their segmentation results with those obtained from the existing K-nearest neighbour (KNN) classification algorithm in BIANCA.

Our fourth strategy aims to improve the thresholding stage. We propose LOCally Adaptive Thresholds Estimation (LOCATE), a supervised method to determine optimal local thresholds for binarising the subject-level lesion probability map, taking into account the spatial heterogeneity in lesion probabilities. In this work, we present LOCATE with specific applicability to BIANCA, however, in principle it can be applied to the lesion probability maps obtained from a different method, provided the availability of training data with manual lesion masks. We evaluated LOCATE on various clinical datasets to observe the performance of LOCATE on data acquired under different protocols, from subjects with different lesion loads, distributions and patterns.

4.2 Dataset details and preprocessing

4.2.1 Datasets

In this work we used five datasets to study the effect of variations in lesion characteristics on the WMH segmentation. The datasets are diverse in terms of acquisition, population and WMH load and characteristics. A brief overview of the datasets are provided below. For more details regarding MRI acquisition parameters, refer to appendix A.

1. **Neurodegenerative cohort (NDGEN):** The dataset includes MRI data from 9 subjects with probable Alzheimer's Disease, 5 with amnestic mild cognitive impairment and 7 cognitively healthy control subjects recruited from the Oxford Project to Investigate Memory and Ageing (OPTIMA) study and from

the Memory Assessment Clinic at the John Radcliffe Hospital in Oxford [334]. Age range 63 - 86 years, mean age 77.1 ± 5.8 years, median age 77 years, F:M = 10:11. The total brain volume range: 1189282 - 1614799 mm^3, median: 1424669 mm^3. Manual segmentations are available with the WMH load range: 1878 - 89259 mm^3, median: 20772 mm^3.

2. **Vascular cohort - Oxford Vascular Study (OXVASC):** The dataset consists of 18 participants in the OXVASC study [323], who had recently experienced a minor non-disabling stroke or transient ischemic attack. Age range 50 - 91 years, mean age 73.27 ± 12.32 years, median age 75.5 years, F:M = 7:11. The total brain volume: brain volume range: 1290926 - 1918604 mm^3, median: 1568233 mm^3. Manual segmentations are available with WMH load range: 3530 - 83391 mm^3, median: 16906 mm^3.

3. **Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) cohort**: The dataset consists of 15 patients with CADASIL (age range 33 - 70 years, mean age 53.73 ± 11.31 years, median age 55.5 years, F:M = 11:4. The total brain volume range: 1226777 - 1603417 mm^3, median: 1396526 mm^3. Manual segmentations are unavailable for this dataset. However, in order to obtain binary lesion mask from the lesion probability map from BIANCA, a lower threshold of 0.2 was empirically chosen (based on visual inspection of the lesion probability map) by an expert neurologist as the optimal global threshold for this specific dataset.

4. **Healthy controls (HC)**: This dataset consists of 19 healthy controls, age-matched with the CADASIL dataset subjects with age range 29 - 70 years, mean age 54.58 ± 11.25 years, median age 57 years, F:M = 6:13. The total brain volume range: 1313929 - 1679837 mm^3, median: 1446443 mm^3. Manual segmentations are not available for this dataset.

5. **MICCAI WMH segmentation challenge Dataset (MWSC)**: The dataset consists of 60 subjects from three different sources (20 subjects each) provided

as training sets for the challenge (`http://wmh.isi.uu.nl/`): UMC Utrecht, NUHS Singapore and VU Amsterdam. The brain volumes are 1257820 - 1844920 mm^3 (median 1473389 mm^3) for UMC Utrecht, 1147248 - 1532268 mm^3 (median: 1351325 mm^3) for NUHS Singapore and 1219614 - 1787321 mm^3 (median: 1441201 mm^3) for VU Amsterdam. Manual segmentations are available for all three datasets, with an additional exclusion label provided for other pathologies. In the challenge, the masks with exclusion labels were ignored during performance evaluation. However, we included these masks as parts of non-lesion tissue, during the calculation of the performance metrics, for a more stringent evaluation in the presence of other pathologies. The WMH volume ranges (excluding other pathologies) are 845 - 74991 mm^3 (median: 26240 mm^3) for UMC Utrecht, 786 - 61332 mm^3 (median: 17795 mm^3) for NUHS Singapore and 1522 - 43528 mm^3 (median: 6015 mm^3) for VU Amsterdam. For more details regarding MRI acquisition parameters and the image dimension details, refer to `http://wmh.isi.uu.nl/`. Also, in the challenge the performance metrics were calculated on independent unseen test datasets. Since we did not participate in the challenge, we used publicly available training datasets for evaluation, using leave-one-out cross validation. Therefore, our results are not directly comparable with those published after the outcome of the challenge.

4.2.2 Image preprocessing

We skull stripped the images (used for intensity features) using FSL BET [335], followed by bias field correction using FSL FAST [336]. Diffusion-weighted images in OXVASC, CADASIL and HC datasets were preprocessed as described in [86] to extract mean diffusivity (MD) map, used as an additional intensity feature. We registered the image modalities to the base modality (in our case, FLAIR for all datasets) using linear rigid-body registration with FSL FLIRT [337].

4.2.3 BIANCA features and training options

Currently, the default options in BIANCA are: spatial weighting (sw) = 1, no patch, subjects included in the training set = any WMH load training subjects, location of training points = any location for non-WMH training points, number of training points = Fixed + equal with 2000 training points. For more details regarding the BIANCA options, refer to [53].

For NDGEN, we used FLAIR + T1 as features. Other than the default options, we used the following non-default options: location of training points = no border, number of training points = Fixed + unbalanced with 2000 lesion points and 10,000 non-lesion points. The 'no border' option and fixed + unbalanced lesion and non-lesion points have been shown to provide the best results during initial validation of BIANCA [53] and also in our initial tests on the same data. Also, we found that 2000/10,000 lesion and non-lesion points combination provided the best results. For OXVASC we used FLAIR + T1 + MD as features. The non-default options used were: sw = 2, 3D patch with patch size of 3. Due to the anisotropic nature of voxels, additional intensity features obtained by averaging over a smaller 3D patch provided better results during the initial tests [53]. For CADASIL and HC, we used the same features and BIANCA options used for OXVASC except spatial weighting. We excluded spatial features by setting sw = 0 for these two datasets, since their lesion pattern does not follow those of other datasets (especially in the anterior temporal lobes, which is typical of the CADASIL subjects). For MWSC dataset, we trained BIANCA separately on the three datasets (UMC Utrecht, NUHS Singapore and VU Amsterdam), using the same features and BIANCA options used for NDGEN.

4.3 Feature extraction

4.3.1 Use of population-level parametric lesion probability map (PPLPM)

Population-level parametric lesion probability maps (PPLPMs) describe the pattern of WMH distribution with respect to a specific parametric factor. Therefore, they

could provide useful additional information to improve WMH segmentation. For our experiments in this work, we considered age as our parametric factor of interest, since the relationship has been established between age and WMH distribution [54, 328]. In chapter 3, we modelled the PPLPM with respect to age within a population consisting of 474 subjects from the OXVASC dataset. We used the modelled PPLPM here in BIANCA in two ways: either as an additional feature to the KNN classifier, or post-multiplying the PPLPM with the subject's lesion probability map that is obtained using BIANCA with existing features.

For the first experiment, we implemented the PPLPM in BIANCA as follows: (1) we included the PPLPM (a 4D volume in the MNI space with the 4^{th} dimension representing different age groups) as an additional input, (2) we included age as an extra input in BIANCA and used it to select the appropriate 3D map from the PPLPM corresponding to the specific age group and (3) we transformed the 3D map from the MNI space to the subject's native space non-linearly using FSL FNIRT [338] before extracting its probability values as features to be used by the KNN classifier. We also tested the effect of the average 3D map across all age groups instead of the age-specific one. For the second experiment, we multiplied the age-specific 3D map from PPLPM (in the subject's native space) with lesion probability map obtained from BIANCA (with the existing options and no additional features).

We performed both experiments on the NDGEN dataset (keeping all the other options constant, as optimised in [53] and specified in section 4.2.3), and evaluated the performance with respect to the manual masks as described in section 4.6.

4.3.2 Use of non-linear registration for estimating spatial features

Currently in BIANCA, all the intensity features are calculated in native space, and a linear transformation is required as input to extract the spatial features (spatial coordinates in MNI space). However, a non-linear registration is, in general, more suitable for registering single-subject images to MNI, especially for pathological brains. Hence,

we performed experiments with non-linear registration for getting spatial coordinates in MNI space, to test the improvement in BIANCA performance.

4.4 Classification: comparison of alternative classifiers within BIANCA

Currently, BIANCA is based on the KNN algorithm. We assessed the performance of four other classifiers: random forest (RF), neural network (NN), support vector machine (SVM) and adaboost (AB). A brief overview of the above classifiers has been provided in appendix B. First we optimised the parameters for each classifier based on the area under the curve values from ROC curves, and then compared their results with KNN. Table 4.1 provides the list of available parameters (from the python *scikit-learn* package) and the parameters that were considered for tuning in the optimisation step.

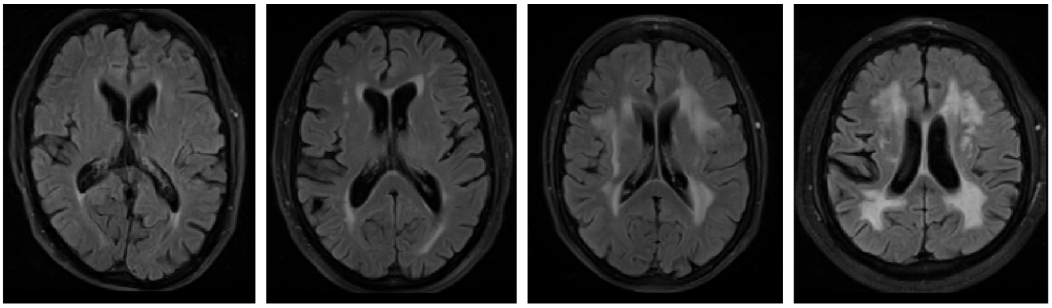


Figure 4.2: Examples of subjects with different WMH loads used for parameter tuning of the classifiers: (a) low, (b) medium, (c) high and (d) very high lesion load. The manually segmented lesion volumes for low, medium, high and very high lesion loads are 5409 mm^3, 21005 mm^3, 50585 mm^3 and 89259 mm^3 respectively.

We performed the initial tuning of parameters (from the python *scikit-learn* package) for the classifiers, experimenting with their values exhaustively within a predetermined range (refer to the table 4.1). We determined this range of parameters by running initial tests on a selection of four subjects from the NDGEN dataset with four different lesion loads: low, medium, high and very high (ranging from 5409 - 89259 mm^3, see figure 4.2). After optimisation, we applied the four classifiers on the remaining 17 subjects from the NDGEN dataset (keeping all the other options

Classifier	**Available main parameters and default values (in python *scikit-learn* package)**	**Range of parameter values considered for tuning**
Random forest (RF)	No. of trees - 10 Criteria for splitting - MSE Maximum depth of trees - Till the leaves are pure Minimum no. of samples required for a split - 2 Minimum samples at each leaf node - 1 Maximum number of features for each tree - No. of features Minimum impurity measure at each split - 1×10^{-7} Bootstraping done - True Out-of-bag error score - False	No. of trees - 20, 30, 40, 50, 100, 1000 Minimum samples at each leaf node - 200, 400, 500, 600, 800, 1000 Minimum impurity measure at each split - 1×10^{-7}, 1×10^{-6}, 1×10^{-5}, 1×10^{-4}, 1×10^{-3}, 1×10^{-2} Maximum depth of trees - 25, 50, 100
Neural network (NN)	No. of neurons in a hidden layer - 100 Activation function - RELU Output activation function - Logistic sigmoid Solver - ADAM ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1 \times 10^{-8}$) No dropout L2 regularisation, Alpha - 1×10^{-4} Change in learning rate - Constant Initial learning rate - 1×10^{-3} Maximum no. of iterations - 200 Tolerance - 1×10^{-4} Momentum - 0.9 Fraction of samples for validation - 0.1	No. of neurons in a hidden layer - 100, 120, 150 Solver - LBFGS, ADAM Initial learning rate - 1×10^{-4}, 1×10^{-3} Maximum no. of iterations - 1000, 1500, 1800 Tolerance - 1×10^{-4}, 1×10^{-3}, 1×10^{-2}
Support vector machine (SVM)	Type of kernel - Radial basis function (RBF) γ - 1/(no.of features) Tolerance - 1×10^{-3} C-value - 1.0 Epsilon - 0.1 Maximum no. of iterations - Until convergence	γ - 0.001, 0.01, 0.1, 1.0, 10 C-value - 0.1, 1.0, 10 Maximum no. of iterations - 1000, 2000, 3000, 5000, 6000
AdaBoost classifier (AB)	Base estimator - Decision Tree Regressor No. of trees in base estimator - 1 No. of estimators - 50 Learning rate - 1.0 Loss function - Linear	No. of trees in base estimator - 2 No. of estimators - 30, 40, 50 Learning rate - 0.5, 0.75, 1 Loss function - linear, exponential, square

Table 4.1: Parameters considered for tuning different classifiers.

constant, as specified in section 4.2.3) and evaluated the performance with respect to the manual masks as specified in section 4.6. In order to test the consistency of the results across various cohorts, we tested the performance of the best performing alternative classifiers in the three cohorts from the MWSC dataset.

4.5 Thresholding: LOCally Adaptive Threshold Estimation (LOCATE)

In order to overcome the impact on spatial heterogeneity of lesion probabilities, we propose a method that determines spatially adaptive thresholds at different regions of the lesion probability map.

4.5.1 Proposed method

LOCATE takes as input the subject-level lesion probability map obtained from a lesion segmentation algorithm (in our case, BIANCA run with the options specified in section 4.2.3) and estimates local thresholds in three steps by dividing the lesion probability map into sub-regions (using Voronoi tessellation), extracting local characteristics (features) within those sub-regions and estimating the optimal local threshold values based on the extracted features using a supervised learning method.

4.5.1.1 Voronoi tessellation

Firstly, we detected local maxima points M_i, where $i = 1...N$, on the lesion probability map to identify the plausible lesion locations (indicated by red dots in figure 4.3b). In order to avoid spurious local maxima points due to isolated voxels, we applied a small amount of spatial smoothing to the lesion probability map with a Gaussian filter, prior to local maxima detection. The FWHM size of the Gaussian kernel was empirically chosen to be 1.2 voxels (with standard deviation, $\sigma = 0.5$ voxels).

We then tessellated the lesion probability map based on local maxima M_i into N Voronoi polygons V_i, (figure 4.3c) around the maxima M_i. In order to ensure that our Voronoi polygons are within the region of interest (in our case, the brain white

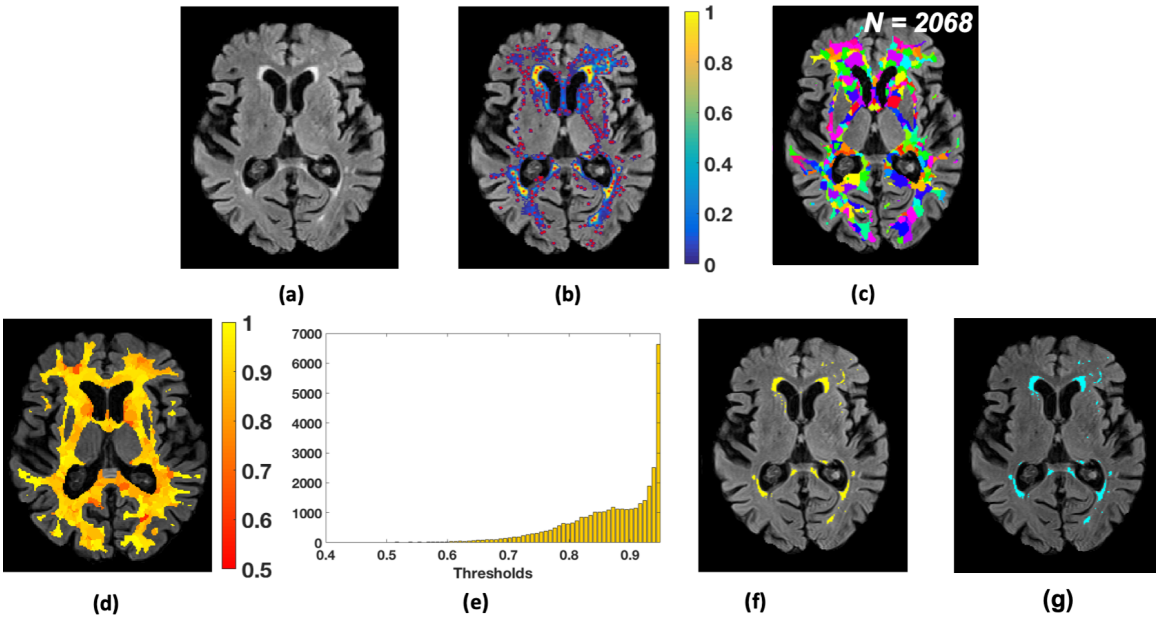


Figure 4.3: LOCally Adaptive Threshold Estimation (LOCATE). (a) FLAIR image; (b) detection of local maxima (M_i, red dots) on the smoothed lesion probability map (blue-yellow); (c) Voronoi spaces V_i around local maxima points (indicated by random colours); (d) threshold map showing the local threshold values obtained from the regression model; (e) histogram of thresholds Th_opt; (f) final binary lesion map obtained by applying the thresholds (d) on the lesion probability map (b); (g) binary lesion map obtained by applying a global threshold of 0.9.

matter), we constrained the Voronoi polygons within a white matter mask obtained from a dilated and inverted cortical CSF tissue segmentation, combined with other deep grey exclusion masks, as described in [53].

Effect of tessellation method on LOCATE performance: In order to test the robustness of LOCATE performance with respect to the tessellation method, we also used simple linear iterative clustering (SLIC) [339] to form sub-regions. For this experiment, we set the number of superpixels to 500, since this was comparable to the number of regions formed with Voronoi tessellation. We compared the performance of LOCATE between Voronoi tessellation and SLIC using performance metrics described in section 4.6.

4.5.1.2 Feature extraction

Within each Voronoi polygon V_i, we applied different levels of thresholds (Th) from 0 to 0.9 with incremental steps of 0.05, and extracted the following features at each Th:

1. Mean greyscale intensity of the image used to identify the lesions (in our case, FLAIR) within the thresholded region.

2. Distance between the ventricles and the centre of gravity of the thresholded region. Lateral ventricles were segmented from T1 images as described in [340] and the distance from the ventricle mask was calculated using the FSL command *distancemap*).

3. Volume of the thresholded region

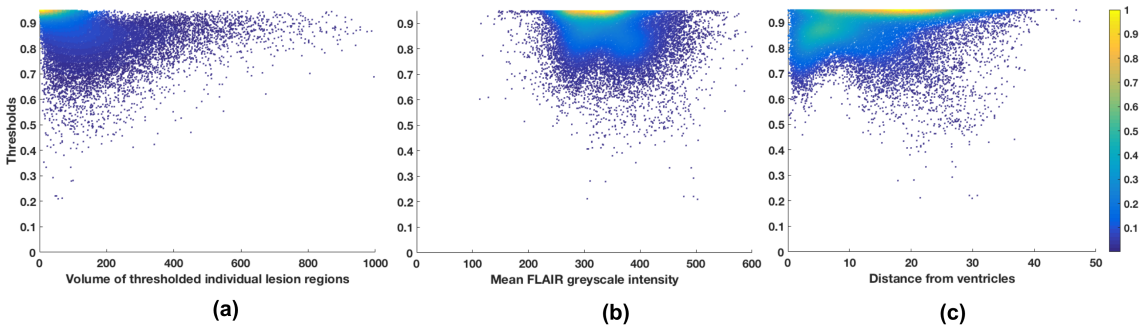


Figure 4.4: Features extracted from each Voronoi polygon plotted against the threshold values for 21 subjects from the NDGEN dataset: (a) volume of individual lesion region, (b) mean greyscale intensity and (c) distance of the centre of gravity from the ventricles mask. A simple straightforward relationship was not found between the features and the thresholds, indicating the need for a machine learning method.

Due to lack of a simple relationship between the features and the thresholds (figure 4.4), we determined the optimal local threshold for each V_i using a random forest [341] regression model with 1000 trees and min leaf size of 5.

4.5.1.3 Classification

Training phase: For each V_i, among the set of thresholds Th we determined the highest value Th_{max} at which we obtained the best similarity index with respect to the manually segmented binary lesion mask. We then trained the random forest regression model with the above features against Th_{max}.

Testing phase: We applied the trained regression model to get the optimal threshold Th_opt for each V_i in the test images. Figure 4.3d shows an instance of Th_opt

map. Note that the periventricular region shows higher threshold values compared to the deep region, indicating the variation in lesion probabilities. Moreover, the histogram of thresholds Th_opt (figure 4.3e) obtained for an individual image shows a wider range of local threshold values depending on the local lesion characteristics and spatial distribution (refer to appendix F for more threshold histograms representative of various datasets). As a final step, we thresholded the lesion probability map within each V_i using the corresponding Th_opt to get the final binary lesion map (figure 4.3f).

4.5.2 LOCATE evaluation and validation

We initially evaluated LOCATE using leave-one-out testing independently on the NDGEN, OXVASC and MWSC datasets on the outputs from BIANCA obtained using features and options specified in section 4.2.3. The performance of LOCATE was compared with that of the optimal global threshold against the manual segmentations using the metrics described in section 4.6. In order to evaluate LOCATE performance in lesions with different characteristics, we additionally calculated the performance metrics in periventricular and deep lesions separately. We also performed ANOVA on the performance metrics to observe the main and interaction effects on the measures. We adopted the 10 mm distance rule [33] as a criteria for classification of PWMHs and DWMHs (refer to the section 1.2.1 of chapter 1).

We used CADASIL and HC datasets to further validate LOCATE in terms of robustness with respect to lesion load and of flexibility with respect to the training dataset. In fact, they represent two very different scenarios from the datasets tested so far. The subjects in the CADASIL dataset have a very different lesion pattern and, on average, a higher lesion load than the other datasets, while HC have negligible lesion loads. Moreover, since manual segmentation was not available for these datasets, we used the OXVASC dataset for training (both for obtaining the lesion probability map with KNN using the same features/options, and for LOCATE) since the images were acquired using the same MRI protocol. The output obtained with

LOCATE on these datasets was qualitatively evaluated and quantitatively compared with the lesion mask obtained by applying the optimal global threshold (0.9), determined with leave-one-out on the training dataset. In addition, on the CADASIL subjects, we compared LOCATE output with the mask obtained by applying a lower global threshold (0.2, the optimal global threshold for this specific dataset empirically chosen by an expert neurologist). Using this comparative analysis, we explored the possibility of using LOCATE on different cohorts and various lesion loads, having a very different optimal global threshold from the pre-established value of 0.9. Also, by using CADASIL and HC datasets, we aimed to observe the effect of the difference in the lesion characteristics between patient group (CADASIL) and controls (HC) on thresholds obtained from LOCATE.

4.6 Performance evaluation metrics

We evaluated the lesion segmentation results using the following metrics:

- Dice Similarity Index (SI)
- True positive rate (TPR)
- False positive rate (FPR)
- Cluster-level TPR in deep and periventricular white matter

Definitions of the above evaluation metrics are provided in appendix C.

4.7 Results

4.7.1 Use of population-level parametric lesion probability map (PPLPM)

Figure 4.5 shows the ROC curves and SI plots at different thresholds for our experiments using the PPLPM with respect to age on the NDGEN dataset. Using PPLPM as an additional feature did not improve the performance of BIANCA, although the use of the age-specific 3D map (figure 4.5a) still gave better results than the use of

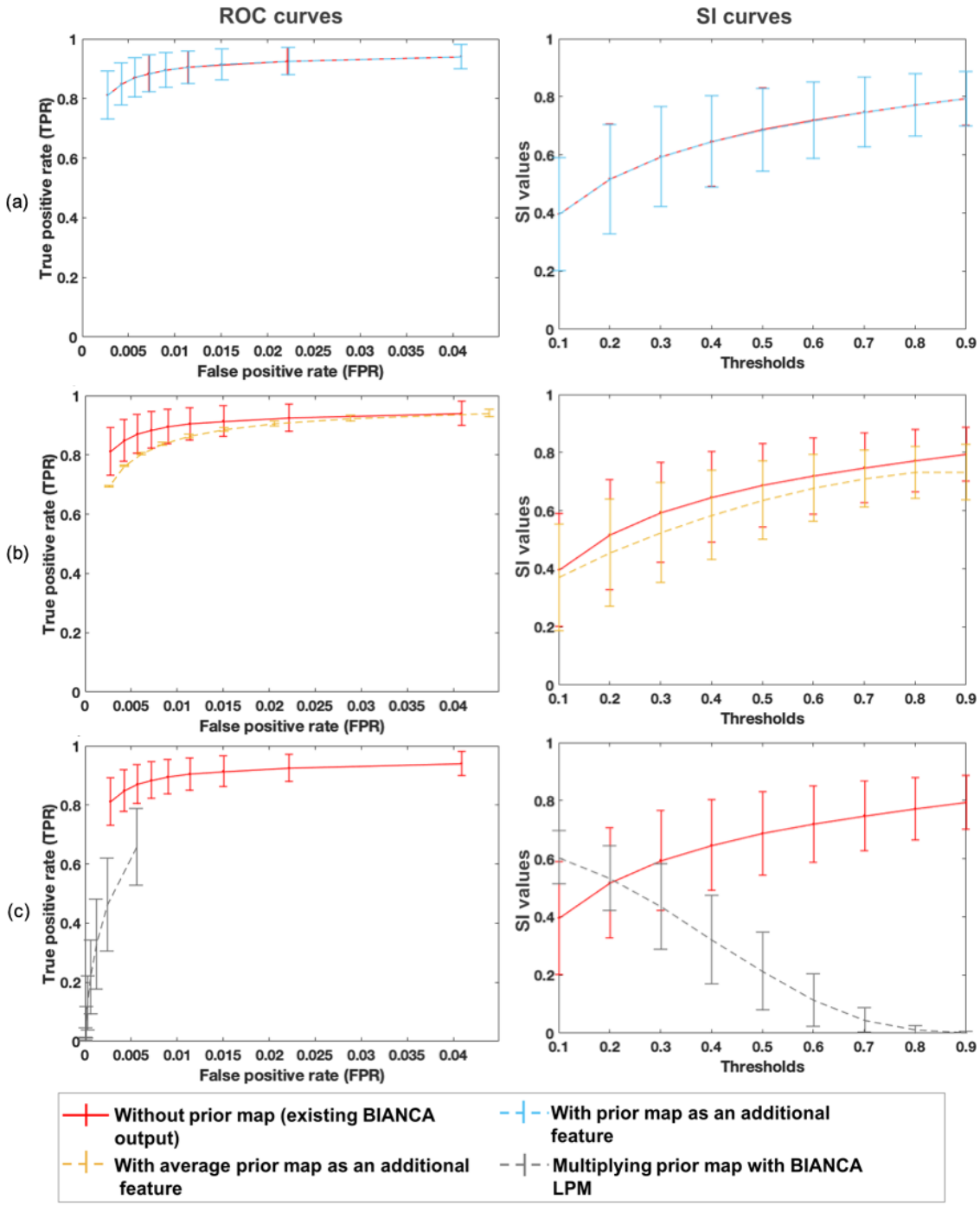


Figure 4.5: ROC (left) and SI (right) curves for the NDGEN dataset in the case of use of PPLPM when (a) the age-specific 3D map is added as an additional feature (light blue), (b) the average 3D map is added as an additional feature (yellow), and (c) when age-specific 3D map is post-multiplied with BIANCA lesion probability map (grey). The curves corresponding to the existing BIANCA results [53] are shown in red for comparison in all the three cases.

an average map across all age groups (figure 4.5b). When the PPLPM was post-multiplied with the subject's lesion probability map (figure 4.5c), the TPR and SI values were very low, especially at higher thresholds.

The optimal threshold for the existing version of BIANCA for this dataset was found to be 0.9 [53]. After adding the age-specific 3D map as a new feature, we still found 0.9 to be the optimal threshold for both age-specific and average cases, while for the post-multiplication case the optimal threshold was 0.1. The paired t-test showed significantly higher SI values for existing BIANCA (at the threshold of 0.9) compared to all cases using the parametric map ($p = 0.02$ for age-specific map as feature, threshold 0.9; $p = 0.002$ for average map as feature, threshold 0.9; $p<0.001$ for the post-multiplication, threshold 0.1).

4.7.2 Use of non-linear registration for estimating spatial features

Figure 4.6 shows the ROC and SI curves for the comparison of linear and non-linear registration on the NDGEN dataset, which would contain more cases of atrophic brains and therefore benefit the most from a non-linear registration. However, it can be seen from the figure that performing non-linear registration, with or without using PPLPM, resulted in negligible differences in the BIANCA performance.

4.7.3 Comparison of alternative classifiers within BIANCA

Table 4.2 lists the best performing set of parameters that were obtained for each classifier on the four subjects selected from the NDGEN dataset in the optimisation phase, when we tested using the area under the curve values for all possible combinations of the parameter values listed in table 4.1.

Figure 4.7 shows the ROC curves and SI plots for the four classifiers (RF, NN, SVM, AB) using the best and the default set of parameters (in the python *scikit-learn* package), with respect to KNN, on the remaining 17 subjects from the NDGEN dataset. From the ROC curves and SI indices, KNN performed better than other

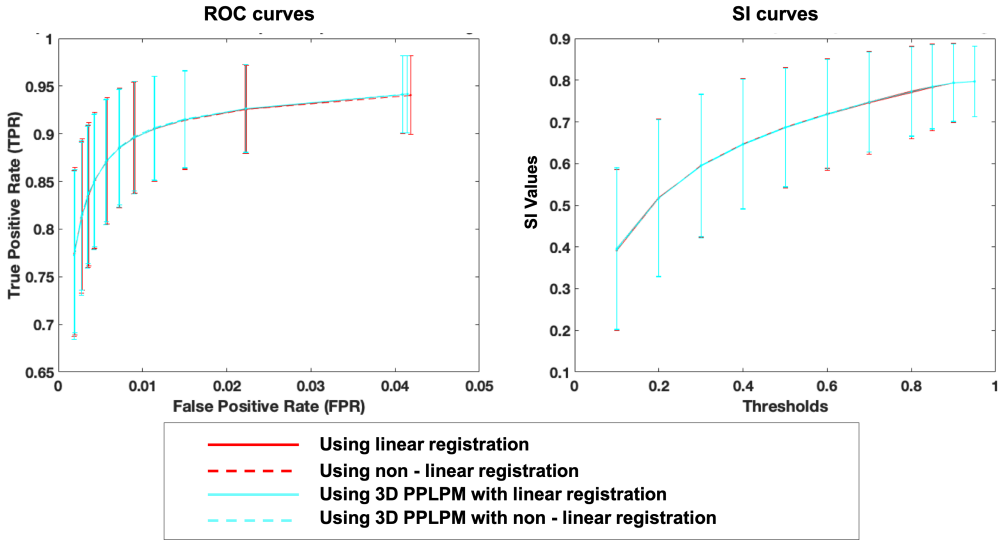


Figure 4.6: ROC (left) and SI (right) curves for the cases with linear and non-linear registration for extracting spatial features, with and without using PPLPM on the NDGEN dataset.

classifiers, even for their best set of parameters and using the optimal threshold for each classifier. Paired t-test results showed that SI values from KNN (mean SI = 0.77 ± 0.10 at a threshold of 0.9) were significantly higher than those of RF ($p = 0.02$, mean SI for RF = 0.76 ± 0.10 at a threshold of 0.9), SVM ($p < 0.001$, mean SI

Classifier	**Best performing set of parameters**
Random forest (RF)	No. of trees - 50 Min. samples at each leaf node - 400 Min. impurity measure at each split - 1×10^{-3} Max. depth of trees - 25
Neural network (NN)	No. of neurons in a hidden layer - 100 Solver - LBFGS, Initial learning rate - 1×10^{-3} Max. no. of iterations - 1800 Tolerance - 1×10^{-3}
Support vector machine (SVM)	γ - 1.0 C-value - 1.0 Max. no. of iterations - 5000
AdaBoost classifier (AB)	No. of trees in base estimator - 2 No. of estimators - 30 Learning rate - 1.0 Loss function - exponential

Table 4.2: Best performing set of parameters for different classifiers.

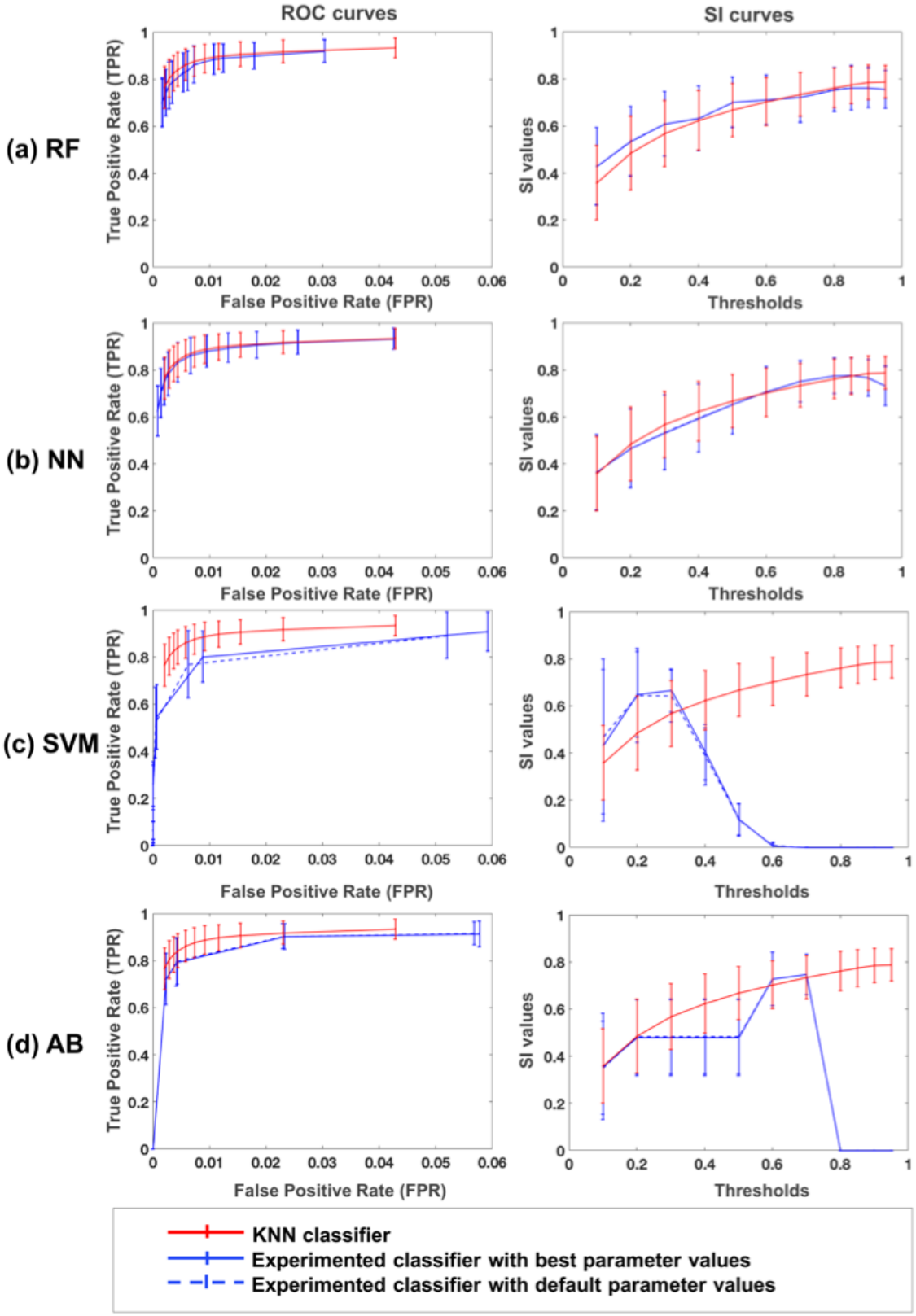


Figure 4.7: ROC curves (left) and SI curves (right) for alternative classifiers for the NDGEN dataset: (a) random forest (RF), (b) neural network (NN), (c) support vector machine (SVM) and (d) adaboost classifier (AB). Results are shown for each classifier's best (blue solid line) and default parameters (blue dashed lines) along with the results for KNN classifier, currently used in BIANCA (red solid line).

for SVM = 0.75 ± 0.10 at a threshold of 0.3) and AB ($p < 0.001$, mean SI for AB = 0.65 ± 0.12 at a threshold of 0.7), while SI values using NN were not significantly different from KNN ($p = 0.32$, mean SI for NN = 0.75 ± 0.09 at a threshold of 0.8).

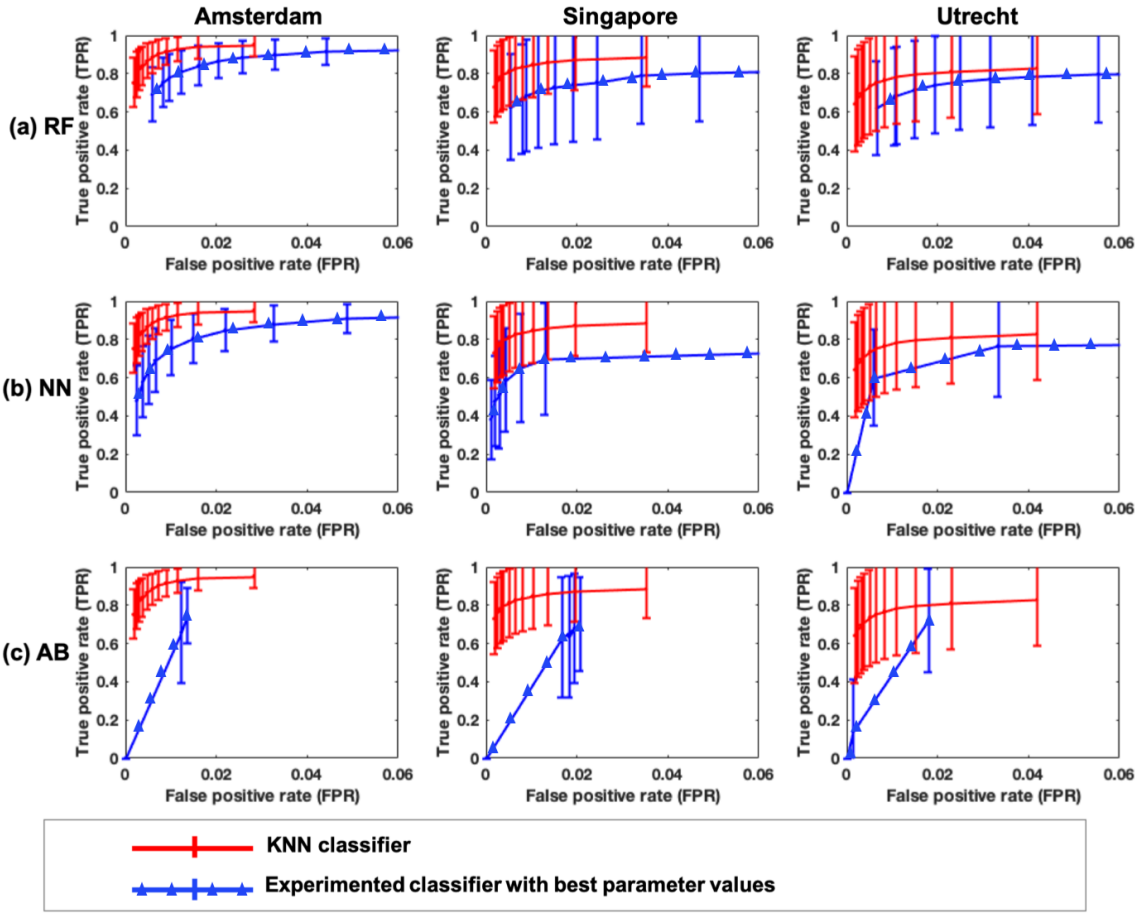


Figure 4.8: ROC curves for alternative classifiers: (a) random forest (RF), (b) neural network (NN) and (c) adaboost classifier (AB) for the VU Amsterdam, NUHS Singapore and UMC Utrecht cohorts of the MWSC dataset. Results are shown for each classifier's best parameters (blue solid line) along with the results for KNN classifier, currently used in BIANCA.

Figures 4.8 and 4.9 show the ROC and SI curves for RF, NN and AB against KNN on the 3 MWSC cohorts. The trend of ROC and SI curves for all the classifiers across all cohorts remained similar to those of NDGEN, although the NN classifier gave slightly lower SI values than the KNN classifier in the Singapore and Utrecht cohorts.

The paired t-test results on the SI values between KNN and alternative classifiers show that on the VU Amsterdam cohort, SI values obtained using KNN were not

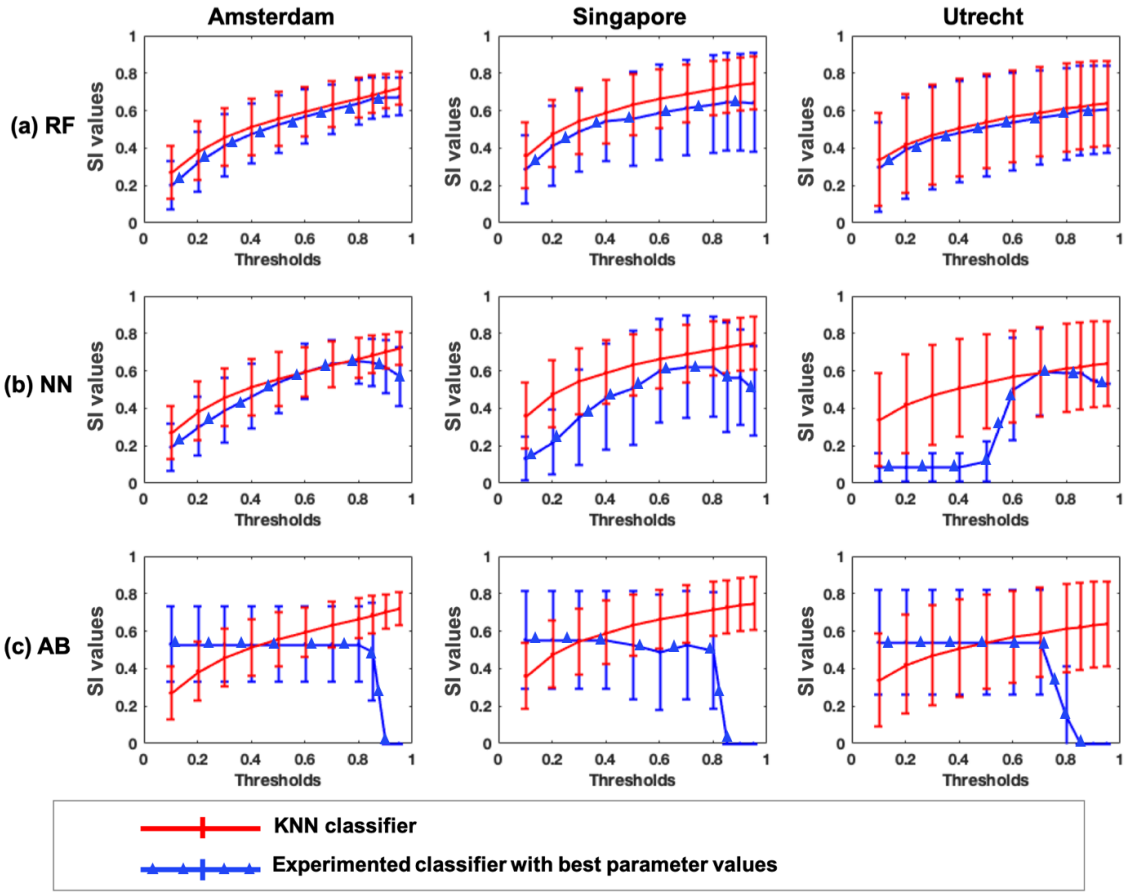


Figure 4.9: SI curves for alternative classifiers for the MWSC dataset: (a) random forest (RF), (b) neural network (NN) and (c) adaboost classifier (AB). Results are shown for each classifier's best parameters (blue solid line) along with the results for KNN classifier, currently used in BIANCA.

significantly different from those obtained from alternative classifiers (RF: mean SI = 0.67 ± 0.10 at a threshold of 0.9, p = 0.56; NN: mean SI = 0.65 ± 0.12 at a threshold of 0.8, p = 0.41; AB: mean SI = 0.53 ± 0.20 at a threshold of 0.4, p = 0.10). On the NUHS Singapore cohort, SI values obtained from KNN were not significantly different from those obtained from the RF classifier (mean SI = 0.65 ± 0.26 at a threshold of 0.9, p = 0.15) but they were significantly higher than SI values obtained using NN (mean SI = 0.62 ± 0.27 at a threshold of 0.8, p = 0.009) and AB (mean SI = 0.55 ± 0.26 at a threshold of 0.4, p = 0.005). Similarly, on the UMC Utrecht cohort, SI values obtained from KNN were not significantly different from those obtained from the RF classifier (mean SI = 0.60 ± 0.23 at a threshold of 0.9, p = 0.5), but were significantly higher than SI values obtained using NN (mean SI = 0.59 ± 0.23 at a

threshold of 0.7, p = 0.01) and AB (mean SI = 0.54 ± 0.28 at a threshold of 0.4, p = 0.02).

Therefore, SI values obtained using KNN were never significantly less than any of the alternative classifiers on any of the datasets. However, SI values of RF and NN classifiers were not significantly different from those obtained with KNN.

4.7.4 LOCally Adaptive Threshold Estimation (LOCATE)

4.7.4.1 Effect of tessellation method on LOCATE performance

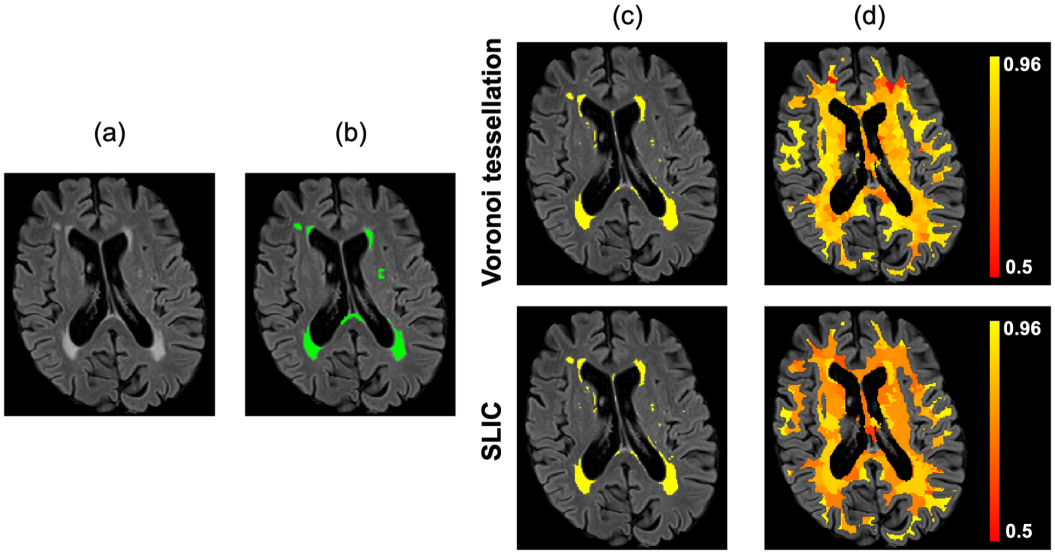


Figure 4.10: Results of LOCATE using Voronoi tessellation and SLIC for forming sub-regions. FLAIR (a) and manual segmentation (b) shown with LOCATE results (c) and thresholds map (d) using Voronoi tessellation (top row) and SLIC (bottom row).

Figure 4.10 illustrates examples of LOCATE outputs using SLIC and Voronoi tessellation compared to the manual segmentation. Figure 4.11 shows ROC curves and SI curves for BIANCA using various global thresholds and LOCATE using Voronoi tessellation and SLIC on NDGEN dataset. From the figure, LOCATE gave similar TPR and SI values for both the tessellation methods, with negligible difference in the performance. This shows that LOCATE is robust with respect to the method used for forming sub-regions.

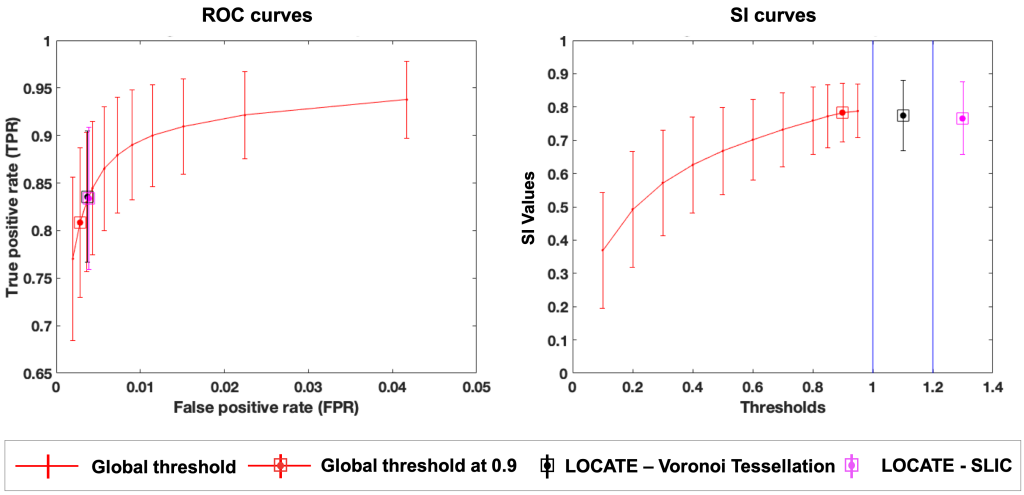


Figure 4.11: Comparison of ROC curves (left) and SI plots (right) for global thresholding (red) with LOCATE using Voronoi tessellation (black) and SLIC (magenta) on the NDGEN dataset.

The paired t-test results show that the SI values obtained with LOCATE using Voronoi tessellation were significantly higher than those obtained using SLIC (LOCATE Voronoi tessellation SI = 0.77 ± 0.10, LOCATE SLIC SI = 0.76 ± 0.11, p = 0.003). Therefore, we have used Voronoi tessellation method in LOCATE henceforth for all the experiments and evaluation. However, note that while we observed that LOCATE performance does not depend on the number of regions in Voronoi tessellation, SLIC method was not explored as thoroughly. Hence SLIC, with appropriate number of superpixels, could be considered as a valid alternative to Voronoi tessellation for creating sub-regions.

4.7.4.2 Evaluation of LOCATE on NDGEN, OXVASC and MWSC datasets

Figures 4.12 and 4.13 illustrate examples of BIANCA results with LOCATE outputs compared to the global thresholded outputs and the manual segmentations in various datasets. LOCATE detected more deep lesions (figures 4.12c, d and 4.13b, and frontal region of d) and segmented periventricular lesions better (figures 4.12a, b, d and 4.13a) with respect to the manual segmentation. The additional lesions detected by LOCATE with respect to manual segmentation in 4.12b, c, d did not represent false positives as they were labelled as lesions in the manual segmentation in the

adjacent slices. In fact, in MWSC datasets LOCATE segmented the lesions in the temporal regions with better accuracy (figure 4.13c) compared to global threshold, which missed a few well-defined lesions in that region.

Figure 4.14 shows the ROC curves and SI plots for BIANCA at various global thresholds and using LOCATE for NDGEN, OXVASC and MWSC datasets. The SI values obtained with LOCATE were not significantly different according to the paired t-test to those obtained with global thresholding values of 0.9 for NDGEN (LOCATE SI = 0.77 ± 0.10; global threshold SI = 0.77 ± 0.08, $p = 0.94$), OXVASC (LOCATE SI = 0.75 ± 0.14; global threshold SI = 0.74 ± 0.11, $p = 0.94$) and MWSC (UMC Utrecht: LOCATE SI = 0.64 ± 0.23, global threshold SI = 0.63 ± 0.22, $p = 0.29$, NUHS Singapore: LOCATE SI = 0.73 ± 0.13, global threshold SI = 0.74 ± 0.14, $p = 0.93$, VU Amsterdam LOCATE SI = 0.70 ± 0.09, global threshold SI = 0.70 ± 0.09, $p = 0.46$) datasets. However, in terms of detection rates, LOCATE gave an increase in voxel-wise TPR (0.03 for the NDGEN and 0.10 for the OXVASC) with a negligible increase in FPR (0.001 for the NDGEN and 0.002 for the OXVASC) compared to the global threshold of 0.9. For instance, in the NDGEN dataset a TPR of 0.03 corresponds to ≈1600 mm^3, while FPR of 0.001 corresponds to a negligible amount of ≈90 mm^3. In the case of the MWSC, the voxel-wise TPR showed negligible difference between LOCATE and global thresholding results, despite a small increase in FPR. It is worth noting that for MWSC datasets, we obtained the above results on the publicly available training datasets (using leave-one-out validation), while in the challenge the evaluation was done on the held out test dataset and the performance of our method could change depending on the dataset characteristics.

4.7.4.3 Comparison of LOCATE performance in the deep and the periventricular regions

We calculated the performance metrics such as TPR, FPR, SI and cluster-wise TPR in the deep and the periventricular regions separately and compared them. Figures 4.15, 4.16 and 4.17 show the ROC, SI and cluster-wise TPR plots respectively for the deep and the periventricular regions on NDGEN, OXVASC and the 3 cohorts of MWSC.

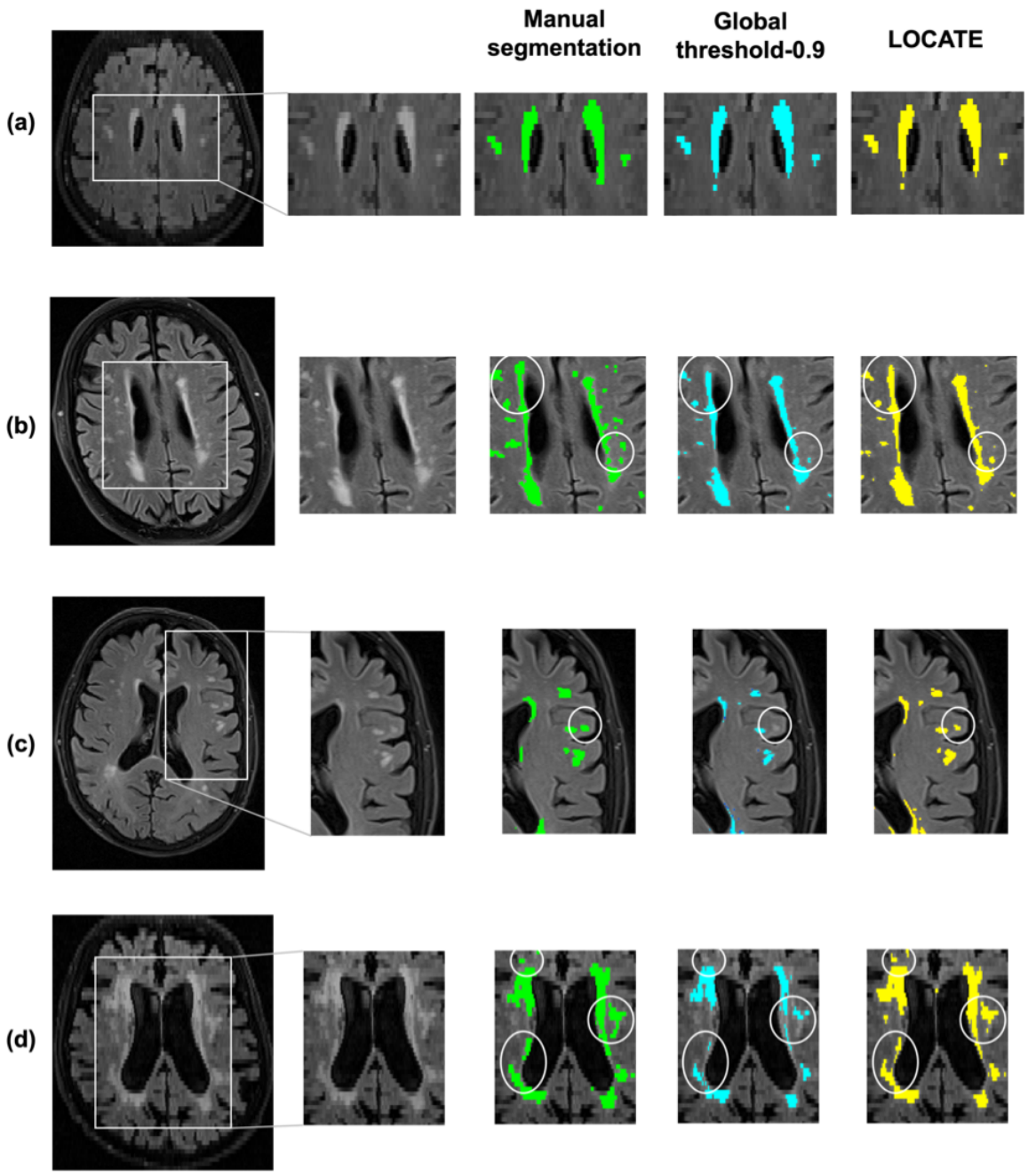


Figure 4.12: Example results of LOCATE for (a) low, (b) medium, (c) high and (d) very high lesion load (a,d from the OXVASC; b,c from the NDGEN datasets). Manual segmentations (green) are shown along with outputs of global thresholding at 0.9 (light blue) and LOCATE (yellow). LOCATE provides better segmentation in the deep (c, d) and the periventricular white matter lesions (a, b, d) when compared to the global thresholding.

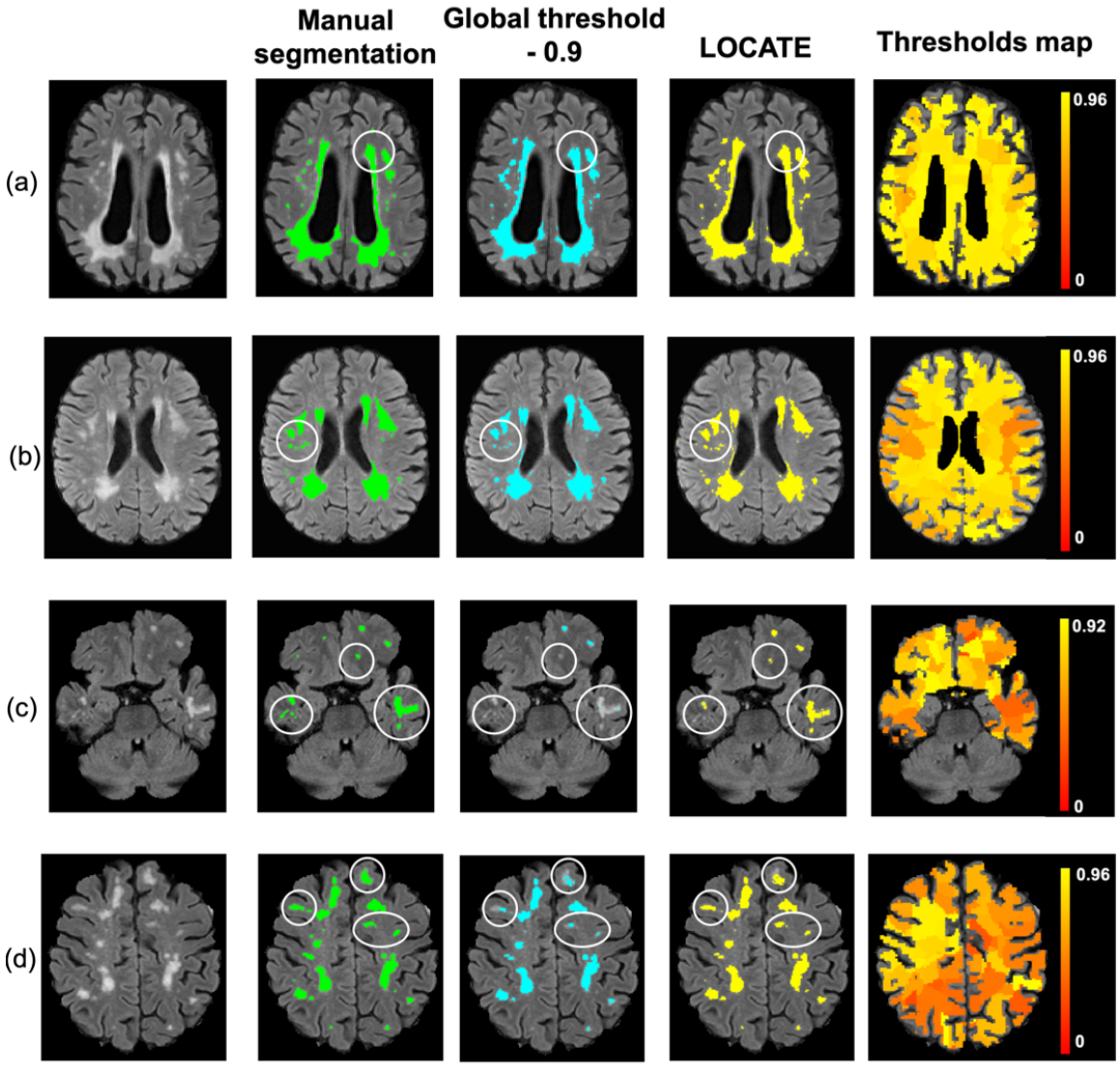


Figure 4.13: Example results of LOCATE and BIANCA - manual segmentations (green) compared to global thresholding (blue) and LOCATE (yellow). The corresponding threshold maps are shown in red-yellow (note the heterogeneity of lesion probabilities shown in the brain white matter).

We also performed 2-way repeated measures ANOVA considering region (deep and periventricular) and method (LOCATE and Global thresholding at 0.9) as independent factors and the performance metric (TPR, FPR, SI and cluster-wise TPR) as the dependent measure. We determined the main effect of region and method, along with the effect of their interaction, on the metrics. Tables 4.3, 4.4, 4.5 and 4.6 report the descriptive statistics and ANOVA results for comparison of TPR, FPR, SI and cluster-wise TPR values respectively. Briefly, the main findings of this analysis are:

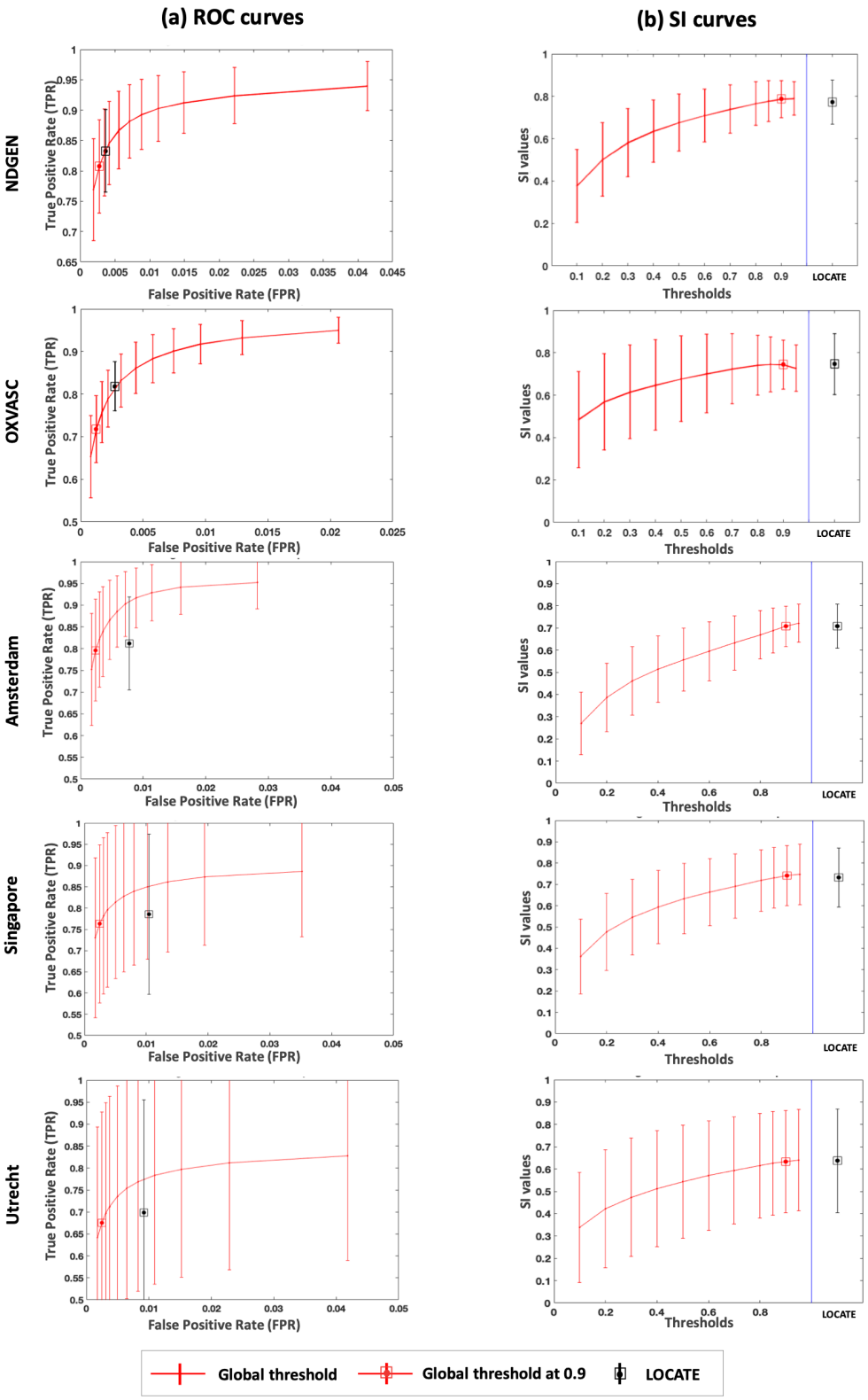


Figure 4.14: Comparison of ROC curves, SI plots for LOCATE and global thresholding. ROC curves (a), SI indices (b) for the NDGEN, OXVASC and MWSC (VU Amsterdam, NUHS Singapore, UMC Utrecht) datasets.

- For TPR values, regarding the main effect of method, LOCATE provided significantly higher TPR values when compared with the global thresholding. As a main effect of the region, the TPR values were significantly higher in the periventricular regions when compared to the deep regions. In almost all datasets, the interaction effect was significant in favour of the deep regions, indicating that the increase in TPR using LOCATE with respect to a global threshold was significantly higher in the deep regions compared to the periventricular regions.

- Even though FPR values were negligible (around 90 - 100 mm^3, which is only a 5% of increase in TPR) in general, they were higher for LOCATE when compared to global thresholding and were higher in the periventricular regions for the NDGEN, NUHS Singapore and UMC Utrecht datasets. We observed that the interaction effect was also significant in most of the MWSC cohorts and the NDGEN dataset (increase in FPR using LOCATE was significantly higher than global threshold in the periventricular regions compared to the deep).

- The main effect of method was not significant on SI values except for the NDGEN dataset. However, SI values were significantly higher in periventricular regions when compared with the deep regions for all the datasets.

- We observed a significant main effect of method for all datasets with cluster-wise TPR always being higher for LOCATE than global thresholding. The main effect of region was significant only for the MWSC cohorts: cluster-wise TPR was not significantly different between the periventricular and the deep regions in NDGEN and OXVASC, while it was significantly higher in the periventricular than the deep regions for the MWSC cohorts. In these cohorts we also observed a significant interaction effect: the increase in cluster-wise TPR using LOCATE with respect to global threshold was higher in the deep regions compared to the periventricular regions.

Overall, we observed significant main effects of both method and region on the detection of true positive lesions. Additionally, in most of the datasets we observed a

significant interaction effect. This shows that the interactive effect of improvement in detection of true positive lesions using LOCATE with respect to global thresholding was significantly higher in the deep regions compared to the periventricular regions.

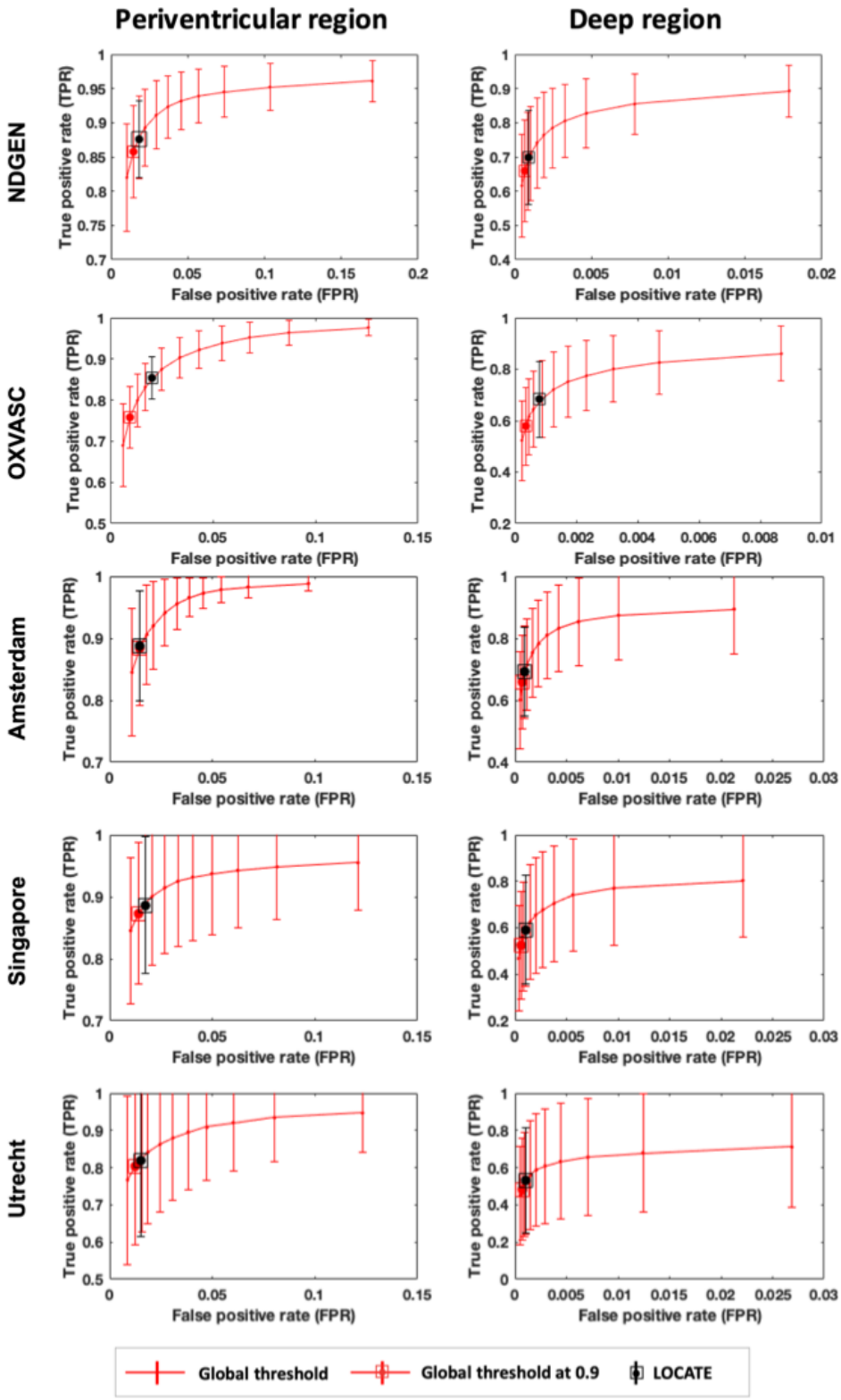


Figure 4.15: ROC curves in periventricular (left) and deep (right) regions for LOCATE (black) and global thresholding (red) on the NDGEN, OXVASC and MWSC (VU Amsterdam, NUHS Singapore and UMC Utrecht) datasets.

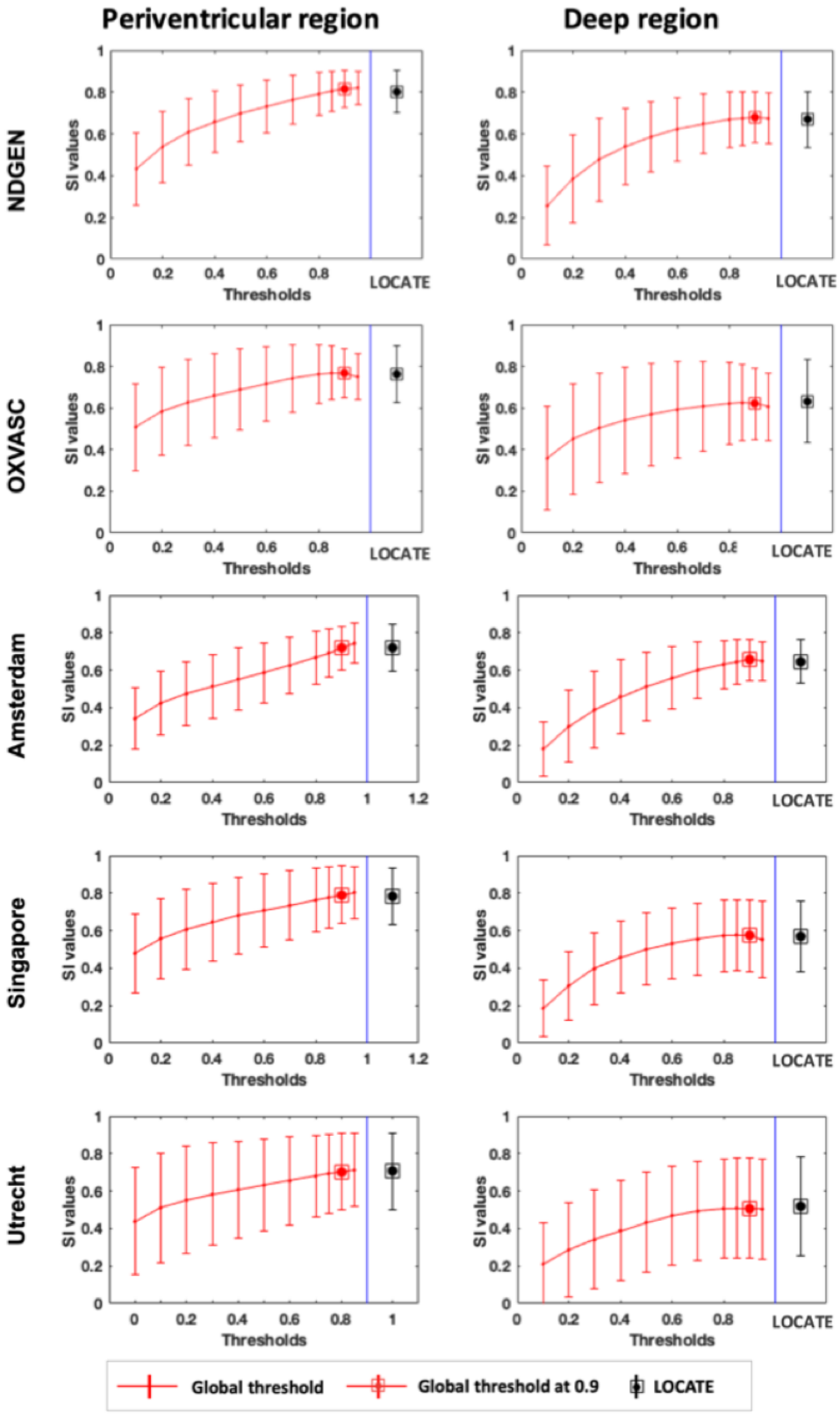


Figure 4.16: SI curves in the periventricular (left) and the deep (right) regions for LOCATE (black) and global thresholding (red) on the NDGEN, OXVASC and MWSC (VU Amsterdam, NUHS Singapore and UMC Utrecht) datasets.

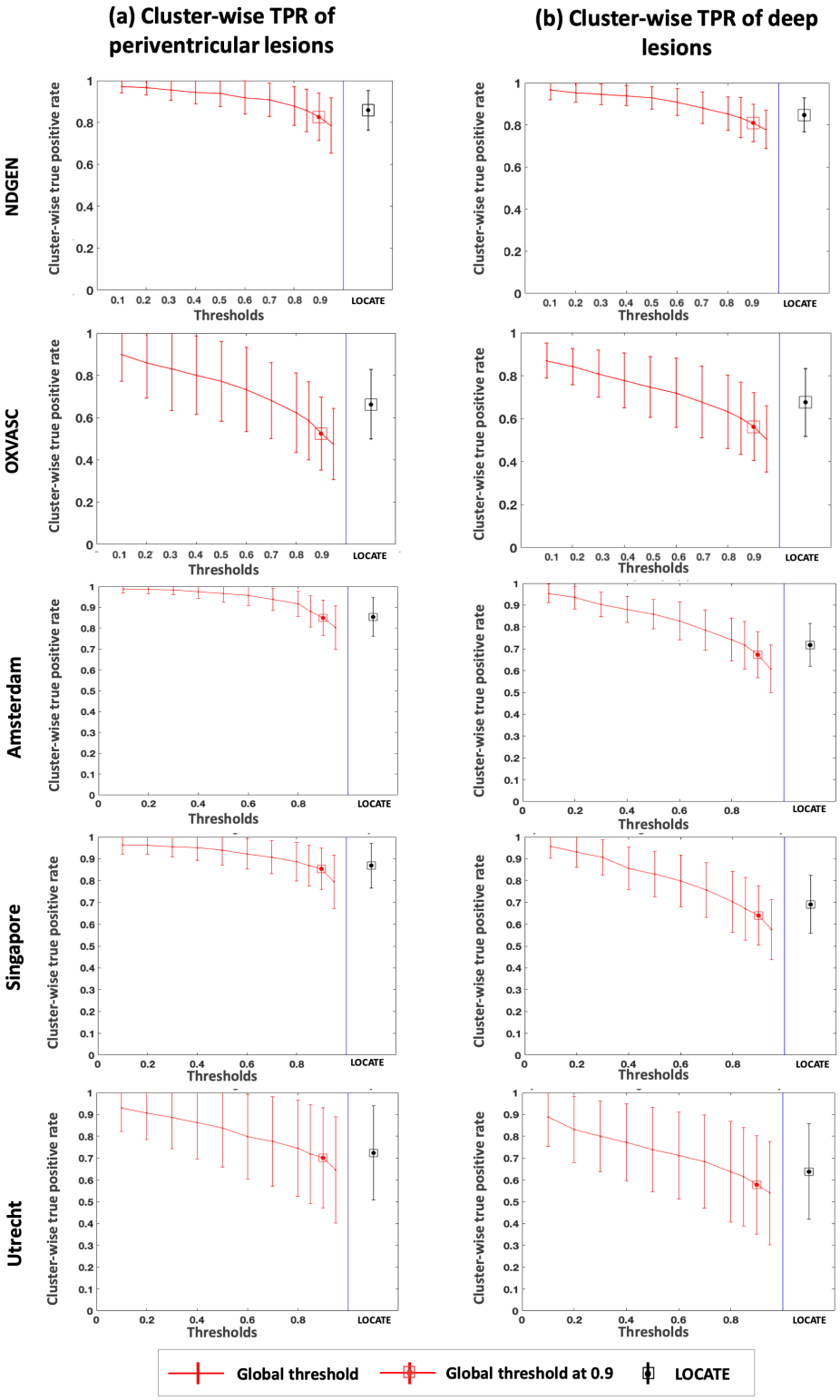


Figure 4.17: Comparison of cluster-wise true positive rates in the periventricular and the deep region for LOCATE and global thresholding. Cluster-wise true positive rates in the periventricular (a) and the deep (b) regions shown for the NDGEN, OXVASC and MWSC (VU Amsterdam, NUHS Singapore, UMC Utrecht) datasets.

			NDGEN	OXVASC	VU Amsterdam	NUHS Singapore	UMC Utrecht
Descriptive statistics	Deep region	LOCATE	0.70 ± 0.14	0.68 ± 0.15	0.69 ± 0.14	0.59 ± 0.23	0.53 ± 0.28
		Global 0.9	0.66 ± 0.15	0.57 ± 0.15	0.65 ± 0.15	0.52 ± 0.23	0.48 ± 0.27
	Peri-venticular region	LOCATE	0.88 ± 0.06	0.85 ± 0.05	0.89 ± 0.08	0.89 ± 0.11	0.82 ± 0.20
		Global 0.9	0.86 ± 0.07	0.76 ± 0.07	0.88 ± 0.09	0.87 ± 0.11	0.80 ± 0.21
ANOVA	Main effect of region (deep/PV regions)		**F(1,19)=43.1 p<0.001 η_p^2=0.694**	**F(1,19)=41.4 p<0.001 η_p^2=0.685**	**F(1,16)=19.3 p<0.001 η_p^2=0.547**	**F(1,19)=50.8 p<0.001 η_p^2=0.728**	**F(1,19)=46.7 p<0.001 η_p^2=0.711**
	Main effect of method (LOCATE/Global 0.9)		**F(1,19)=15.3 p=0.001 η_p^2=0.446**	**F(1,19)=61.1 p<0.001 η_p^2=0.763**	**F(1,16)=116.7 p<0.001 η_p^2=0.879**	**F(1,19)=54.1 p<0.001 η_p^2=0.740**	**F(1,19)=23.6 p<0.001 η_p^2=0.554**
	Interaction (region * method)		**F(1,19)=49.2 p<0.001 η_p^2=0.663**	**F(1,19)=49.3 p<0.001 η_p^2=0.721**	F(1,16)=0.51 p=0.48 η_p^2=0.031	**F(1,19)=27.8 p<0.001 η_p^2=0.594**	**F(1,19)=12.2 p=0.002 η_p^2=0.391**

Table 4.3: Descriptive statistics and ANOVA results for comparison of TPR values for global threshold of 0.9 and LOCATE in the periventricular and the deep regions for the MWSC dataset.

			NDGEN	OXVASC	VU Amsterdam	NUHS Singapore	UMC Utrecht
Descriptive statistics	Deep region	LOCATE	10.00×10^{-4}	11.3×10^{-4}	9.07×10^{-4}	10.00×10^{-4}	11.00×10^{-4}
		Global 0.9	7.00×10^{-4}	8.40×10^{-4}	7.00×10^{-4}	6.00×10^{-4}	7.00×10^{-4}
	Peri-venticular region	LOCATE	1.9×10^{-2}	2.0×10^{-2}	1.5×10^{-2}	1.7×10^{-2}	1.5×10^{-2}
		Global 0.9	1.5×10^{-2}	0.9×10^{-2}	1.5×10^{-2}	1.4×10^{-2}	0.13×10^{-2}
ANOVA	Main effect of region (deep/PV regions)		**F(1,18)=42.5 p<0.001 η_p^2=0.703**	F(1,16)=0.55 p=0.47 η_p^2=0.033	**F(1,18)=24.3 p<0.001 η_p^2=0.574**	**F(1,16)=113.2 p<0.001 η_p^2=0.876**	**F(1,16)=34.3 p<0.001 η_p^2=0.682**
	Main effect of method (LOCATE/Global 0.9)		**F(1,18)=28.9 p=0.001 η_p^2=0.616**	F(1,16)=1.12 p=0.30 η_p^2=0.065	F(1,18)=0.07 p=0.80 η_p^2=0.004	**F(1,16)=34.6 p<0.001 η_p^2=0.684**	**F(1,16)=12.6 p=0.003 η_p^2=0.440**
	Interaction (region * method)		**F(1,18)=23.9 p<0.001 η_p^2=0.570**	F(1,16)=0.34 p=0.57 η_p^2=0.021	F(1,18)=2.39 p=0.139 η_p^2=0.118	**F(1,16)=23.2 p=0.001 η_p^2=0.592**	**F(1,16)=12.1 p=0.003 η_p^2=0.430**

Table 4.4: Descriptive statistics and ANOVA results for comparison of FPR values for global threshold of 0.9 and LOCATE in the periventricular and the deep regions for the MWSC dataset.

			NDGEN	OXVASC	VU Amsterdam	NUHS Singapore	UMC Utrecht
Descriptive statistics	Deep region	LOCATE	0.67 ± 0.14	0.63 ± 0.20	0.65 ± 0.11	0.57 ± 0.19	0.52 ± 0.26
		Global 0.9	0.68 ± 0.13	0.62 ± 0.18	0.65 ± 0.11	0.57 ± 0.20	0.50 ± 0.27
	Peri-venticular region	LOCATE	0.80 ± 0.10	0.76 ± 0.14	0.72 ± 0.12	0.79 ± 0.14	0.70 ± 0.20
		Global 0.9	0.81 ± 0.09	0.76 ± 0.12	0.72 ± 0.12	0.79 ± 0.15	0.70 ± 0.20
ANOVA	Main effect of region (deep/PV regions)		**F(1,19)=57.9 p<0.001 η_p^2=0.753**	**F(1,19)=22.4 p<0.001 η_p^2=0.583**	**F(1,16)=4.45 p=0.048 η_p^2=0.190**	**F(1,19)=28.9 p<0.001 η_p^2=0.604**	**F(1,19)=30.1 p<0.001 η_p^2=0.613**
	Main effect of method (LOCATE/Global 0.9)		**F(1,19)=5.4 p=0.031 η_p^2=0.222**	F(1,19)=0.09 p=0.76 η_p^2=0.006	F(1,16)=0.82 p=0.38 η_p^2=0.041	F(1,19)=0.72 p=0.41 η_p^2=0.037	F(1,19)=0.76 p=0.40 η_p^2=0.039
	Interaction (region * method)		F(1,19)=0.24 p=0.63 η_p^2=0.013	F(1,19)=2.52 p=0.13 η_p^2=0.136	F(1,16)=1.14 p=0.30 η_p^2=0.056	F(1,19)=0.32 p=0.58 η_p^2=0.016	F(1,19)=2.66 p=0.12 η_p^2=0.123

Table 4.5: Descriptive statistics and ANOVA results for comparison of SI values for global threshold of 0.9 and LOCATE in the periventricular and the deep regions for the MWSC dataset.

			NDGEN	OXVASC	VU Amsterdam	NUHS Singapore	UMC Utrecht
Descriptive statistics	Deep region	LOCATE	0.85 ± 0.08	0.67 ± 0.16	0.71 ± 0.09	0.69 ± 0.13	0.64 ± 0.22
		Global 0.9	0.80 ± 0.09	0.56 ± 0.16	0.67 ± 0.10	0.64 ± 0.14	0.58 ± 0.23
	Peri-venticular region	LOCATE	0.86 ± 0.09	0.66 ± 0.16	0.86 ± 0.08	0.87 ± 0.09	0.72 ± 0.21
		Global 0.9	0.83 ± 0.11	0.52 ± 0.18	0.85 ± 0.08	0.85 ± 0.09	0.70 ± 0.23
ANOVA	Main effect of region (deep/PV regions)		F(1,19)=0.22 p=0.64 η_p^2=0.011	F(1,16)=0.44 p=0.51 η_p^2=0.027	**F(1,19)=48.7** **p<0.001** η_p^2**=0.719**	**F(1,19)=31.5** **p<0.001** η_p^2**=0.623**	**F(1,19)=9.3** **p=0.007** η_p^2**=0.327**
	Main effect of method (LOCATE/Global 0.9)		**F(1,19)=38.3** **p<0.001** η_p^2**=0.668**	**F(1,16)=83.5** **p<0.001** η_p^2**=0.839**	**F(1,19)=31.9** **p<0.001** η_p^2**=0.627**	**F(1,19)=50.1** **p<0.001** η_p^2**=0.725**	**F(1,19)=34.6** **p<0.001** η_p^2**=0.646**
	Interaction (region * method)		F(1,19)=0.22 p=0.65 η_p^2=0.011	F(1,16)=0.69 p=0.42 η_p^2=0.041	**F(1,19)=17.7** **p<0.001** η_p^2**=0.482**	**F(1,19)=16.3** **p=0.001** η_p^2**=0.462**	**F(1,19)=10.8** **p=0.004** η_p^2**=0.362**

Table 4.6: Descriptive statistics and ANOVA results for comparison of cluster-wise TPR values for global threshold of 0.9 and LOCATE in the periventricular and the deep regions for the MWSC dataset.

4.7.4.4 Validation of LOCATE on CADASIL and HC datasets

Figure 4.18 illustrates the results of lesion segmentation by applying LOCATE on the CADASIL dataset. Since CADASIL patients have different lesion characteristics compared to those in the OXVASC dataset (different lesion location and very high lesion loads), applying the global threshold of 0.9 that was determined in the training phase on the OXVASC dataset yields poorly segmented binary lesion maps (figure 4.18b). On visual inspection, the results of global threshold at 0.2 (determined as optimal by an expert neurologist) showed increased false positives in some cases (especially in the third case of figure 4.18c). On the other hand, we can observe that LOCATE provided a much better segmentation and detected the lesions in the temporal lobe, typical of the pathology (figure 4.18d). A similar visual comparison of LOCATE results with those of global thresholding at 0.9 on the HC dataset is shown in figure 4.19.

When the binary lesion maps obtained using the optimal global threshold value of 0.2 is used as the reference segmentation, the paired t-test results show that the SI values obtained for CADASIL with LOCATE (SI = 0.79 $\pm$ 0.01) were significantly higher than those obtained using a global threshold at 0.9 (SI = 0.57 $\pm$ 0.10, p<0.001). For CADASIL, using LOCATE increases the voxel-wise TPR by 0.48 for a constant FPR of 5×10^{-4}.

4.7.4.5 Comparison of lesion volumes between LOCATE and global thresholding

The Bland-Altman plots in figure 4.20 show the agreement in lesion volumes for global threshold (a) and LOCATE (b) with respect to the reference segmentation for the above datasets. From the figure, it can be seen that there is not much scope for improvement in NDGEN and VU Amsterdam, however LOCATE shows improvement for the OXVASC, NUHS Singapore and UMC Utrecht datasets both in terms of slope and limits of agreement.

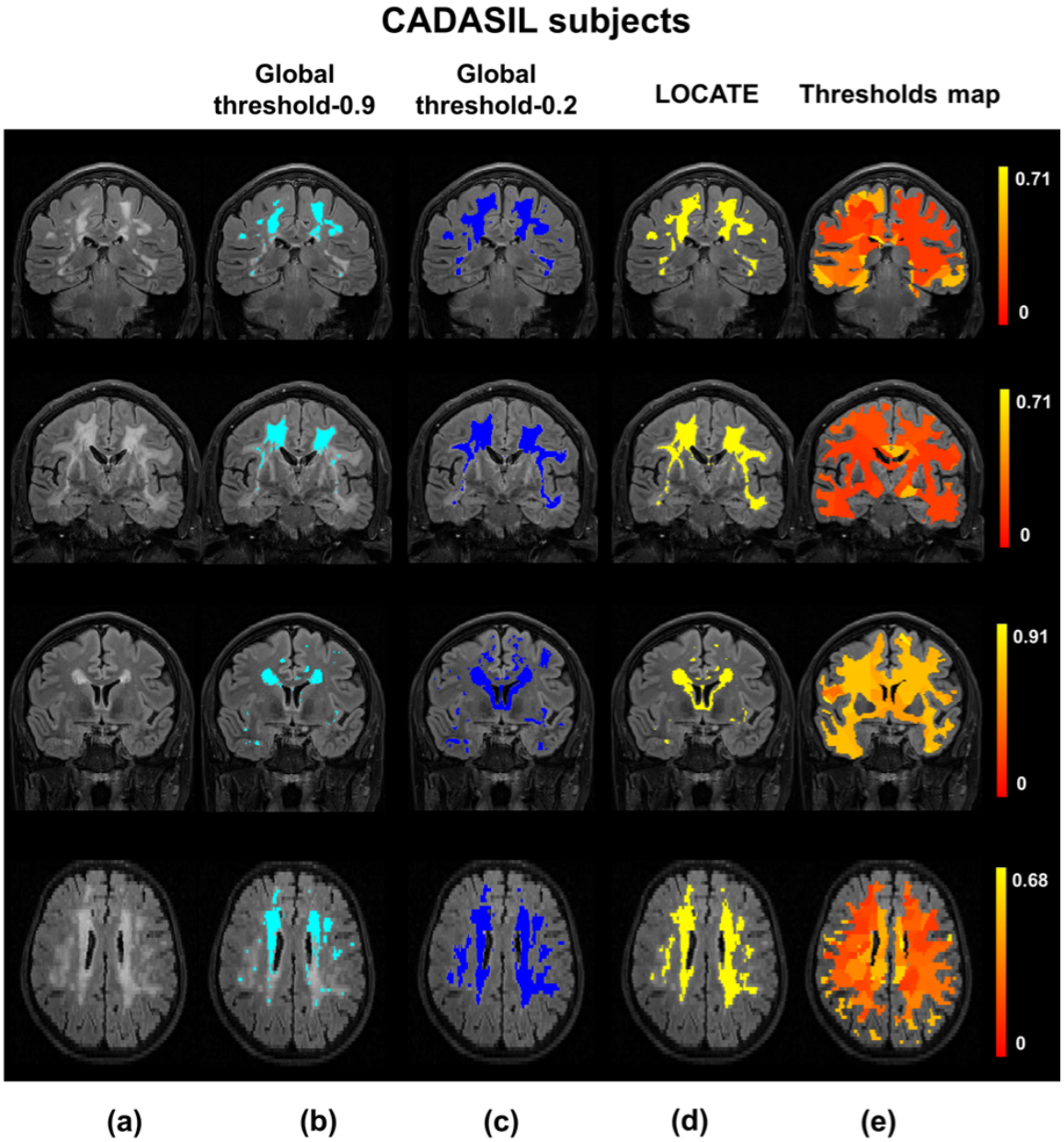


Figure 4.18: Example results of LOCATE on the CADASIL data. (a) FLAIR image, (b) global thresholding at 0.9 (light blue), (c) global thresholding at 0.2 (dark blue), (d) LOCATE results (yellow) and (e) threshold maps obtained from LOCATE (red-yellow). Note that the threshold maps show the heterogeneity in the lesion probabilities in various regions of the white matter.

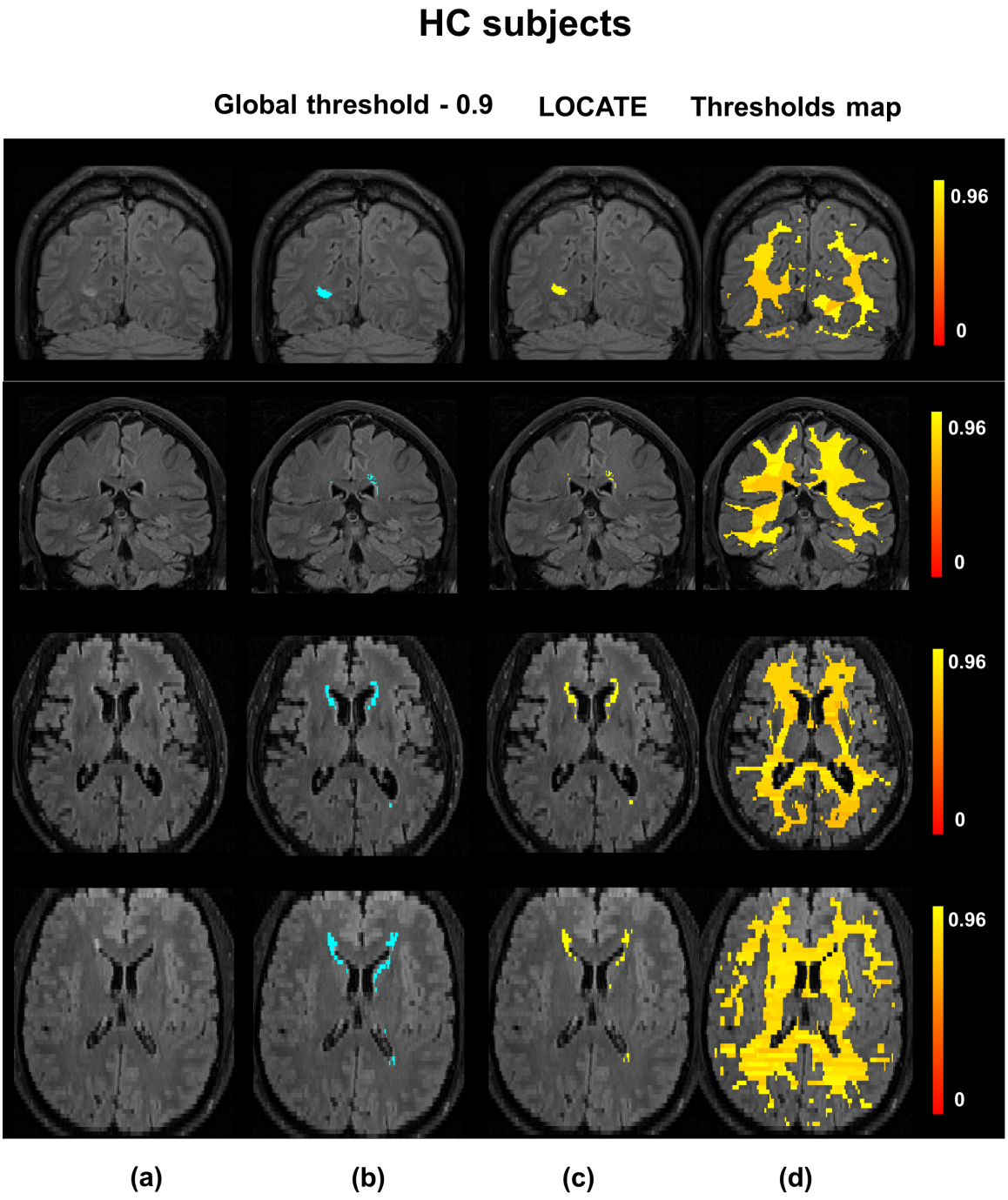


Figure 4.19: Example results of LOCATE on the HC data. (a) FLAIR image, (b) global thresholding at 0.9 (light blue), (c) LOCATE results (yellow) and (d) threshold maps obtained from LOCATE (red-yellow). Note that the threshold maps show the heterogeneity in the lesion probabilities in various regions of the white matter.

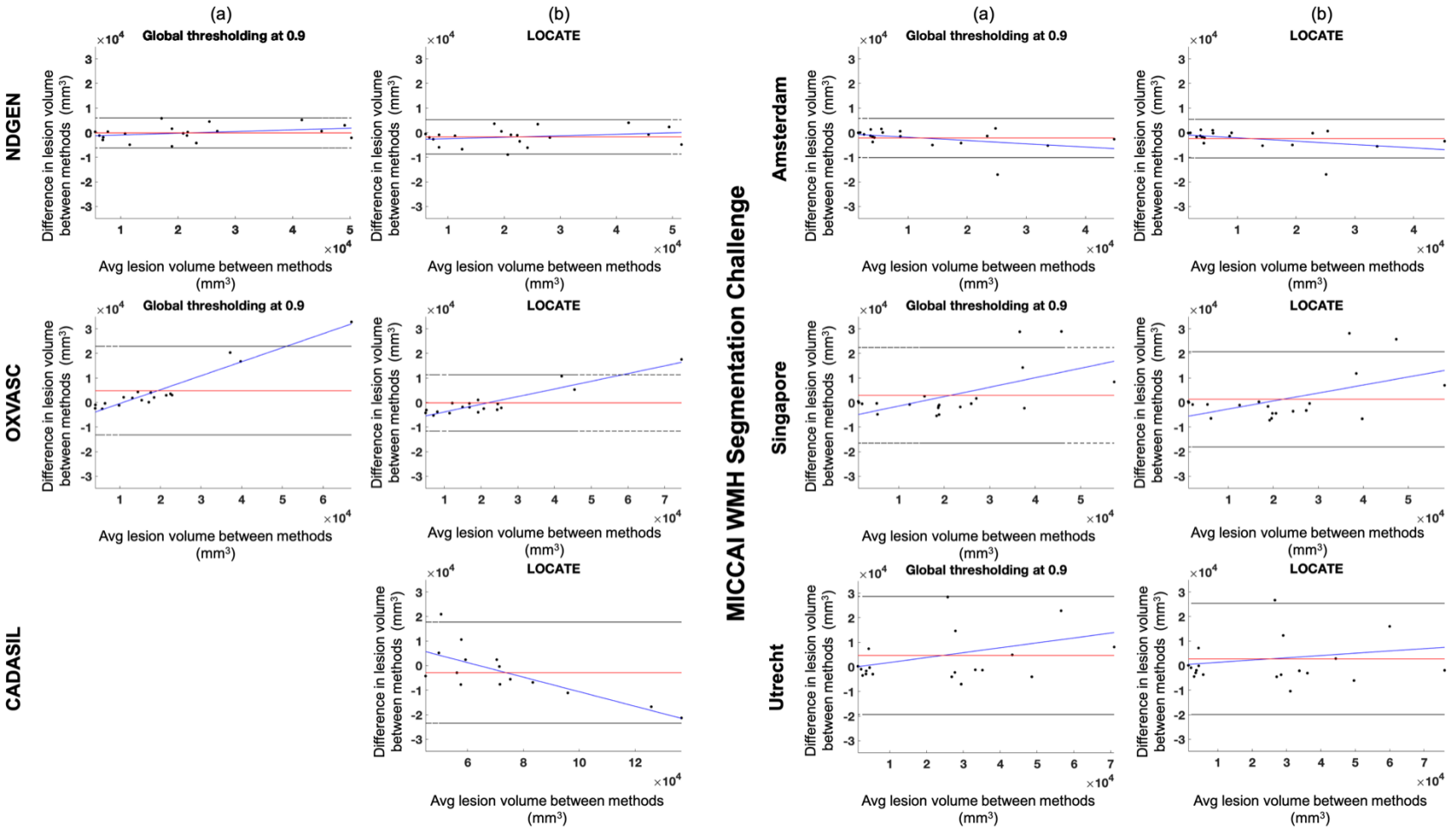


Figure 4.20: Bland-Altman plots of WMH volumes on results obtained from reference segmentation and global thresholding (a) and reference segmentation and LOCATE (b) on all datasets.

4.7.4.6 Analysis of LOCATE thresholds

To summarise the effect of LOCATE on different datasets, figure 4.21 shows boxplots of local threshold values determined by LOCATE for the five datasets (also individually for the three MWSC datasets), along with the optimal global threshold values for each dataset. The median threshold values of the NDGEN, MWSC and HC datasets are higher than other two datasets, especially the CADASIL dataset. LOCATE assigned higher thresholds for datasets with lower lesion load (HC and NDGEN, which contains 30% of healthy subjects) and lower thresholds for datasets with higher lesion load (OXVASC and especially CADASIL). This is also observable from the thresholded maps shown in figure 4.18e and figure 4.19d. We have also shown LOCATE threshold histograms for various subjects within a specific dataset and across various datasets in appendix F.

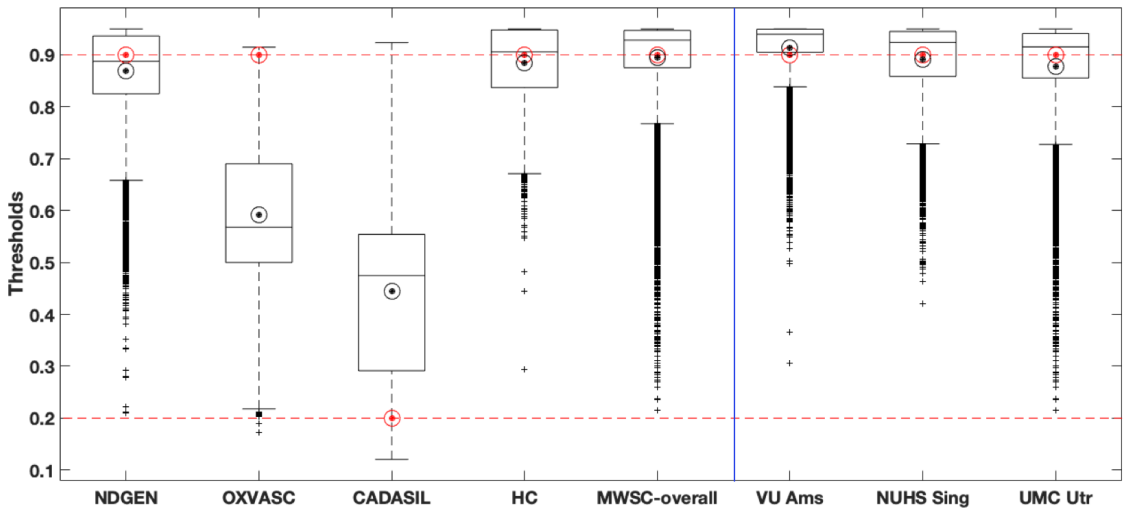


Figure 4.21: Boxplots of local thresholds obtained from the random forest regression model in LOCATE. The optimal global thresholds determined manually are shown in red circles. The mean values of local thresholds obtained from LOCATE are shown in black circles.

In line with the results of the paired t-test, the largest improvement in SI corresponds to the largest difference between the LOCATE threshold and the global threshold. In the case of the HC dataset, the median of local thresholds obtained from LOCATE is very close to the optimal threshold of 0.9, thus showing only a slight improvement in the lesion segmentation. This is also evident from figure 4.19,

where LOCATE results appear very similar to the results of global thresholding at 0.9 (figure 4.19b and 4.19c).

4.8 Discussion and conclusions

In this work, we studied the effect of single-subject and population-level heterogeneity in amount, location and characteristics of WMHs on the performance of our recently developed WMH segmentation method BIANCA, with the aim of improving feature extraction, classification and thresholding steps. At the feature extraction step, we analysed the effect on the subject-level lesion probability maps of using population-level lesion probability map (PPLPM, with age as factor of interest) and non-linear registration. In the classification step, we explored the effect of using alternative classifiers, which have been previously used for this task in the literature. For thresholding the lesion probability map, as an alternative to selecting a global threshold, we proposed LOCATE, a method to determine local thresholds that are sensitive to spatial differences in the lesion probabilities, giving more accurate binary lesion masks.

We observed the effect of using the PPLPM and found that post-multiplying the age-specific 3D map with the lesion probability map performed worse than adding the PPLPM as an additional feature. This is because PPLPM indicates the likelihood of finding a lesion at a given location in a population, informed primarily by the anatomy of disease signs. On the other hand, the subject-level probability map reflects the probability of identifying a lesion versus the normal tissue and is primarily determined by the intensity contrasts and noise in the image. For instance, in our previous work [78] we showed that in PPLPM the lesion probabilities typically increase more in the periventricular regions than in the deep regions in the elderly population (which is in line with the existing literature [54, 342, 343]). Therefore, post-multiplying the PPLPM with the subject-level lesion probability map can bias the subject-level lesion probabilities more towards the periventricular lesions, irrespective of the local spatial characteristics in the images from the individual subject. However, when we provide

the PPLPM as an additional feature, the classifier assigns sensible lesion probability values taking into account the PPLPM as well as the image intensity contrasts for the different modalities and the spatial coordinates. Particularly, we observed that adding an age-specific map as an additional feature gave better results than the average map, since the former models the lesion probabilities for the subject's age group specifically and hence is more accurate. Overall we observed that, at present, using the PPLPM does not significantly improve the performance of BIANCA (which is also evident from the p-values of our t-test results). Hence, our current recommendation is that PPLPM (with age as parameter of interest) is not necessary for improving BIANCA segmentation. However, the way this option will be implemented in a future release of BIANCA is quite flexible and will allow the user to use any 4D PPLPM with respect to any other factor of interest, which still could improve the segmentation in their specific dataset. Also, it is worth noting that before using this option, the user needs to ensure whether the demographics of the test dataset are comparable to those of the population used to model the PPLPM.

On comparing the use of linear and non-linear registration for extracting spatial features, we observed negligible difference in the performance. However, since the use of non-linear registration could still be beneficial in more extreme cases, in a future release of BIANCA users will be able to provide non-linear warp fields as an alternative to the linear transformation matrix for getting spatial coordinates.

The results of the comparison between KNN and other classifiers showed that KNN provided better segmentation. For the same set of features (excluding the PPLPM), we observed that SVM and AB classifiers provided significantly lower SI than KNN, especially at higher thresholds. On the other hand, SI values for RF and NN followed a similar trend to KNN with relatively similar SI values (figure 4.7). In the NDGEN dataset, SI values for RF were still significantly lower than KNN, while NN gave SI values that were not significantly different from KNN. In the MWSC dataset, SI values obtained using KNN did not differ significantly from those of RF in all cohorts and from those of NN in the VU Amsterdam cohort. Moreover, for a given

FPR, KNN has a higher TPR compared to any other classifier, indicating that KNN detects more true lesions than other classifiers. However, since NN and RF showed the potential to perform as well as KNN, they will be included as alternative classifiers in a future release of BIANCA. At this point, it is worth noting that irrespective of the classifier used, there is a certain degree of uncertainty associated with the manual segmentation - they are an approximation to the real ground truth, meaning that very fine discrimination may reflect errors in the manual segmentation rather than errors in the automated classification results.

Regarding the improvements at the thresholding step with LOCATE, the initial leave-one-subject-out results of LOCATE on NDGEN, OXVASC and MWSC datasets showed improvement in BIANCA segmentation for all datasets. LOCATE detected more true positive WMHs, especially in the deep white matter, and provided better delineation of WMHs with respect to the manual segmentation as indicated by the improved SI values and the visual results shown in figures 4.12 and 4.13. Since BIANCA is currently not optimised for segmentation of (juxta)cortical, cerebellar and subcortical lesions, we could not further evaluate LOCATE performance in these areas, since they are not included in the BIANCA output.

LOCATE provided results comparable with those obtained using a global threshold, irrespective of changes in the acquisition protocols and image characteristics in the MWSC dataset. The top ranking methods for the MICCAI WMH segmentation challenge used deep learning models such as Deep Stack Networks and Ensemble Learning [344], Multi-dimensional Gated Recurrent Units [238] and a modified location sensitive deep convolutional neural network [227]. The methods were ranked based on the set of performance metrics described in `http://wmh.isi.uu.nl/`, including SI values (Dice) and cluster-wise TPR (Sensitivity of individual lesions). We achieved an average SI value of 0.70 for both LOCATE and global thresholding, while the top ranking methods achieved SI values of 0.80, 0.78 and 0.77. Also, we achieved cluster-wise true positive rate of 0.75 with LOCATE and 0.71 with global thresholding, while the top ranking methods achieved 0.87, 0.83 and 0.73. Therefore,

LOCATE gives a better cluster-wise TPR when compared to global threshold and the third ranking method in the challenge. However, a direct comparison would not be possible since the results of the top ranking methods were determined on the independent test datasets, while our results were obtained on leave-one-out validation on the MWSC training datasets that are publicly available. Also, we retained the regions corresponding to other pathologies for a more stringent evaluation, while these regions were ignored during evaluation in the MICCAI challenge.

Our validation of LOCATE on CADASIL and HC datasets proves that LOCATE is robust with respect to variations in lesion load and location without the need to retrain both BIANCA and LOCATE. In fact, LOCATE can be trained on any data having the same modalities, containing examples of same (or similar) pathologies, acquired with the same sequence as the test dataset. For instance, we used the OXVASC dataset to train LOCATE for testing it on the CADASIL dataset since both datasets were acquired with the same sequences. The pattern of lesions for CADASIL subjects differs from that of vascular subjects in the OXVASC dataset, as CADASIL subjects, in addition to the normal distribution of WMHs seen in SVD, typically have lesions within the external capsule and anterior temporal lobes [40]. Due to this variability in lesion characteristics and pattern, training BIANCA on the OXVASC dataset yields very low lesion probabilities in the temporal regions. Reflecting these lower lesion probabilities, the thresholds map from LOCATE (figure 4.18) showed much lower threshold values, especially in the temporal regions in the CADASIL images. Therefore LOCATE yielded much better segmentation performance (SI = 0.79) compared to a global threshold of 0.9 (SI = 0.57), when the lesion probability map thresholded at 0.2 was considered as the reference segmentation. Moreover, LOCATE results on CADASIL showed a substantial increase of 0.48 in TPR without an increase in FPR, due to the detection of more lesions in the temporal region. This indicates that LOCATE takes into account different lesion patterns and hence can be used in different pathologies. Also, compared with the approaches proposed by [330],

LOCATE overcomes the need for manually determining thresholds for individual subjects on the CADASIL dataset, while still providing similar SI values and comparable limits of agreement with the gold standard on Bland-Altman plots (figure 4.20 versus [330], figure 7, col. 3).

The threshold maps obtained from LOCATE for the HC dataset showed higher thresholds when compared to the CADASIL dataset. This is due to the fact that HCs have mostly periventricular lesions (that are common in healthy ageing) that are usually assigned higher probabilities. Hence by using higher threshold values, LOCATE gave results that appear visually similar to the global thresholding at 0.9, detecting mostly periventricular lesions. While the threshold maps from LOCATE (figure 4.18 and 4.19) provide information regarding the spatial heterogeneity of lesion probabilities, boxplots of thresholds shown in figure 4.21 show the overall characteristics of the datasets. On analysing the LOCATE threshold maps and histograms for various cases (refer to appendix F), we found that distribution of LOCATE thresholds is fairly consistent within a specific population and varies across different populations. The improvement in segmentation performance is more prominent in the OXVASC and CADASIL datasets when compared to the NDGEN, MWSC and HC datasets. This is reflected in their corresponding optimal global thresholds being very different from the median of their LOCATE thresholds. Interestingly, the ranges of thresholds obtained from LOCATE on the OXVASC and CADASIL datasets are wider (excluding the outliers), when compared to the other datasets. This could be due to the different sequence characteristics (e.g. lower resolution) and pathological characteristics (high amount of small vessel disease) of both these datasets with respect to the others. LOCATE results on the five datasets, especially the HC dataset (which has negligible lesion load compared with the CADASIL subjects) show that LOCATE is more adaptive to the variation in the global lesion load in addition to the spatial heterogeneity in lesion probabilities, compared to global thresholding (shown in figure 4.18 and 4.19).

Regarding the number of subjects needed for training, we used a minimal training set of 18 subjects from OXVASC to get good performance on both OXVASC (on 17 subjects, with leave-one-out) and CADASIL. Hence we would recommend using 15 - 20 subjects for training LOCATE. However, it is worth noting that the number of training subjects needed to achieve good results can vary across datasets (for more details on the optimal number of training subjects, refer to [53]). For this reason, we advise checking the segmentation accuracy on each dataset, comparing the output with the manual masks whenever available and, if needed, increasing the number of subjects included in the training or consider re-training on the specific dataset. So far, evaluation of LOCATE has been performed on relatively small datasets (containing 15 - 20 subjects) which limits the power of our results. Also, the experiments done so far show that LOCATE is generalisable when trained with dataset-specific images or when using a training dataset generated with images acquired with the same scanner and sequences. We are currently working towards improving robustness across scanners and sequences. In its current MATLAB implementation, the running time of LOCATE to determine a binary lesion map from an individual lesion probability map is approximately 15 minutes per subject, when run on an iMac with a 2.9 GHz Intel Core i5processor.

LOCATE will be included as an additional option for generating binary lesion maps, as an alternative to the faster but less accurate global thresholding, in a future release of BIANCA. The MATLAB implementation of LOCATE is currently available to use (see `https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BIANCA/Userguide` under the *post-processing* section for more details).

A potential future development of LOCATE could be to use the estimated thresholds to normalise the lesion probabilities locally in order to obtain a more uniform lesion probability map that is robust to the diffuse nature of the lesions in different areas of the brain. Another interesting future direction for BIANCA could be to investigate the amount and distribution of lesions in various cerebral lobes, similar to the analysis of lesions in periventricular and deep regions done in [36].

4.9 Summary

1. We explored different strategies in three steps of BIANCA: feature extraction, classification and thresholding, to improve WMH segmentation.

2. In the feature extraction step, using PPLPM as an additional feature and using non-linear registration (instead of linear registration) for extracting spatial features provided negligible difference in BIANCA performance.

3. In the classification step, out of all the four classifiers, RF and NN classifiers provided results comparable with KNN and might have the potential for improvement in different datasets.

4. The proposed algorithm for the thresholding step, LOCATE, provided considerable improvement in the segmentation. In particular, the improvement in performance with LOCATE is significantly higher in deep regions.

Chapter 5

Deep learning for segmentation of white matter hyperintensities

The work described in this chapter aims to provide an improved and robust segmentation of WMHs, across various datasets, using a deep learning method.

Section 5.1 defines the premise of our experiments and introduces various deep learning concepts that will be explored in this chapter. Section 5.2 lists the datasets used for our experiments and describe our preprocessing pipeline. Section 5.3.1 explains the proposed model architecture, its novel aspects and the post-processing step, while the experimental setup is specified in section 5.4. Sections 5.5 and 5.6 describe our experiments and the corresponding results respectively. Section 5.7 consists of the discussion of our results, conclusions and possible future directions of our work.

The preliminary results of this work have been submitted to the 2020 ISMRM conference (Sundaresan et al., ISMRM 2020).

5.1 Introduction

Automated WMH segmentation in a robust manner across datasets is challenging due to variations in the lesion load and image characteristics. For instance, a supervised learning algorithm trained on a high lesion load dataset, could perform poorly on healthy controls due to the bias towards the higher lesion load learnt by the algorithm. Hence we need a method that avoids over-fitting and is more flexible towards variations in lesion characteristics, load and distribution.

Deep learning (DL) methods such as end-to-end trainable encoder-decoder networks [217, 218] and *fully convolutional networks* (FCNs) [222, 210, 208] have been shown to perform better in the segmentation tasks [58, 238, 227, 59, 345, 232, 226] than the conventional machine learning methods (refer to chapter 2). The increase in computational resources such as GPUs and improvements in pattern recognition and computer vision techniques have enabled the use of neural networks for improved and robust segmentation results. Especially, U-nets [224] and its modified versions [345, 247, 58, 346], including 3D U-Nets [226, 247, 225], have been successfully used for medical image segmentation. In fact, in MICCAI WMH Segmentation Challenge (MWSC) 2017 [216], the top ranking method used ensemble U-net model [58] on FLAIR and T1-weighted images. The architecture of a U-Net (figure 1 in [224]) follows the encoder-decoder framework [217], by creating high-dimensional features by combining multi-scale information. In a U-Net, the encoder arm captures the finer contextual information through a stack of convolutional and pooling layers. The decoder arm consists of deconvolution layers which impart localisation information. There are skip-connections from the encoder to the decoder between the layers at the same levels of abstraction, enabling the aggregation of contextual and spatial information for quicker convergence, making U-Nets computationally efficient.

Various techniques have been used to improve the segmentation performance. Attention gates (AG) aim to do so by selective enhancement of features. In AGs, additional modules or *gates* are added to a layer to weigh the *local* features at that layer with more *global* or coarser features. This weighing would selectively enhance salient features and filters that would lead to better coarse features for localisation. These AGs, called *soft trainable attention gates* were used in end-to-end trainable networks [245] initially for classification tasks, and later for fetal echocardiography plane detection [345], pancreas segmentation [247] and lesion segmentation [347].

Improving model generalisability to reduce over-fitting on the training data is an important aspect of model optimisation. Data augmentation [204], efficiently improves model generalisability by inflating the amount of training data artificially

[348], introducing variation in image characteristics. Ensemble networks have been shown to provide better performance [207, 208, 210, 232, 349] and proven to be resistant to over-fitting, which can occur with more complex models [350]. Ensemble models could consist of models with different initialisations [58] or from different planes [351].

In addition to the above techniques, domain adaptation (DA) methods address the issue of discrepancies in the data distributions obtained from various *domains* (e.g. different scanners), ensuring robust performance of the model irrespective of the domain [249]. DA methods aim to transfer the knowledge from a *source* domain to a *target* domain by leveraging the invariant features across various domains [268, 266]. Various DA techniques used so far include, for example, minimising a distance metric of domain variance [352, 260, 261, 353], using transferable features [262, 263] for creating intermediate feature representation between domains, transfer learning [266, 262] and so on. In the transfer learning (TL) [266] paradigm, the knowledge from the source domain is transferred to the target domain by fine-tuning the source-trained model using target domain data. During fine-tuning, the initial convolutional layers (domain invariant low-level features) are generally kept constant or *frozen*, while the final layers (task/domain specific high-level features) are fine-tuned on the target datasets. Fine-tuning of the pre-trained models has shown to improve the performance [267], especially when image characteristics between domains are more similar [262, 268]. TL has been applied for medical image segmentation tasks [269, 270, 267], including lesion segmentation [271]. In TL, the training on different domains occurs separately and hence does not enable us to mix data from different domains. Also, the performance relies on the amount of target training data [267].

Another DA strategy, *domain adversarial training of neural networks* (DANN) [272, 273] relies on domain invariant features to achieve good domain adaptation. The DANN model consists of a feature extractor network with two discriminative classifiers: (1) domain predictor, which predicts the data domain, and (2) label predictor that predicts the class labels. The model framework performs adversarial

training at domain predictor using a *gradient-reversal layer*, placed between the feature extractor and the domain predictor, optimising the features and shared weights. The gradient-reversal layer maximises the domain prediction loss, thus minimising the shift between the domains, while making the model discriminative towards the main task of label prediction. Hence, domain transfer is achieved by simultaneously using source and target domain datasets.

In this work, we propose *Triplanar U-Net ensemble network* (TrUE-Net), an ensemble of U-Nets applied to three planes of structural brain MR images, using loss functions that take into account the anatomical distribution of WMHs, for accurate WMH segmentation (section 5.3). We study the effect of the various aspects and parameters of the proposed model on the segmentation results (sections 5.5.2 and 5.5.3). We perform the initial leave-one-out evaluation of the model on various datasets with different acquisition characteristics and pathological signs (section 5.5.4). We also compare the performance of TrUE-Net and its modified version (with the addition of AG) with BIANCA [53] (section 5.5.5). The corresponding results are reported in sections 5.6.1 - 5.6.4. We later test the generalisability of the model and experimented with several test cases that explore the possibility of DA using two domain transfer techniques: TL and DANN (sections 5.5.6 and 5.6.5).

5.2 Dataset preparation

5.2.1 Datasets used

For the experiments described in this chapter, we used the same datasets used in chapter 4. For more details, see section 4.2.1 and appendix A.

1. Neurodegenerative cohort (NDGEN)
2. Vascular cohort - Oxford Vascular Study (OXVASC)
3. MICCAI WMH segmentation challenge Dataset (MWSC) consisting of three cohorts from Utrecht, Singapore and Amsterdam

4. Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) cohort

5. Healthy controls (HC)

5.2.2 Data preprocessing

We used both T1-weighted and FLAIR images as inputs for the model. We reoriented the T1-weighted and FLAIR images to match the orientation of the standard MNI space, to aid in further processing. We skull stripped the images using FSL BET [335], followed by bias field correction using FSL FAST [336]. We registered the T1-weighted image to FLAIR using linear rigid-body registration with FSL FLIRT [337].

Secondly, we cropped the above skull stripped FLAIR and T1-weighted images closer to the brain edges to make the FOV tighter, thus avoiding the non-brain voxels. We then extracted the 2D slices from the volumes for training and testing from all three planes. Note that, rather than ignoring the top/bottom slices as done in [58] (the top ranking method in MWSC [216]), we do not omit any slice in the 3D volume, since there is a possibility of lesion occurrence, albeit slim, in those regions. We then apply Gaussian normalisation on the cropped images to normalise the intensity values. While preparing 2D slices in the axial plane, we cropped the slices to a dimension of 128 $\times$ 192 voxels for all datasets. For sagittal and coronal slices, we extracted the slices from the tightly cropped 3D volumes and resized them from their various original dimensions to 192 $\times$ 120 and 128 $\times$ 80 voxels respectively, using bilinear interpolation.

5.3 Triplanar U-Net Ensemble Network (TrUE-Net)

5.3.1 Network architecture

The choice of the model dimension plays a crucial role in extracting the useful spatial information, and hence for accurate segmentation. This choice is affected by various

factors such as size and distribution of lesions and amount of data available. As described in chapter 2, 3D CNNs have been successfully used in the segmentation of variety of lesions [232, 233, 247], since they utilise the contextual information along all dimensions. However, images with WMHs can include small lesions in the deep regions, with low contrast and poor context. Also, one of the dimensions, usually the z-dimension, often has poor resolution, leading to inconsistencies in the detection of WMH boundary voxels using 3D CNNs. From the implementation point of view, 2D models use far fewer parameters when compared to 3D models. However, using 2D models can cause discontinuities in segmentation across the z-dimension since individual slices are considered separately.

The proposed triplanar architecture overcomes the disadvantages of both 2D and 3D models. The network consists of three U-Nets, each detecting WMHs from 3 different planes of an image. The triplanar network is well-suited for this task, since the model (1) avoids discontinuities in WMH segmentation, especially in images with larger slice thickness - since each U-Net views the image from a different plane, it captures even small DWMHs with less spatial context, in a continuous manner; (2) provides better and comprehensive edge detection, since the edges in a single plane correspond to whole lesions in another plane; (3) gives better context across slices with far fewer parameters when compared to a 3D CNN.

We combined three 2D U-Nets in parallel within an ensemble model. We chose to use an ensemble model because (1) different models in an ensemble can recognise subtle differences in feature maps that other models might miss, (2) using an ensemble of models avoids the variation in the resulting probability maps, due to noise or spurious structure, since final probabilities are combined with equal weights (usually averaging).

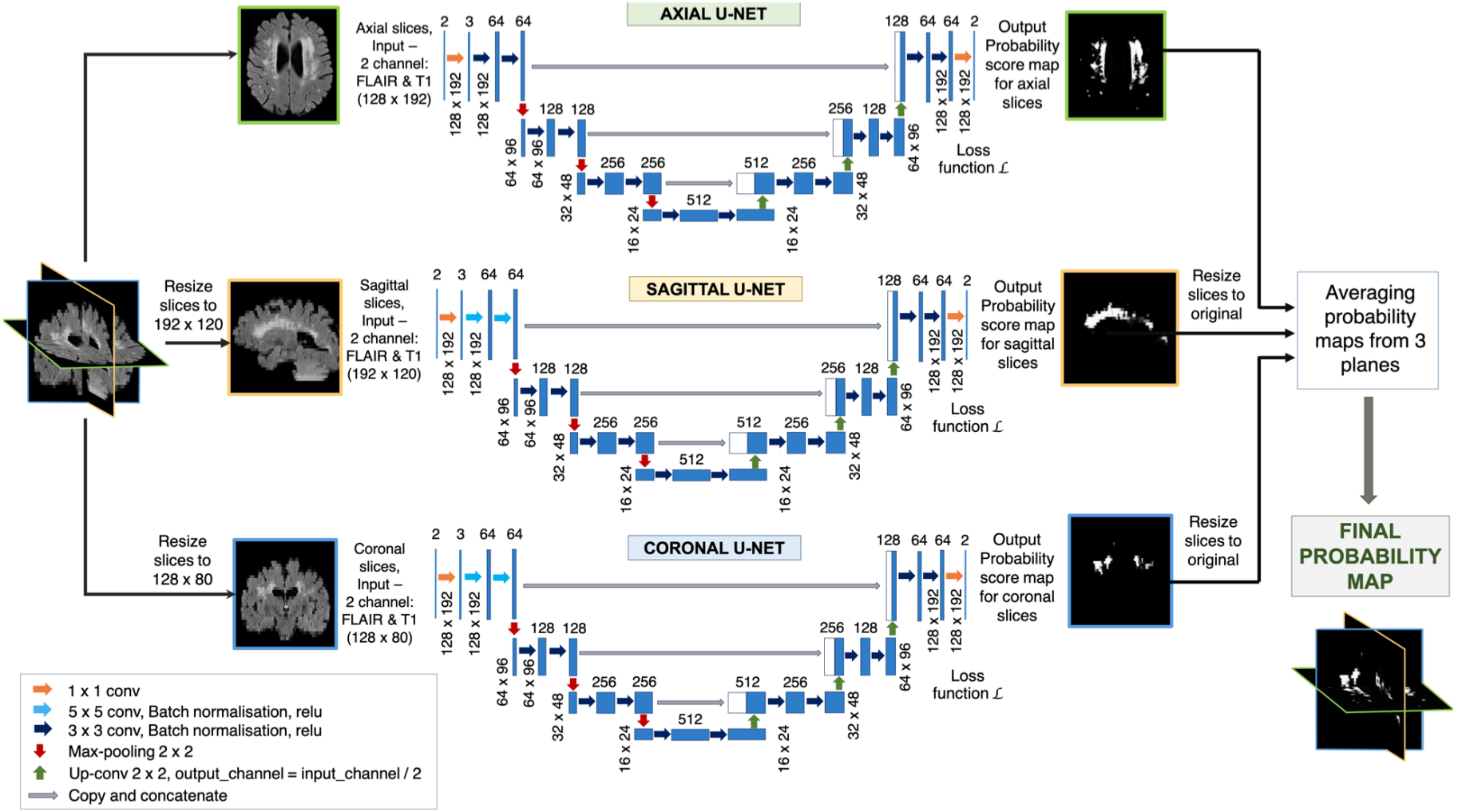


Figure 5.1: Triplanar U-Net ensemble network (TrUE-Net) architecture.

Figure 5.1 shows the architecture of the proposed ensemble network. For each plane, the 2D U-Net model takes FLAIR and T1-weighted slices as input channels and provides the probability map in the corresponding plane. We trimmed the depth of the classic U-Net [224] to obtain the 3-layer deep U-Net model to reduce the computational load and to improve the model sensitivity towards small lesions. The encoder arm of U-Net pools (*downsamples*) the feature maps, while the decoder arm expands the features (*upsamples*), with skip connections that concatenate features between the encoder and decoder arm. At each level of abstraction in the encoder and decoder, the model consists of two consecutive 3×3 convolutional layers followed by the 2×2 max-pooling layer (for encoder) or 2×2 upsampling layer (for decoder). Each convolution layer is followed by batch normalisation and $ReLU$ (rectified linear unit) layers. Our model mostly uses 3×3 convolutional layers, except for the initial 5×5 convolutional layers in the cases of sagittal and coronal U-Nets, since larger receptive fields could adapt well to variations in lesion characteristics and lower resolution along the z-direction. We added a 1×1 convolutional layer finally before the softmax layer for predicting the probability maps. In the ensemble model, training of U-Nets in the individual planes occurs independently using the slices extracted from the corresponding planes from the resized training images. During testing, we predicted the probabilities for all slices in the individual planes and then assemble these into 3D probability maps in the corresponding planes. We then obtained the final probability map by averaging the individual 3D probability maps from the 3 planes, after resizing them back to their original dimensions.

5.3.2 Loss functions

We used the sum of voxel-wise cross-entropy (CE) loss function and Dice loss (using Dice overlap measure between the lesion and the ground truth) as the total cost function. The CE loss function has been widely used for voxel-wise semantic segmentation [222, 224]. In our case, the CE loss aims towards the detection of larger periventricular WMHs, due to the imbalance between WMH and background class samples, especially in deep regions. Hence we weighted the CE loss function to up-weight

the areas that are more likely to contain the less represented class (i.e. WMH). In addition to that, we added Dice loss, which has stronger sensitivity to missing small WMHs when compared to the CE loss. Dice loss has also been shown to work well with unbalanced classes [226, 58].

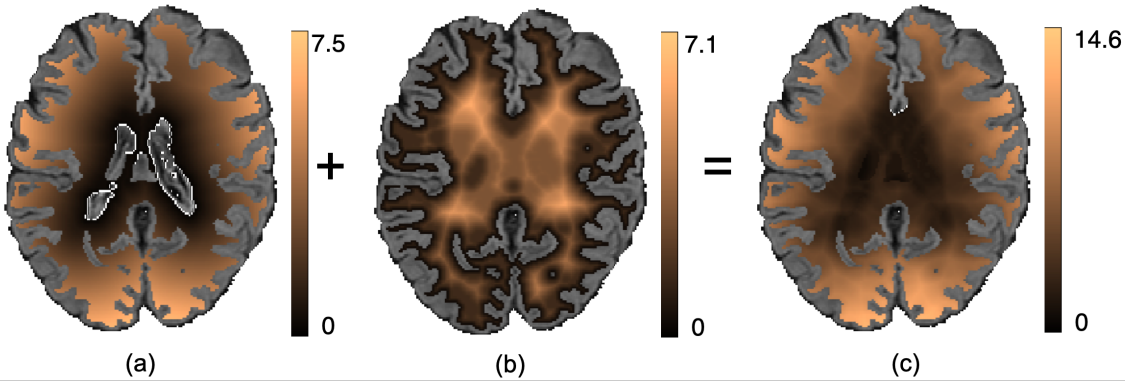


Figure 5.2: Example weight maps for weighing voxel-wise cross-entropy loss function. Maps showing the sum of (a) Square-root of distance from ventricles D_{vent} and (b) Square-root of distance from grey matter (GM) D_{GM} to get (c) Final weight map D_{wei}.

Figure 5.2 shows an example of weight map for weighting CE loss function. We determined the square-root distance from ventricles, $0 \leq D_{vent} \leq \sqrt{N_D},\ N_D \in \mathcal{R}$ (figure 5.2a), and the distance from the brain grey matter (GM), $0 \leq D_{GM} \leq \sqrt{M_D},\ M_D \in \mathcal{R}$ (figure 5.2b) and determined the final weight map as the sum of both, $D_{wei} = D_{vent} + D_{GM}, 0 \leq D_{wei} \leq (\sqrt{N_D} + \sqrt{M_D})$ (figure 5.2c), where N_D and M_D are maximum distances from ventricles and GM respectively (all distances in mm). Hence the weighted CE loss is given by,

$$CE = -\sum_{x=1}^{C} y(x)\ D_{wei}(x)\ \log(p(x)) \tag{5.1}$$

where $p(x)$ denotes the output of the soft-max layer, C is the number of classes and $y(x) \in \{0, 1\}$ indicates the binary value at voxel x. Given the manual segmentation M and binary map P_{th} obtained by thresholding predicted probability maps, the Dice loss for N voxels is given by,

$$DL = -\frac{2 \times \sum_{x=1}^{N} M(x) \cdot P_{th}(x)}{\sum_{x=1}^{N} M(x) + \sum_{x=1}^{N} P_{th}(x)} \tag{5.2}$$

Hence the total loss function $\mathcal{L}$ is given by,

$$\mathcal{L} = CE + DL = -\sum_{x=1}^{C} y(x)\ D_{wei}(x)\ \log(p(x)) - \frac{2 \times \sum_{x=1}^{N} M(x) \cdot P_{th}(x)}{\sum_{x=1}^{N} M(x) + \sum_{x=1}^{N} P_{th}(x)} \quad (5.3)$$

5.3.3 Post-processing

The post-processing consists of three steps: firstly, as a reversal of the tight cropping done in preprocessing step, we padded the probability maps with zeros to get the images back to their original dimensions. Secondly, we masked the probability map with WM obtained from a dilated and inverted cortical CSF tissue segmentation, combined with other deep grey exclusion masks [53] (used in chapter 4). Finally, we thresholded the masked probability maps at a global threshold of 0.5.

5.4 Experiment setup

5.4.1 Optimisation parameters

We trained the network using the Adam Optimiser [354] with epsilon (ϵ) value of 1×10^{-4}. We used a batch size of 8, with an initial learning rate of 1×10^{-3} and reducing it by a factor 1×10^{-1} every 2 epochs, until it reaches 1×10^{-5}, after which we maintain the fixed learning rate value. We chose the above parameters empirically based on the model convergence (refer to section 5.5.2 for more details). We also empirically set the total number of epochs to 100, while we observed that the model usually converged before 80 epochs. We used the truncated normal variable (with σ = 0.1) for weights initialisation and the biases were initialised as constants with a value of 0.1.

5.4.2 Data augmentation

The commonly used data augmentation techniques include geometric transformations [250] such as translation, rotation, shearing, noise injection, filtering and using neural network models such as GANs (e.g. style transfer) [251, 252]. In this work, for all the three planes the augmentations on the slices were performed using randomly chosen

transformations in random combinations: translation, rotation, adding random noise and Gaussian filtering. For the above operations, we randomly chose the hyperparameters from a specific range of values (listed in table 5.1). The values were set empirically by experimenting to introduce sufficient variance in the generated data and to mimic the clinical difference observed between various subjects. For axial slices, we augmented the data by a factor of 10, while for sagittal and coronal slices we augmented by a factor of 6. Despite the high augmentation factor, the random combinations of transformation avoid redundant information. We augmented the axial slices by a larger factor than the other two, in order to compensate for the smaller number of axial slices, especially after cropping the volumes to tighten the FOV (refer section 5.2.2 for details). After augmentation, we obtained $\approx$ 900-950 slices per subject.

Transformations	**Parameter ranges**
Translation (voxels)	x-offset: [-10, 10], y-offset: [-10, 10]
Rotation (degrees)	θ = [-10, 10]
Random noise injection (after Gaussian normalisation)	Distribution: Gaussian, $\mu = 0$, σ^2 = [0.01, 0.09]
Gaussian filtering (voxels)	σ = [0.1, 0.3]

Table 5.1: Transformations used for data augmentation and their parameter ranges.

5.4.3 Implementation details

We implemented the network in Python 3.6 using TensorFlow 1.11.0. The network was trained on a NVIDIA Tesla K40c, taking an total training time of $\approx$ 125 minutes (for 3 planes) for 15,000 training samples. For testing, each subject with an average of 45 slices took an average testing time of $\approx$ 6.2 seconds (on GPU) for the final prediction. The training and validation split was set to 90% and 10% of samples respectively.

5.5 Experiments

This section explains our experiments to determine the performance of the proposed model and to explore its robustness with respect to variations in dataset characteristics.

5.5.1 Performance evaluation metrics

We used the same performance measures as those used in chapter 4 as follows:

- Dice Similarity Index (SI)
- Voxel-wise true positive rate (TPR)
- Voxel-wise false positive rate (FPR)
- Cluster-wise TPR

In addition to the above measures, we used 3 additional measures that were used in MICCAI WMH Segmentation Challenge for evaluating model performance and for comparison analysis (between regions/methods) in sections 5.6.3 and 5.6.4.

- Absolute volume difference (AVD) in percentage
- Cluster-wise F1-measure
- 95^{th} percentile of Hausdorff distance measure in mm

Definitions of the above evaluation metrics are provided in appendix C.

5.5.2 Effect of training parameters on model performance

We used the NDGEN dataset for the initial optimisation of network parameters, and for determining the number of epochs. We explored the effect of batch-size, learning rate and epsilon (ϵ) value of the Adam optimiser on model convergence. Batch size is the number of training data used by the optimiser at a given instance for updating model parameters. The parameter ϵ is used in the denominator of the weight

update function in the optimiser to avoid the divide-by-zero error. Learning rate determines the extent of change/adjustment of weights with respect to the gradient. We experimented with three batch sizes: 8, 16 and 32, and three ϵ values: 1×10^{-2}, 1×10^{-4} and 1×10^{-6}. For learning rate, we tested the effect of following 3 settings:

1. **Higher:** Initially 1×10^{-2}, reducing by a factor 1×10^{-1} every 2 epochs, till it reaches 1×10^{-4} and keeping it constant afterwards.

2. **Medium:** Initially 1×10^{-3}, reducing by a factor 1×10^{-1} every 2 epochs, till it reaches 1×10^{-5}.

3. **Lower:** Initially 1×10^{-4}, reducing by a factor 1×10^{-1} every 2 epochs, till it reaches 1×10^{-6}.

5.5.3 Ablation study: effect of the model dimension and loss function on the segmentation performance

We performed an ablation study to determine the effect of model dimension (2D/3D) and loss function on the segmentation results from the NDGEN dataset. We removed one aspect of our model dimension at a time (from 3 planes to single plane, from 2D to 3D) and compared the performance with three different models: 3D U-Net, 2D U-Net on axial slices only and the proposed TrUE-Net. We maintained the same training and model parameters for all three options. In the case of the loss function, we removed one component of the loss function at a time and compared the performance of TrUE-Net with three cases of loss functions: cross-entropy loss, weighted cross-entropy loss and weighted cross-entropy loss + Dice loss.

5.5.4 Leave-one-out evaluation of WMH segmentation

We performed leave-one-out (LOO) evaluation of our proposed model on the 3 cohorts of the MWSC dataset (Utrecht, Singapore and Amsterdam), NDGEN and OXVASC datasets, based on the performance metrics specified in section 5.5.1. We also determined the performance of the model in the deep and the periventricular regions separately, and performed paired t-tests to compare the results between regions. As

in chapter 4, we adopted the 10 mm distance rule [33] for classification of PWMH and DWMH (section 1.2.1, chapter 1).

5.5.5 Comparison with BIANCA

We compared the performance of our proposed model with BIANCA. Further, we also studied the effect of adding attention gates (AGs) to the coarsest layer of abstraction in TrUE-Net on the segmentation performance, when compared to the original model and BIANCA. We performed this comparative analysis on MWSC cohorts, NDGEN and OXVASC datasets, based on the performance metrics specified in section 5.5.1.

For BIANCA, we used the same features and training options specified in section 4.2.3 of chapter 4, in addition to the default options [53]. Also, for all the above datasets, we applied a global threshold value of 0.9 on the BIANCA lesion probability maps to obtain the final binary lesion maps.

5.5.5.1 Effect of adding AG at the coarsest level of TrUE-Net (TrUE-Net-Att)

We used the layer at the coarsest scale at the bottom of the U-Net for adding our trainable AG. By adding the AG to the coarsest scale, we motivate the model to learn the overall lesion distribution better. We used the AG architecture proposed in [245]. Let $f^l = \left[f_i^l\right]_{i=1}^{i=n}$ be the feature map in a layer l where f^l is the voxel vector of length n with C channels. AG determines the attention weights $a^l = \left[a_i^l\right]_{i=1}^{i=n}$, where $a_i \in [0, 1]$. The weights a^l enhance the regions with salient features and down-weight the false regions, when multiplied with the layer feature to provide an attention output $A_f^l = \left[a_i^l f_i^l\right]_{i=1}^{i=n}$.

Architecture: Figure 5.3 illustrates the TrUE-Net model with the AG (TrUE-Net-Att) and the architecture of the AG layer for the axial U-Net (sagittal and coronal U-Nets were modified in a similar manner). We used the coarsest layer of the decoder after the maximum number of pooling layers for deriving our global feature, since this layer has the largest spatial context/support. We applied 2 $\times$ 2 max-pooling layers on the coarsest layer consecutively until it reached the dimensions of 1 $\times$ 1 $\times$

512. This is the global feature $\mathfrak{g}$, which has minimum dimension but has the largest spatial context of the input images. AG has two inputs (figure 5.3b): the global feature $\mathfrak{g}$ ($1 \times 1 \times 512$) and the feature representation $\mathcal{F}$ ($32 \times 48 \times 256$) at the layer that needs refining. We applied a projection layer (convolution using 1×1 kernels, [210]) on the feature $\mathcal{F}$ to match the channel dimensions of $\mathfrak{g}$. The global feature $\mathfrak{g}$ was then added voxel-wise in the corresponding channels to the projected output. The sum ($\mathcal{S}_{\mathcal{F}}$, $32 \times 48 \times 512$) was projected again into a single channel and normalised to obtain the attention weights. The attention weights were multiplied with the $\mathcal{S}_{\mathcal{F}}$ channel-wise to obtain the output of AG, $\mathcal{A}_{\mathcal{F}}$ ($32 \times 48 \times 512$). The output $\mathcal{A}_{\mathcal{F}}$ replaces the skip connection output from the encoder and was concatenated with the upsampled decoder layer at that level, as done normally in U-Nets.

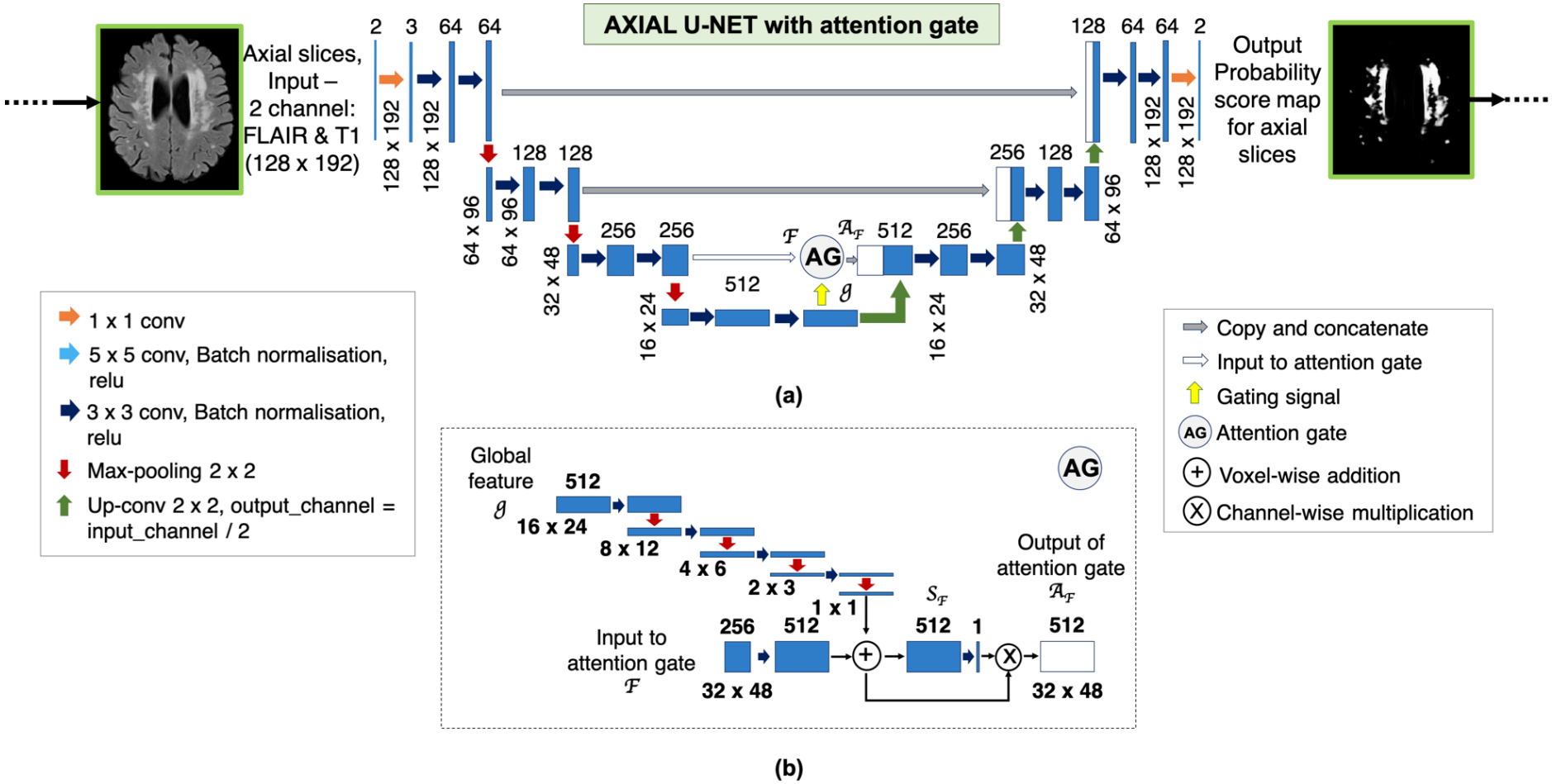


Figure 5.3: Triplanar U-Net architecture with attention gate (AG) at the coarsest scale (TrUE-Net-Att). (a) The U-Net architecture with AG shown for the axial U-Net only. Note that the sagittal and the coronal U-Nets were modified in a similar manner. (b) Structure of AG with input feature $\mathcal{F}$ and output $\mathcal{A}_\mathcal{F}$.

Implementation details: We used the same optimisation parameters and training/validation splits as those used for TrUE-Net. TrUE-Net-Att model was implemented on an NVIDIA Tesla P100 GPU and took $\approx$ 130 minutes for training in all 3 planes for 15,000 training slices. During testing, each subject with an average of 45 slices took $\approx$ 6.3 seconds for the final prediction.

5.5.6 Harmonisation and domain adaptation: transfer learning and domain adversarial training

One of the most desirable attributes of a model is its generalisability across various datasets. We have datasets from different domains, acquired using different scanners and having different intensity characteristics. Figure 5.4 shows the t-distributed Stochastic Neighbour Embedding (tSNE) plot [355] of datasets based on their FLAIR and T1-weighted intensity values (after bias field correction, cropping and Gaussian normalisation - refer to section 5.2.2). From the plot, it is evident that the datasets are quite dissimilar from each other, forming individual domains, except for two similarities (e.g. OXVASC and CADASIL, Singapore and Amsterdam) based on the intensity values alone. In addition to intensities, there could be variations in the acquisition protocols, for instance, OXVASC and CADASIL have different voxel dimensions, with maximum spatial resolution in a different plane (coronal instead of axial) with respect to the other datasets. Given these variations between domains, we tested the model's generalisability and explored the possibility of data harmonisation using various test cases. For these experiments, we considered the combination of Utrecht and Singapore datasets as our *source* domain datasets, and the combination of OXVASC and CADASIL as our *target* domain datasets, as shown in figure 5.5. We started with the simplest case where the source model is directly applied to the target data. We then performed domain adaptation (DA) using two techniques: transfer learning (TL) and domain adversarial training of neural networks (DANN). Finally, we trained the model from scratch on the target domain and tested it on the target data. The final case is generally expected to perform better (if trained on sufficient data) since the model is trained on similar data.

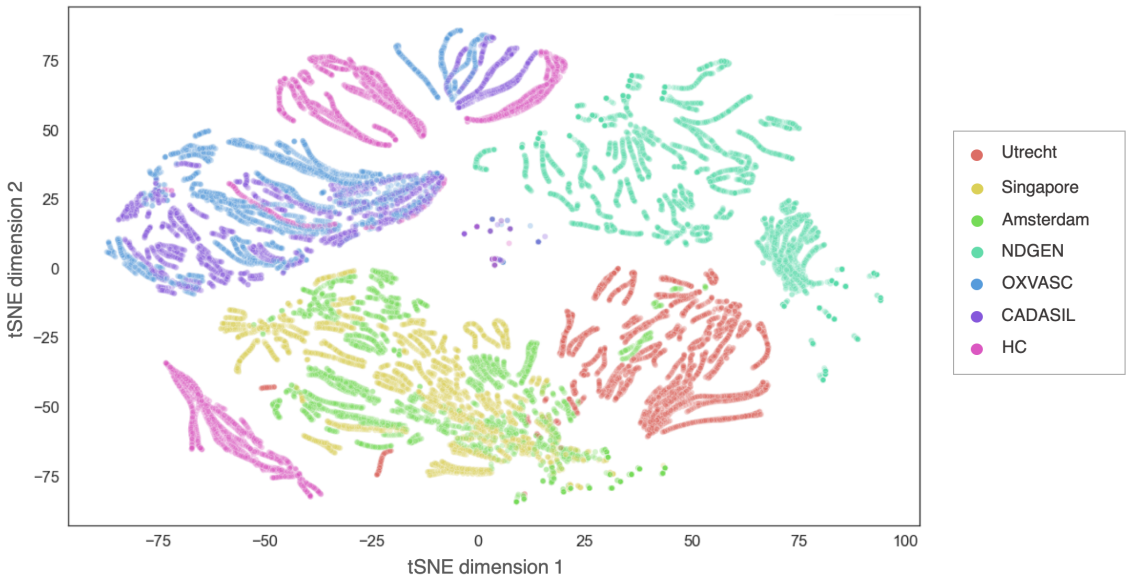
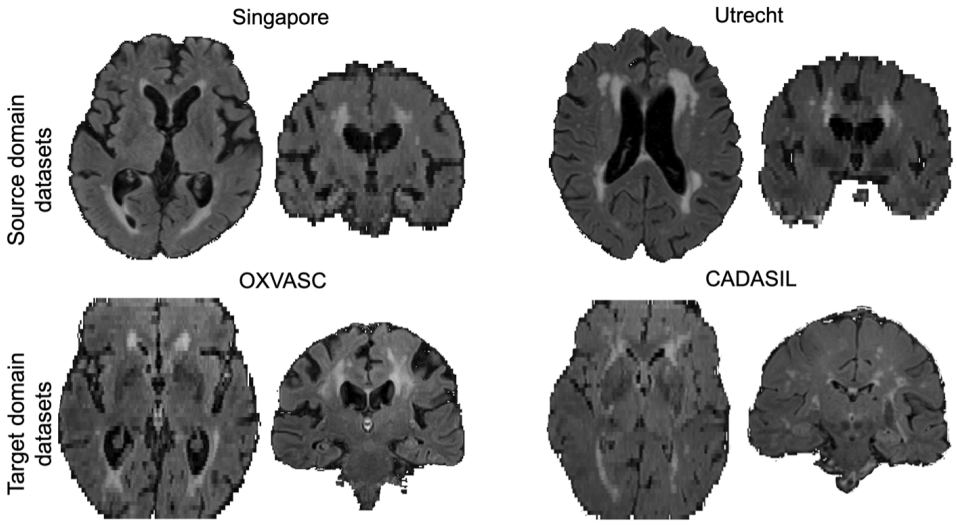


Figure 5.4: T-distributed Stochastic Neighbour Embedding (tSNE) plot showing the combined distribution of FLAIR and T1-weighted intensities for various datasets. The intensity values were obtained after bias field correction, cropping and Gaussian normalisation, specified in section 5.2.2

We chose OXVASC and CADASIL for the target domain, since they are different from other datasets in terms of their intensity profiles and acquisition characteristics, as shown in the bottom panel of the figure 5.5. For the cases requiring training data from the target domain, we used the CADASIL dataset for training. Since we did not have manual segmentations for the CADASIL dataset, we used the LOCATE outputs (from chapter 4) for training purposes. For reporting the results of all the cases on the target domain, we used the OXVASC dataset as the target domain test dataset (the table in figure 5.5) since it has manual segmentations and enables us to determine the performance measures. Also, the OXVASC and CADASIL datasets were quite similar to each other (suggested by their overlap in the tSNE plot in figure 5.4) compared to other datasets and hence did not require to be combined together before the train/test split, as done with the source domain datasets.

5.5.6.1 Case 1: Train on a source domain and apply directly to a target domain

In order to determine the inherent generalisability of TrUE-Net, we trained the model on the source domain datasets and tested the model directly on the OXVASC dataset



	Source	Target
Training dataset	Singapore (10) + Utrecht (10)	CADASIL (15)
Test dataset	Singapore (10) + Utrecht (10)	OXVASC (18)

Figure 5.5: Top panel: sample images from axial and coronal planes shown from the source (top row) and the target (bottom row) domains. Bottom panel: the splits for training and test datasets in the source and the target domain are provided in the table.

in the target domain.

5.5.6.2 Case 2: Transfer the model trained on a source domain to a target dataset with fine-tuning

We trained TrUE-Net on the source domain datasets to get the source pre-trained model. We later fine-tuned the model by training the model on the CADASIL dataset (using LOCATE outputs as ground truth) from the target domain, starting with specific layers. For fine-tuning we used a smaller learning rate (lower setting used in section 5.5.2). Figure 5.6a shows the layer numbers for the U-Net model, with an arrow at bottom indicating order of fine-tuning. Given $\mathcal{L}$ layers in total, 'fine-tuning i layers' means that $\mathcal{L} - i$ layers before i were frozen and the layers from i towards the decoder end were fine-tuned. For each fine-tuning we increased the number of training subjects in the CADASIL dataset and measured the performance of each fine-tuned model. Finally, we determined the best layer to start fine-tuning the model, and

determined the effect of training data on the performance of the transferred model applied to the OXVASC dataset.

5.5.6.3 Case 3: Unsupervised domain adversarial training on the source and target domains

We implemented DANN as shown in figure 5.6b by adding a domain predictor to the proposed TrUE-Net model. We added the domain predictor to the coarsest level of abstraction after maximum levels of pooling, since it has high-level features. After another 2×2 max-pooling layer, we added two fully connected layers (with 1024 and 100 nodes) at the end, followed by a softmax layer for domain prediction. We added a gradient reversal layer between the feature extractor and the domain predictor. This reverses the direction of estimated gradients during back propagation, leading to adversarial training with respect to domain prediction. In this model, the domain predictor requires domain labels (i.e. source/target) for both source and target datasets. On the other hand, the label predictor requires manual segmentations from source datasets only, making the model unsupervised with respect to the target domain. Hence, the domain predictor makes the model domain invariant by considering data from both domains, while the label predictor optimises the model for accurate WMH segmentation in both domains. We tested the DANN model on the test datasets from the source and target domains (specified in the table in figure 5.5), and measured model performance in each domain individually.

Implementation details: For the label predictor of the DANN model, we maintained the same parameters that were used for the original TrUE-Net model. For the domain predictor, we trained with the Momentum optimiser [356] with a momentum value of 0.9. We used a batch size of 16, with the same learning rate and number of epochs as that of the label predictor, with values set empirically using a trial-and-error method. The model was implemented on an NVIDIA Tesla V100 GPU and took $\approx$ 138 minutes for training in all 3 planes for 14,000 training samples. During testing, each subject, with an average of 50 slices, took an average time of $\approx$ 8 seconds.

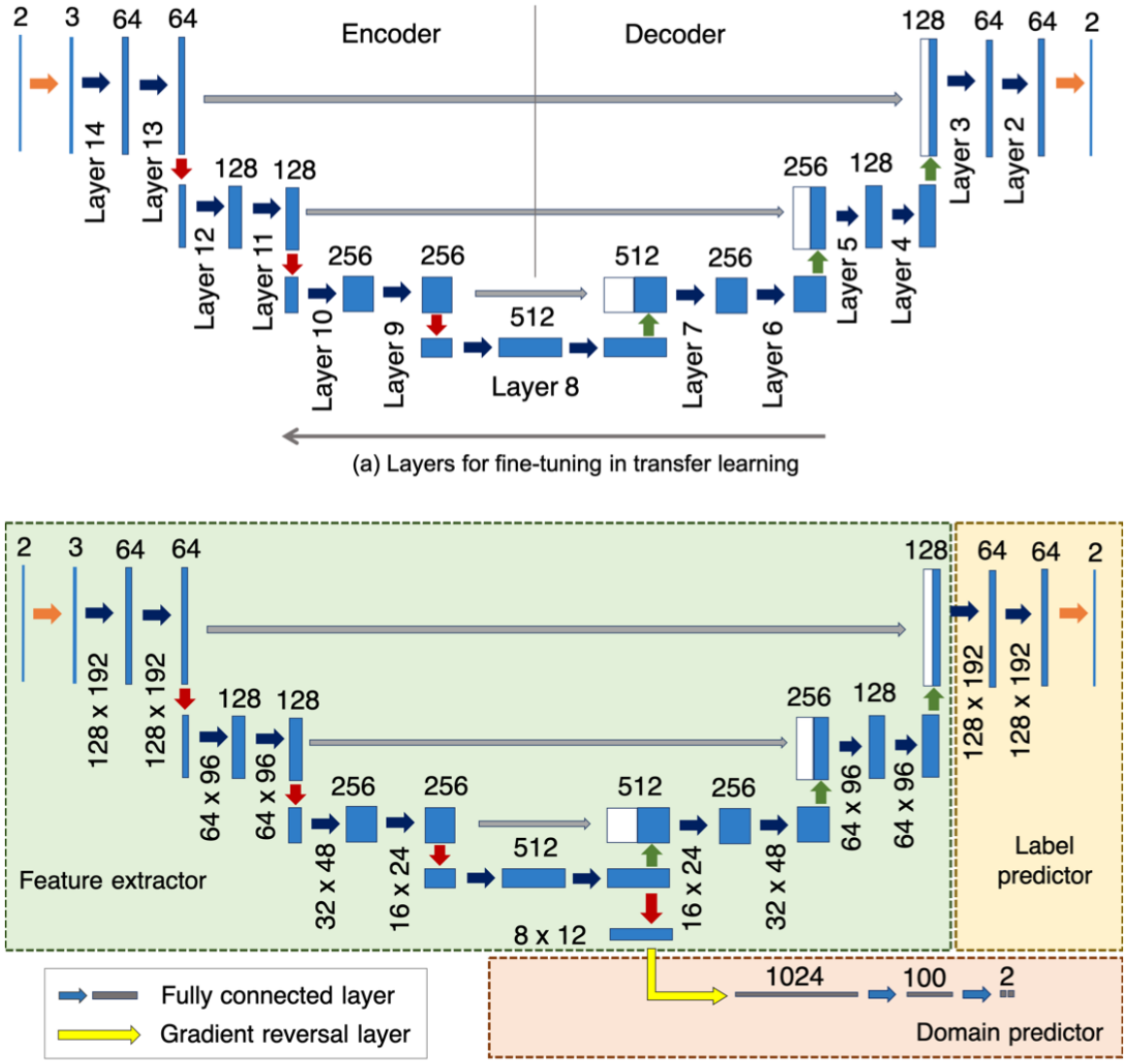


Figure 5.6: Transfer learning (TL) framework and domain adversarial neural network (DANN) architectures. (a) Layer numbers indicated for the TL case. The arrow indicates order in which the layers were unfrozen, (b) DANN architecture with TrUE-Net illustrating feature extractor (green), label predictor (yellow) and domain predictor (red). For both (a) and (b), only the axial U-Net is shown; note that the sagittal and the coronal models were modified similarly.

5.5.6.4 Case 4: Semi-supervised domain adversarial training on source and target domains

For this experiment, we trained the DANN model (shown in figure 5.6b) in a semi-supervised manner, wherein manual segmentations from a fraction of target training data were used in addition to source training datasets for training the label predictor. The remaining target training data is used only for domain prediction. Hence, the

only change from the previous case is that we used a small proportion (25%, chosen empirically, which amounts to 4 subjects) of the target training data, CADASIL, for training the label predictor, in addition to the source training dataset. As mentioned before, for CADASIL, we used BIANCA outputs thresholded using LOCATE (derived in chapter 4) as ground truth segmentations for training. We tested the model on the source and target domain test datasets (refer to the table in figure 5.5) to determine source and target performance of the model individually.

5.5.6.5 Case 5: Train on the target domain from scratch and apply to the target domain

For this case, we trained the TrUE-Net model on the CADASIL dataset and tested the model on the OXVASC dataset, both of them from the target domain. This differs from case 1 due to fact that the OXVASC and CADASIL datasets in target domain, though not identical, are similar in several aspects: both of them are acquired with the same scanner with the same protocol (hence similar voxel resolutions along various planes) and have similar intensity characteristics, when compared to source datasets. Hence, this case is expected to provide better results compared to case 1.

We studied the performance using DA techniques (case 2 to 4), compared to applying the source domain model directly to the target domain (case 1). We also compared the segmentation performance between transfer learning (case 2) and domain adversarial training (cases 3 and 4). Finally, we observed the improvement in the segmentation performance on target datasets when updating the model from unsupervised to semi-supervised DANN (between case 3 and 4). For these comparisons, we performed two-tailed paired t-tests between the individual pairs of test cases.

5.6 Results

5.6.1 Effect of training parameters on model performance

Figure 5.7 shows the loss function decay with different batch sizes, learning rates and epsilon (ϵ) values. In all the cases, the model started to converge at approximately

80 epochs.

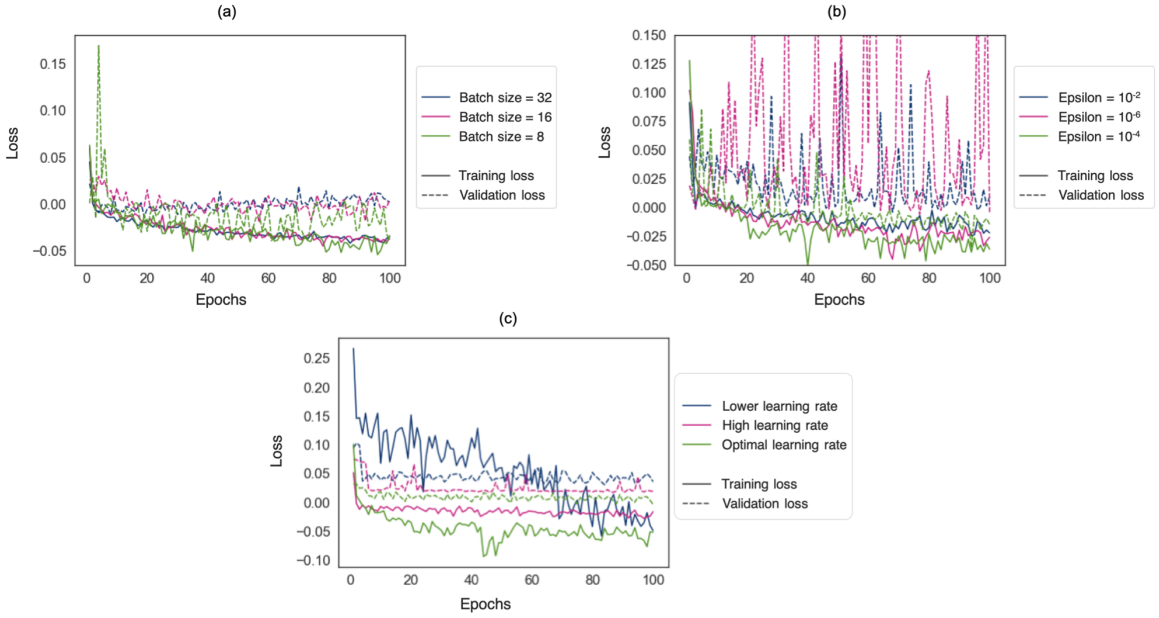


Figure 5.7: Effect of training parameters on model convergence on the NDGEN dataset. Training and validation loss decays have been shown for different (a) batch sizes, (b) epsilon (ϵ) values and (c) learning rates. The plots shown in green correspond to the optimal values chosen for each parameter.

At a lower batch size of 8, the loss function became more noisy, but had lower values. It has been shown that lower batch size has a regularisation effect during optimisation, perhaps due to the noise they add to the training loss [357, 358]. The increase in batch size often results in over-fitting, as evident from higher validation loss values for batch sizes of 16 and 32 (figure 5.7a). Therefore, we chose a batch size of 8 for our experiments henceforth.

As the ϵ value gets smaller in the denominator, the optimiser makes larger weight updates leading to unstable optimisation, as shown in figure 5.7b for an ϵ value of 1×10^{-6}. Having a reasonably large value of ϵ could make weight updates smaller although at the cost of slower training. Since an ϵ value of 1×10^{-2} provided higher loss values (with slightly unstable validation loss), we set the ϵ value to the optimal value of 1×10^{-4} for the further experiments.

From figure 5.7c, for a lower learning rate, the loss decay was slower and hence required more epochs for convergence. On the other hand, for a higher learning rate

the loss values values converged before 20 epochs, although with higher loss values. At the optimal learning, the loss decay was reasonably slow and converged to a much lower loss value and also showed lower validation loss. Hence, we chose the optimal learning rate (1×10^{-3} to 1×10^{-5}) for our experiments henceforth.

5.6.2 Ablation study: effect of the model dimension and loss functions on segmentation performance

Figure 5.8 shows a sample prediction, from all three planes, of our proposed model compared with 3D U-Net and 2D axial U-Net.

The 3D U-Net detected the most straightforward bright PWMHs, which have high contrast. The model missed most of the DWMHs and provided under-segmentation or incorrect delineation of PWMHs (circled regions in figure 5.8c). We found that the 3D model missed more WMHs as the lesion load decreased, giving a poor segmentation in low lesion load subjects. The 2D axial U-Net gave a better segmentation when compared to the 3D model, but still lacked the contextual information across slices. Hence, it missed parts of WMHs in some of the contiguous slices, leading to discontinuities in WMH detection. An instance of this is shown in figure 5.8d, where the discontinuity is clearly visible from the sagittal and coronal planes. The proposed TrUE-Net model predicted WMHs using contextual information from all three planes and provided a continuous and more accurate segmentation (figure 5.8e) using fewer parameters compared with the 3D U-Net.

Figure 5.9 shows the effect of CE and Dice loss functions on TrUE-Net segmentation performance for a subject with medium lesion load. While the effect of model dimension is clearly visible, the effect of loss function is quite subtle and was observable only along edges and very faint DWMHs. Since CE loss function aggregates the loss values at individual voxels into an image-level value, it is generally biased towards larger PWMHs, rather than DWMHs. As a result, the PWMHs were over-segmented (figure 5.9c), and in some cases even extended to the normal ventricle lining. Using weighted CE loss controls this behaviour, since it relies on the weight

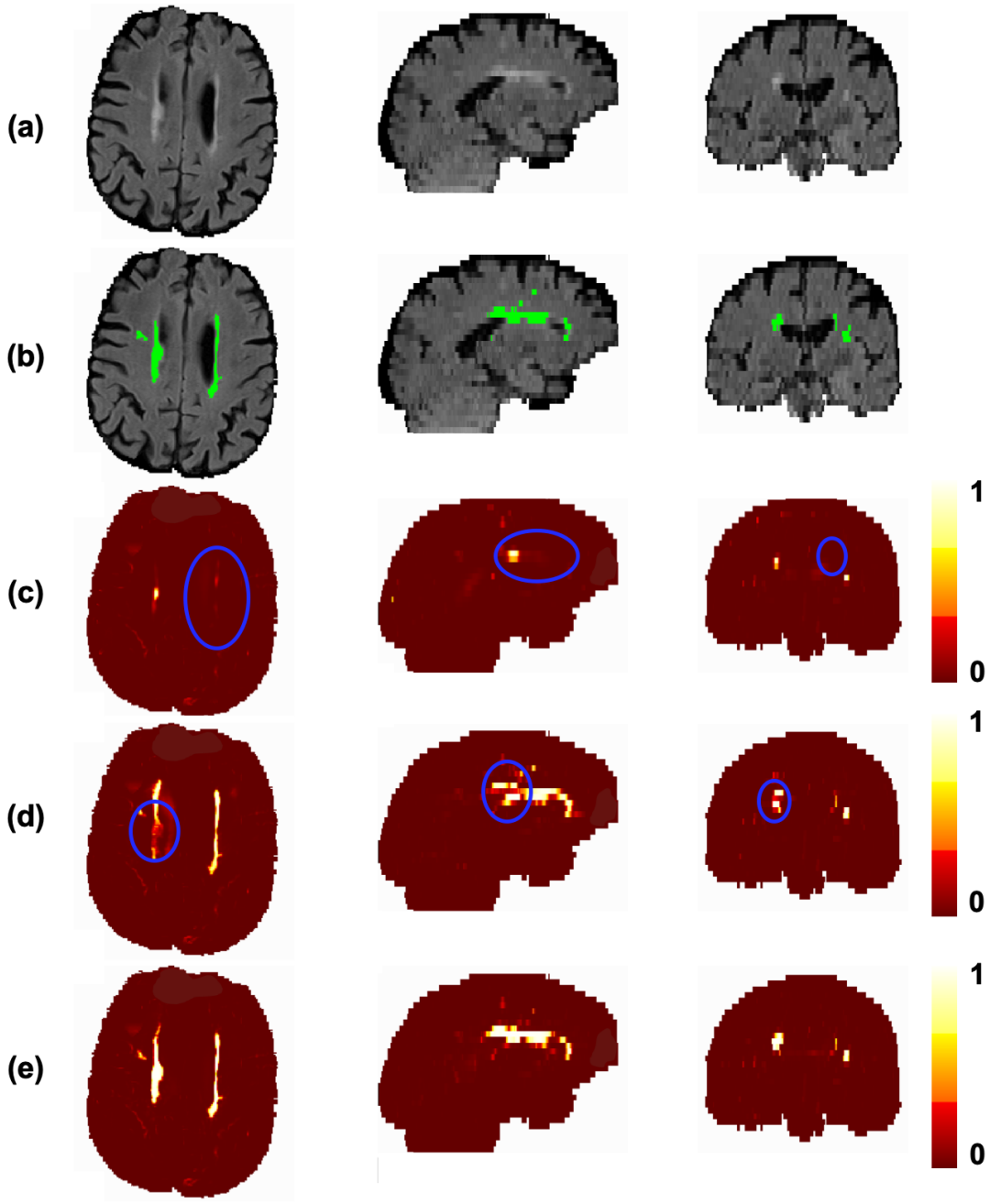


Figure 5.8: Effect of model dimension on the segmentation in the NDGEN dataset. An example showing (b) manual segmentations against the results of (c) 3D U-Net, (d) 2D axial U-Net and (e) TrUE-Net. Differences in lesion segmentations indicated by blue circles. The colour bar indicates the WMH probability value at each voxel.

maps (described in section 5.3.2) prepared with prior anatomical information. Therefore, DWMHs were given weights that were relatively similar to the ones given to PWMHs, detecting more DWMHs and avoiding over-segmentation in periventricular

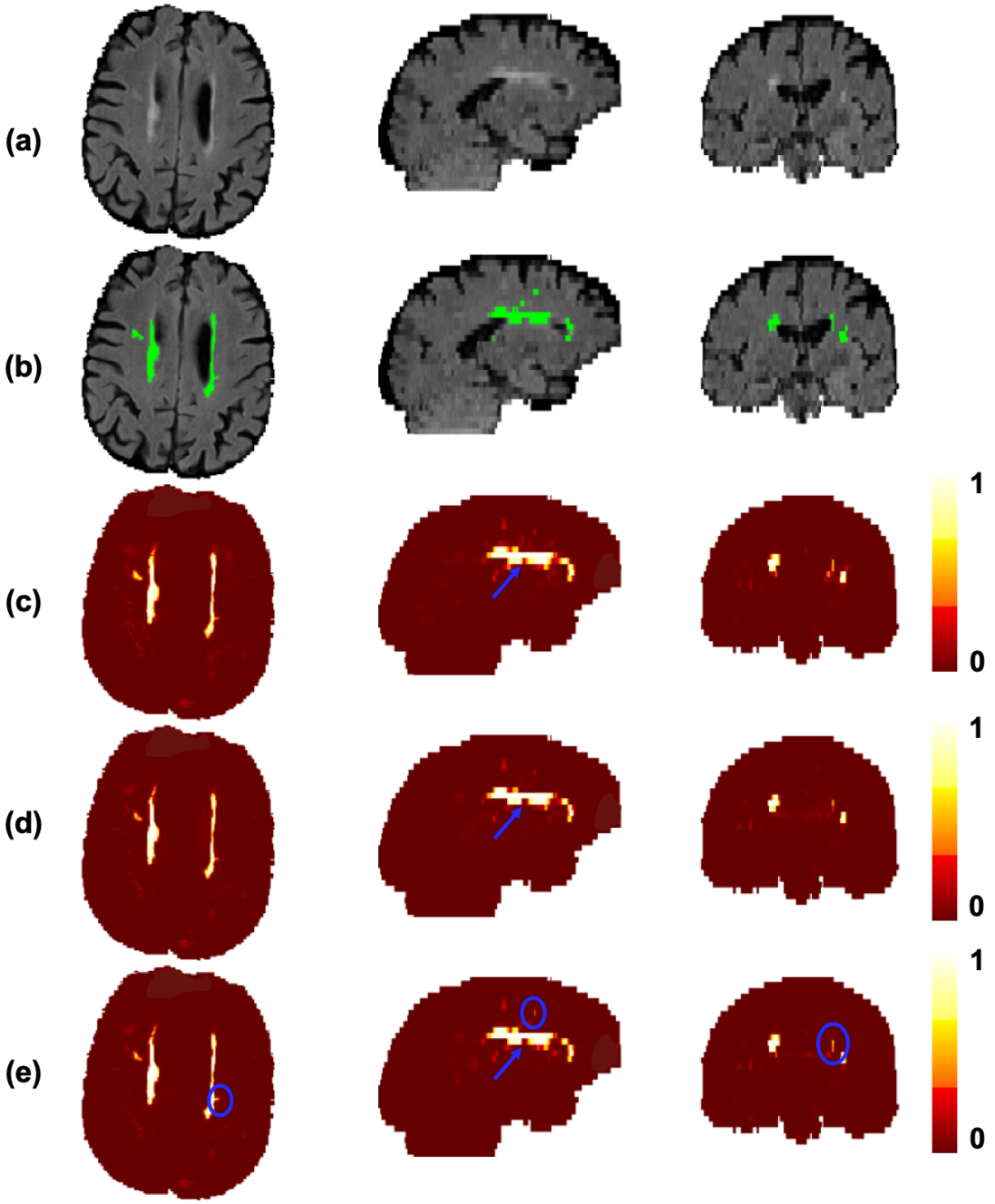


Figure 5.9: Effect of loss functions on the segmentation performance in the NDGEN dataset. An example showing (a) FLAIR images, (b) manual segmentations against the results of (c) cross-entropy (CE) loss, (d) weighted cross-entropy loss and (e) weighted cross-entropy loss + Dice loss. Differences in lesion segmentations indicated by blue circles and arrows. The colour bar indicates the WMH probability value at each voxel.

regions (figure 5.9d). In general, we have a large imbalance between foreground and background and aim to focus on detecting more true WMHs. Since missing the small WMHs would make more difference to the Dice than to CE, our expectation would

be that training the network with the Dice loss function is going to favour finding small lesions and reducing false negatives. Therefore, combining the advantages of both, adding the Dice loss to weighted CE loss function provided better segmentation along PWMH borders and detection of faint, small DWMHs (figure 5.9). The effect was particularly observable in low lesion load subjects, where the addition of the Dice loss provided more precise segmentation, with fewer false positives. Figure 5.10 shows the effect of lesion load on SI (Dice) values of model segmentation using different loss functions. Both loss functions showed positive correlation with lesion loads (weighted CE: $\rho = 0.51$, p=0.02; weighted CE + Dice loss: $\rho = 0.33$, p=0.16). However, the combination of weighted CE and Dice loss showed less correlation compared to the weighted CE loss alone (though the difference is not significant, p = 0.46), and provided higher SI values for low lesion load subjects.

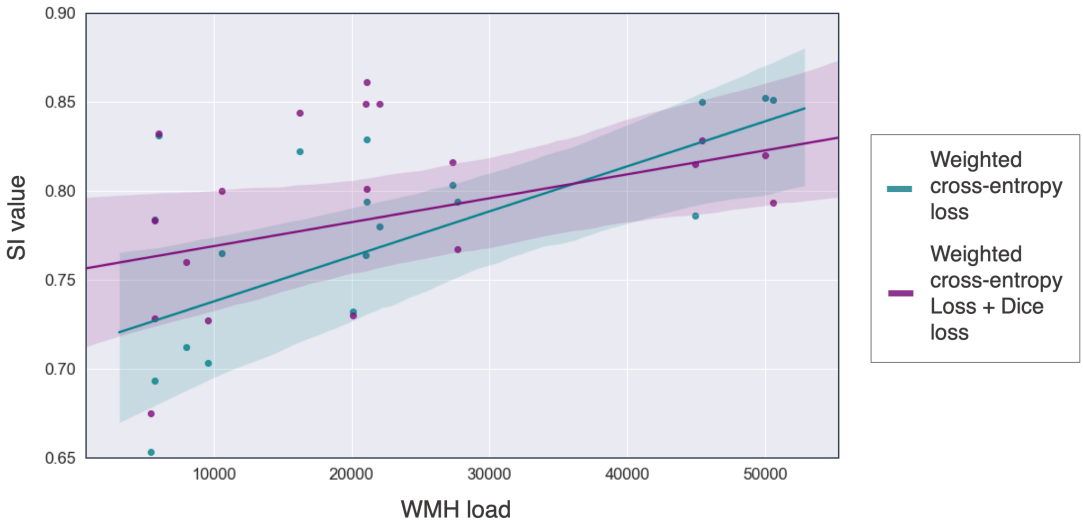


Figure 5.10: Regression plot of SI values with respect to lesion loads for weighted cross-entropy loss (blue) and weighted cross-entropy loss + Dice loss (magenta), on the NDGEN dataset.

5.6.3 Evaluation of WMH segmentation performance

The results of leave-one-out (LOO) evaluations on the MWSC cohorts, NDGEN and OXVASC datasets are illustrated by a few examples shown in figures 5.11 and 5.12. In both high and low lesion load cases, TrUE-Net provided an accurate segmentation with respect to manual segmentation without detecting many false positives. In

particular, TrUE-Net detected the deep lesions in the low lesion load cases, including subtle ones (figure 5.11b and 5.12a). It is also worth noting that there is often a possibility of inconsistency in the manual segmentation, at times even resulting in faulty delineation of lesions. For instance, in a high lesion load example from the Utrecht cohort in figure 5.11, the posterior ventricles were included in the manual segmentation (indicated by a circle), while TrUE-Net successfully excluded the ventricles by providing lower probability values in those regions.

Figure 5.13 shows the boxplots of performance metrics for the LOO validation on the MWSC cohorts, NDGEN and OXVASC datasets. Table 5.2 shows the corresponding performance values. TrUE-Net achieved its best performance on the Singapore cohort, providing the best SI (0.81 ± 0.09), AVD (16.1 ± 15.2) and voxel-wise TPR values (0.78 ± 0.10), while it provided the best cluster-wise TPR (0.74 ± 0.12) on the NDGEN dataset. The model achieved the lowest FPR for the Amsterdam cohort, resulting in a higher cluster-wise F1-measure value, thereby achieving higher sensitivity and precision.

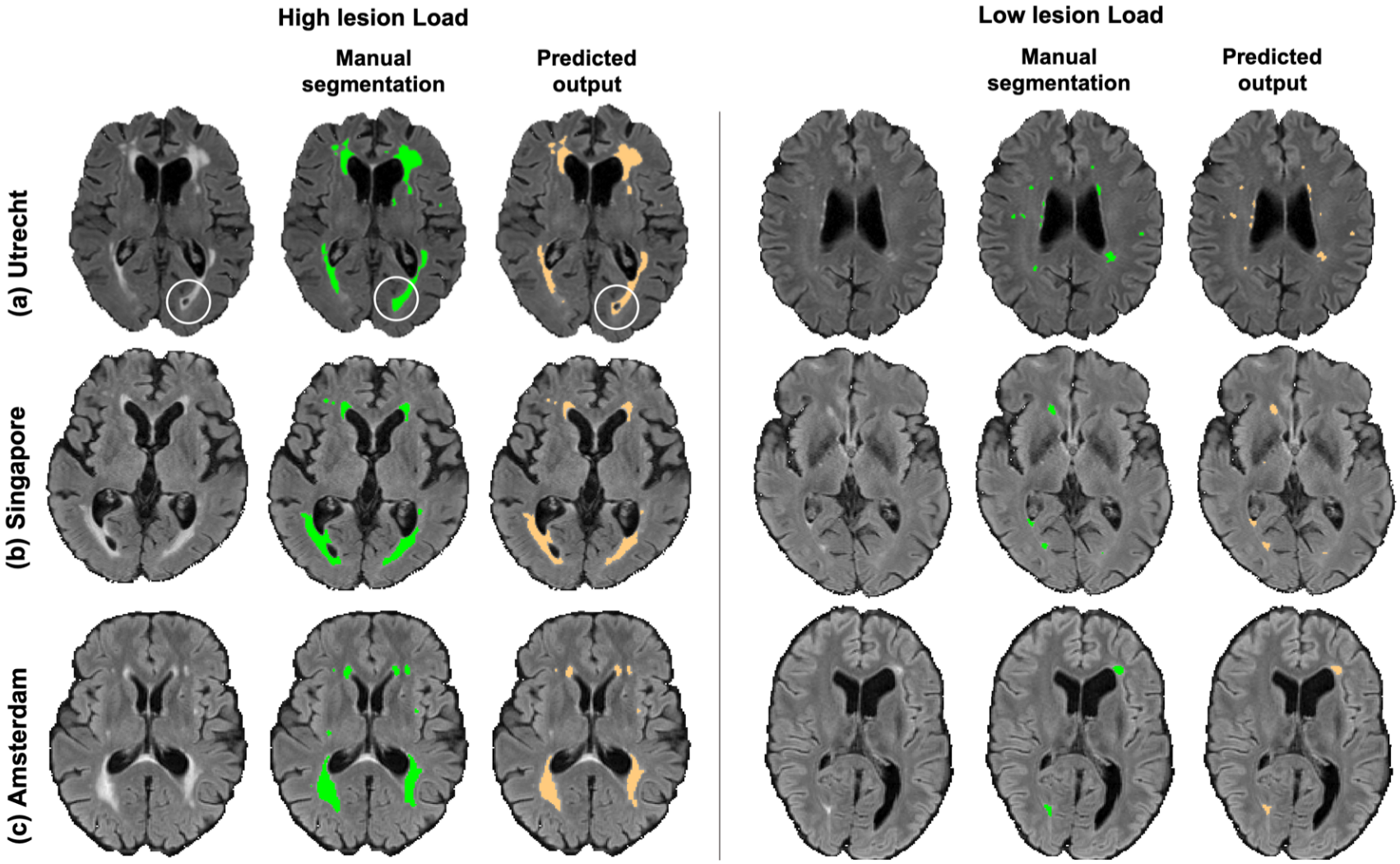


Figure 5.11: Sample results of TrUE-Net segmentation on the (a) Utrecht, (b) Singapore and (c) Amsterdam cohorts of the MWSC dataset for high and low lesion load subjects. Green indicates the manual segmentations and pale orange indicates the predicted outputs.

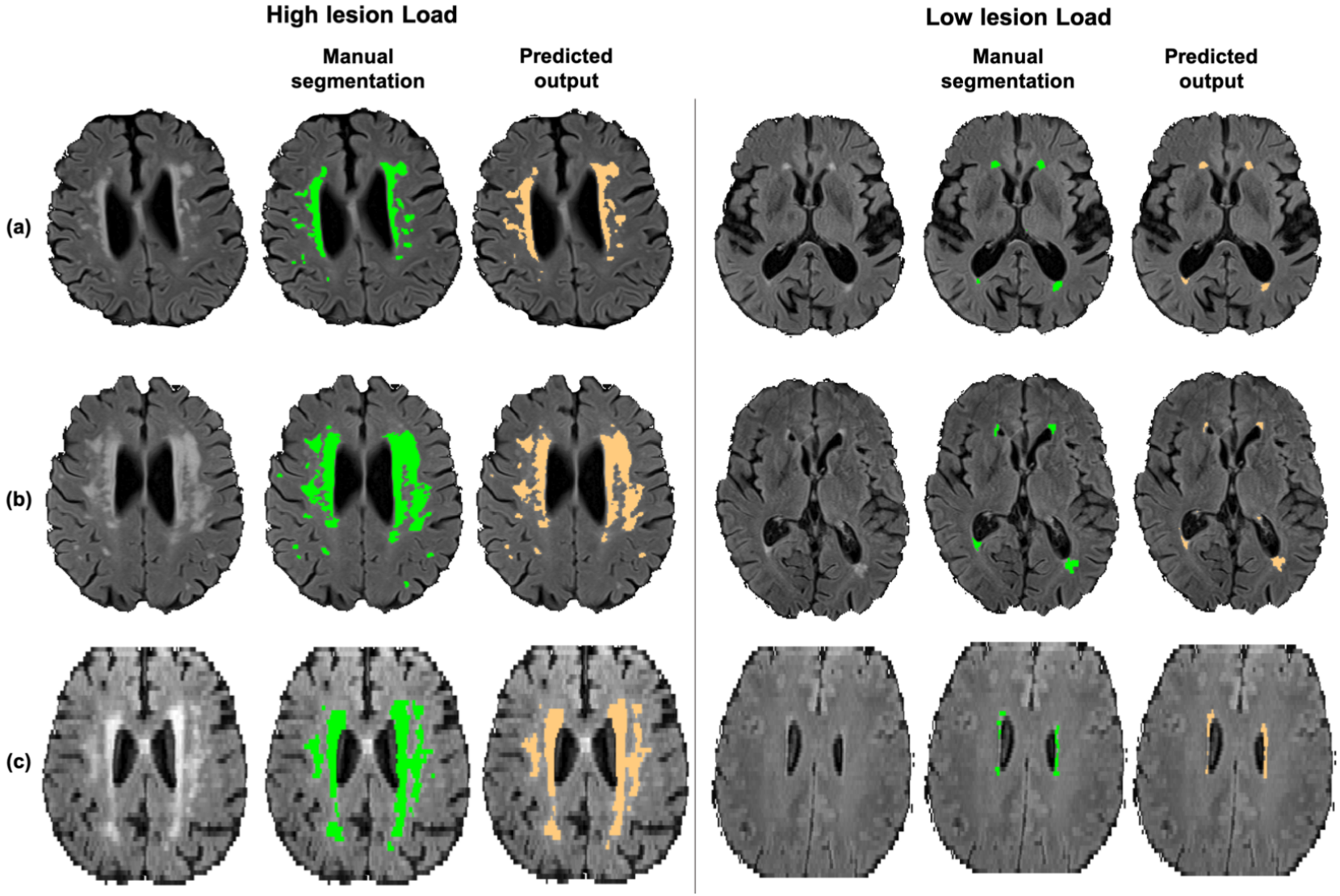


Figure 5.12: Sample results of TrUE-Net segmentation on the (a, b) NDGEN and (c) OXVASC datasets for high and low lesion load subjects. Green indicates the manual segmentations and pale orange indicates the predicted outputs.

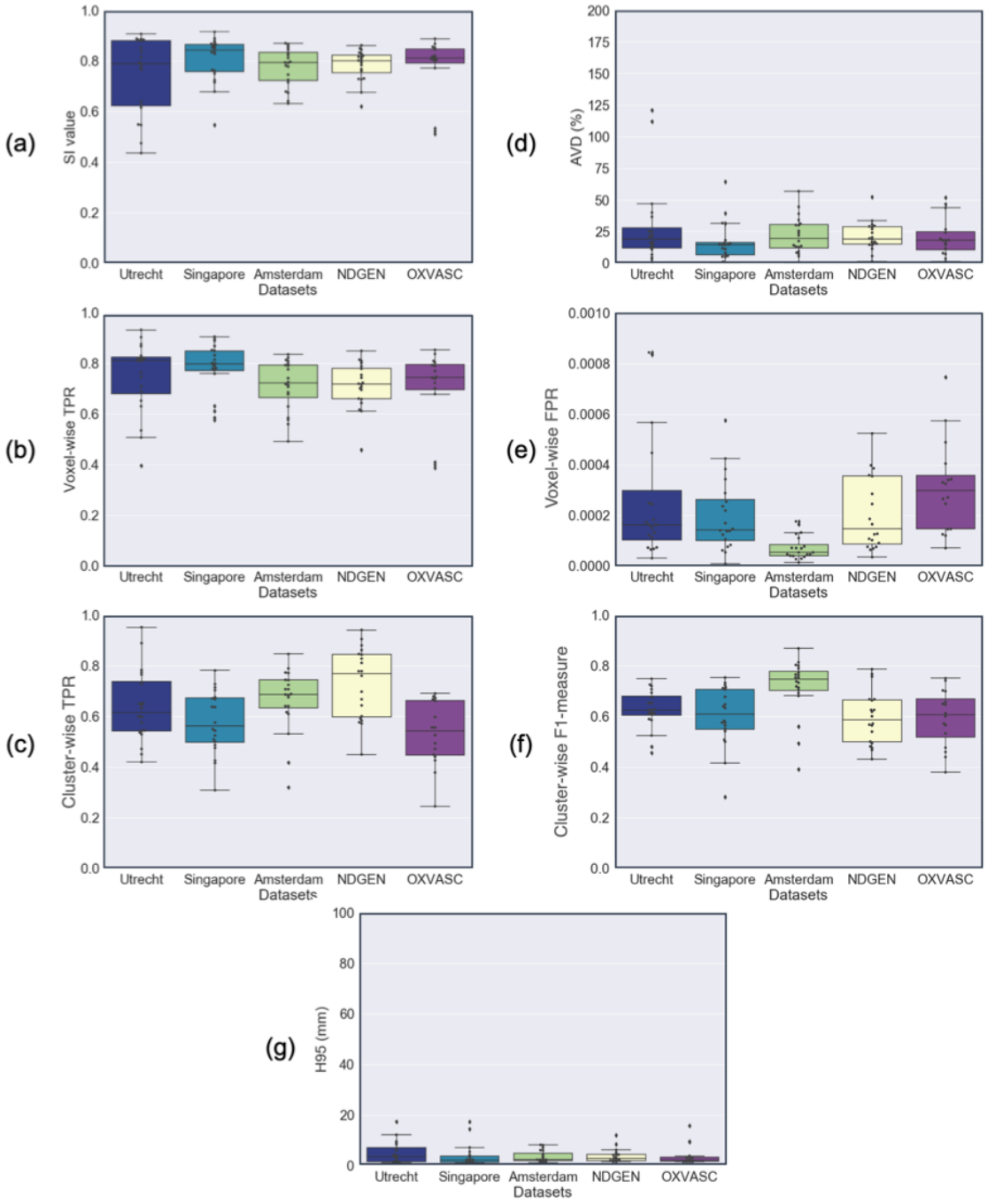


Figure 5.13: Boxplots of performance measures obtained from LOO evaluation on the Utrecht, Singapore, Amsterdam, NDGEN and OXVASC datasets - (a) SI value, (b) voxel-wise TPR, (c) cluster-wise TPR, (d) Absolute volume difference (AVD), (e) voxel-wise FPR, (f) cluster-wise F1-measure and (g) 95^{th} percentile of Hausdorff distance.

Perf. metrics	**Utrecht**	**Singapore**	**Amsterdam**	**NDGEN**	**OXVASC**
SI	0.74 ± 0.15	**0.81 ± 0.09**	0.78 ± 0.08	0.78 ± 0.06	0.78 ± 0.11
AVD (%)	28.4 ± 32.4	**16.1 ± 15.2**	21.9 ± 14.5	21.1 ± 11.2	20.4 ± 15.1
Voxel-wise TPR	0.75 ± 0.14	**0.78 ± 0.10**	0.71 ± 0.1	0.72 ± 0.10	0.70 ± 0.15
Voxel-wise FPR	3.8×10^{-4}	2.0×10^{-4}	$\mathbf{0.7 \times 10^{-4}}$	3.5×10^{-4}	3.1×10^{-4}
Cluster-wise TPR	0.64 ± 0.14	0.58 ± 0.12	0.67 ± 0.13	**0.74 ± 0.12**	0.54 ± 0.13
Cluster-wise F1 measure	0.63 ± 0.08	0.60 ± 0.12	**0.72 ± 0.11**	0.60 ± 0.10	0.60 ± 0.11
H95 (mm)	4.9 ± 4.4	3.7 ± 4.4	3.5 ± 2.3	**3.4 ± 1.8**	3.9 ± 4.0

Table 5.2: LOO evaluation of TrUE-Net on the Utrecht, Singapore and Amsterdam cohorts of the MWSC, NDGEN and OXVASC datasets (the best value for each measure across the datasets is highlighted in bold).

5.6.3.1 Performance in deep and periventricular lesions

Figure 5.14 shows boxplots of the performance metrics for DWMHs and PWMHs for the experimented datasets. Table 5.3 reports the corresponding descriptive statistics of the performance measures in the deep and periventricular regions, along with the p-values of the paired t-tests performed between DWMHs and PWMHs. The main observations of the analysis were:

- Even though SI values in the periventricular region were higher in all datasets, the differences were significant only for the Singapore cohort, NDGEN and OXVASC datasets.
- AVD values were generally slightly higher for DWMHs when compared to PWMHs for all datasets (and the value was significantly higher for OXVASC). This is because of lower lesion volumes and smaller lesions in the deep region, when compared to the periventricular region. For instance, AVD values of 25% in the deep region corresponds to a lesion volume of around 600 mm^3 while the same value corresponds to a much higher value of around 3800 mm^3 in the periventricular region.

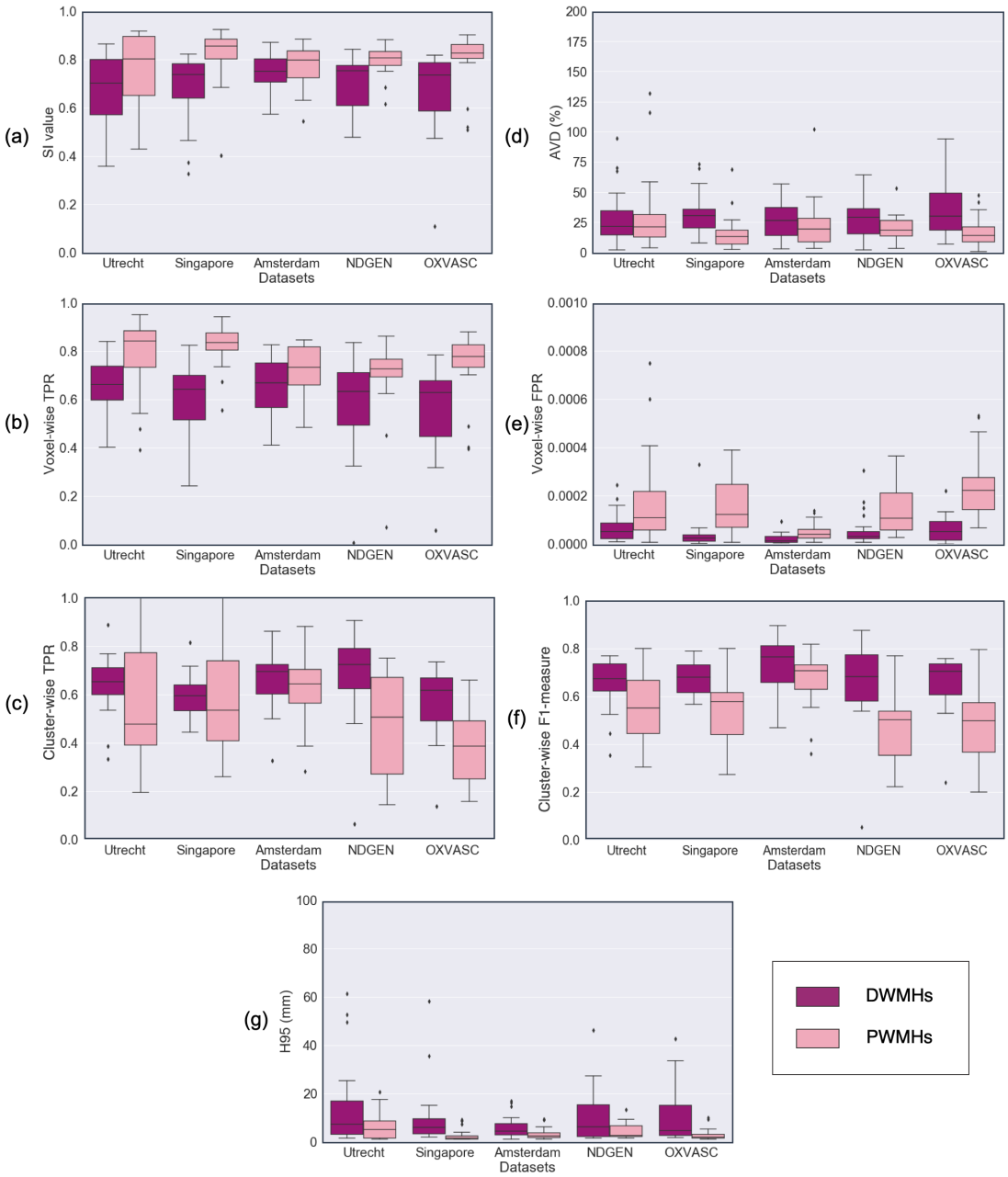


Figure 5.14: Boxplots of performance measures obtained in the deep and the periventricular regions for LOO evaluation on the Utrecht, Singapore, Amsterdam, NDGEN and OXVASC datasets - a) SI value, (b) voxel-wise TPR, (c) cluster-wise TPR, (d) Absolute volume difference (AVD), (e) voxel-wise FPR, (f) cluster-wise F1-measure and (g) 95^{th} percentile of Hausdorff distance.

- Voxel-wise TPR values were higher in the periventricular regions, when compared to the deep regions, in all datasets. This difference was significant in the Utrecht, Singapore and OXVASC datasets.

- Similarly, voxel-wise FPR values were also higher in the periventricular regions

when compared to the deep regions, with the difference significant for the Singapore, Amsterdam and OXVASC datasets.

- Interestingly, the difference in cluster-wise TPR values between DWMHs and PWMHs did not show a consistent pattern across the datasets. For the NDGEN and OXVASC datasets, cluster-wise TPR values were significantly higher in the deep regions than those in the periventricular regions, indicating that fewer lesions were missed in the deep regions. In general, the overall values were comparable for the deep and the periventricular regions. However, the range is wider for the periventricular regions compared to the deep regions. This could be due to heterogeneity in the segmentation performance resulting from inconsistencies in manual segmentation (e.g. inclusion/exclusion of ventricle lining) of WMHs around ventricles.

- Cluster-wise F1 measures were significantly higher in the deep regions when compared to the periventricular regions for all datasets, indicating that both cluster-wise TPR and precision of WMH segmentation were higher in the deep regions.

- The 95^{th} percentile of Hausdorff distance values were higher in the deep regions, when compared to the periventricular regions, the difference is significant for all datasets except the Utrecht dataset.

Overall, we observed that the voxel-wise measures were higher in the periventricular region, while the cluster-wise/lesion-level measures and SI values were higher in the deep regions. This shows that the model provides better delineation of PWMHs, while reducing the number of false positives in the deep regions.

5.6.4 Comparison with BIANCA

Figure 5.15 illustrates a few example cases, ordered according to decreasing lesion load, showing segmentations obtained with TrUE-Net, TrUE-Net-Att and BIANCA, with respect to manual segmentation.

		Utrecht	Singapore	Amsterdam	NDGEN	OXVASC
SI	DWMHs	0.68 ± 0.15	0.68 ± 0.15	0.75 ± 0.08	0.70 ± 0.10	0.66 ± 0.18
	PWMHs	0.76 ± 0.15	0.82 ± 0.12	0.70 ± 0.09	0.80 ± 0.06	0.79 ± 0.13
	p-value	0.13	**0.002**	0.40	**0.001**	**0.02**
AVD (%)	DWMHs	29.5 ± 24.1	36.4 ± 18.1	26.9 ± 14.9	28.7 ± 15.9	36.05 ± 23.8
	PWMHs	32.6 ± 34.4	25.1 ± 45.1	22.9 ± 22.0	20.0 ± 11.2	17.5 ± 13.6
	p-value	0.75	0.51	0.49	0.5	**0.01**
Voxel-wise TPR	DWMHs	0.65 ± 0.13	0.59 ± 0.15	0.65 ± 0.12	0.58 ± 0.18	0.55 ± 0.19
	PWMHs	0.77 ± 0.15	0.82 ± 0.09	0.72 ± 0.10	0.69 ± 0.17	0.72 ± 0.15
	p-value	**0.01**	**<0.001**	0.08	0.06	**0.01**
Voxel-wise FPR	DWMHs	6.6×10^{-5}	4.1×10^{-5}	1.9×10^{-5}	5.9×10^{-5}	6.2×10^{-5}
	PWMHs	3.1×10^{-4}	1.6×10^{-4}	0.5×10^{-4}	2.3×10^{-4}	2.4×10^{-4}
	p-value	0.09	**<0.001**	**0.004**	0.09	**<0.001**
Cluster-wise TPR	DWMHs	0.64 ± 0.12	0.60 ± 0.09	0.65 ± 0.12	0.69 ± 0.15	0.57 ± 0.15
	PWMHs	0.55 ± 0.24	0.56 ± 0.21	0.63 ± 0.15	0.48 ± 0.20	0.39 ± 0.16
	p-value	0.16	0.57	0.48	**0.002**	**0.003**
Cluster-wise F1-measure	DWMHs	0.65 ± 0.11	0.68 ± 0.11	0.74 ± 0.17	0.66 ± 0.17	0.68 ± 0.13
	PWMHs	0.55 ± 0.15	0.53 ± 0.15	0.66 ± 0.11	0.46 ± 0.15	0.47 ± 0.16
	p-value	**0.02**	**<0.001**	**0.04**	**<0.001**	**0.002**
H95 (mm)	DWMHs	14.7 ± 18.3	10.1 ± 13.5	5.9 ± 4.9	10.5 ± 11.6	11.3 ± 12.17
	PWMHs	6.1 ± 5.5	2.6 ± 2.6	3.3 ± 2.4	4.2 ± 3.2	3.0 ± 2.8
	p-value	0.05	**0.02**	**0.03**	**0.02**	**0.01**

Table 5.3: Comparison of performance measures between PWMHs and DWMHs, along with p-values of two-tailed paired t-tests on the Utrecht, Singapore and Amsterdam cohorts of MWSC, NDGEN and OXVASC datasets (the best performance values and significant p-values highlighted in bold).

TrUE-Net provided more accurate segmentations when compared to BIANCA, especially in the low lesion load subjects. From the figure, it can be observed that, as the lesion load decreases, BIANCA detected more false positive regions, particularly

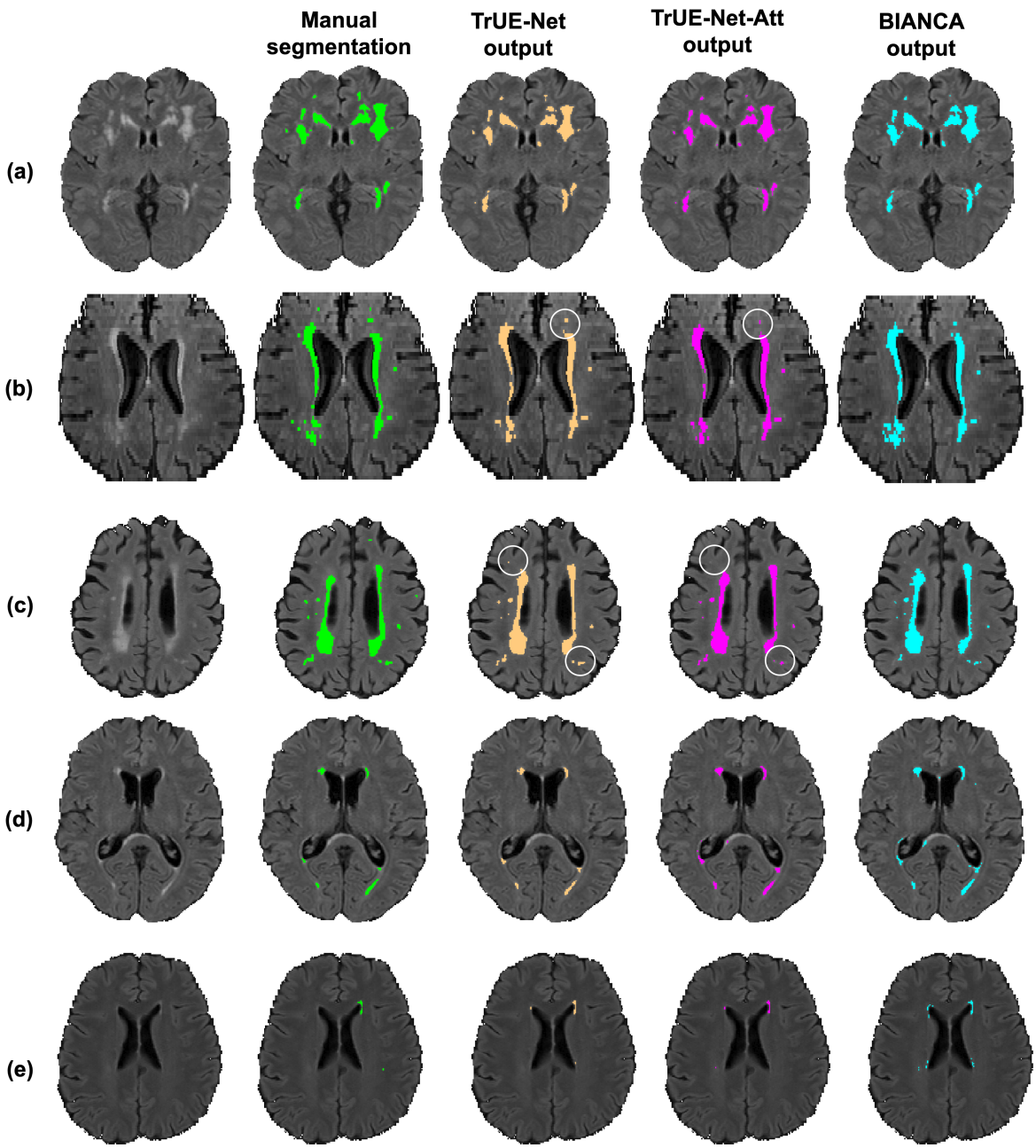


Figure 5.15: Sample results for comparison of TrUE-Net segmentation with those of TrUE-Net-Att and BIANCA. Manual segmentations (green) shown along with TrUE-Net (pale orange), TrUE-Net-Att (magenta) and BIANCA (cyan). Top to bottom: decreasing order of lesion load from the Utrecht (a,d), Singapore (c, e) and OXVASC (b) datasets.

around the ventricles (shown in figure 5.15a, d and e). Moreover, BIANCA oversegmented the lesions in the periventricular regions (figure 5.15b and c). TrUE-Net and TrUE-Net-Att provided highly similar results with very subtle differences. In high le-

sion load subjects, TrUE-Net-Att provided a slight undersegmentation of deep lesions (circled regions in figure 5.15b and c), while as lesion load decreases TrUE-Net-Att provided more precise segmentation.

Figure 5.16 shows the boxplots comparing the performance measures between TrUE-Net, TrUE-Net-Att and BIANCA. The corresponding performance values are reported in table 5.4. The results of paired t-tests between the individual pairs of methods (TrUE-Net, TrUE-Net-Att, BIANCA) are shown in table 5.5.

TrUE-Net provided the best SI values for all the datasets when compared to BIANCA (although none of the differences are significant). The AVD values of TrUE-Net were lower when compared to BIANCA but the difference was significant only on the OXVASC dataset. Even though BIANCA showed higher TPR values for most datasets, none of the differences were significant. Voxel-wise FPR values of TrUE-Net were lower compared to BIANCA for all datasets, with significant differences (both $\approx 85\text{mm}^3$) in the Singapore and Amsterdam datasets. While the cluster-wise TPR values were higher for BIANCA in the Singapore, Amsterdam and NDGEN datasets, the difference was significant only for the Singapore dataset. The higher values of both TPR and FPR values indicates that while BIANCA detected more WMHs, it also detected more false positives, thus leading to lower SI values, when compared to the TrUE-Net model. This was more evident by higher cluster-wise F1-measures achieved by TrUE-Net when compared to BIANCA in all datasets (all of them significant except the OXVASC dataset). Also, TrUE-Net provided the best Hausdorff distance values, while the distance values were higher for BIANCA (especially in the Utrecht and OXVASC datasets), indicating oversegmentation of WMHs.

Addition of AGs did not provide much improvement over the TrUE-Net model, and the performance was comparable in most cases. From tables 5.4 and 5.5, it can be seen that while there were significant differences between the TrUE-Net models and BIANCA, none of the performance values showed significant differences between TrUE-Net and TrUE-Net-Att.

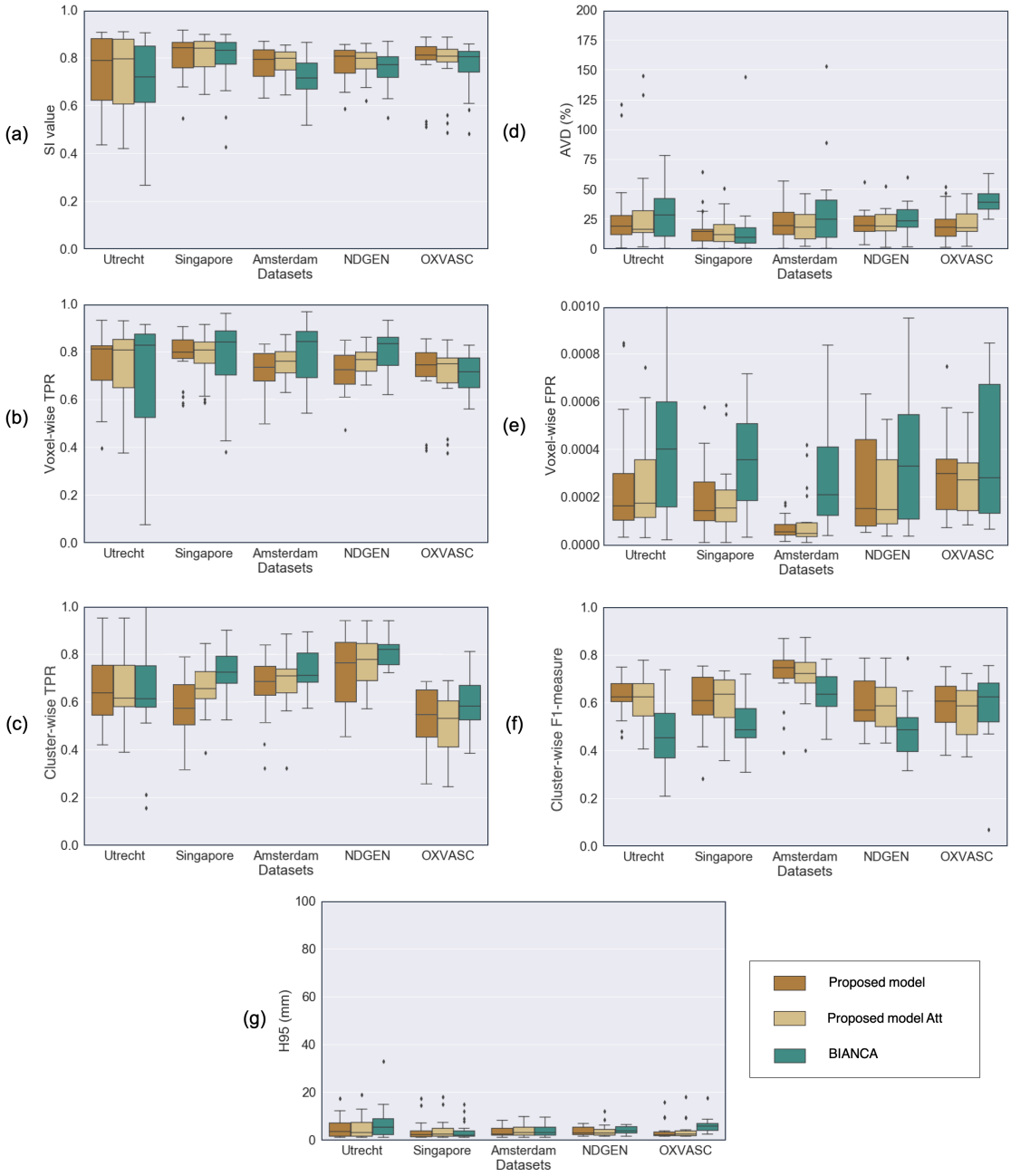


Figure 5.16: Boxplots of performance measures obtained for TrUE-Net, TrUE-Net-Att and BIANCA on the Utrecht, Singapore, Amsterdam, NDGEN and OXVASC datasets - a) SI value, (b) voxel-wise TPR, (c) cluster-wise TPR, (d) Absolute volume difference (AVD), (e) voxel-wise FPR, (f) cluster-wise F1-measure and (g) 95^{th} percentile of Hausdorff distance.

Figure 5.17 shows the regression plot of SI values against lesion loads of the datasets. BIANCA showed a positive correlation with respect to lesion loads (ρ = 0.57, p <0.001) and provided lower SI values in low lesion load subjects, due to

		Utrecht	Singapore	Amsterdam	NDGEN	OXVASC
SI	TrUE-Net	**0.74 ± 0.15**	**0.81 ± 0.09**	**0.78 ± 0.08**	**0.78 ± 0.06**	**0.78 ± 0.11**
	TrUE-Net-Att	0.74 ± 0.16	0.81 ± 0.07	0.77 ± 0.07	0.78 ± 0.06	0.76 ± 0.12
	BIANCA	0.63 ± 0.22	0.74 ± 0.14	0.70 ± 0.09	0.77 ± 0.08	0.74 ± 0.11
AVD (%)	TrUE-Net	**28.4 ± 32.4**	16.4 ± 15.2	21.9 ± 14.5	**21.1 ± 11.2**	**20.4 ± 15.1**
	TrUE-Net-Att	31.6 ± 39.0	**15.4 ± 13.8**	**20.3 ± 12.8**	21.1 ± 11.8	21.7 ± 13.8
	BIANCA	60.7 ± 59.9	17.7 ± 30.9	32.3 ± 35.7	25.0 ± 13.1	40.8 ± 10.3
Voxel-wise TPR	TrUE-Net	**0.75 ± 0.14**	**0.78 ± 0.10**	0.71 ± 0.10	0.72 ± 0.10	0.70 ± 0.15
	TrUE-Net-Att	0.74 ± 0.15	0.78 ± 0.10	0.75 ± 0.06	0.76 ± 0.05	0.69 ± 0.15
	BIANCA	0.67 ± 0.26	0.76 ± 0.19	**0.80 ± 0.12**	**0.80 ± 0.08**	**0.71 ± 0.08**
Voxel-wise FPR	TrUE-Net	$\mathbf{3.8 \times 10^{-4}}$	2.0×10^{-4}	$\mathbf{0.7 \times 10^{-4}}$	3.5×10^{-4}	3.1×10^{-4}
	TrUE-Net-Att	3.9×10^{-4}	$\mathbf{1.9 \times 10^{-4}}$	0.9×10^{-4}	$\mathbf{3.1 \times 10^{-4}}$	$\mathbf{2.8 \times 10^{-4}}$
	BIANCA	4.3×10^{-4}	3.6×10^{-4}	3.5×10^{-4}	4.6×10^{-4}	5.3×10^{-4}
Cluster-wise TPR	TrUE-Net	0.64 ± 0.14	0.58 ± 0.12	0.67 ± 0.13	0.74 ± 0.12	0.54 ± 0.13
	TrUE-Net-Att	**0.66 ± 0.15**	0.65 ± 0.10	0.68 ± 0.11	0.76 ± 0.11	0.50 ± 0.13
	BIANCA	0.63 ± 0.20	**0.73 ± 0.09**	**0.73 ± 0.08**	**0.80 ± 0.06**	**0.59 ± 0.13**
Cluster-wise F1-measure	TrUE-Net	**0.63 ± 0.08**	**0.60 ± 0.12**	**0.72 ± 0.11**	**0.60 ± 0.10**	**0.60 ± 0.11**
	TrUE-Net-Att	0.62 ± 0.10	0.60 ± 0.11	0.71 ± 0.09	0.59 ± 0.10	0.56 ± 0.11
	BIANCA	0.46 ± 0.15	0.51 ± 0.10	0.63 ± 0.09	0.51 ± 0.12	0.58 ± 0.06
H95 (mm)	TrUE-Net	**4.9 ± 4.4**	**3.7 ± 4.4**	**3.5 ± 2.3**	**3.4 ± 1.8**	**3.9 ± 4.0**
	TrUE-Net-Att	5.0 ± 4.8	4.0 ± 4.6	3.8 ± 2.4	3.7 ± 2.6	4.2 ± 4.4
	BIANCA	7.1 ± 7.3	3.8 ± 3.9	3.8 ± 2.6	3.7 ± 1.5	6.1 ± 3.6

Table 5.4: Comparison of TrUE-Net performance with TrUE-Net-Att and BIANCA on the Utrecht, Singapore and Amsterdam cohorts of the MWSC, NDGEN and OXVASC datasets (the best performance values highlighted in bold).

the increased false positives. TrUE-Net and TrUE-Net-Att provided slightly lower positive correlations (TrUE-Net: $\rho = 0.45$, p<0.001 and TrUE-Net-Att: $\rho = 0.44$, p-value<0.001) with respect to lesion loads. None of the correlations were significantly

		Utrecht	Singapore	Amsterdam	NDGEN	OXVASC
SI	TrUE-Net vs TrUE-Net-Att	0.90	0.82	0.97	0.90	0.80
	TrUE-Net vs BIANCA	0.28	0.56	0.06	0.39	0.65
	TrUE-Net-Att vs BIANCA	0.34	0.42	**0.04**	0.30	0.86
AVD (%)	TrUE-Net vs TrUE-Net-Att	0.77	0.83	0.73	0.93	0.80
	TrUE-Net vs BIANCA	0.21	0.86	0.23	0.29	**<0.001**
	TrUE-Net-Att vs BIANCA	0.27	0.76	0.17	0.31	**<0.001**
Voxel-wise TPR	TrUE-Net vs TrUE-Net-Att	0.91	0.99	0.82	0.48	0.85
	TrUE-Net vs BIANCA	0.29	0.74	0.15	0.17	0.80
	TrUE-Net-Att vs BIANCA	0.33	0.74	0.18	0.06	0.62
Voxel-wise FPR	TrUE-Net vs TrUE-Net-Att	0.97	0.84	0.30	0.81	0.65
	TrUE-Net vs BIANCA	0.78	**0.006**	**0.001**	0.54	0.23
	TrUE-Net-Att vs BIANCA	0.84	**0.004**	**0.006**	0.40	0.18
Cluster-wise TPR	TrUE-Net vs TrUE-Net-Att	0.90	0.94	0.73	0.72	0.06
	TrUE-Net vs BIANCA	0.76	**0.03**	0.14	0.08	0.96
	TrUE-Net-Att vs BIANCA	0.69	**0.02**	0.07	0.12	0.05
Cluster-wise F1-measure	TrUE-Net vs TrUE-Net-Att	0.79	0.93	0.90	0.98	0.42
	TrUE-Net vs BIANCA	**<0.001**	**0.01**	**0.02**	**0.002**	0.81
	TrUE-Net-Att vs BIANCA	**<0.001**	**0.007**	**0.01**	**0.002**	0.68
H95 (mm)	TrUE-Net vs TrUE-Net-Att	0.93	0.81	0.72	0.67	0.83
	TrUE-Net vs BIANCA	0.26	0.96	0.67	0.48	0.11
	TrUE-Net-Att vs BIANCA	0.30	0.83	0.94	0.90	0.19

Table 5.5: Results of two-tailed paired t-tests (after multiple comparison correction) on the performance measures between individual pairs of methods on the Utrecht, Singapore and Amsterdam cohorts of the MWSC, NDGEN and OXVASC datasets ($p < 0.05$ highlighted in bold).

different (TrUE-Net vs BIANCA: p = 0.27, TrUE-Net-Att vs BIANCA: p = 0.23 and TrUE-Net vs TrUE-Net-Att: p = 0.93).

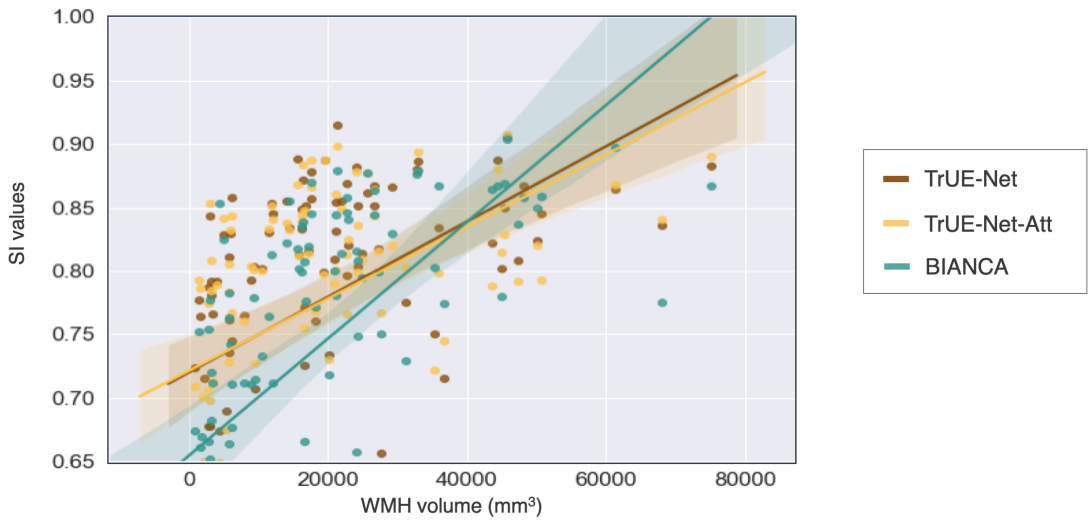


Figure 5.17: Regression plot of SI values with respect to lesion loads for TrUE-Net (brown), TrUE-Net-Att (orange) and BIANCA (green) on all datasets combined together. The shaded region represents the 95% confidence interval of the regression estimates.

5.6.5 Domain adaptation: transfer learning and domain adversarial training

Figures 5.18 and 5.19 show some sample results for the test cases of domain adaptation experiments for high and low lesion subjects from the OXVASC dataset, which is part of the target domain. Table 5.6 reports the performance values of the test cases, while figure 5.20 shows the boxplots of the measures. Table 5.7 reports the p-values of two-tailed t-tests performed between the individual pairs of test cases on the OXVASC dataset.

5.6.5.1 Case 1: Train on the source domain and apply directly to the target domain

As shown in figures 5.18a and 5.19a, the source-trained model detected most PWMHs while missing some DWMHs. However, even in the detected PWMHs, the model undersegmented the boundaries. The segmentation was more precise in the low lesion load cases. From the boxplots in figure 5.20 and the performance measures reported

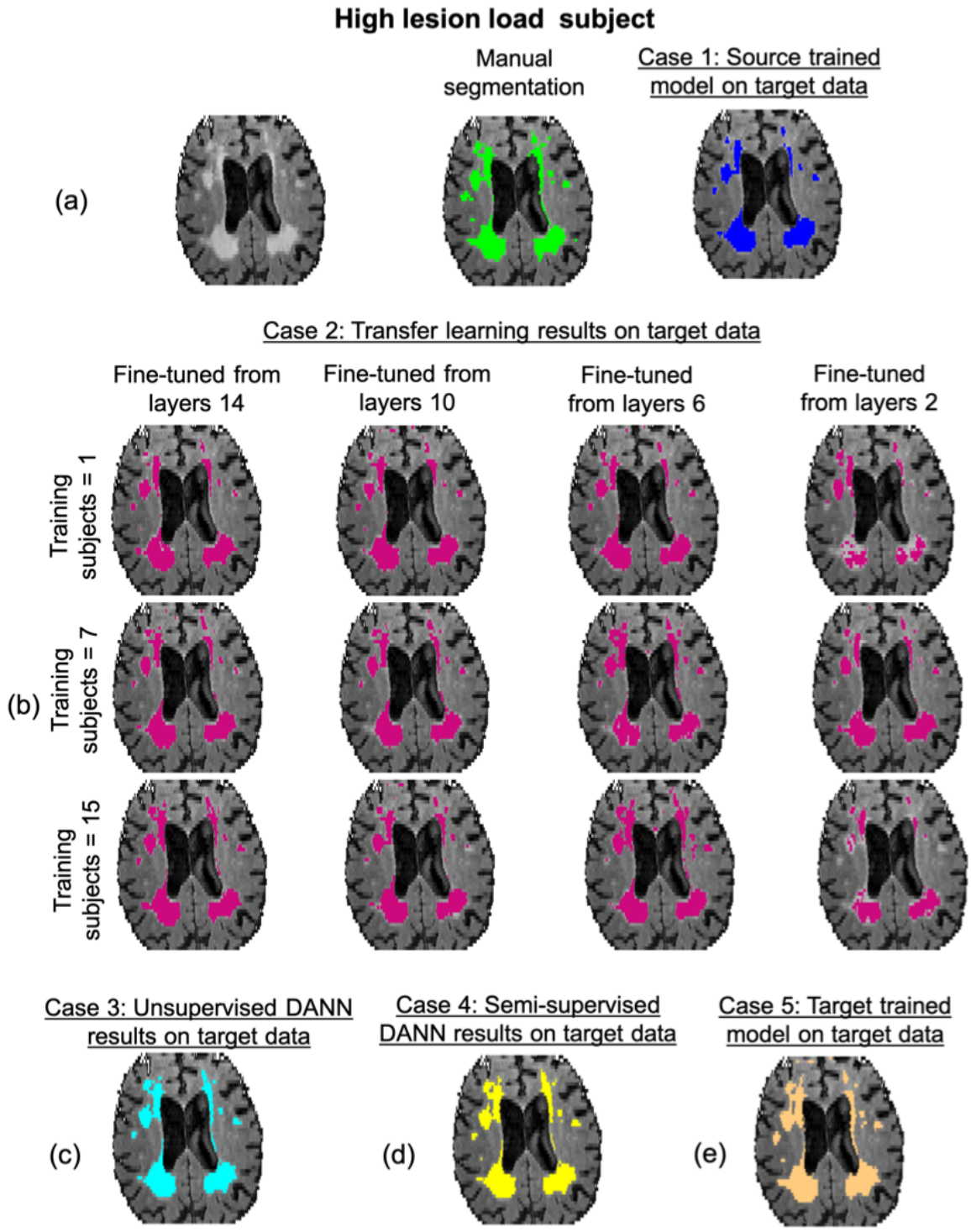


Figure 5.18: Sample results of domain adaptation experiment test cases on a high lesion load subject from the OXVASC dataset (target domain), along with manual segmentation (green).

in table 5.6, this case showed the worst performance among all the cases. This case showed significantly lower voxel-wise TPR and significantly higher FPR when

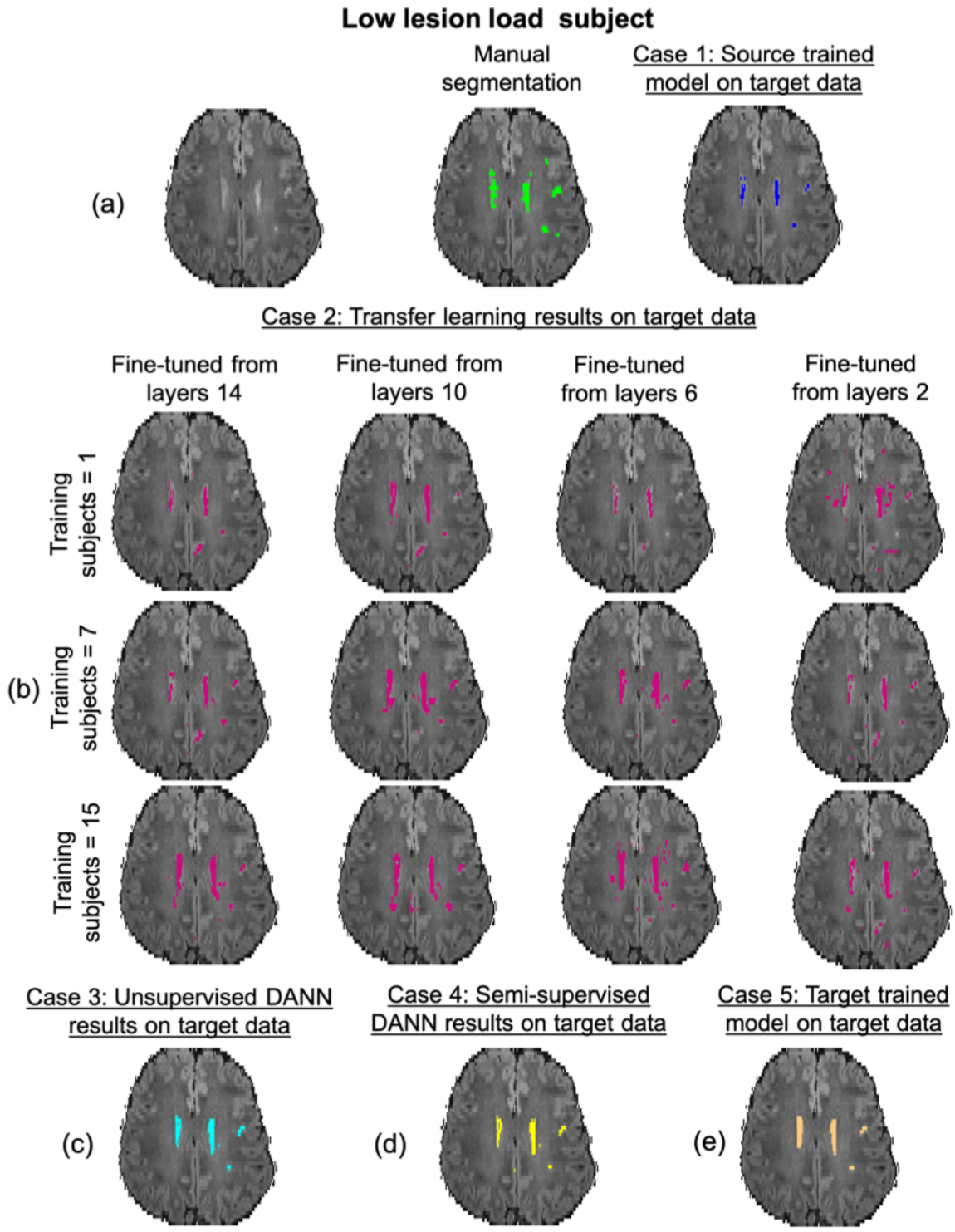


Figure 5.19: Sample results of domain adaptation experiment test cases on a low lesion load subject from the OXVASC dataset (target domain), along with manual segmentation (green).

compared to other cases (table 5.7).

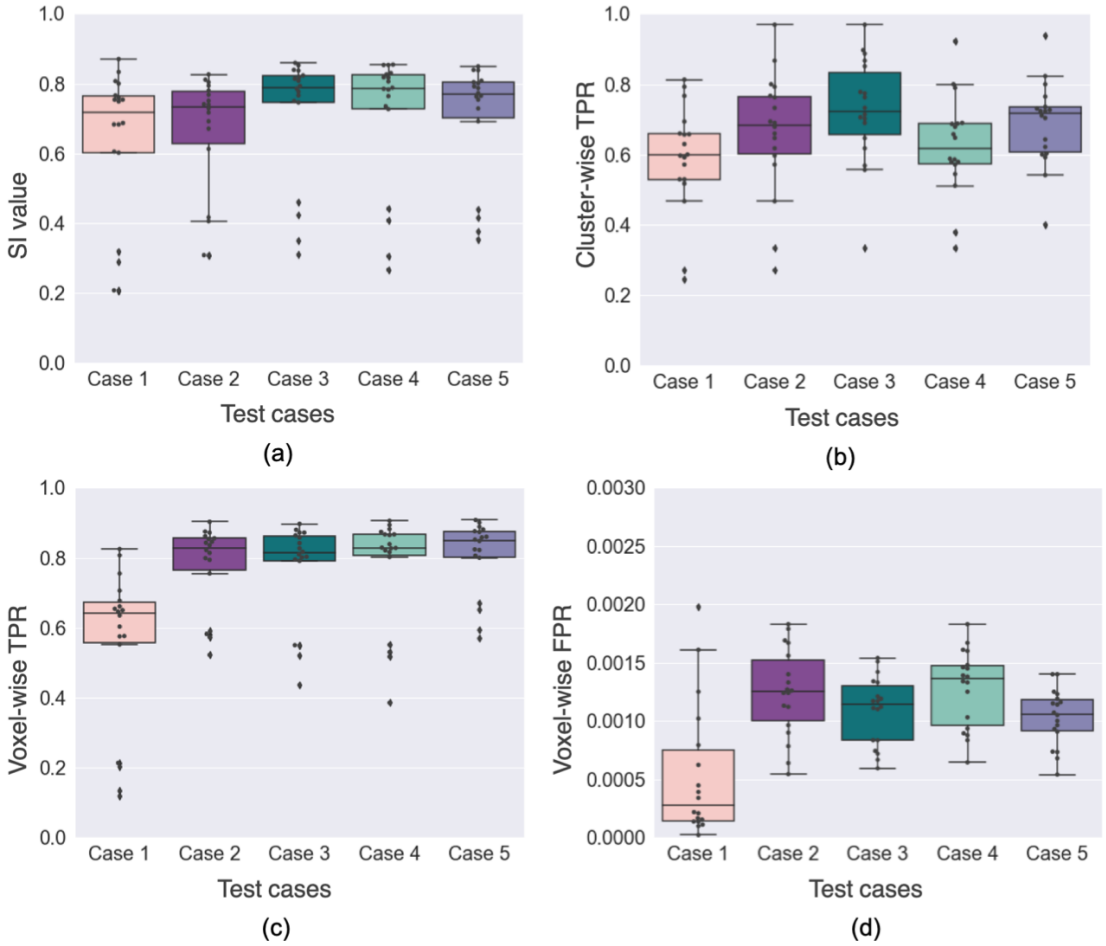


Figure 5.20: Boxplots of performance measures obtained for various test cases of the domain adaptation experiment applied to the OXVASC dataset from the target domain - (a) SI values, (b) cluster-wise TPR, (c) voxel-wise TPR and (d) voxel-wise FPR values.

Perf. metrics	**Case 1**	**Case 2**	**Case 3**	**Case 4**	**Case 5**
SI	0.63 ± 0.21	0.67 ± 0.18	0.71 ± 0.17	0.71 ± 0.19	$\mathbf{0.72 \pm 0.18}$
Voxel-wise TPR	0.55 ± 0.22	0.77 ± 0.12	0.76 ± 0.14	0.77 ± 0.15	$\mathbf{0.80 \pm 0.10}$
Voxel-wise FPR	$\mathbf{5.3 \times 10^{-4}}$	12.4×10^{-4}	6.9×10^{-4}	12.7×10^{-4}	10.4×10^{-4}
Cluster-wise TPR	0.60 ± 0.15	0.66 ± 0.17	$\mathbf{0.72 \pm 0.15}$	0.64 ± 0.14	0.69 ± 0.12

Table 5.6: Comparison of performance measures for 5 test cases on the OXVASC dataset from the target domain (the best value for each performance measure is highlighted in bold).

	Case 2	**Case 3**	**Case 4**	**Case 5**
Case 1	SI: p = 0.68 Voxel-wise TPR: **p < 0.001** Voxel-wise FPR: **p < 0.001** Cluster-wise TPR: p = 0.18	SI: p = 0.25 Voxel-wise TPR: **p = 0.002** Voxel-wise FPR: **p < 0.001** Cluster-wise TPR: **p = 0.01**	SI: p = 0.31 Voxel-wise TPR: **p = 0.002** Voxel-wise FPR: **p < 0.001** Cluster-wise TPR: p = 0.47	SI: p = 0.32 Voxel-wise TPR: **p < 0.002** Voxel-wise FPR: **p = 0.002** Cluster-wise TPR: **p = 0.03**
Case 2		SI: p = 0.39 Voxel-wise TPR: p = 0.79 Voxel-wise FPR: p = 0.18 Cluster-wise TPR: p = 0.25	SI: p = 0.49 Voxel-wise TPR: p = 0.91 Voxel-wise FPR: p = 0.76 Cluster-wise TPR: p = 0.47	SI: p = 0.50 Voxel-wise TPR: p = 0.49 Voxel-wise FPR: p = 0.05 Cluster-wise TPR: p = 0.61
Case 3			SI: p = 0.90 Voxel-wise TPR: p = 0.90 Voxel-wise FPR: p = 0.08 Cluster-wise TPR: **p = 0.04**	SI: p = 0.83 Voxel-wise TPR: p = 0.36 Voxel-wise FPR: p = 0.51 Cluster-wise TPR: p = 0.42
Case 4				SI: p = 0.94 Voxel-wise TPR: p = 0.48 Voxel-wise FPR: **p = 0.02** Cluster-wise TPR: p = 0.15

Table 5.7: Results of two-tailed paired t-test results (after multiple comparison correction) between individual combinations of test cases on the OXVASC dataset ($p < 0.05$ highlighted in bold).

5.6.5.2 Case 2: Transfer the model trained on the source domain to the target dataset with fine-tuning

For both high and low lesion load cases in the OXVASC dataset, segmentation results improved with increased amounts of training data (figures 5.18b and 5.19b). For instance, segmentation results with fine-tuning using 1 training subject showed the worst results for both lesion loads, and improved with 7 and 15 training subjects. This is also evident from the heatmaps of SI values, voxel-wise TPR and FPR values shown in figure 5.21. The heatmaps show the values of various performance measures (SI, TPR and FPR) for different numbers of training subjects and numbers of fine-tuned layers. Between the number of training subjects and the number of fine-tuned

layers, the effect of training data size was observed to be larger than the effect of fine-tuned layers. All the performance measures showed best values for the largest training size (15 subjects).

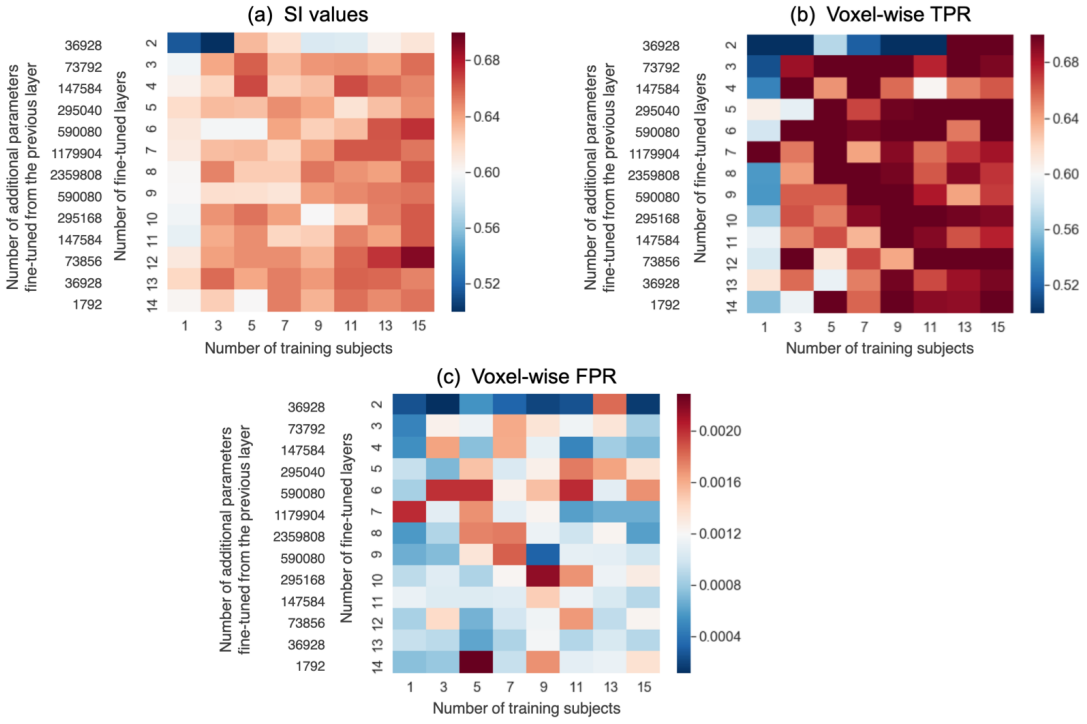


Figure 5.21: Heatmaps of performance measures corresponding to the number of training subjects and the number of fine-tuned layers. The maps are shown for (a) SI values, (b) voxel-wise TPR and (c) voxel-wise FPR values. Note that layers are numbered from the decoder to the encoder end (figure 5.6). Layers 2 7 indicate the decoder arm and the encoder starts from layer 8 and goes until layer 14. Given a number of fine-tuned layers, the layers prior to them in the encoder end were frozen, and the remaining layers towards the decoder end were fine-tuned. The number of parameters associated for individual layers has been reported (only for a single plane). For example, if the final 5 layers are fine-tuned, the sum of top 4 values in the left column denotes the total number of parameters fine-tuned per plane.

In figure 5.21a, the SI values were, generally, slightly lower when fewer layers in the decoder arm were fine-tuned, and also when there was less training data. This is because the fine layers at the top of the decoder contain low-level features that are consistent across domains and do not change much with fine-tuning. However, the SI values really started to increase from the case when fine-tuning was done in the intermediate coarser layers (layers 5,6 and layers 10-12), with even a small amount of training data. This is due to the rich domain-specific information at the intermediate

layers. The increase in SI values were noticeably higher when encoder layers (layers 10-12) were fine-tuned because of the higher context in the encoder features. Therefore the visual results were better for the first 3 columns of figures 5.18b and 5.19b when compared to their corresponding results in the last column (fine-tuned from layer 2). The best performance measures were obtained when the source model was fine-tuned starting from layer 12 with 15 training subjects (table 5.6 and figure 5.20). In this case, TL provided better performance than case 1 but gave lower SI values and cluster-wise TPR values when compared to unsupervised DANN (however, none of them are significantly different, as reported in table 5.7).

5.6.5.3 Case 3: Unsupervised domain adversarial training on source and target domains

In the case of unsupervised DANN, the segmentation was better than both cases 1 and 2 as shown in figures 5.19c and 5.18c on the OXVASC dataset. The DANN model detected more lesion voxels along the ventricles and lesion edges when compared to the results obtained with a source trained model, and less false positives when compared to the TL case involving more training data (especially in low lesion load case, figure 5.19c and the third row of figure 5.19b). Even without using target labels for training, the model provided better delineation of PWMHs, indicating the ability of the model to learn domain-invariant features and provided better segmentation in the target domain. The unsupervised DANN model achieved an SI value of 0.71 $\pm$ 0.17, a voxel-wise TPR of 0.76 $\pm$ 0.14 with an FPR of 6.9×10^{-4} and a cluster-wise TPR of 0.72 $\pm$ 0.15 on the OXVASC dataset as shown in table 5.6 and figure 5.20. Applying the unsupervised DANN model after adaptation on a source test dataset (50% each of the Singapore and Utrecht), we achieved an SI value of 0.71 $\pm$ 0.15, a voxel-wise TPR of 0.80 $\pm$ 0.14 with an FPR of 7.1×10^{-4} and a cluster-wise TPR of 0.70 $\pm$ 0.12. Unsupervised DANN gave the best cluster-wise TPR, and the values were significantly higher than those of case 1 (table 5.7). It also provided improvement in SI, voxel-wise FPR and cluster-wise TPR compared to TL (case 2), however, the differences in the performance measures were not significant.

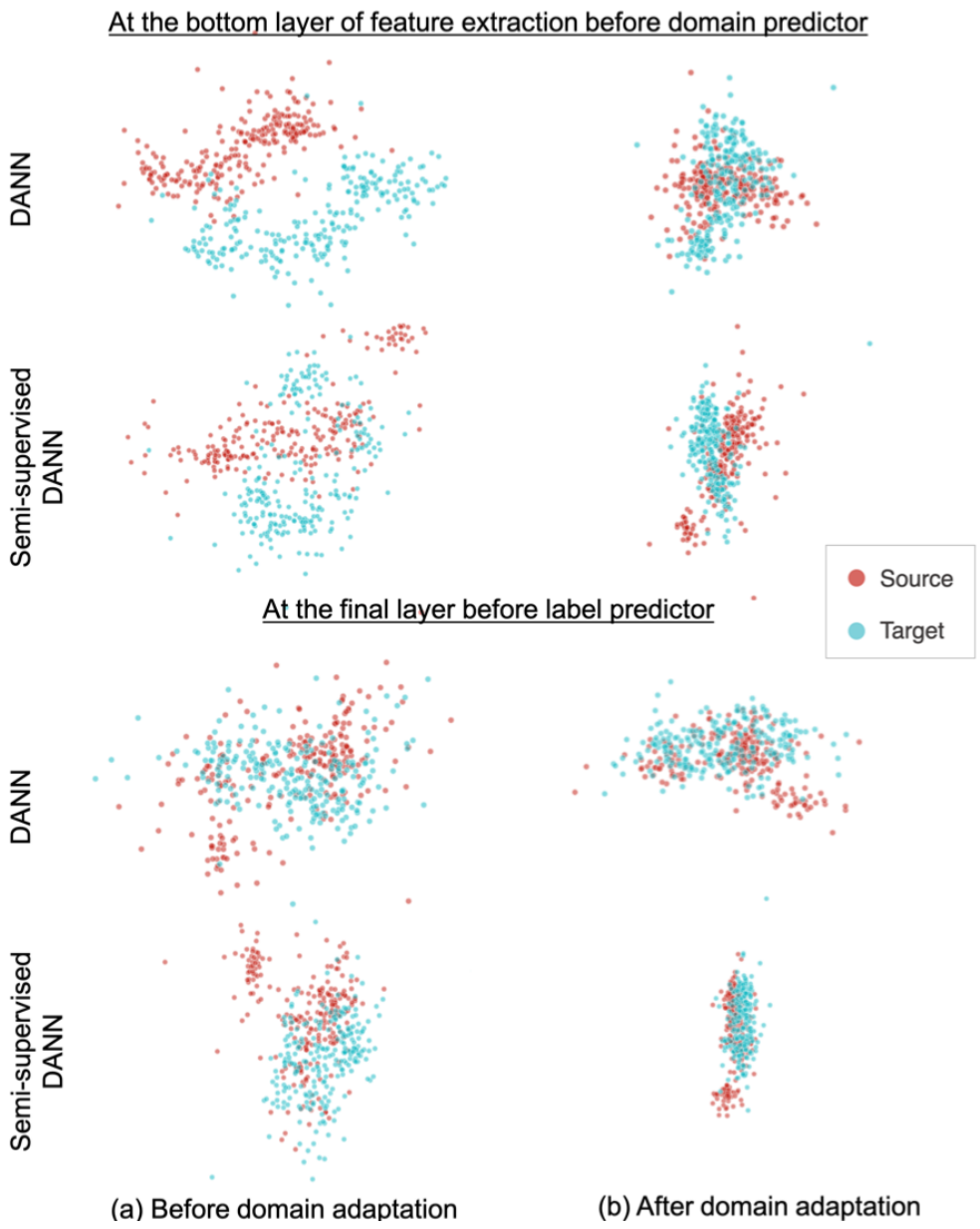


Figure 5.22: The effect of domain adaptation on the extracted feature distributions for source (red) and target domains (blue). Top row: tSNE plots of the feature map values at the layer before the domain predictor for the unsupervised DANN and the semi-supervised DANN (a) before and (b) after adaptation. Bottom row: Similar tSNE plots at the layer before the label predictor.

Figure 5.22 shows the visualisation of the feature embedding values at different layers within the unsupervised DANN, for the source and the target domains. From the figure, it can be seen that the model brings the features from the two domains much closer, both at the layer before domain prediction and the layer before label prediction (figure 5.22a, b).

5.6.5.4 Case 4: Semi-supervised domain adversarial training on the source and target domains

In the case of the semi-supervised DANN, the addition of a fraction of the labelled target data to label prediction did not provide noticeable improvement in the segmentation performance over the unsupervised DANN. The improvement was subtle on visual assessment, and observable mainly along the boundaries of PWMHs, as shown in figure 5.19d and figure 5.18d. The semi-supervised DANN model achieved performance measures similar to those of the unsupervised DANN: an SI value of 0.71 ± 0.19, a voxel-wise TPR of 0.77 ± 0.15 with an FPR of 12.7×10^{-4} and a cluster-wise TPR of 0.64 ± 0.14 as provided in table 5.6 and figure 5.20. On the source dataset the semi-supervised DANN model achieved an SI value of 0.65 ± 0.20, a voxel-wise TPR of 0.82 ± 0.13 with an FPR of 10.1×10^{-4} and a cluster-wise TPR of 0.70 ± 0.11. The overall performance of semi-supervised DANN was not better than the unsupervised case, despite providing some labelled target data for training. In fact, the cluster-wise TPR values were significantly lower than that of unsupervised DANN (case 3) - see table 5.7. Similar to the unsupervised case, we observed that the semi-supervised DANN also brings the distribution of extracted features from the source and target domains close together as shown in figure 5.22.

5.6.5.5 Case 5: Train on the target domain from scratch and apply to the target domain

Out of all the test cases, the result from the target-trained model (trained on the CADASIL dataset) showed the best segmentation performance, as shown in figure 5.19e and figure 5.19e. Specifically this case performs better than case 1, since in case 1 the model trained on the source datasets (i.e. Utrecht and Singapore) would have learnt very different features with respect to the target test dataset (OXVASC). However, in this case even though the model is trained on CADASIL (target training dataset), the features learnt would be more similar to OXVASC (target test dataset) compared to the source datasets. But even in this case, the results showed false positive voxels in the high lesion load case, while the delineation of PWMHs was

better than all the other cases. The target-trained model achieved the best SI value as reported in table 5.6 and shown in figure 5.20 (although the difference is not significantly different from the other cases, table 5.7). Unsupervised DANN (case 3) provided comparable SI, slightly lower voxel-wise TPR, significantly lower FPR and higher cluster-wise TPR values (tables 5.6 and 5.7).

5.7 Discussion and conclusions

In this work, we proposed a DL model using an ensemble of U-Nets, TrUE-Net, for accurate WMH segmentation. We studied the effect of training parameters on model optimisation and the effect of model dimension, loss functions and the addition of attention gate on the segmentation performance. We evaluated the overall segmentation performance and the individual performances in deep and periventricular regions on various datasets. In addition, we also performed comparative analysis between the proposed model and BIANCA. Finally, we studied the generalisability of our proposed model across data from different domains, and tested various domain adaptation techniques.

We observed that using a lower batch size results in noisy but lower loss values. The lower batch size could be advantageous because of two reasons: lower computation load per batch (hence higher speed) and faster convergence due to the regularisation effect of noisy gradient estimation. Firstly, using smaller batches reduces the gradient estimation time per iteration. However, this would increase the overall number of iterations per epoch, which brings us to the second advantage. Due to the noisy nature of gradient estimation, during subsequent iterations, the cost function fluctuates, and hence can get out of some spurious local minima and converge to a better local minima quickly. This is not possible in larger batches, since the accumulated weights force the optimiser to take larger steps, leading to sharper minimisers and overshoot of the gradient values [357, 358]. Hence, smaller batch sizes lead to quicker convergence in fewer epochs, despite more training steps per epoch. Additionally, they avoid over-fitting, as evident from the lower validation loss for a batch

size of 8, compared to 16 and 32 (figure 5.7a). A batch size of 8 also requires less computation memory, which is crucial in training ensemble models.

Regarding the parameter ϵ, we chose the value 1×10^{-4} as an optimal value for further experiments, since it showed lower loss values. Since epsilon occurs in the denominator of the weight update step, having a very low value results in higher gradient updates, leading to unstable optimisation as shown in the case of 1×10^{-6} (figure 5.7b). In the case of learning rate, choosing a higher value results in earlier convergence with higher loss values. On the other hand, lower values require more epochs to converge due to smaller steps of the updates in weights. Hence, we chose an optimal learning rate value between 1×10^{-3} and 1×10^{-6}, figure 5.7c, for our further experiments.

The model dimension has a very noticeable effect on the segmentation performance. The 3D U-Net model provides better results for larger objects with rich context. For instance, the 3D U-Net model detects most of the PWMHs lesions, since they are larger and have more contextual information, but it failed to detect smaller DWMHs (figure 5.8c) and does not handle well the discontinuities in manual segmentations. The 2D model is more suitable for those cases, but the 2D model in single plane alone is not sufficient to capture contextual information across slices and hence often misses lesions in contiguous slices, resulting in discontinuous segmentation (figure 5.8d). Hence, the triplanar ensemble of 2D U-Nets are better than a single plane 2D U-Net since the ensemble model has lower bias towards spurious detections compared to its constituent single-plane models. Moreover, the triplanar model sees data from different planes, thereby avoiding discontinuities in WMHs across consecutive slices and attenuating noise voxels. Finally, triplanar models are efficient in detecting edge information, since the model *views* each WMH in a comprehensive manner from all 3 planes. Hence using the triplanar model leads to both sensitive and precise detection (figure 5.8e), with fewer parameters than the 3D U-Net model.

The effect of loss function on segmentation is quite subtle by visual assessment when compared with the effect of model dimension. In general, we observed that

CE loss is responsible for identification of the majority of lesions against the background voxels, while the Dice loss is more sensitive to smaller lesions and edges. We observed that weighting the CE loss function with values based on distances from ventricles and GM imparts more anatomical information, and additionally addresses the class imbalance. This avoids over-segmentation of highly bright PWMHs and under-segmentation of DWMHs (figure 5.9c,d). In addition, using the Dice loss also results in the detection of more subtle deep lesions, giving more accurate segmentation (figure 5.9e) and makes the segmentation more precise (higher SI values in the low lesion load subjects in figure 5.10).

The LOO evaluation of TrUE-Net on the MWSC cohorts, NDGEN and OXVASC datasets shows that the overall performances are better for the Singapore and Amsterdam cohorts of the MWSC dataset when compared to other datasets. This could be due to the fact that the above two datasets have more subjects with low lesion loads. Although our model performs well in all cases, the performance is particularly good in low lesion load subjects since the segmentation is more precise. This is also consolidated by the fact that the Amsterdam cohort with the lowest mean lesion load ($\approx$ 10762mm^3) shows the highest cluster-wise F1-measure (0.72 $\pm$ 0.11, table 5.2). TrUE-Net detects more lesions, especially on the NDGEN dataset, showing the highest cluster-wise TPR value (0.74 $\pm$ 0.12, table 5.2). On the MWSC dataset, we achieved mean performance measures of: SI = 0.77 $\pm$ 0.11, AVD = 22.01 $\pm$ 20.1, cluster-wise TPR = 0.63 $\pm$ 0.13, cluster-wise F1-measure = 0.65 $\pm$ 0.11 and H95 = 4.0 $\pm$ 3.8. However, it is worth noting that, as in the case of any LOO evaluation used to estimate performance, for each subject the training data used for the model is slightly different and hence they cannot be summarised as the behaviour of a single model. In the MICCAI WMH segmentation challenge [216], on unseen test datasets, the 3 top performing methods achieved SI = 0.80, 0.78, 0.77; AVD = 28.7, 21.7, 18.3; cluster-wise TPR = 0.84, 0.83, 0.73, cluster-wise F1-measure = 0.76, 0.70, 0.78 and H95 = 6.3, 6.8, 7.2. Our model achieves comparable SI, AVD and H95 values with respect to the challenge results, with a lower cluster-wise F1-measure and cluster-wise

TPR values. However, this is not a direct comparison since the results reported by the challenge were determined on independent test datasets, while our results were obtained using leave-one-out validation on the MWSC training datasets that are publicly available. Moreover, we retained the regions corresponding to other pathologies for a more stringent evaluation, while these regions were ignored during evaluation in the challenge.

When comparing the performance of our model in the deep and the periventricular regions, the trend of performance measures for DWMHs and PWMHs remained consistent across datasets. In general, we observed that SI values and voxel-wise TPR values are higher for PWMHs when compared to DWMHs. In the deep regions, cluster-wise TPR is higher that in the periventricular regions. This shows that while more DWMHs are detected correctly, they might not be delineated correctly, which is also evident from higher cluster-wise F1-measure values (indicating more precise detection) and higher Hausdorff distance in the deep regions (figure 5.14 and table 5.3). At this point, it is worth remembering that manual segmentation is our gold standard but not necessarily the absolute truth. Therefore these errors in the delineation of DWMHs could be due to inconsistencies in manual segmentation (due to low contrast and other confounders) rather than lower model sensitivity.

On comparing TrUE-Net with TrUE-Net-Att and BIANCA, we found that both the ensemble models outperform BIANCA with significant differences in cluster-wise F1-measure values in almost all datasets (except the OXVASC dataset). Overall, we found that TrUE-Net achieves the highest SI values and the lowest Hausdorff distance values for all datasets, and provides better segmentation in various lesion load cases (figure 5.16). Both TrUE-Net and TrUE-Net-Att perform relatively well in subjects with low lesion load, detecting less false positives than BIANCA. The addition of AG slightly improves the lesion segmentation in low lesion load subjects. Hence, for lesion segmentation in datasets with more healthy controls, AGs could be added to TrUE-Net at the coarsest level for more precision, with a trade-off of higher computational cost. For instance, addition of AG in the coarsest scale in our case

resulted in $\approx$ 10,624,000 additional parameters, which is around twice the amount of total parameters of the U-Net in a single plane. However, we observed from the overall measures that TrUE-Net outperforms both the other methods, providing more precision and better delineation of lesions.

In the experiments looking at model generalisability, for the case where we applied the source trained model directly on the OXVASC dataset, we observed that the model gave the lowest performance compared to the DA techniques. This is due to the fact that source and target datasets are quite different from each other in image resolution, pathology, intensity characteristics and lesion distribution, as shown in figures 5.5 and 5.4.

For the TL case, as shown in the heatmaps in figure 5.21, we observed an improvement in the performance when more intermediate features from both encoders and decoders (layers 4-6 and layers 10-12) were fine-tuned. This is because the intermediate layers in a network have been shown to capture domain-specific features regarding small objects during visualisation [359, 303]. This is especially true for fine-tuning layers 10-12, since the intermediate layers in the encoder arm provide better context than those in the decoder arm. Generally, finer layers at the top often contain low-level features (e.g. edges) that are domain invariant, and hence fine-tuning this layer does not improve the target performance much. On the other hand, the coarsest features at the bottom contain higher-level information such as lesion pattern, that are slightly domain specific but are still fairly consistent across the domains.

On comparing the segmentation performance of the model transferred from the source to target domain (case 2) and the DANN model, we observed that the unsupervised DANN performed better than TL. From figures 5.18b, c, d and 5.19b, c, d, even with a higher number of training subjects for TL, the DANN model provided better visual results with less noise and more precise segmentation of boundary voxels. This was also reflected in the performance measures in table 5.6 and figure 5.20. Fine-tuning in TL would require more data than we currently have in the CADASIL dataset, even with augmentation, since the source and target datasets have different

distributions. Another important factor to consider is the difference in lesion distribution. We chose CADASIL and OXVASC as our target domain datasets, since they have similar intensity profiles, voxel resolutions and were acquired with similar protocols. However, typically CADASIL subjects have lesions within the external capsule and anterior temporal lobes [40], while this pattern is absent in vascular subjects in the OXVASC subjects. Hence, we observed more false positives in the anterior temporal lobes for the transfer learning case on the OXVASC dataset. Unsupervised DANN, on the other hand, does not require lesion labels from CADASIL dataset and hence does not specifically capture the CADASIL lesion distribution. DANN extracts domain invariant features (e.g. contrast between lesion and background, distribution of PWMHs and so on) from both source and target datasets, resulting in more precise segmentation and less false positives (especially in low lesion load subjects, figure 5.19).

One of the main advantages of the unsupervised DANN model is that it does not require manually labelled data from the target dataset. Even then, our main aim for exploring the semi-supervised DANN, despite the additional manual labelling effort, was to observe if there was any significant improvement over using an unsupervised DANN on the OXVASC dataset. From figure 5.20 and tables 5.6 and 5.7, we observed that the semi-supervised DANN does not provide any significant improvement over the unsupervised DANN on source or target test datasets. In fact, the unsupervised DANN provides better cluster-wise TPR values compared to the semi-supervised DANN. This could be due to the use of limited number of the target training data (25% of 15 subjects) used for label prediction. Also, by adding labelled CADASIL to the source end, we might be slightly amplifying the impact of the lesion distribution in anterior temporal lobes, leading to more false positives (almost similar to the transfer learning case). From the data harmonisation viewpoint, we observed from the tSNE plots that the unsupervised and semi-supervised DANNs bring the source and target features closer together, resulting in a good domain adaptation. Hence, it is necessary to weigh carefully the trade-off between improvement in segmentation performance

and the amount of manual effort involved while choosing between unsupervised and semi-supervised DANNs. Our current recommendation for domain adaptation would be to use an unsupervised DANN when the amount of data is limited and there is a higher variation within the target datasets. This is because we observed that unsupervised DANNs handles both inter- and intra-domain variations better than semi-supervised DANNs.

The unsupervised DANN and the model trained on the target dataset (CADASIL) provided comparable performance on the OXVASC dataset (tables 5.6 and 5.7). Models trained on the same domain (or dataset) are expected to give better results, provided that there is sufficient amount of training data. For instance, we observed the best SI value of 0.78 ± 0.11 (with lower FPR of 3.1×10^{-4}, table 5.2) on the OXVASC dataset during LOO evaluation done with a model trained on the same dataset, rather than the CADASIL dataset. However with insufficient or slightly different data (e.g. OXVASC vs CADASIL), domain transferred models have been shown to provide better segmentation than the target model (trained from scratch) with careful fine-tuning or using more data [271]. DANNs provide results on par with case 5, likely due to 2 reasons: 1. variations in the lesion distribution between CADASIL and OXVASC datasets, 2. uncertainties in the labels on the CADASIL datasets. As we discussed earlier, lesions in the anterior temporal lobes in CADASIL subjects lead to false detections in these regions in the OXVASC dataset. Secondly, we used LOCATE outputs as our 'manual' segmentations for training the model. In chapter 4, we showed that LOCATE provides better segmentation when compared to global thresholding values considered optimal (0.9). Also, on visual assessment, LOCATE results were better than the results of BIANCA obtained using a global threshold deemed 'optimal' by an expert neurologist (0.2). However, there might still be uncertainties, especially along the lesion boundary voxels in the 'manual' segmentation used for training.

Concluding, we found that our proposed model provided accurate segmentation, with better performance than BIANCA, when evaluated on various dataset with different characteristics. We explored various DA techniques and found that DANN

models perform better than TL even with limited data and differences in their distributions. For further improving the robustness of the model, a potential future avenue would be to explore various data harmonisation techniques that could work well with a limited amount of data (e.g omni supervised learning, where predicted outputs are used to train the model in a recurrent manner [360]). Also in the semi-supervised DANN, a trial-and-error method could be used to determine the best proportion of the target training data to be used for label prediction at the source end, towards improving WMH segmentation in both domains. Another future direction could be to further improve the segmentation performance of the model by applying LOCATE to the probability maps obtained from the proposed model.

5.8 Summary

1. We proposed a triplanar U-Net ensemble network (TrUE-Net) for WMH segmentation.

2. Our model performs better than BIANCA, as it is less noisy and more accurate, especially in subjects with low lesion load.

3. Our comparison analysis between deep and periventricular regions show that the model provided better delineation of PWMHs and detected more DWMHs precisely.

4. On exploring various domain adaptation techniques to adapt the model from a source to a target domain, we observed that unsupervised DANNs perform better with limited data and provide results comparable to the model trained on the target dataset.

Chapter 6

Detection of cerebral microbleeds in a large population

The work described in this chapter aims to provide a comprehensive end-to-end automated analysis of cerebral microbleeds (CMBs), starting with a pipeline for the preselection of CMB candidate subjects, proceeding to a deep learning-based method for CMB detection and finally an automated characterisation/rating of CMBs in individual subjects.

Section 6.1 explains the need for the development of an automated method for the subject-level preselection and the object-level CMB detection, and describes various deep learning (DL) concepts/approaches to be used in this work. Section 6.2 lists various datasets used in this work and their preprocessing. Our proposed CMB subject preselection pipeline and the relevant experiments are explained in section 6.3. Section 6.4 provides an overview of our proposed DL-based method for the detection of CMBs in individual subjects, and section 6.5 explains our method for automated CMB rating. The results of our experiments are provided in section 6.6, followed by the discussion of our results, conclusions and the potential future directions in section 6.7.

The preliminary results of this work were presented at the 2019 OHBM conference (Sundaresan et al., OHBM 2019).

6.1 Introduction

CMBs are the areas of haemosiderin deposits due to micro-haemorrhages in the brain. They appear as small, circular, well defined hypointense lesions ranging from 2 to 10 mm in size on T2*-weighted gradient recalled echo (GRE) images and are more prominent on susceptibility weighted images (SWI). CMBs are associated with hypertension and cerebrovascular disease, and also with Alzheimer's disease and cerebral amyloid angiopathy (CAA). However, due to their small size, contrast variations and other mimics (e.g. blood vessels), CMBs are highly challenging to detect automatically. On the other hand, exhaustive manual labelling of CMBs is difficult and time consuming, especially in large datasets (e.g. the UK Biobank dataset). Hence it would be useful to preselect candidate subjects in order to focus on those for manual labelling, which is essential for training and testing automated detection tools on these datasets. In this work, we propose an automated learning-based pipeline (section 6.3) for the preselection of CMB candidate subjects from large datasets. We evaluate the pipeline on datasets with different image modalities and in the presence of other pathological signs. We then apply the pipeline on the UK Biobank dataset for CMB candidate subject preselection.

At the single subject level, automated detection of CMBs to obtain voxel-wise CMB maps would enable us to characterise CMBs (e.g. determining their count, size and distribution). Moreover, detection of CMBs would aid in understanding the relationship between their distribution and various clinical factors. Chapter 2 reports various semi-automated and automated methods proposed for CMB segmentation. CMB detection methods in the literature follow a well-defined pattern with feature extraction, CMB candidate detection (and additional classification step at times) and false positive (FP) reduction/removal. Some of the best performing methods in the literature have used DL-based convolutional neural network (CNN) models for feature extraction [300], candidate detection [301, 302, 278] and FP removal [301, 278], occasionally with additional morphological operations for FP removal [278].

In general, networks used features derived at multiple scales [309, 310, 314] for more accurate object detection. Probabilistic networks [310] and weighted sum approach [311] have been used to combine features extracted at these multiple scales. Another approach involved feature extraction from a lower scale and 'zooming-out' to obtain information from voxel- to image-level [314]. Additionally, several patch-based approaches [311, 310, 217, 312] have been used for localisation of small objects in images. One of those approaches, *region based recognition* [315, 316] has been widely adopted in region-based CNN (R-CNN) models [303] and in their faster versions [304, 305, 307] for the detection of finer details in the images. In the initial R-CNN framework [303], a selective search algorithm was used to select candidate patches at different scales for feature extraction. Later, in the faster version [305], CNNs were used for both automatic proposal of candidate patches (using a region proposal network (RPN)) and for classification. However, the use of CNNs in the RPN makes the network computationally intensive. Moreover, unlike the generic object detection task, we have additional anatomical information regarding CMBs including the fact that they are localised, small, hypointense structures. Hence, those information could be used for extracting the ROIs, thereby avoiding the extra computation required for RPNs. In general, multiscale feature extraction and patches have been shown to provide accurate localisation of small objects [305, 311, 313, 314]. Hence, in this work, we propose an automated 3D multiscale patch-based DL method for the detection of CMBs at object-level (section 6.4), following the commonly adopted 3-step approach (candidate detection, classification and FP reduction) in the literature. Our method can be applied to any suitable input modality (T2*-weighted GRE or SWI), and uses appearance based characteristics of CMBs for efficient ROI extraction and initial candidate detection. For the first time to the best of our knowledge, we evaluate our method on a pathological dataset consisting of other vascular signs with characteristics similar to those of CMBs.

Once CMBs are detected, their characterisation in terms of size, count and spatial distribution is a crucial step towards understanding their clinical impact. For

instance, subcortical (deep) CMB are associated with cerebrovascular diseases, while lobar CMBs are associated with CAA [63, 64, 68]. In clinical settings, CMBs are often reported using rating scales, based on count and distribution of CMBs. As mentioned in chapter 1, two manual rating scales are widely used: microbleed anatomical rating scale (MARS) [73] and the brain observer microbleed scale (BOMBS) [74]. However, similar to manual segmentation, manual rating of CMBs can be time consuming, especially in larger population studies. Hence, providing an automated rating according to these scales would be useful to efficiently provide a clinically useful metric. Our third aim was therefore to propose an atlas-based approach for the automated rating of CMBs (section 6.5) according to the MARS scale. We initially evaluate our automated rating method using manual segmentation maps, and later validate our method using the results of object-level CMB detection, on a pathological dataset. Finally, we use our automated rating to further evaluate the performance of our object-level CMB detection method.

6.2 Dataset details

In this work we used three datasets to evaluate the preselection pipeline and the detection of CMBs. The datasets are diverse in their intensity characteristics and are acquired using different protocols. A brief overview of the datasets is provided below. For more details regarding MRI acquisition parameters, refer to appendix A.

- **The Oxford Vascular Study (OXVASC) dataset:** The dataset consists of T2*-weighted GRE images from 40 participants from the OXVASC study [323], who had recently experienced a minor non-disabling stroke or transient ischemic attack (TIA). Age range 35.6 - 94.8 years, mean age 68.7 $\pm$ 15.5 years, median age 67.4 years, female to male ratio F:M = 15:25. Out of 40 subjects, 20 subjects have CMBs and the corresponding manual segmentations available. Total number of CMBs: 267, mean: 10.7 $\pm$ 1.13 CMBs/subject, median: 2 CMBs/subject. The median number of CMBs/subject is much lower compared to the total number, since almost all subjects have very low CMB count (< 10)

except one subject having 202 CMBs (shown in histogram of subject-level CMB counts, figure 6.1a). A zoomed-in version of the histogram for the subjects with CMB count < 10 is shown in figure 6.1b.

- **The Tranexamic acid for IntraCerebral Haemorrhage 2 (TICH-2) trial dataset:** We obtained a subset of the dataset used in [80] through collaboration with the TICH2 trial group [80, 81] at University of Nottingham, UK. The dataset used in this work consists of 50 SWI from the subjects with spontaneous intracerebral haemorrhage. Age range 29 - 88 years, mean age 64.76 $\pm$ 15.5 years, median age 66.5 years, female to male ratio F:M = 24:26. Out of 50 subjects, 25 subjects have CMBs and manual segmentations for CMBs are available for all 25 subjects. Total number of CMBs: 505, mean: 20 $\pm$ 32.6 CMBs/subject, median: 5 CMBs/subject. Additionally, MARS rating scales were provided for the CMB subjects for both '*definite*' and '*possible*' cases (shown in a sample MARS rating form in appendix D). For training and testing purposes, we obtained the final manual segmentation maps by combining the manually segmented CMB from both categories, for all our experiments. Similar to OXVASC, the histogram of subject-level CMB counts in the dataset (figure 6.1c) shows that most of the subjects have less than 10 CMBs (zoomed-in histogram shown for subjects with CMB count < 10 in figure 6.1d).

- **The Hong Kong dataset (HK):** The dataset [301] consists of 320 SWI images in total, out of which 126 are subjects with stroke (mean age: 67.4 $\pm$ 11.3) and 194 are from normal ageing subjects (mean age: 71.2 $\pm$ 5.0). In this work, we have used a subset of 20 subjects that is publicly available from this dataset. Manual annotations in the form of CMB coordinates are available for these 20 subjects along with the dataset. Hence, another rater independently provided the manual segmentations on the dataset, and we considered the combination of both the information of the coordinates and the manual masks as our final ground truth. Total number of CMBs: 89, mean: 4.5 $\pm$ 4.4 CMBs/subject,

median: 2 CMBs/subject. The histogram of subject-level CMB counts for this dataset has been shown in figure 6.1e.

Additionally, we also used the UK Biobank (UKBB) dataset for applying our subject-level preselection pipeline (described in section 6.3). Since the UKBB dataset does not have manual segmentations available, we initially used a subset of images from UKBB to evaluate the performance of our preselection pipeline by testing the relationship between CMB count per subject and clinical factors known to be related to CMBs (i.e. age and blood pressure). Details of data subsets and characteristics are described in section 6.3.4.3. Later, we apply our final pipeline on a larger set of UKBB data to obtain manual segmentations for further analyses. Details of data subsets and characteristics are described in section 6.6.1.4.

6.2.1 Data preprocessing

We reoriented the T2*-weighted GRE and SWI images to match the orientation of standard MNI space, in order to aid in further processing. We skull stripped the images using FSL BET [335], followed by bias field correction using FSL FAST [336].

6.3 Pipeline for preselection of CMB candidate subjects in big datasets

The automated pipeline for CMB candidate subject preselection takes T2*-weighted GRE images or SWI as input and provides a subject-level decision of CMB/non-CMB subjects. The pipeline consists of three steps: 1. removal of blood vessels and sulci, 2. voxel-wise detection of initial CMB candidates, 3. filtering of initial candidates using shape-based attributes.

6.3.1 Removal of blood vessels and sulci

In the first step, we removed the blood vessels, sulci and other elongated dark structures in the input image to reduce the appearance of CMB mimics. Figure 6.2 shows

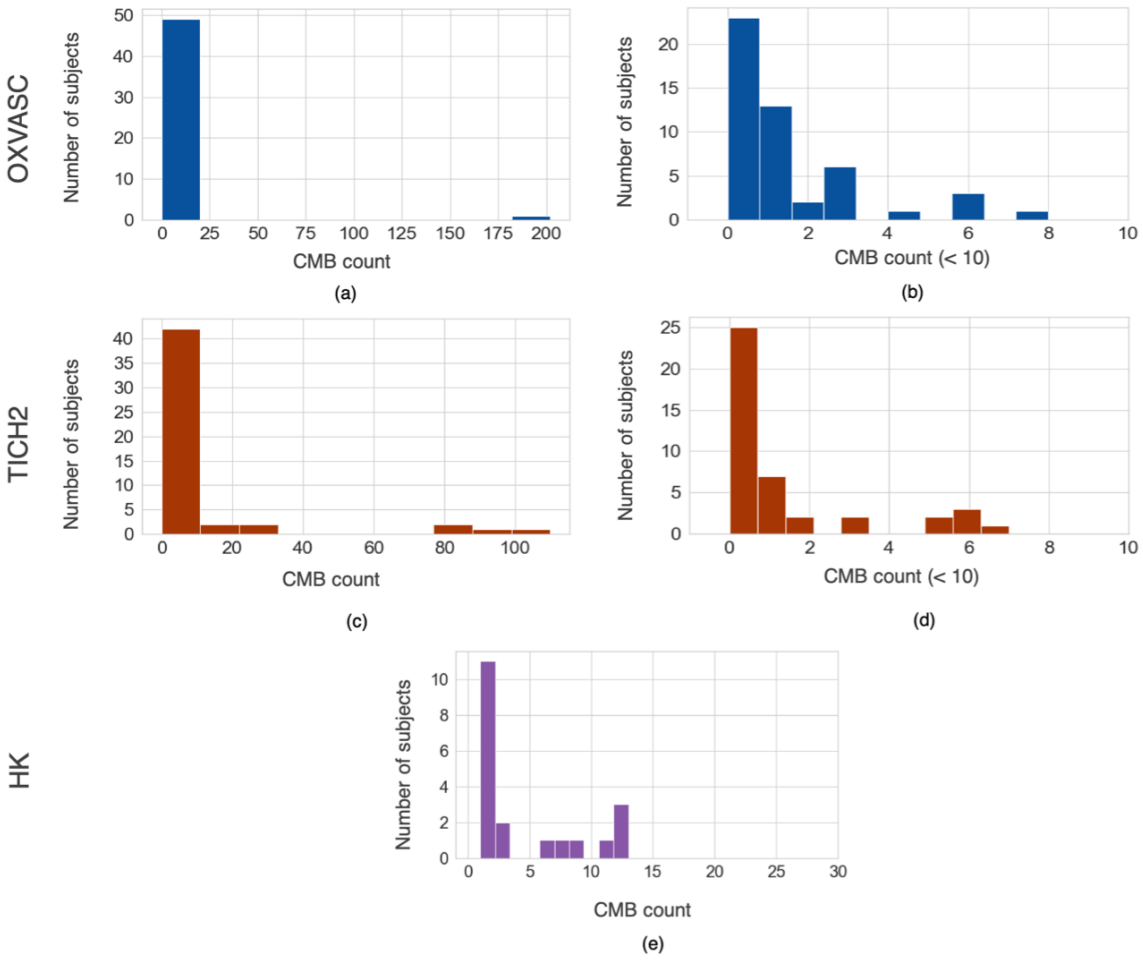


Figure 6.1: Histograms of the subject-level CMB counts based on manual segmentations for the (a) OXVASC, (c) TICH2 and (e) HK datasets. Since most of the subjects from OXVASC and TICH2 have CMB counts < 10, zoomed-in histograms are shown for those subjects with CMB counts < 10 for the OXVASC (b) and TICH2 (d) datasets.

the extracted features and a few samples of the vessel/sulci removed image. We exploited the elongated tubular structure of vessels/sulci to extract the following edge and orientation-based features:

1. **Frangi filters [361]:** Frangi filters are based on multiscale second order local structure of the image and denote the degree of vessel-like/edge characteristics at each voxel. We extracted voxel-wise intensity values of Frangi filter output as features (figure 6.2b).

2. **The eigenvalues of structure tensor:** Structure tensor is the covariance matrix consisting of partial derivatives of the gradients [362]. The eigenvalues λ_1

and λ_2 at each voxel indicate the edge strength at the voxel. We considered the combination of principal eigenvalue λ_1 and the linearity measure $\ell = |\lambda_1 - \lambda_2|/2$ (figure 6.2c) at each voxel as features.

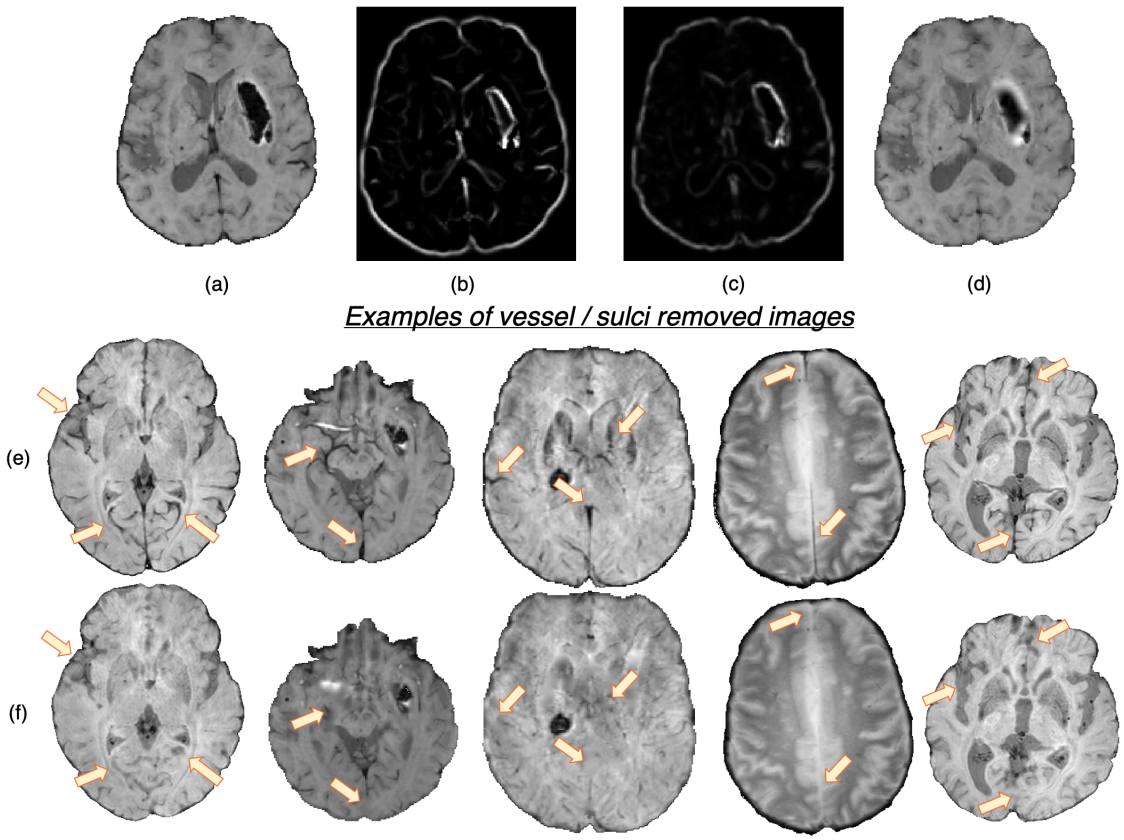


Figure 6.2: Features used for the vessel/sulci removal, along with a few samples of vessel/sulci removed images. Top panel: (a) Input image, (b) Frangi filter output, (c) Structure tensor output and (d) vessel removed image. Note that even the edges of the haemorrhages have been smoothed out, aiding in false positive reduction. Bottom panel: A few instances of images shown in (e), along with their vessel removed results in (f). The arrows indicate the areas of noticeable vessel/sulci removal.

We used K-means clustering, an unsupervised learning algorithm on the above features to classify voxels into 2 classes (vessel vs background). The detected voxels with vessel-like structures and sulci are used as masks for inpainting. We filled the masked regions with the intensity values similar to those in the non-masked neighbourhood to remove linear dark structures from the input image.

6.3.2 Voxel-wise detection of initial CMB candidates

In the second step, as shown in figure 6.3, we extracted the following 7 features at each voxel on the vessel-removed image for the initial CMB candidate detection:

1. **Intensity transformations (3 features):** At each voxel, we extracted its intensity value (figure 6.3a). Additionally, to obtain contrast enhanced intensity features, we extracted the exponential of the intensity, setting the scale parameter value to 1 (figure 6.3b) and applied 'contrast limited adaptive histogram equalisation' (CLAHE) with a clip limit of 0.01 on the input image to obtain the CLAHE output at each voxel (figure 6.3c).

2. **Fast radial symmetry transform (FRST) [363]:** We performed FRST with four radii of 2, 3, 4 and 6. We extracted the mean value of the outputs of the above radii at each voxel as a feature (figure 6.3d).

3. **The eigenvalues of structure tensor:** We considered the principal eigenvalue λ_1 of the structure tensor at each voxel (figure 6.3e).

4. **Gaussian filter:** We smoothed the vessel removed image I with a Gaussian filter with $\sigma = 1.5$ (empirically determined to roughly match the size of CMBs) to get a smoothed images I_S. The difference $I_\delta = I - I_S$ removes the background and provides the sharper objects and blobs, and hence can be interpreted as 'blobness' measure. We obtained the above blobness measure I_δ (figure 6.3f) at each voxel.

5. **Laplacian of Gaussian filter (LoG):** Since CMBs have well-defined edges, we applied LoG, a second order derivative filter for edge detection, on the vessel removed image and obtained the filtered output (figure 6.3g) at each voxel.

We normalised the above features individually by dividing by their maximum value and use the normalised features for training a support vector machine (SVM) classifier at the voxel-level to obtain an initial probability map P_{CMB}. We then thresholded it at a global threshold th_{prob} of 0.8. The threshold value was determined empirically,

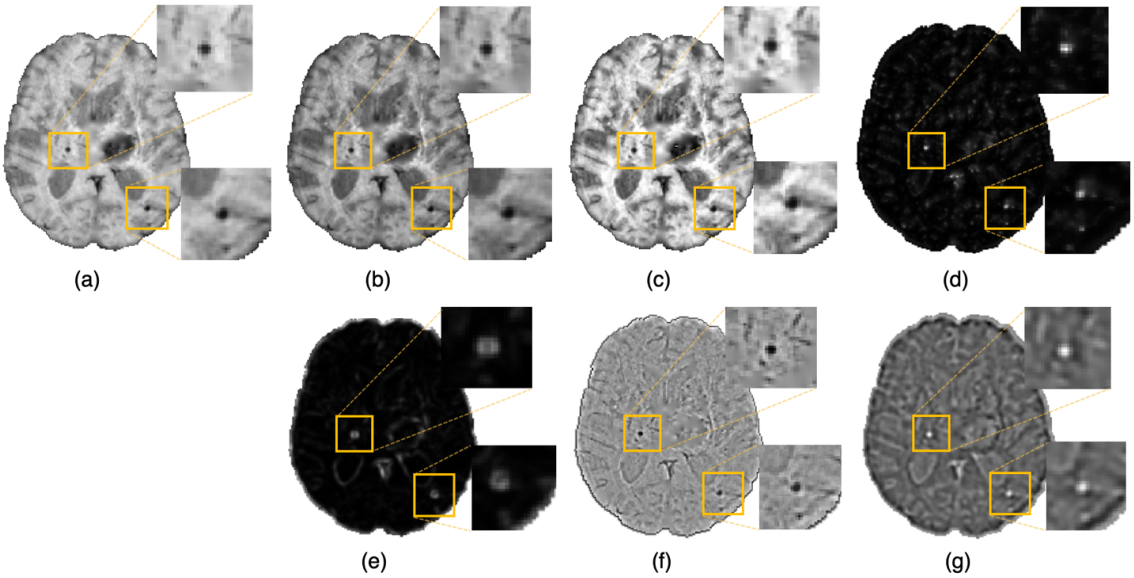


Figure 6.3: Features extracted for the voxel-wise CMB candidate detection. Images showing (a) Input intensity, (b) exponential transformed intensity, (c) CLAHE output, (d) fast radial symmetry transform (FRST) output, (e) structure tensor output, (f) Gaussian filtered output and (g) Laplacian of Gaussian (LoG) output. Inset figures show the magnified versions of the regions indicated in the boxes.

based on the performance values (refer to the section 6.6.1.1 for more details and results).

6.3.3 Filtering of initial candidates using shape-based attributes

The small dark structures in the image other than CMBs (including noise and remaining fragments of blood vessels) were detected as false positives (FPs) in voxel-level classification. In this step, we used the following shape- and location-based object-level attributes for reducing FP candidates as shown in figure 6.4:

1. **Volume** $\mathcal{V}_c$ of the candidate in mm^3.

2. **Ellipticity** ε_c**:** The measure of elongated nature of an object. The value ranges between [0,1] and a sphere has a value of 0.

3. **Solidity** $\mathcal{S}_c$**:** The ratio between volume of the candidate and its convex volume. The value ranges within [0,1] and a solid object (of any shape) has a value of

1; this criterion removes the fragments of blood vessels that survived the vessel removal step, since they would have low solidity (e.g. curved fragments and intersection of vessels).

4. **Diameter** $\mathcal{D}_c$ of the candidate in mm.

We considered a candidate as CMB if: 5 mm^3 $< \mathcal{V}_c <$130 mm^3, $\varepsilon_c < 0.3$, $\mathcal{S}_c > 0.5$, 2mm $< \mathcal{D}_c <$ 10mm. The values for above criteria were chosen empirically by a trial-and-error method based on the best performance using the measures specified in section 6.3.5.

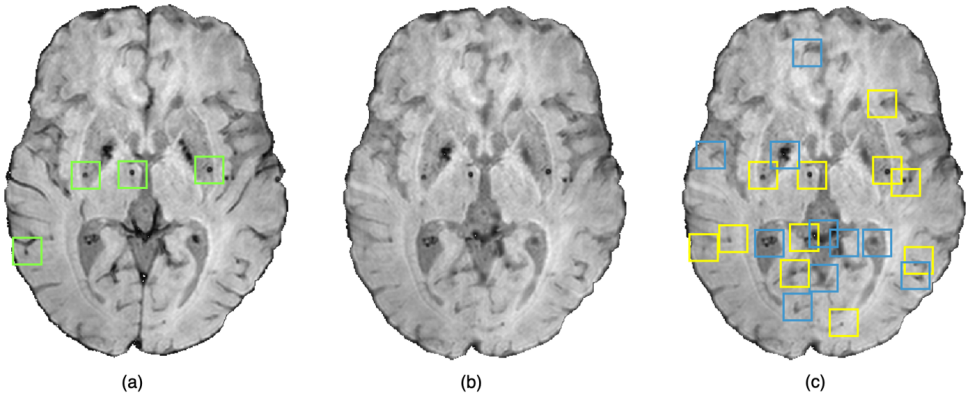


Figure 6.4: An instance of the shape-based filtering of initial candidates. (a) Input image with green boxes indicating manually segmented CMBs, (b) vessel removed image and (c) result of the shape-based filtering. In the output (c), blue boxes indicate the rejected initial candidates and yellow boxes indicate the final CMB candidates predicted by the pipeline.

Preselection criterion: Finally, a subject is classified as having CMBs if the count of the detected CMB candidates (including true positive and false positive candidates) obtained after applying our pipeline exceeds an empirically set threshold value Th_{NCMB}. We experimented with various threshold values and set Th_{NCMB} to 35 CMBs (for more details on the experiment and results, refer to sections 6.3.4 and 6.6.1).

6.3.4 Evaluation of CMB candidate preselection pipeline

We evaluated the proposed pipeline by training it using different datasets to study the effect of training dataset characteristics on its performance.

6.3.4.1 Case 1: Pipeline trained on OXVASC and evaluated on TICH2 data

We trained the SVM classifier of the pipeline on T2*-weighted GRE images from 40 subjects of the OXVASC dataset. We later evaluated it on the SWIs from 50 subjects of the TICH2 dataset. We determined the performance measures (specified in section 6.3.5) at different values of the threshold Th_{NCMB}, ranging from 25 to 75 CMBs in steps of 5, for plotting the ROC curve.

6.3.4.2 Case 2: Pipeline trained on TICH2 and evaluated on OXVASC data

We trained the pipeline on SWI images from the TICH2 dataset and evaluated it on the OXVASC dataset. We evaluated the model on T2*-weighted GRE images from the OXVASC dataset using the metrics specified in section 6.3.5, applying different values of threshold Th_{NCMB}, ranging from 15 to 55 CMBs in steps of 5. We used a lower range compared to case 1 due to the lower number of CMBs in the manual segmentations of the OXVASC dataset.

6.3.4.3 Case 3: Application to the UKBB dataset

For the initial evaluation on the UKBB dataset, we considered a subset of UKBB, **UKBB-I** consisting of 180 randomly selected subjects with age range 45 - 75 years, mean age 58.5 $\pm$ 8.8 years, median age 60.1 years, female to male ratio F:M = 85:95 and blood pressure (BP) range: 94/58 - 202/114 mmHg, mean BP 141.7/86.6 $\pm$ 20.1/12.3 mmHg, median 147/91 mmHg. We applied the above two instances of our candidate selection pipelines, one trained on 40 subjects from the OXVASC dataset (case 1) and another on 50 subjects from TICH2 dataset (case 2), on these 180 subjects chosen from UKBB-I dataset. We evaluated the pipelines by studying the relationship between subject-level CMB counts and clinical/demographic factors. We chose age and BP since they have been more commonly associated with CMBs [98, 101, 63, 64] than other factors. We categorised subjects into three age-groups (younger: age $\leq$ 50, mid: 50 $<$ age $\leq$ 65 and older: age $>$ 65) and two levels of BP

(low: BP $\leq$ 120/80, high: BP $>$ 120/80). We performed a one-way ANOVA test with post-hoc pair-wise comparisons to determine the significance of the difference in subject-level CMB counts among age groups. Similarly, we performed two-sample t-tests for determining the significance of the difference in subject-level CMB counts between two levels of BP.

6.3.5 Performance evaluation metrics for CMB candidate subjects preselection pipeline

For the CMB candidate subject preselection pipeline, we used the following measures (refer to the appendix C for more details) for evaluation:

- Subject-level true positive rate (TPR)
- Subject-level specificity
- Subject-level accuracy

We plotted subject-level ROC curves using TPR and false positive rate (FPR = 1 - specificity) values.

6.4 Detection of CMBs on individual subjects

Figure 6.5 provides the block diagram of steps involved in the object-level detection of CMBs. As an input for the proposed method, we considered the preprocessed (section 6.2.1) and vessel removed (T2*-weighted GRE or SWI) 3D volumes (section 6.3.1) and cropped the FOV of the volume close to the brain. As shown in the figure, our CMB detection method consists of 4 main steps: 1. initial local maxima identification for ROI extraction, 2. voxel-level CMB initial candidate screening with 3D multiscale encoder-decoder network, 3. object-level CMB candidate discrimination using a 3D classification network and 4. postprocessing for FP reduction.

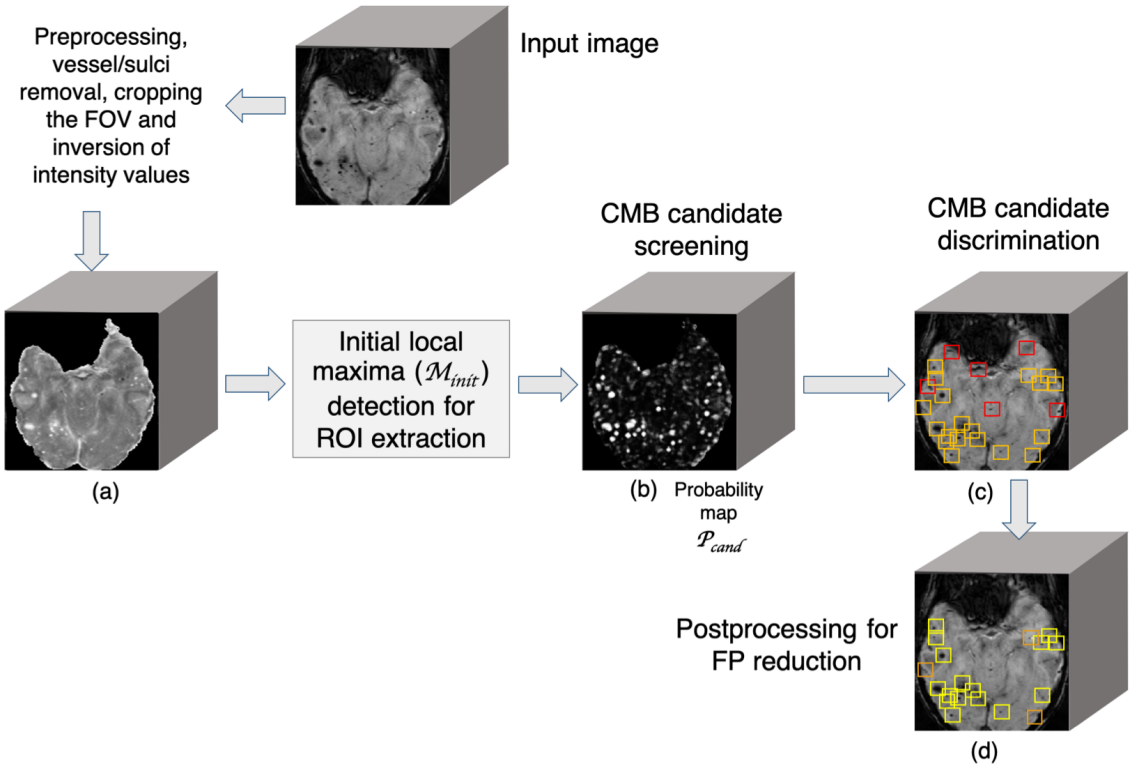


Figure 6.5: Overview of the CMB detection method. The original image shown along with (a) the preprocessed image, (b) CMB candidates from the screening step, (c) candidates discriminated into CMBs and non-CMBs and (d) the final candidates after FP reduction. The candidates rejected in the object-level candidate discrimination step are shown in red, while the candidates that passed this step are shown in orange. The final detected CMBs are shown in yellow.

6.4.1 Initial local maxima identification for ROI extraction

As widely used in the *region-based recognition* approach (refer to section 2.3.1.2) for detecting smaller objects in an image, the first step is to identify small ROIs (that are more likely to contain the objects of interest) in the image. We achieved this by inverting the intensity values of the input volume (by subtracting the intensity-normalised image from 1) so that CMBs have maximum intensities (to facilitate easier computation in the later steps) and then identifying the local maxima M_{init} on the intensity-inverted input volume. This avoids the use of more computationally intensive methods such as RPNs [304, 305, 307]. The obtained local maxima points M_{init} represents the centres of 3D ROIs to be extracted in the next step.

6.4.2 Voxel-level CMB initial candidate screening with 3D multiscale encoder-decoder network

In this step, we refined the initial local maxima points M_{init} to get the final local maxima M_{final} and screened the image for obtaining the initial CMB candidate map P_{cand}.

Figure 6.6 shows the block diagram of this step and the architecture of the 3D encoder-decoder model used in the screening. We used a multiscale framework for the local maxima refinement and detection of initial candidates. Our multiscale approach consists of ROIs at 4 scales: 24 × 24 × 24, 32 × 32 × 32, 48 × 48 × 48 and 64 × 64 × 64 voxels. We initially extracted ROIs at the largest scale (64 × 64 × 64) from the detected local maxima M_{init} points. We applied a 3D encoder-decoder model (explained in section 6.4.2.1) for the detection of initial CMB candidates within the patches. We chose the largest scale, since the patches are large enough to overcome the effect of local noise and assign higher probabilities to CMB-like regions. After replacing the estimated probability map patches back into the 3D volume, we extracted the local maxima on the probability maps to obtain the final, more refined set of local maxima M_{final}.

We then extracted ROI patches at three smaller scales (24 × 24 × 24, 32 × 32 × 32 and 48 × 48 × 48) from the final maxima M_{final}. By extracting patches at different scales, we 'zoom-out' the features at a given maximum point, therefore simultaneously obtaining comprehensive information regarding both context and local neighbourhood. We therefore applied the same 3D encoder-decoder model at the corresponding 3 scales, obtaining three probability maps. We further reduced the noise at different scales by averaging the probability maps at the two lower scales (24 × 24 × 24 and 32 × 32 × 32) and two higher scales (48 × 48 × 48 and 64 × 64 × 64) separately. Finally, to extract the CMBs of different sizes, we obtained the maximum (provides higher sensitivity to CMB candidates compared to averaging) of the two averaged probability maps to get the candidate screening probability map P_{cand}.

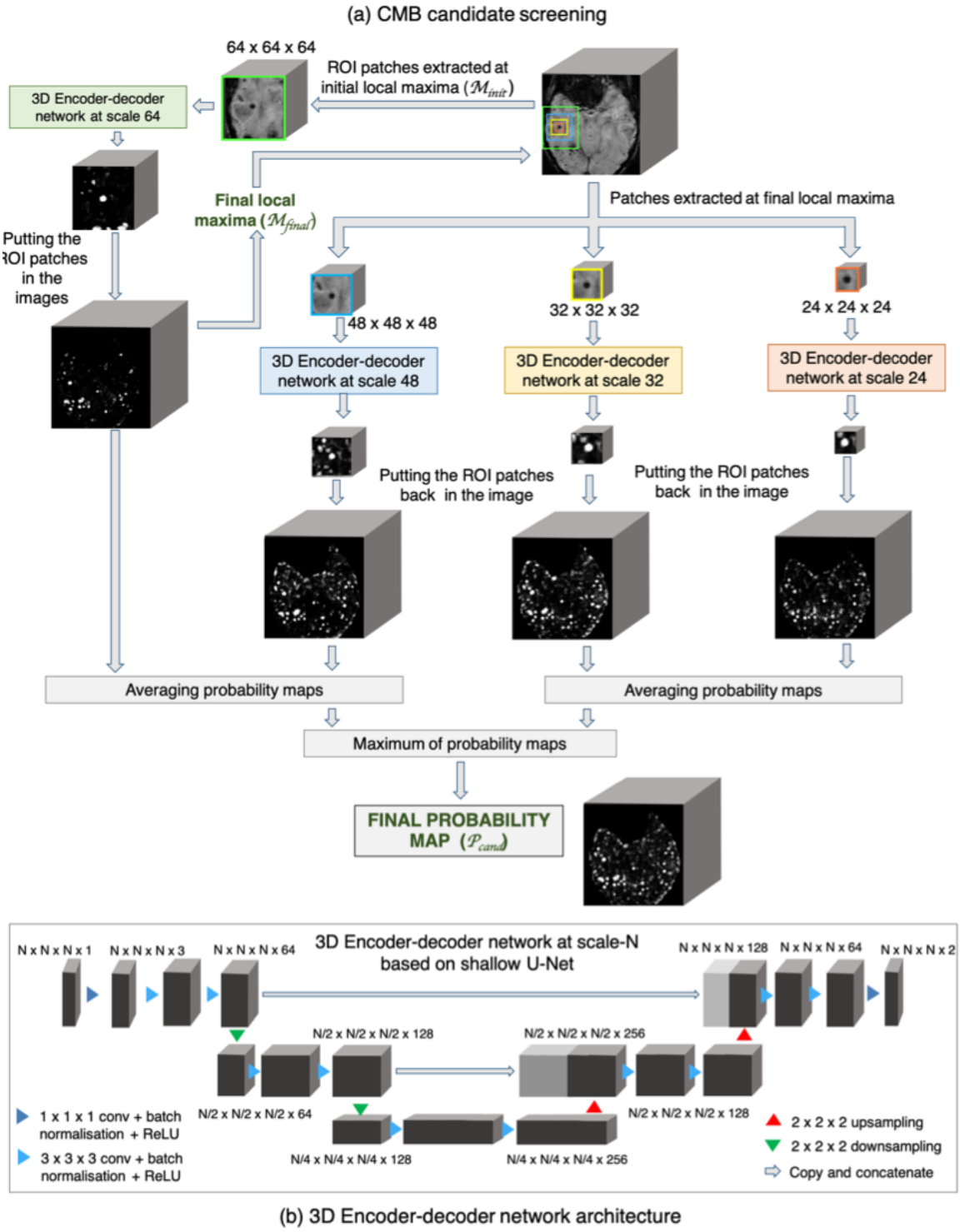


Figure 6.6: CMB candidate screening step. (a) Block diagram illustrating the multiscale CMB candidate screening step, with (b) the architecture of the 3D encoder-decoder network based on shallow U-Net, used at multiple scales.

6.4.2.1 3D Encoder-decoder architecture

As shown in figure 6.6b, the architecture of the 3D encoder-decoder network at a scale N is based on a shallow U-Net. We trimmed the U-Net to a shallow architecture with 2 pooling layers, since CMBs are small and depend more on local neighbourhood context, rather than image-level context. The single channel input is converted into 3 channels by the initial projection layer with $1 \times 1 \times 1$ kernel, followed by $3 \times 3 \times 3$ convolution to get the initial filter channel depth of 64. The architecture consists of two consecutive $3 \times 3 \times 3$ convolutional layers followed by the $2 \times 2 \times 2$ max-pooling layer (in the case of encoder) or $2 \times 2 \times 2$ upsampling layer (in the case of decoder). We added a $1 \times 1 \times 1$ convolutional layer before the final softmax layer for predicting the probability maps.

6.4.2.2 Loss functions

Similar to WMH segmentation in chapter 5, we used a combination of cross-entropy (CE) and Dice loss functions as the total loss function. In the CE loss function, for higher scale 3D patches ($48 \times 48 \times 48$ and $64 \times 64 \times 64$), we weight the CMB voxels 10 times than that of the non-CMB voxels during training to compensate for the imbalance in the classes. Dice loss is based on the voxel-wise Dice similarity measure and aids in the accurate detection of edges and small CMBs in the patches.

6.4.3 Object-level CMB candidate discrimination using a 3D classification network

Since we aimed for high sensitivity in the initial screening step, we also detected several CMB mimics (e.g. penetrating blood vessels and small haemorrhages in the case of TICH2 dataset). Hence this step needs to be highly specific without reducing the number of true CMBs. We have shown the workflow of this step in figure 6.7a. As shown in the figure, we thresholded the probability map P_{cand} obtained from the screening at the threshold Th_{cand} of 0.7 (since the Dice loss function was trained for this threshold value). We extracted 3D patches of size $S_p = 24 \times 24 \times 24$ mm, by selecting an appropriate number of voxels based on the resolution. Note that, rather

than extracting patches with a specified number of voxels, as we did in the previous step, we selected the patches with a constant dimension in mm to standardise the patch dimensions (irrespective of voxel resolutions) to CMB size. This would help in rejecting the larger mimics (e.g. small haemorrhages, ends of the central sulcus and so on) and vessels (especially in the z-dimension) and would also make the model resistant to variations in voxel resolutions. We then predicted the probability P_{disc} of the patches containing CMBs using a 3D classification network. We applied a threshold Th_{disc} of 0.5 on P_{disc} for classifying CMB and non-CMB patches. We set the threshold empirically based on the performance (refer to the results in section 6.6.2.2 for more details). Due to the high number of FPs in the previous step, there was an extreme imbalance between the classes (number of FP patches at least 5-6 times that of true CMB patches). In order to account for this imbalance, we sampled twice as many non-CMB patches (sampled randomly from non-CMB brain areas) as the number of true manually segmented CMB patches for training.

6.4.3.1 3D classification network architecture

The architecture of the 3D CMB classification network is shown in figure 6.7b. Our classification network consists of a 3-layer deep encoder with two consecutive 3 $\times$ 3 $\times$ 3 convolutional layers followed by a pooling layer in each level of abstraction. The output of the third layer of the encoder is fed into dense fully connected layers (FC). Three fully connected layers with 256, 128 and 32 nodes are then followed by a softmax layer. We added dropout layers with *keep probabilities* of 80% and 50% before FC-256 and FC-32 respectively. We used cross entropy loss function for training.

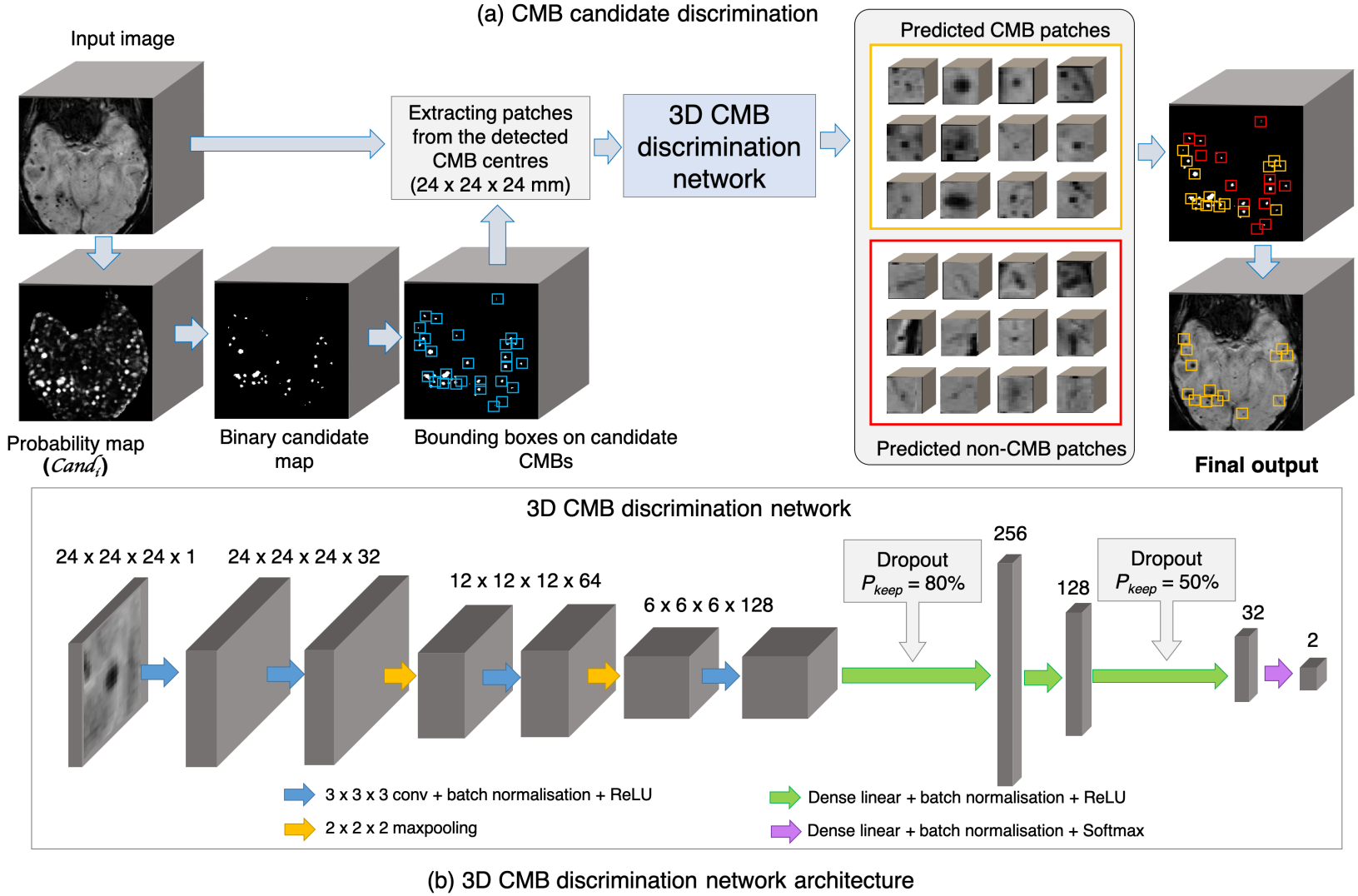


Figure 6.7: CMB candidate discrimination step. (a) Block diagram of the discrimination step and (b) 3D CMB discrimination network architecture.

6.4.4 Postprocessing for false positive reduction

We further reduced FPs by removing the stray noisy voxels and regions along edges of the brain. On the final output, we rejected candidates that are closer to the skull (< 3 voxels) and the spurious noise regions with volume less than a volume threshold V_{th} (set empirically based on the performance, values for various datasets specified in the results in section 6.6.2.2).

6.4.5 Experiment setup

6.4.5.1 Optimisation parameters

CMB screening network: We trained the network using the Adam Optimiser [354] with epsilon (ϵ) value of 1×10^{-4}. For all the scales, we used a batch size of 8, with an initial learning rate of 1×10^{-3} and reducing it by a factor 1×10^{-1} every 2 epochs, until it reaches 1×10^{-6}, after which we maintain the fixed learning rate value. We also empirically set the total number of epochs to 100. We observed that while the two higher scales ($48 \times 48 \times 48$ and $64 \times 64 \times 64$) usually converged in less than 30 epochs, the lower scales ($24 \times 24 \times 24$ and $32 \times 32 \times 32$) converged at around 50 epochs. We used the truncated normal variable (with $\sigma = 0.05$) for weights initialisation and the biases were initialised as constants with a value of 0.1. We chose the above parameters empirically using a trial-and-error method, starting from the best performing values observed for the WMH segmentation architecture (refer to section 5.4.1).

CMB classification network: We used the same optimisation parameters as those used for CMB screening network, and set the total number of epochs to 150. Also in this case, we chose the above parameters empirically based on model convergence.

6.4.5.2 Data augmentation

Due to the small size of CMBs, transformations such as rotation and down-scaling could result in the loss of CMBs in the augmented data patches. Hence, we chose our data augmentation carefully, with minimal interpolation of intensity values and

potential to inject variations in the data. We performed augmentation on the patches, increasing the dataset size by a factor of 10, using random combinations of the following transformations: translation, random noise injection and Gaussian filtering (with a small σ value). The parameters for the above transformations were chosen randomly from the ranges mentioned in table 6.1. We used similar augmentation for the discrimination step, increasing training data by a factor of 5.

Transformations	**Parameter ranges**
Translation	Lower two scales: x-offset: [-15, 15], y-offset: [-15, 15] Higher two scales: x-offset: [-25, 25], y-offset: [-25, 25]
Random noise injection	Distribution: Gaussian, $\mu = 0$, $\sigma^2 = [0.01, 0.04]$
Gaussian filtering	$\sigma = [0.1, 0.2]$

Table 6.1: Transformations used for data augmentation and their parameter ranges.

6.4.5.3 Implementation details

We implemented the networks in Python 3.6 using TensorFlow 1.11.0. The CMB candidate screening network was trained on a NVIDIA Tesla V100 GPU, taking around 25 mins, 35 mins, 60 mins and 120 mins for training scales 24, 32, 48 and 64 respectively, requiring a total training time of $\approx$ 250 mins for $\approx$ 25,000 training patches. For testing, each subject took an average testing time of $\approx$ 7 seconds (on GPU) for the prediction of P_{cand} map. The CMB candidate discrimination network was trained on NVIDIA GeForce GTX 1080 Ti GPU, taking $\approx$ 10 mins for 20,000 training patches. The average prediction time was $<$ 1sec for a subject. For both the networks, the training and validation split was set to 90% and 10% of samples respectively.

6.4.6 Ablation study: effect of individual scales in the CMB candidate screening step

We performed an ablation study on the TICH2 dataset to determine the effect of the individual scales on CMB detection in the candidate screening step. Hence,

we removed one scale (28, 32, 48) at a time and compared the performance of our candidate screening step. Finally, since the scale 64 is also involved in the refinement of local maxima, we removed the scale 64 and the associated local maxima refinement. Therefore in this case, we extracted other lower scales at M_{init}.

6.4.7 Initial evaluation on the TICH2, HK and OXVASC datasets

We performed the initial evaluation on a subset of images from different datasets to evaluate the performance of our method as a proof of concept, with limited computational resources. We performed 5-fold cross validation on 25 subjects from TICH2 (total CMBs: 185), 20 subjects from OXVASC (total CMBs: 214) and 20 subjects from HK (total CMBs: 89). We evaluated the performance based on the metrics specified in section 6.4.8. Additionally, we compared our performance with those of existing CMB detection methods in the literature.

6.4.8 Performance evaluation metrics for CMB detection

We evaluated the CMB detection results using the following metrics for total number of CMBs over the entire dataset, as done in existing work [294]-[297] (refer to appendix C):

- Cluster-wise TPR
- Average number of FPs per subject (FPavg)
- Cluster-wise precision

We used TPR and FPavg values for plotting a free-response ROC (FROC) curve, which is a plot of TPR versus the average number of false positives per image/subject.

6.5 Automated CMB characterisation and rating

6.5.1 CMB size estimation

From the clinical point of view, it is desirable to obtain the size of CMBs (in mm) and distribution of CMBs in the individual regions of the brain. For example, determining the size of CMBs would help us in discriminating microbleeds (< 5 mm) from macrobleeds (larger haematomas of size 5-10 mm) [27]. Microbleeds and macrobleeds are shown to represent two separate categories of haemorrhagic events, with thicker amyloid-positive vessel fragments leading to high microbleed counts in subjects with CAA [364]. However, this is still an open area of research, and determining the size of CMBs would help us in understanding the clinical impacts of microbleeds and macrobleeds in large datasets such as UKBB. To obtain the exact size of CMBs, within the bounding box of the detected CMBs, we determined the major-axis length (the longest diameter of an ellipsoid within the bounding box) of the individual CMBs (in voxels) and later multiplied them with voxel dimensions to obtain the final diameters in mm.

6.5.2 Automated MARS rating of CMBs

In table 6.2, we list the regions used in the MARS and BOMBS scales. As can be observed from the table, the scales are quite similar, providing CMB ratings in the deep, lobar and infratentorial regions of the brain. The only differences are the fact that the MARS scale also provides CMB ratings in the individual lobes and that the BOMBS scale classifies subcortical white matter as part of the lobar region, while in the MARS scale, the white matter is classified as a part of the deep region. Out of the two rating scales, we focus on the MARS rating scale for this work, since we have manual MARS ratings available for the TICH2 dataset. Given the similarity of the two scales, the equivalent ratings according to the BOMBS scale would be quite straightforward to obtain. Our method for automatic CMB rating consists of two steps: (1) extracting the regions (e.g. infratentorial) and structures (e.g. brainstem) from MRI atlases and (2) counting the number of CMBs within each structure/region.

BOMBS scale		MARS scale		Atlases
Infra-tentorial	Brainstem	Infra-tentorial	Brainstem	Harvard-Oxford subcortical structural atlas [365, 366] (thr - 50)
	Cerebellum		Cerebellum	MNI structural atlas [367, 368]
Deep	Basal ganglia*	Deep	Basal ganglia*	Harvard-Oxford subcortical structural atlas [365, 366] (thr - 50)
	Thalamus		Thalamus	
	Internal capsule		Internal capsule	JHU ICBM-DTI-81 white-matter labels atlas [369, 370, 371]
	External capsule		External capsule	
	Corpus callosum		Corpus callosum	
Lobar	Sub-cortical white matter		White matter (deep)	Harvard-Oxford subcortical structural atlas [365, 366] (thr - 60)
	Cortical lobes	Lobar	Frontal lobe	MNI structural atlas [367, 368]
			Parietal lobe	
			Temporal lobe	
			Occipital lobe	
			Insular lobe	

Table 6.2: CMB rating scales BOMBS and MARS, with the atlases used for the individual structures. In case a probabilistic atlas was used, the threshold (thr) applied is reported in brackets. *Basal ganglia = caudate + lentiform nuclei.

Table 6.2 reports the structures considered in the MARS scale (also shown in the sample MARS rating form provided in appendix D) and the atlases in the MNI space used for obtaining individual structures. While some of the structures were already available as binary masks, we obtained a few structures from probabilistic atlases, which we thresholded (values specified in the table) to obtain the final binary masks for individual structures. We later registered the structures from the MNI space to the subject-space using linear registration (unsuitable for non-linear registration in

FSL due to the lack of a T1-weighted image) with FSL-FLIRT [337] and thresholded them at 0.5 to obtain binary masks in the single-subject space. We performed a morphological dilation operation (using a structuring element with the size of 3 voxels) on the cortical lobes mask, in order to increase the extent of cortex to include the gray-white matter junctions as 'lobar' region. Also, deep structures such as internal/external capsules and corpus callosum were excluded from the white matter mask since they were considered separately. Figure 6.8 shows sample MARS rating regions from the structural atlases in the MNI space and in the subject space.

We then determined the voxel coordinates corresponding to the centroids of CMBs and, within each structure, we counted the number of centroids to obtain the structure-wise count of CMBs. We finally aggregated the number of CMBs to provide region-wise MARS score.

6.5.3 Characterisation and MARS rating of CMBs on the TICH2 dataset

We obtained the distribution of size for the detected CMBs on the TICH2 dataset, and compared it with the one obtained from manually segmented CMBs. Similarly to what we did for CMB segmentation, in this case we also considered '*definite*' and '*possible*' CMBs together, even if they were classified separately in the MARS ratings. We initially obtained the MARS rating of CMBs from the manually segmented CMB maps to evaluate the accuracy of the automated MARS rating method based on the metrics specified in section 6.5.4. Later, we obtained MARS rating values for the CMB detection results, and studied the performance of the CMB detection in terms of outline errors of CMB segmentation (in case of true positive CMBs) and detection errors (in case of false negatives (FNs) and FP CMBs).

6.5.4 Performance evaluation metrics for automated rating of CMBs

For the evaluation of automated MARS rating of CMBs, we used the following metrics:

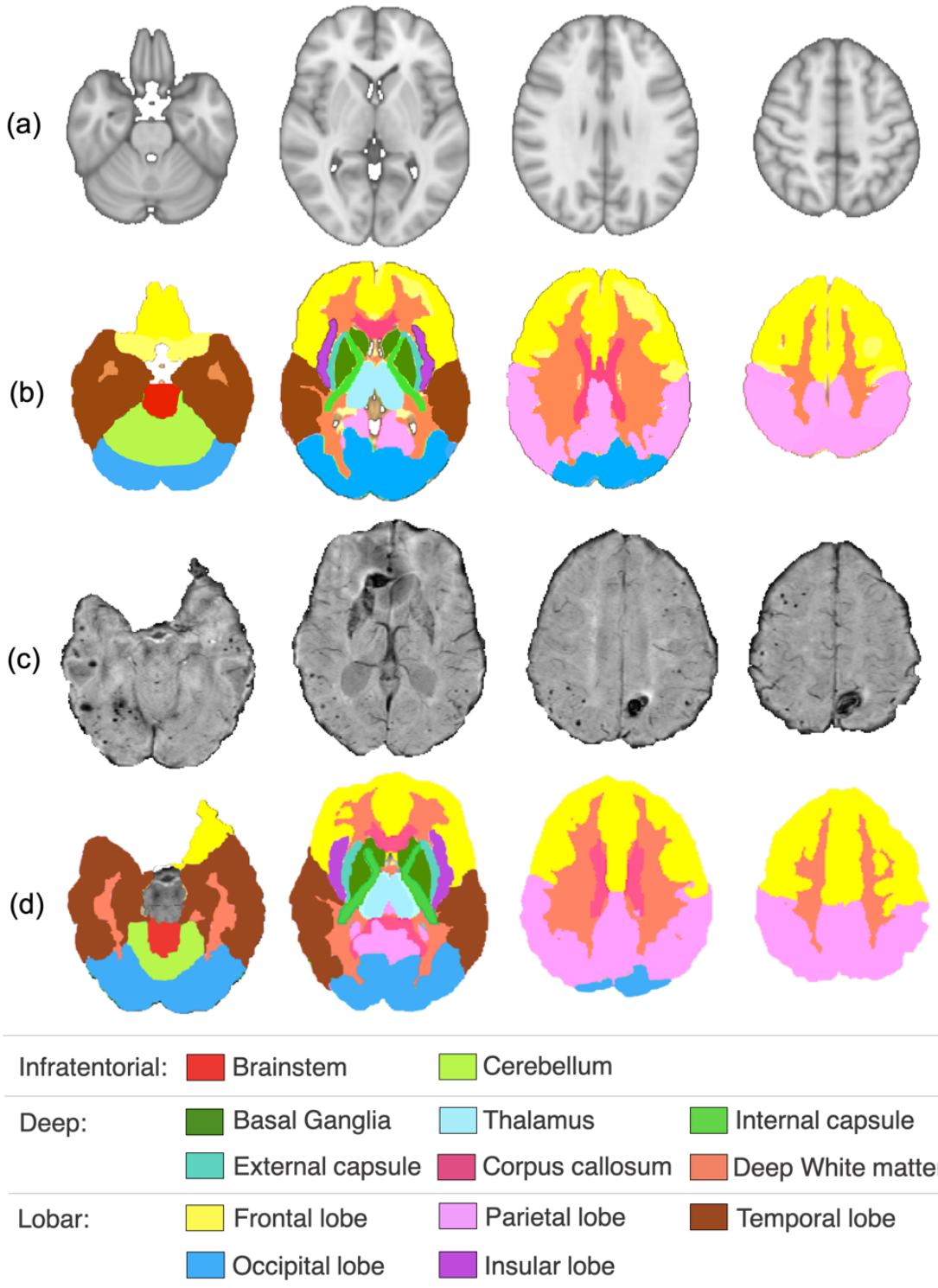


Figure 6.8: MARS rating regions from the structural atlases. (a) MNI T1-weighted structural image, (b) regions for MARS rating formed by using atlases in the MNI space, (c) SWI in the subject space, (d) MARS rating regions formed by registering atlases to the subject space. Legend for various structures shown at the bottom panel (best viewed in colour).

- Region-wise sensitivity
- Region-wise precision
- Region-wise F1-measure
- Accuracy

Definitions of the above evaluation metrics are provided in appendix C.

6.6 Results

6.6.1 Evaluation of CMB candidate subjects preselection pipeline

6.6.1.1 Effect of threshold value for obtaining CMB candidates

Figures 6.9a and b show FROC curves on the TICH2 and OXVASC datasets for determining the threshold value th_{prob} for the voxel-wise initial CMB candidate detection (step 2) of the OXVASC-trained and TICH2-trained pipelines respectively.

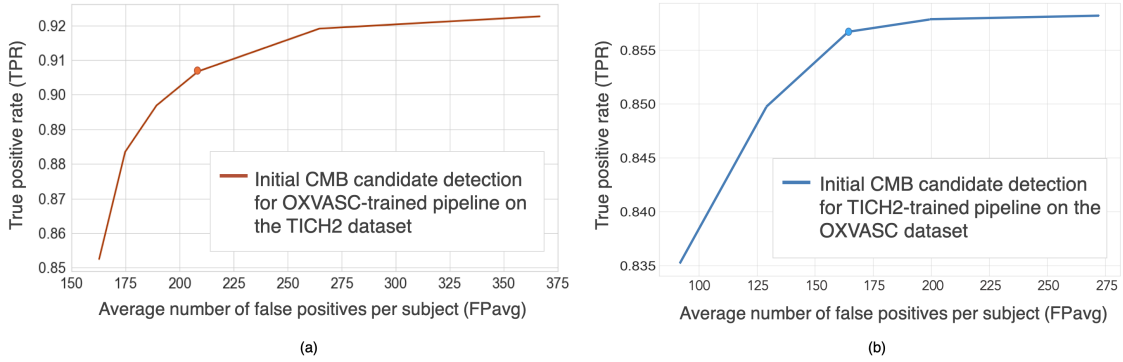


Figure 6.9: Free-response ROC (FROC) curves for initial CMB candidate detection for (a) OXVASC-trained pipeline on the TICH2 dataset and (b) TICH2-trained pipeline on the OXVASC dataset. The circular markers indicate the points at which we achieve the best compromise between TPR and FPs.

Since we will be applying, at a later stage, the shape-based filtering step (step 3) to further reduce the false positives and improve the specificity, we prioritised achieving high cluster-wise TPR in this initial CMB candidate detection (step 2).

For both the cases, we achieved the best performance at th_{prob} value of 0.8. The SVM classifier trained on the OXVASC dataset achieved a cluster-wise TPR of 0.906 with FPavg of 210.6 on the TICH2 dataset, while the SVM trained on the TICH2 dataset cluster-wise TPR of 0.86 with FPavg of 178.7 on the OXVASC dataset.

6.6.1.2 Case 1: Pipeline trained on OXVASC and evaluated on the TICH2 data

Figure 6.10a shows the ROC curve for the prediction of CMB subjects at various values of the threshold Th_{NCMB} on the number of detected CMBs for individual subjects. We achieved the best performance (indicated by the orange circular marker) at Th_{NCMB} value of 35 CMBs. Figure 6.10b shows the confusion matrix for true and predicted CMB subjects at the best performance point of the ROC curve. We achieved a subject-level TPR of 0.82, a subject-level specificity of 0.69 and a subject-level accuracy of 74%.

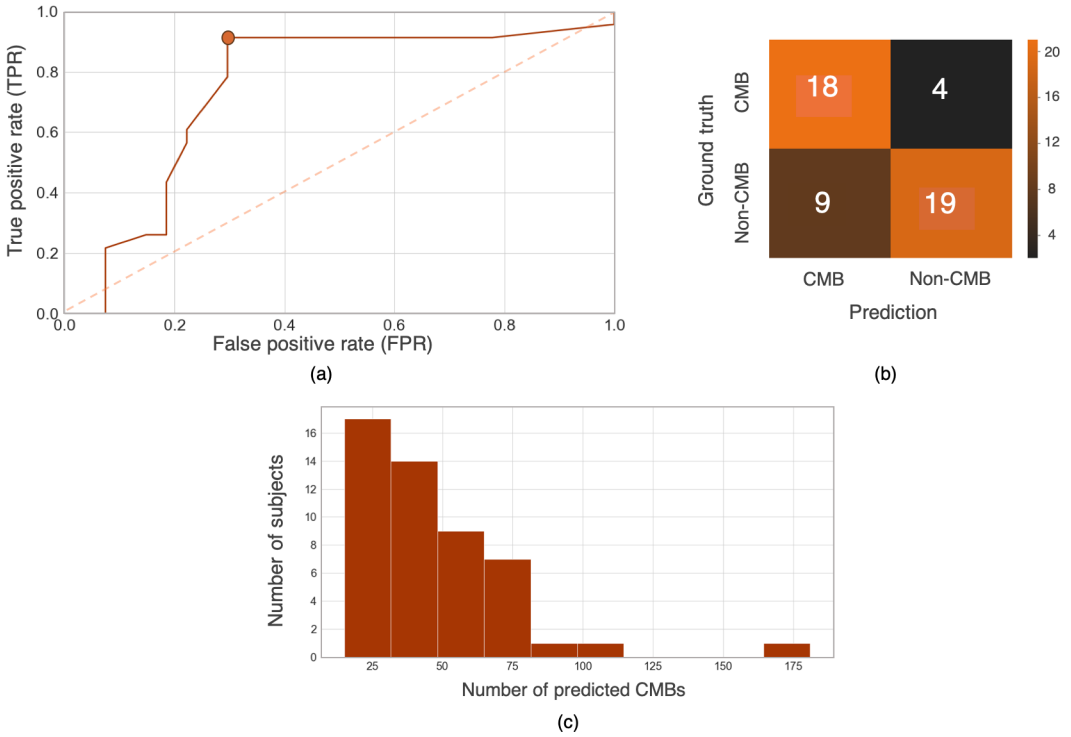


Figure 6.10: Results of the preselection of CMB subjects using OXVASC-trained model on the TICH2 dataset. (a) ROC curve for the preselection of CMB subjects with a circular marker indicating the best performance point, (b) confusion matrix at the best performance point and (c) histogram of CMB counts per subject predicted by the OXVASC-trained pipeline.

Figure 6.10c shows the histogram of CMB counts for individual subjects predicted by the OXVASC-trained model on the TICH2 data. The non-smooth nature of ROC curve, despite the subject-level sensitivity of 0.82, could be due to the non-Gaussian shape of histogram and availability of fewer data (i.e. 50 subjects).

Figure 6.11 shows a few example cases of correct and false subject-level detections on the TICH2 dataset (manually segmented CMBs are indicated with green boxes). Despite the presence of haemorrhagic lesions in all subjects, the pipeline gave a sensitivity of 0.82 and an accuracy of 74%. From the figure, it can be seen that our algorithm correctly predicted the subjects with a high number of CMBs (figure 6.11a,b) as CMB candidate subjects. Similarly, if the subjects had small haemorrhages and did not have CMBs (figure 6.11c, d), the pipeline predicted very low number of CMBs (well under Th_{NCMB} value of 35 CMBs). The border-line cases, for example the subjects with 1-2 CMBs (as shown in figure 6.11e), were at times incorrectly predicted as non-CMB subjects by the pipeline. Regarding the causes of FPs in this dataset (figure 6.11e, f), small fragments of haemorrhages which are not part of the big lesions produced most of the false positive CMBs.

6.6.1.3 Case 2: Pipeline trained on TICH2 and evaluated on the OXVASC data

Figure 6.12a shows the ROC curve for CMB candidate subject preselection by TICH2-trained model on the OXVASC dataset. The confusion matrix corresponding to the best performance point (indicated by blue circular marker) is shown in figure 6.12b. As in the previous case, the model achieved the best performance for the threshold value Th_{NCMB} of 35 CMBs. The model provided a subject-level TPR of 0.74, a subject-level specificity of 0.82 and a subject-level accuracy of 78%.

Compared to the previous case (figure 6.10c), the number of CMBs in the individual subjects is generally lower in this dataset (figure 6.12c). The number of subjects with high CMB counts ($>$ 50 CMBs) in the training dataset (TICH2) is higher than that in OXVASC. Also, most of the images in the OXVASC dataset are borderline cases (containing less than 5 CMBs) and resemble the non-CMB cases of TICH2.

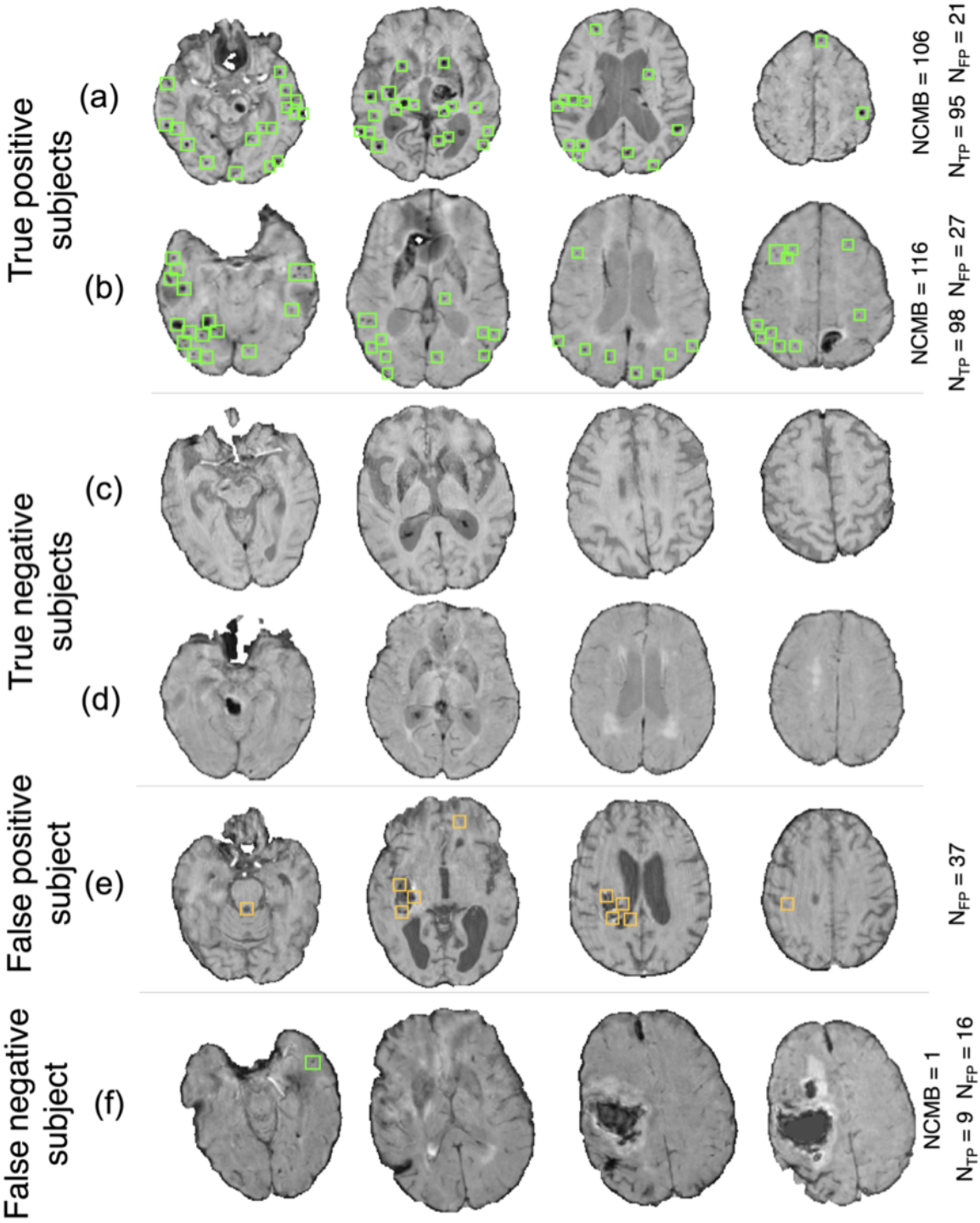


Figure 6.11: Examples of CMB subject preselection OXVASC-trained pipeline results. (a) and (b) true positive CMB subjects, (c) and (d) true negative non-CMB subjects, (e) false positive prediction of non-CMB subject and (f) false negative prediction of CMB subject. The green and orange boxes indicate the manually segmented CMBs and false positives respectively. The CMB counts $NCMB$ are provided, along the number of true positives (N_{TP}) and false positives (N_{FP}).

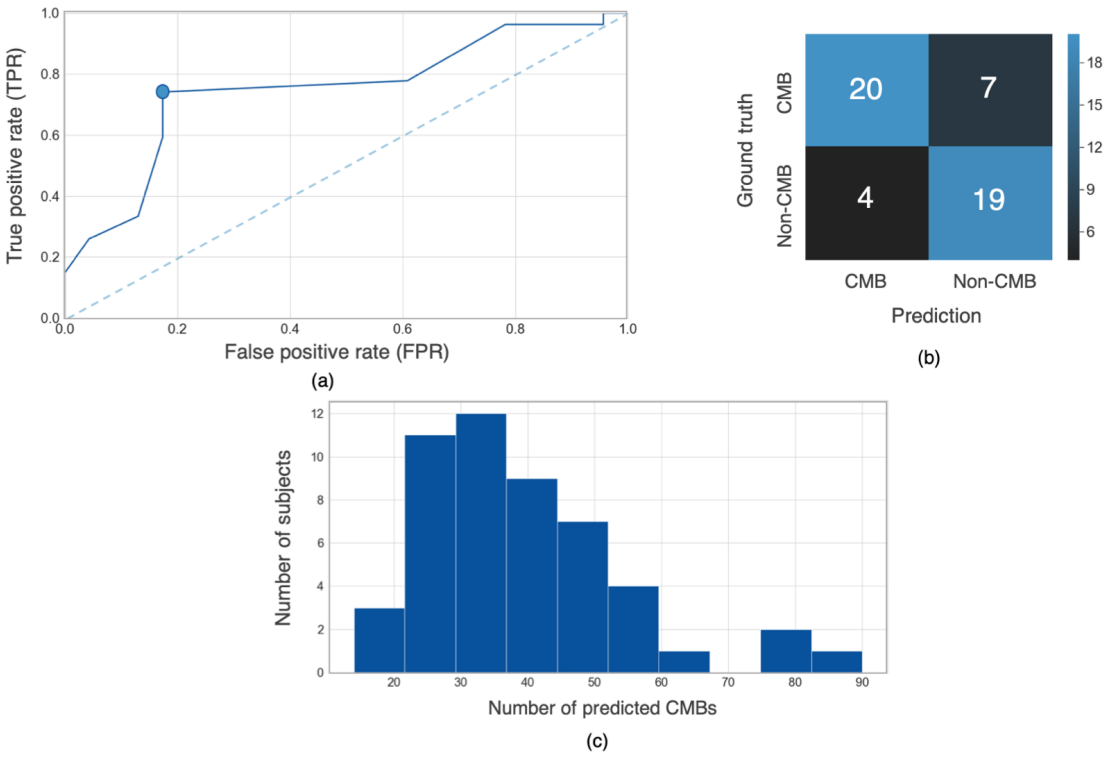


Figure 6.12: Results of the preselection of CMB subjects using TICH2-trained model on the OXVASC dataset. (a) ROC curve for the preselection of CMB subjects with a circular marker indicating the best performance point, (b) confusion matrix at the best performance point and (c) histogram of CMB counts per subject predicted by the pipeline.

Hence, the TICH2-trained model was more specific in detecting CMB subjects on the OXVASC dataset, achieving a higher specificity of 0.82.

We have shown a few examples of true positive and true negative detections and an example each for FPs and FNs in figure 6.13. As in the previous case, subjects with high CMB counts (figure 6.13a and b) were correctly predicted as CMB subjects. Also, the pipeline detected more true positive CMB subjects than the previous case, even though most of them were in the borderline (refer to the CMB distribution histogram in figure 6.1a). The pipeline correctly identified the non-CMB subjects, despite the poor image quality, especially in the top slices (figure 6.13c and d). The false detections were mainly due to the partial volume effect in the tissue-CSF interfaces, intensity variation nearer to the skull and vessel fragments remaining from the vessel removal step.

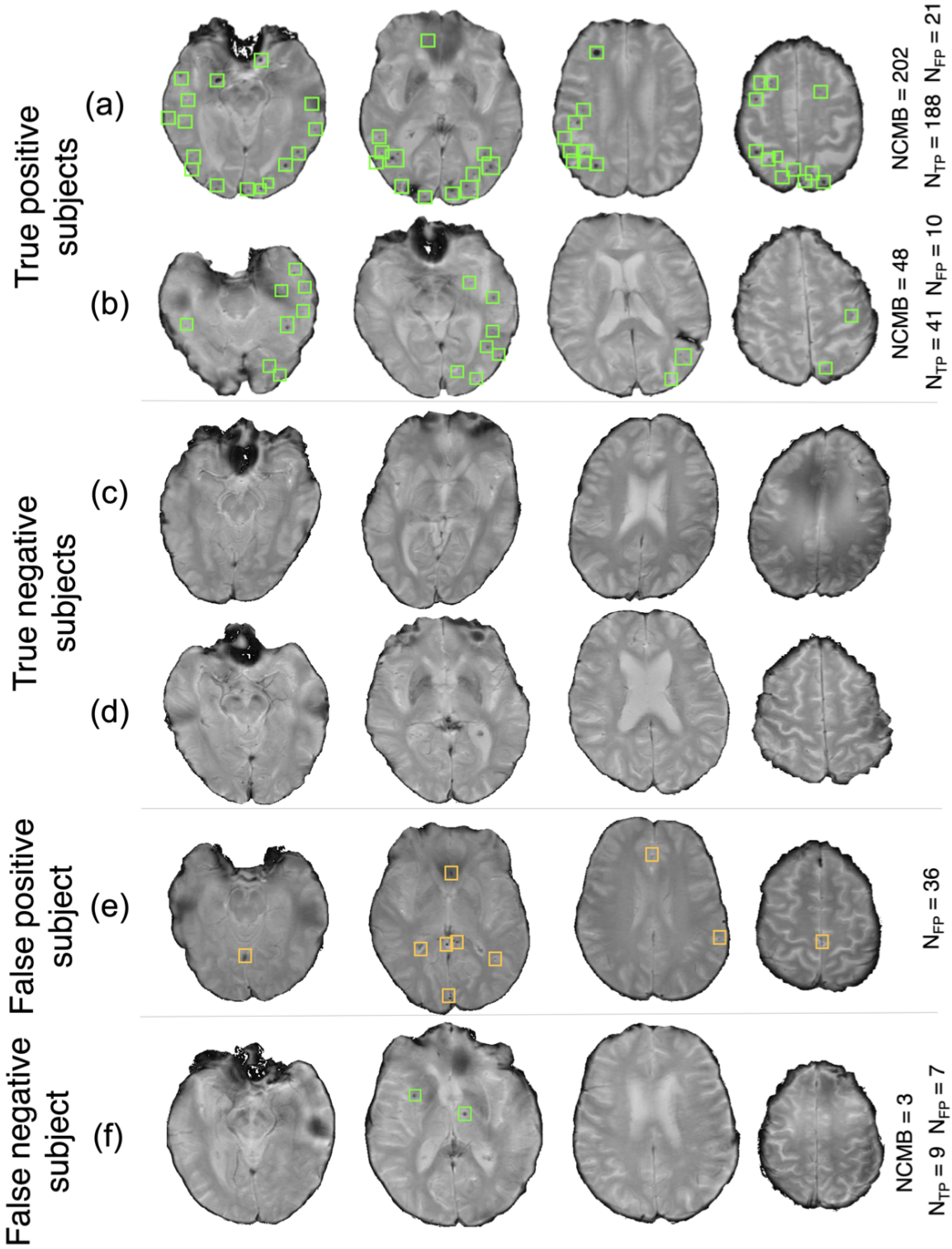


Figure 6.13: Examples of CMB subject preselection TICH2-trained pipeline results. (a) and (b) true positive CMB subjects, (c) and (d) true negative non-CMB subjects, (e) false positive prediction of non-CMB subject and (f) false negative prediction of CMB subject. The green and orange boxes indicate the manually segmented CMBs and false positives respectively. The CMB counts $NCMB$ are provided, along the number of true positives (N_{TP}) and false positives (N_{FP}).

6.6.1.4 Case 3: Application on the UKBB dataset

We applied both the OXVASC-trained and the TICH2-trained models on the UKBB-I dataset and set the value of threshold Th_{NCMB} to 35 CMBs since it gave the best performance values for both cases 1 and 2. Out of the 180 initial subjects, the OXVASC-trained pipeline identified 57 (31%) subjects with CMBs, while the TICH2-trained pipeline identified 52 (28%) subjects, with an overlap of 42 (74% and 81% of those identified in cases 1 and 2 respectively) predicted CMB subjects between both the cases. Figure 6.14 shows boxplots of CMB counts predicted by the OXVASC-trained model for 3 age groups (younger, mid and older) and two levels of BP (low, high). Similar boxplots are shown for the TICH2-trained model in figure 6.15.

As shown in figure 6.14a and b, CMB counts increased with age and BP. On performing a one-way ANOVA across 3 age groups, we found that the CMB counts were significantly different among age groups ($F = 14.46$, $p < 0.0001$). Both mid and older age groups had significantly higher CMB counts when compared to the younger age group ($p < 0.0001$, figure 6.14a). Similarly, the two-sample t-test showed that the CMB counts were significantly different ($p = 0.02$) between the 2 BP levels.

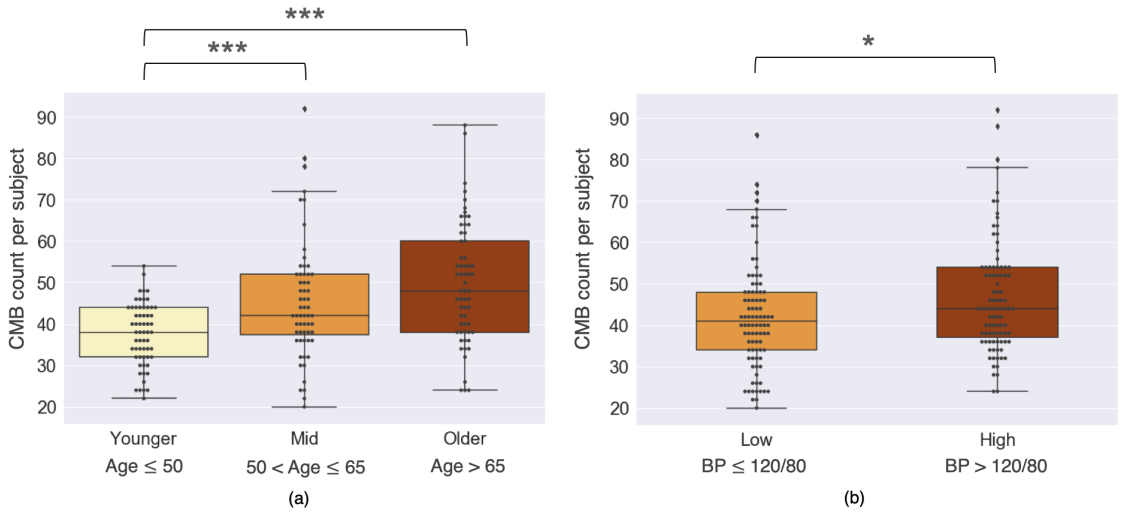


Figure 6.14: Boxplots of CMB counts in individual subjects predicted by the OXVASC-trained model against three age groups (a) and two levels of BP (b) on the UKBB-I dataset. *$p < 0.05$, **$p < 0.001$ and ***$p < 0.0001$ (after Bonferroni correction).

Similarly, when the pipeline was trained using the TICH2 data, the one-way ANOVA test across the 3 age groups shows that the CMB count was significantly different between the age groups ($F = 24.12$, $p < 0.0001$). As in the case of OXVASC-trained model, both mid and older age groups showed significantly higher CMB counts when compared to the younger age group ($p < 0.0001$, figure 6.15a). Also, the CMB counts for high BP were significantly higher than for the low BP subjects ($p = 0.02$). The increase of CMB count with age and BP suggests the potential use of various demographic and clinical factors in shortlisting the candidates. Hence, for the further application of our pipeline, we select a second larger subset (UKBB-II) of UKBB by applying demographic/clinical factors as the shortlisting criteria for our preselection pipeline.

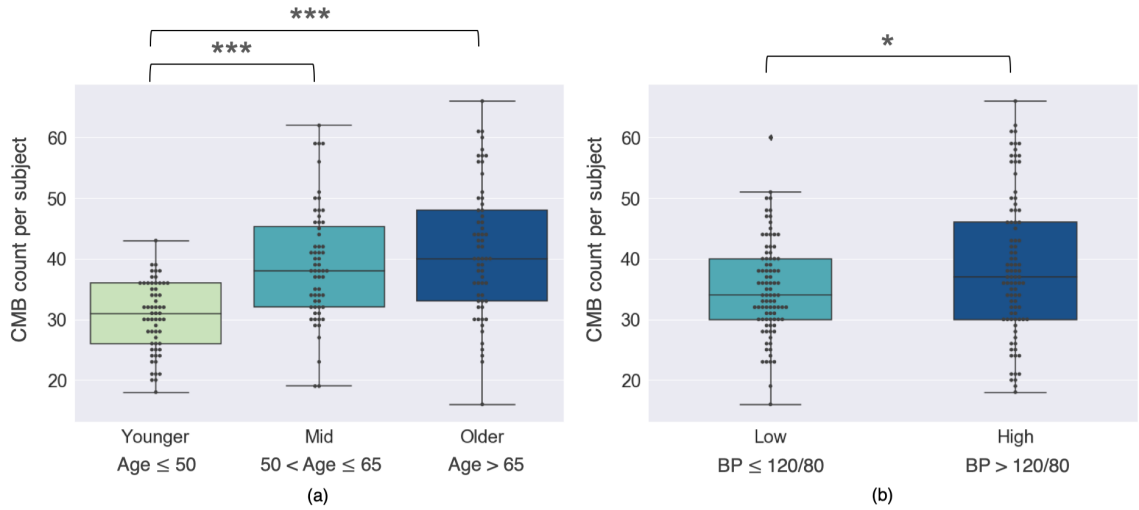


Figure 6.15: Boxplots of CMB counts in individual subjects predicted by the TICH2-trained model against three age groups (a) and two levels of BP (b) on the UKBB-I dataset. *$p < 0.05$, **$p < 0.001$ and ***$p < 0.0001$ (after Bonferroni correction).

For **UKBB-II**, we considered initially 14,521 subjects (out of $\approx$ 20,000 subjects from the January 2018 release of UKBB) having the required data fields. We later selected subjects older than 74 years (corresponding to 95^{th} percentile of subject ages in UKBB, and is also the widely reported age limit above which CMBs occur [25, 97]) and obtained 751 subjects. Out of them, only 656 subjects had SWI scans available, of which 465 subjects were selected after discarding the ones with missing/incomplete

data. We applied the demographic/clinical factors that have been associated with CMBs (specified in table 6.3) on the above 465 subjects and obtained 287 subjects passing the criteria. Hence, UKBB-II consists of 287 subjects with age range 74.1 - 80.3 years, mean age 75.6 ± 1.04 years, median age 75.4 years, female to male ratio F:M = 106:181 and blood pressure (BP) range: 102/50 - 221/106 mmHg, mean BP 141.3/76.6 ± 18.7/10.5 mmHg, median 139/76 mmHg.

Field IDs	**Factors**	**Criteria**
4079	Diastolic blood pressure	$>$ 90 mmHg
4080	Systolic blood pressure	$>$ 140 mmHg
20023	Mean reaction time (cognition)	$>$ 700 ms
20116	Smoking status	Yes
25781	Total volume of WMH	$>$ 10,000 mm^3
42008	Ischemic stroke	Yes

Table 6.3: Demographic/clinical factors and criteria used for shortlisting the potential candidates for applying our subject-level preselection pipeline. The values used as criteria for mean reaction time and total volume of WMHs correspond to $\approx 90^{th}$ percentile in their distribution.

We applied both cases of the pipeline on the UKBB-II dataset. Out of 287 subjects in the UKBB-II dataset, OXVASC-trained and TICH2-trained pipelines predicted 201 (70%) and 195 (68%) CMB candidate subjects respectively, with an overlap of 182 predicted CMB subjects (90% and 93% of those identified in OXVASC-trained and TICH2-trained pipelines respectively).

It can be observed that the proportion of CMB subjects predicted on UKBB-II by the pipeline in both cases (70% and 68%) were higher than those in UKBB-I (31% and 28%). As hypothesised, shortlisting based on age and BP considerably increased the possibility of higher proportion of CMB subjects. In fact, most of the predicted CMB subjects from the UKBB-I dataset were from its older, hypertensive subset. Out of 180 subjects in the UKBB-I dataset, for 60 subjects with high age and BP (age $>$ 65 and BP $>$ 120/80), we observed that OXVASC-trained pipeline and TICH2-trained predicted 41 (68%) and 39 (65%) CMB subjects respectively, with an overlap of 34 predicted CMB subjects (83% and 87% of those identified in OXVASC- and

TICH2-trained pipelines respectively). Hence, with the application of criteria on age and BP, the percentages of CMB candidate subjects preselected remained consistent between the UKBB-I and II datasets. This suggests that utilising the demographic and clinical factors would indeed be useful in further improving the efficiency of CMB candidate subject preselection in the large datasets.

6.6.2 Evaluation of CMB detection

6.6.2.1 Ablation study: effect of individual scales in CMB candidate screening step

Figure 6.16 shows the FROC curves obtained for the cases where individual scales were removed in the ablation study on the TICH2 dataset. Table 6.4 reports the results with the removal of individual scales, compared with the case when all the scales are used to get the CMB candidates at the screening step. From the figure and the table, it can be seen that the number of false positives decreased with the removal of the lowest scale of 24, although with a corresponding decrease in cluster-wise TPR. For the case where scale 64 was removed, the performance was much lower than the rest of the cases with more false positives and low cluster-wise TPR due to the additional removal of local maxima refinement step, which uses scale 64. The initial maxima, M_{init}, were in general less accurate and were prone to detect spurious peaks when compared to M_{final}. Hence we observed more false positives and less sensitivity for this case (without scale 64). Overall, all the cases in the ablation study led to lower performance than the one obtained with the combination of all scales.

6.6.2.2 Evaluation on the TICH2, HK and OXVASC datasets

Figure 6.17 shows FROC curves for the 5-fold validation at various stages of the CMB detection methods on the TICH2, OXVASC and HK data subsets. The corresponding performance values are listed in table 6.5. We achieved the best performance (indicated by markers on the FROC curves) for thresholds of 0.7 and 0.5 on P_{cand} (for the candidate screening step) and P_{disc} (for the discrimination step) respectively, as specified in section 6.4. For the postprocessing step, the best performance was achieved

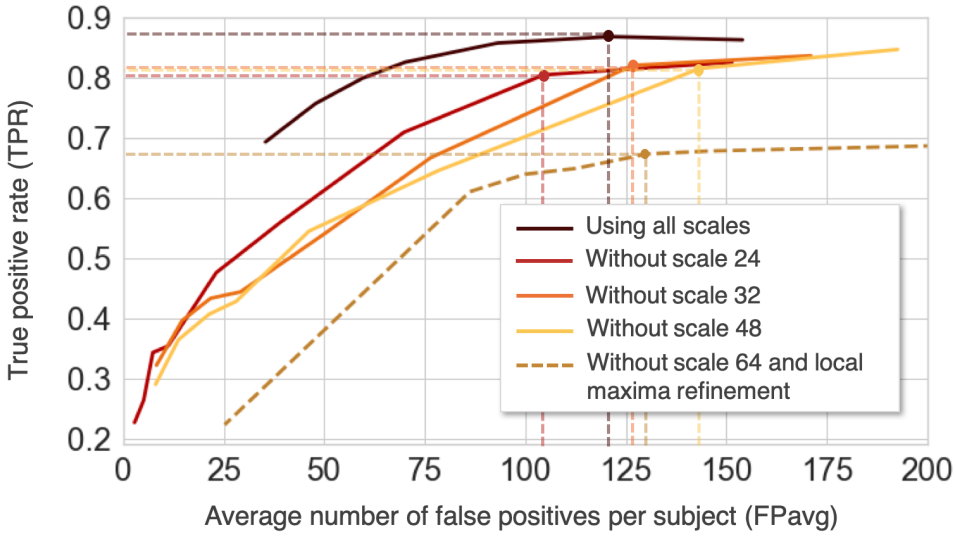


Figure 6.16: FROC curves for the ablation study on the TICH2 dataset by removing individual scales in the CMB candidate screening step.

Test cases	**Cluster-wise TPR**	**FPavg**
Without scale 24	0.80	105.4
Without scale 32	0.82	126.7
Without scale 48	0.82	143.4
Without scale 64 and local maxima refinement	0.67	130.2
CMB candidate screening with all scales	**0.87**	**121.07**

Table 6.4: Results of test cases considered in the ablation study on the TICH2 dataset to observe the effect of individual scales in initial CMB candidate screening step.

for the volume threshold V_{th} of 3 voxels for the TICH2 dataset and 5 voxels for the OXVASC and HK datasets. From the figure and the table, the OXVASC dataset achieved the best performance in all the steps. We obtained the highest number of false positives with low sensitivity for TICH2 at the final stage. For all datasets, with each step we observed a noticeable decrease in the number of false positives, although with a small decrease in the cluster-wise TPR values.

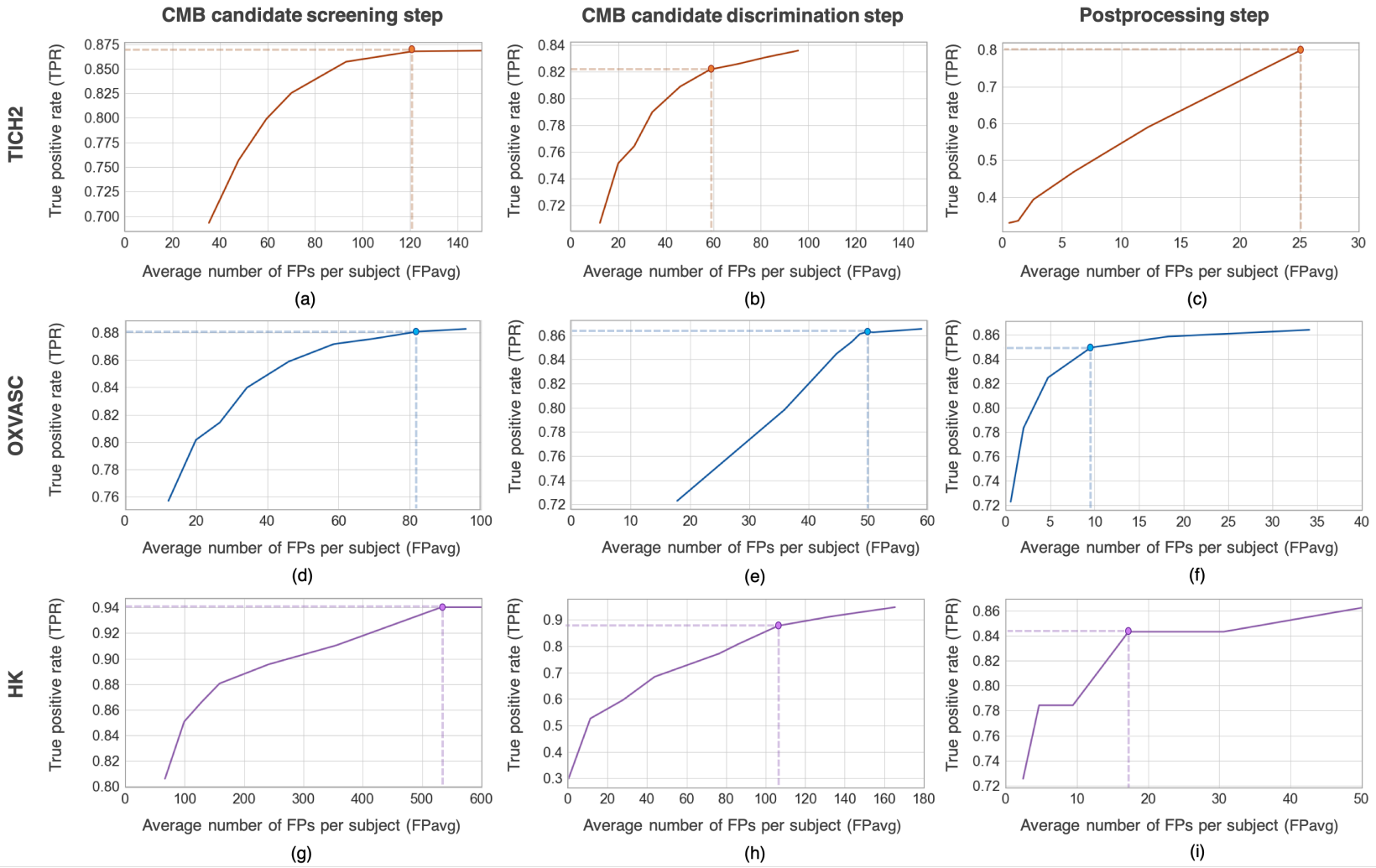


Figure 6.17: FROC curves for the CMB detection on the TICH2, OVXASC and HK datasets. Left to right: FROC curves for candidate screening step (a,d,g), candidate discrimination step (b,e,h) and the postprocessing step (c,f,i) for TICH2 (top), OXVASC (middle) and HK datasets (bottom). Markers shown to indicate the best performance values in each step.

Figure 6.18 shows a few examples of visual results at various steps of CMB detection for TICH2, OXVASC and HK datasets respectively. In general, the FNs were mainly subtle CMBs located very close to either the skull or the vessels (and therefore removed because they were misclassified as vessel-like structures). The OXVASC dataset has the least noisy images (due to less susceptibility effect in T2*-weighted GRE images) out of all the datasets, and hence shows fewer false positives compared to the other two datasets. In the case of TICH2, edges of haemorrhages and small blood vessels were detected as false positives, while in the case of the HK dataset, penetrating blood vessels, troughs of sulci and choroid plexuses in the ventricles were the main source of FPs. It is worth noting that a few false positives were also due to the oversegmented parts of true CMBs present in the contiguous slices above/below (figure 6.18(2)).

Step	**Dataset**	**Cluster-wise TPR**	**FPavg**	**Cluster-wise precision**
CMB candidate screening	TICH2	0.87	121.07	0.05
	OXVASC	0.88	82.04	0.10
	HK	0.94	532.70	0.01
CMB candidate discrimination	TICH2	0.82	58.64	0.09
	OXVASC	0.86	49.60	0.16
	HK	0.88	106.41	0.04
Postprocessing for FP reduction	TICH2	0.80	25.01	0.19
	OXVASC	**0.85**	**9.56**	**0.49**
	HK	0.84	17.25	0.18

Table 6.5: Performance of the CMB detection method at various steps for the TICH2, OXVASC and HK datasets (the best performance values highlighted in bold).

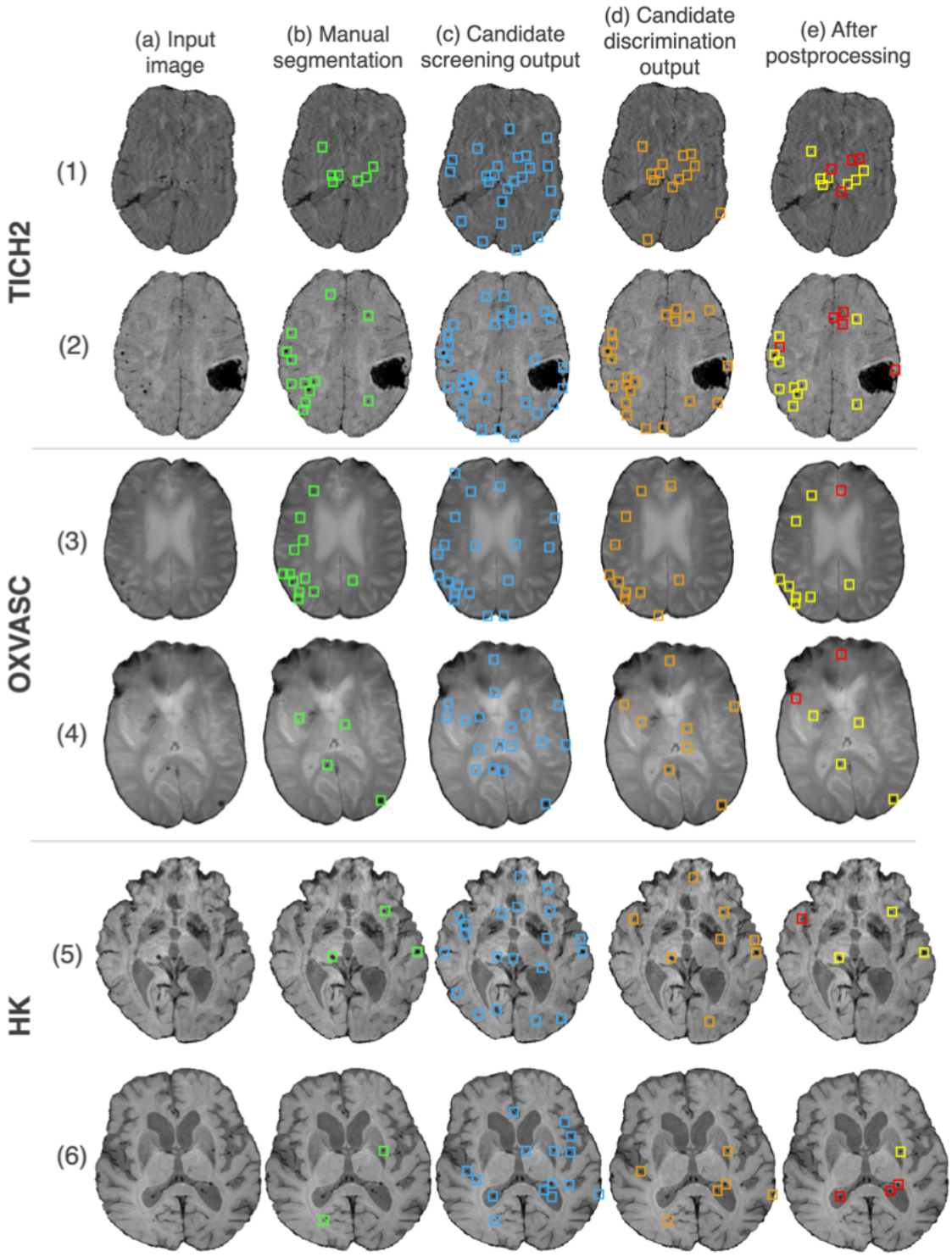


Figure 6.18: Examples of CMB detection results on the TICH2 (1,2), OXVASC (3,4) and HK (5,6) datasets. (a) Input image, (b) manually detected CMBs (shown in green), (c) CMB candidates from the screening step (blue), (d) CMB candidates after discrimination (orange) and (e) the final true positive CMBs (yellow) after postprocessing, along with the FPs (red).

Methods	Datasets		Performance		
	Sequence(s) (# test subjects)	**Total # CMBs**	**CMB screening step**	**CMB discrimination step**	**FP reduction step**
ML methods					
Bian et al., [294]	SWI (10)	304	FPavg - 287.7		Sens - 0.865, FPavg - 44.9
Fazlollahi et al., [295]	SWI (41)	103	Sens - 0.98, FPavg - 695	Sens - 0.92, $FPavg_{CMB}$ - 6.7 $FPavg_{nCMB}$ - 16.8	
Fazlollahi et al., [296]	SWI (66)	231	Sens - 0.97, $FPavg_{D}$ - 669, $FPavg_{P+D}$ - 1373	Sens - 0.87, $FPavg_{D}$ - 10.28, $FPavg_{P+D}$ - 27.8	
Ghafaryasl et al., [298]	T2*-GRE + PD (81)	183	Sens - 0.98, FPavg - 705	Sens - 0.92, FPavg - 19	Sens - 0.91, FPavg - 4.1
Dou et al., [297]	SWI (19)	161	Sens - 1.00, FPavg - 807	Sens - 0.94, FPavg - 190	**Sens - 0.8, FPavg - 7.7, Prec - 0.49**
DL methods					
Chen et al., [300]	SWI (5)	55	Sens - 0.96, FPavg $>$ 800	Sens - 0.89, FPavg - 6.4, Prec - 0.56	
Dou et al., [301]	SWI (50)	117	Sens - 0.98, FPavg - 282	**Sens - 0.93, FPavg - 2.74, Prec - 0.44**	
Chen et al., [302]	3D SWI patches (12)	377			At patch-level: Sens - 0.94, FPavg - 11.7, Prec - 0.72
Liu et al., [278]	Phase + SWI (41)	168	Sens - 0.99, FPavg - 276.8		Sens - 0.96, FPavg - 1.8
Proposed method (5-fold validation)	TICH2 - SWI (25)	185	Sens - 0.87, $FPavg_{P+D}$ - 121.1, Prec - 0.05	Sens - 0.82, $FPavg_{P+D}$ - 58.64, Prec - 0.09	Sens - 0.80, $FPavg_{P+D}$ - 25.01, Prec - 0.19
	OXVASC - T2*-GRE (20)	214	Sens - 0.88, FPavg - 82.04, Prec - 0.10	Sens - 0.86, FPavg - 49.60, Prec - 0.16	**Sens - 0.85, FPavg - 9.6, Prec - 0.49**
	HK - SWI (20)	89	Sens - 0.94, FPavg - 532.7, Prec - 0.01	Sens - 0.88, FPavg - 106.41, Prec - 0.04	**Sens - 0.84, FPavg - 17.2, Prec - 0.18**

Table 6.6: Comparison of the performance of the proposed CMB detection method with the existing methods. Sens - sensitivity, FPavg - Average false positives per image/subject, Prec - precision, $FPavg_{CMB}$ - FPavg for CMB subjects, $FPavg_{nCMB}$ - FPavg for non-CMB subjects, $FPavg_{D}$ - FPavg for '*definite*' CMB subjects, $FPavg_{P+D}$ - FPavg for '*definite and possible*' CMB subjects.

6.6.2.3 Comparison with existing methods

Table 6.6 provides a comparison of our CMB detection method with the existing automated methods. From the table, generally DL-based methods provided better performance when compared with conventional ML methods. While the methods using multiple modalities [278, 298] and patch-wise evaluation [302] provided better results, the proposed method gave comparable results with those of [297], on T2*-weighted GRE images from the OXVASC dataset. Also, our precision values on the OXVASC dataset is better than most of the methods including [301, 294, 295, 296]. Also, we obtained better precision than most of the methods including [301] and [302]. In addition, our method provided a lower FPavg value for 'definite and possible' cases on the TICH2 dataset, when compared with [296].

6.6.3 Characterisation of CMBs in the TICH2 dataset

6.6.3.1 Evaluation of CMB size estimation

Figure 6.19(a-c) shows the histograms of sizes (maximum diameter in mm) for manually segmented CMBs and detected true positive CMBs, along with the histogram of the absolute difference in sizes between them. Figure 6.19d shows the Bland-Altman plot for the CMB size difference between the automated method and manual segmentations. A few CMBs were oversegmented with sizes ranging from 12 to 20 mm. Around 35% of detected true positives differed from the manually segmented CMBs by < 1 mm, and only 6% of CMBs (11 CMBs) had a difference > 6 mm. Most of these cases were due to the nearest small fragments of vessels that were segmented along with the CMBs. We obtained a mean absolute difference value of 2.0 mm between the manually segmented and detected CMBs.

6.6.3.2 Evaluation of automated MARS rating of CMBs

Initial evaluation of atlas registration using manually segmented CMBs: Figure 6.20 shows the confusion matrices between the manual and automated rating obtained from manually segmented CMBs for the individual structures and the three

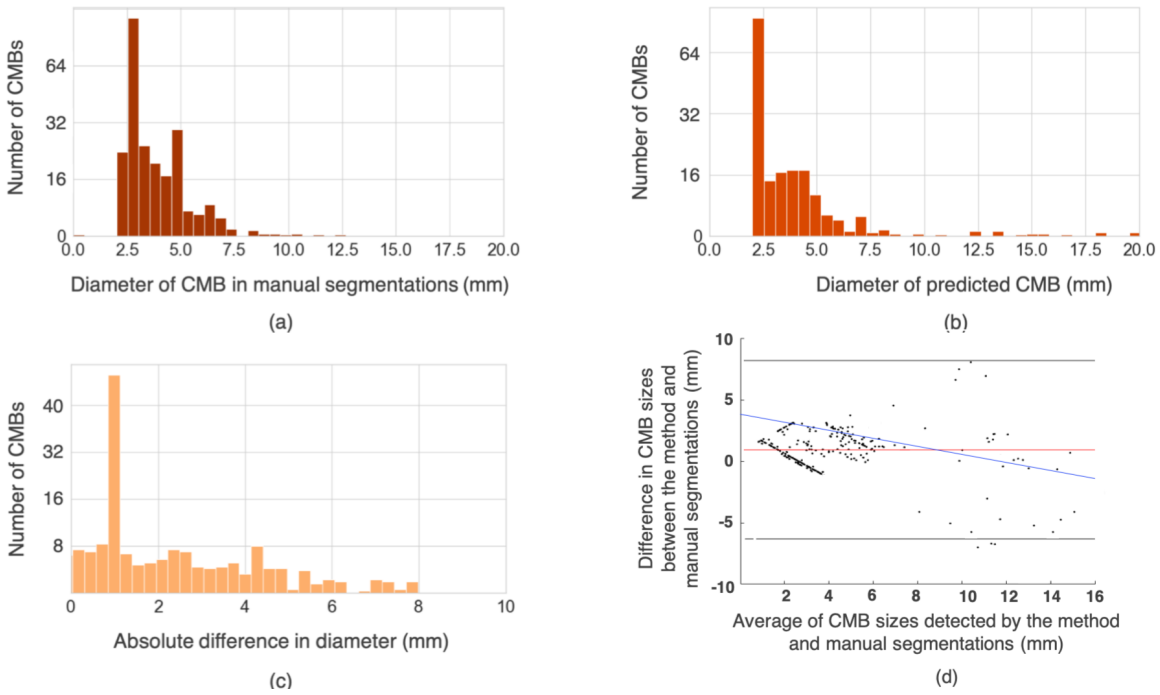


Figure 6.19: Results of CMB size estimation. Histograms of (a) diameters of manually segmented CMBs, (b) diameters of the detected true positive CMBs, (c) difference in the diameters between the manually segmented and the detected true positive CMBs, (d) Bland-Altman plot for difference in diameters between the automated method and manual segmentations. The diameter values are shown in mm.

regions. We obtained a region-wise sensitivity of 1.00, 0.79 and 0.98, a region-wise precision of 1.00, 0.90 and 0.95 and a region-wise F1-measure of 1.00, 0.84, 0.96 for infratentorial, deep and lobar regions respectively. We obtained an overall accuracy of 95%. The misclassifications between regions (e.g. infratentorial, deep) are lower than those between the individual structures (e.g. basal ganglia) indicating that while CMBs at the edges of the structures were misclassified at times, they were correctly classified into the corresponding regions. The high values of sensitivity, precision and F1-measure, especially in the infratentorial and lobar regions indicate that the atlas registration from MNI to subject space was reliably accurate.

Evaluation of automated MARS ratings using the detected CMBs: We determined the counts of detected CMBs for each subject within each structure, and later determined the statistics of the subject-wise detected CMB counts, across all subjects, as reported in table 6.7. Even though it is feasible to obtain the automated

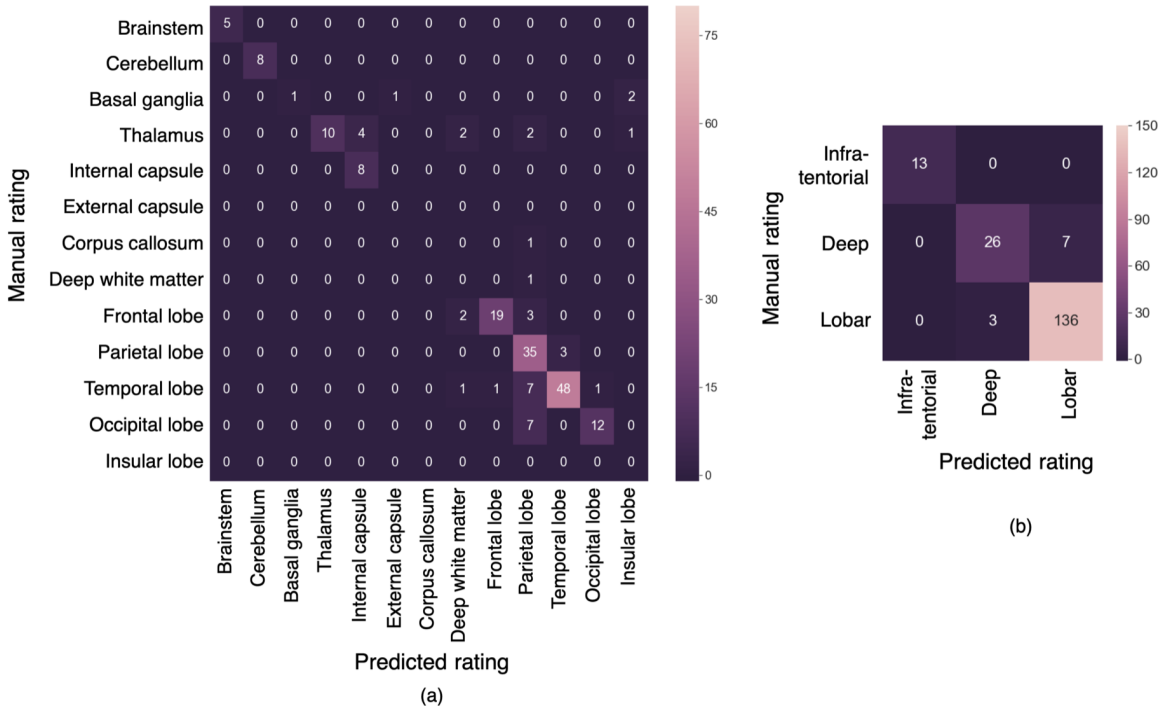


Figure 6.20: Confusion matrices for the automated CMB rating using manually segmented CMBs on the TICH2 dataset for (a) individual structures and (b) the 3 main regions.

rating for all the detected CMBs as reported in the table, for our evaluation using the confusion matrix, we used only true positive CMBs due to the availability of manual ratings. On comparing the manual ratings with the automated rating of true positive CMBs, figure 6.21 shows the confusion matrices for individual structures and the three overall regions. The method achieved a region-wise sensitivity of 0.64, 0.79, 0.96, a region-wise precision of 1.00, 0.85, 0.91 and region-wise F1-measure of 0.78, 0.82, 0.93 for infratentorial, deep and lobar regions respectively, with an overall accuracy of 91%. The F1-measure was high for the lobar region when compared to the other two, indicating a more sensitive and precise rating in that region. The increase in misclassifications compared to the manual segmentation case was due to the over/under-segmentation of predicted CMBs in the region boundaries (hence shifting the centroids to the adjacent regions).

Evaluation of CMB detection using the automated MARS scale rating: Once we verified the reliability of our automated MARS scale rating on manually detected and automatically detected true CMBs, we used this tool as a way to further

Region	Structure	Average of subject-wise count of detected CMBs	Proportion of subjects without any detected CMBs (%)	Maximum of subject-wise count of detected CMBs
Infra-tentorial	Brainstem	0.07	96.3	2
	Cerebellum	1.07	44.4	4
Deep	Basal ganglia	0.8	80.00	1
	Thalamus	1.18	75.2	4
	Internal capsule	1.14	85.8	2
	External capsule	0.08	96.3	1
	Corpus callosum	0.03	96.3	1
	Deep white matter	2.06	62.9	6
Lobar	Frontal	4.65	0.00	11
	Parietal	7.18	0.00	17
	Temporal	5.44	0.00	14
	Occipital	1.87	14.8	11
	Insular	1.5	18.8	8

Table 6.7: Automated CMB rating of detected CMBs the according to the MARS scale. Statistics of subject-wise counts of detected CMBs (including TPs and FPs) are reported for each structure.

evaluate the performance of our CMB detection method. Figure 6.22 shows the location of occurrence of FPs and FNs for the individual structures. From the figure, it is evident that most of FNs occurred in the parietal and frontal lobes, while the FPs occurred generally in the lobar region, mostly in the frontal, parietal and temporal lobes. Most of the FPs in the TICH2 dataset were due to the presence of haemorrhages that were mostly found in the lobar regions in the dataset, in addition to misclassified cortical sulci in the lobar region.

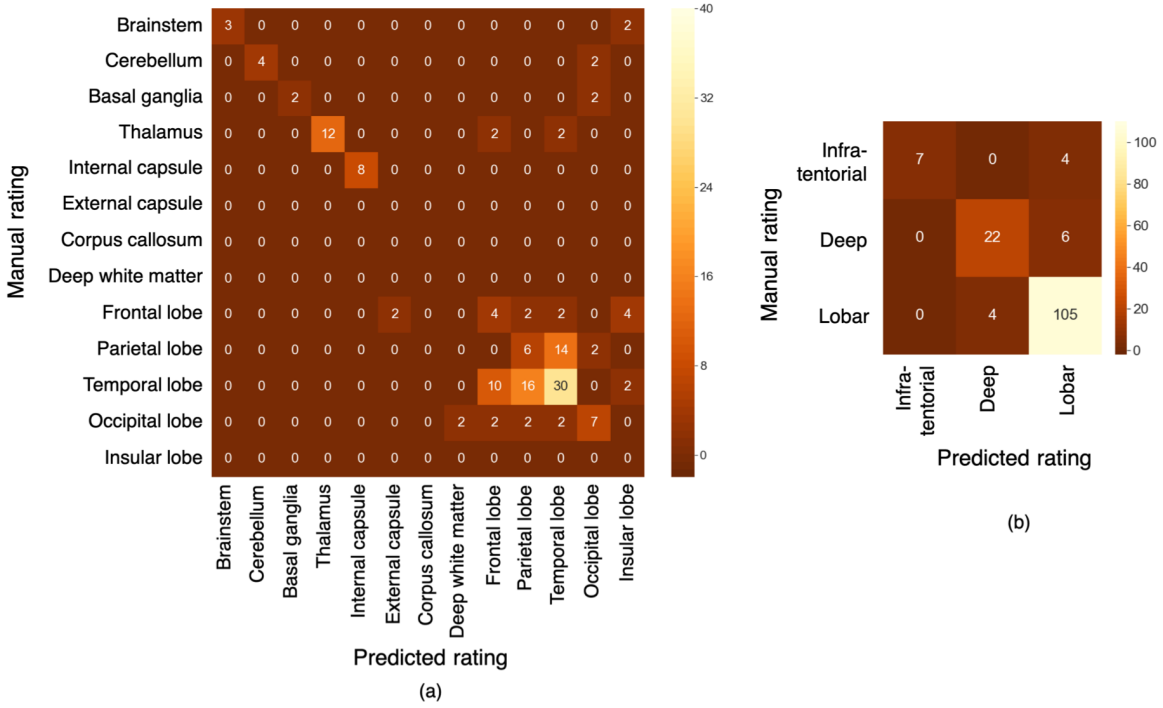


Figure 6.21: Confusion matrices for the automated CMB rating using the detected true positive CMBs on the TICH2 dataset for (a) individual structures and (b) the 3 main regions.

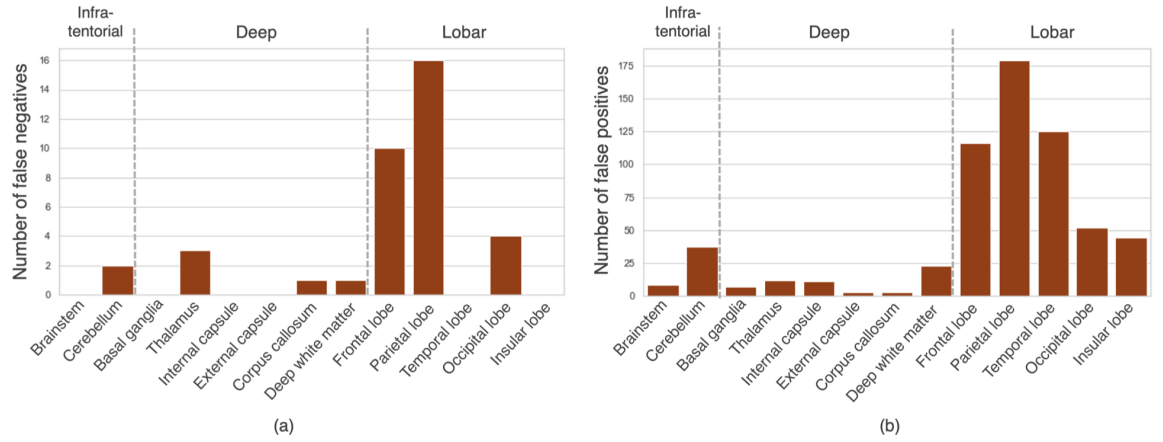


Figure 6.22: Barplots showing the number of (a) false negatives and (b) false positives in the individual structures from the CMB detection on the TICH2 dataset.

6.7 Discussion and conclusions

In this work, we developed an end-to-end automated framework for CMB analysis in large datasets, starting from candidate subject preselection, to CMB detection, to clinically-relevant scores. First, we proposed an automated pipeline for preselecting CMB candidate subjects from the large datasets. We initially evaluated the proposed

pipeline on the OXVASC and TICH2 datasets, and later applied the pipeline to the UKBB dataset for preselecting CMB candidate subjects for manual labelling and further analyses of CMBs. Second, we proposed a DL-based multiscale method for detecting CMBs in individual subjects. We initially studied the effect of individual scales and evaluated different steps of the method on the OXVASC, TICH2 and HK datasets. Finally, to provide clinically-relevant outputs, we determined the sizes of CMBs from the predicted CMB maps, and we utilised the information from structural atlases to provide an automated rating of CMBs according to the MARS scale, both at structure- and region-level.

Our preselection pipeline provided similar results for different training data, with only a slightly better accuracy and specificity on the OXVASC dataset when compared to the TICH2 dataset, while providing a lower subject-level TPR. The prediction of CMB candidate subjects was more precise in the OXVASC dataset probably due to the fact that the pipeline was trained using features extracted from SWI which is highly sensitive to CMB and susceptibility artefacts. At this point, it is worth noting that there are pros and cons with using T2*-weighted GRE or SWI sequences. In the case of T2*-weighted GRE images (the OXVASC dataset), the susceptibility effect is lower than in SWI and hence the number of FPs were low, providing high specificity (0.82) but low cluster-wise TPR (0.74). On the other hand, SWI sequences (the TICH2 dataset) are highly sensitive to CMBs and hence result in detection of more CMBs, albeit with increased FPs as well, resulting in high cluster-wise TPR (0.82) but lower specificity (0.69). In general, the number of subjects with high CMB count is fewer in the OXVASC dataset, when compared to the TICH2 dataset (figure 6.1a and c). Moreover, the presence of haemorrhages increased the number of FPs and reduced the overall subject-level accuracy in TICH2 dataset. However, the similar shape of the histograms of manually segmented CMB count and predicted number of CMBs/subject (figures 6.1c and 6.10c) indicate that the pipeline correctly detected more CMBs on true CMB subjects.

On training the pipeline on OXVASC (case 1) and TICH2 (case 2) datasets and applying them on the UKBB-I dataset, we observed that the predicted number of CMBs/ subject increased with age and BP. These results are in line with the existing knowledge that the risk of CMB is higher at older age and higher BP [285, 27, 64]. In both UKBB-I and II datasets, we observed that in both the cases the pipeline predicted higher number of CMB candidate subjects (68%, 65% in UKBB-I and 70%, 68% in UKBB-II) among the older and hypertensive population. The increase in CMB counts with age and BP demonstrates the potential utility of our pipeline along with demographic factors for CMB candidate subject selection in large datasets, even when manual segmentation is unavailable. While the preselected subjects could be used for manual labelling from methods perspective, from clinical point of view the pipeline could be used to flag the images as possible CMB subjects. This could be further used in determining a preliminary CMB or small vessel disease (SVD) score as done in [372].

In the CMB detection method, we used a 3-step framework for the object-level detection of CMBs. For the screening step, we applied 3D encoder-decoder networks at multiple scales with a local maxima refinement at the largest scale. We used CNNs extracting features at multiple scales instead of hand-crafted features used in the subjects preselection pipeline. This is because, while at the preselection stage the main aim was to identify the possible subjects with CMBs, in the CMB detection we aim for more accurate localisation of individual CMBs. For this aim, CNNs provide more comprehensive information regarding the location and context of CMBs when compared to a limited set of hand-crafted features. We later discriminated the detected CMB and non-CMB patches using a 3D network, followed by a postprocessing cleanup step. The ablation study on the effect of individual scales in the candidate screening shows that the lower scales aided in the detection of more CMBs, but also resulted in more FPs. Since, the lower scales only consider the local neighbourhood, they do not have a wider context and location related information regarding the CMBs. Hence they lead to the detection of more FPs, which is controlled by the averaging step.

The highest scale 64 is the most crucial scale since, apart from reducing FPs, it also helps in the refinement of local maxima.

On the 5-fold initial evaluation of the method on the OXVASC, TICH2 and HK datasets, we observed that the proposed method achieved the best performance on the OXVASC dataset. The highest number of FPavg was obtained on the TICH2 dataset due to the presence of haemorrhages, susceptibility artefacts and the edge artefacts closer to the skull. In general for all datasets, but especially for HK dataset, FPavg are reduced drastically in both candidate discrimination ($\approx$ 5 times) and postprocessing steps ($\approx$ 10 times). This shows that both the later steps, especially the postprocessing cleanup step is really crucial for improving the overall specificity since it removes the stray noise voxels. Using T2*-weighted GRE images, the proposed method gave a performance comparable with [297] and provided better precision than [301]. The best performances have been observed for the methods using SWI alone or in combination with other modalities. However, our method shows that comparable performance could be achieved even for T2*-weighted GRE images (despite their lower sensitivity towards CMBs compared to SWI), without adding other modalities. However, our method provides much lower precision for SWI modality, indicating the detection of more false positives compared to existing methods using SWI images. Also, we observed that our methods provided lower FPavg in the initial candidate screening step than most of the existing methods, indicating a highly precise detection from an earlier stage.

As clinically relevant applications, we extracted the size (diameter) of the CMBs and automatically calculated the MARS score. On comparing the CMB sizes, we obtained the mean absolute difference of 2.0 mm between the diameters of the manually segmented and the predicted true positive CMBs. We also observed that in $\approx$ 35% of CMBs (65 CMBs), the absolute difference was $<$ 1mm. Regarding the automated rating of CMBs based on MARS scale, we initially evaluated the automated rating method on manually segmented CMBs and achieved an overall accuracy of 95%, with higher region-wise F1-measures for infratentorial (1.00) and lobar regions

(0.96). From the confusion matrices in figure 6.20, we observed that the performance measures obtained for the three main regions were higher than the those of individual structures. This suggests that even though misclassifications of CMBs occur between adjacent structures, the method is robust at the region-level. These misclassifications at the structure-level mostly occur near the boundaries of individual structures and could be due to registration errors and size of dilation kernels used in the automated rating method. Achieving a high region-level accuracy is of importance since clinical impacts of CMBs are studied often at the region-level. For instance, lobar CMBs are associated with CAA and Alzheimer's disease, while deep CMBs are associated with cerebrovascular diseases [63, 64, 68, 65]. For the evaluation of the automated rating with detected CMBs, it is possible to obtain the ratings for all the detected CMBs (including TPs and FPs) as shown in table 6.7. However, since our evaluation using the confusion matrix requires manual ratings, we considered only true positive CMBs. In this evaluation, we observed an overall accuracy (of the ratings) of 91%, with higher region-wise sensitivity, precision and F1 measure in the lobar region. This high performance could be due to the fact that we consider the white-gray matter junctions as a part of lobar region, since that aligns more with the clinically provided manual rating. Similar to the case of manually segmented CMBs, the performance for the detected true positive CMBs is more accurate at the region-level than at the individual structure-level. However, the accuracies are in general lower compared to the manually segmented case due to an additional factor of errors in the CMB segmentation. Regarding the FPs obtained from the method, we observed more FPs in the lobar regions (especially in the frontal, temporal and parietal lobes) indicating that the possible source of FPs could be the edge artefacts close to the skull and the troughs of the cortical sulci. FNs were mostly the subtle and small CMBs, especially when they occur closer to the blood vessels.

6.7.1 Future work

The next step for this work would be to perform the preselection pipeline, CMB object-level detection and characterisation (counts, size and rating) on the entire

UKBB dataset. For this purpose, starting from the preselection pipeline, among the predicted CMB and non-CMB subjects, 60 subjects (50 predicted as CMB subjects, 5 predicted as non-CMB subjects and 5 borderline-cases) will be manually labelled by expert neuroradiologists for the validation of our pipeline. The manually labelled images can also be used to train the automated CMB detection method on the UKBB dataset. Another potential option for future work would be to enable adaptive tuning of various criteria values (e.g. volume, ellipticity) in the pipelines based on the dataset characteristics and demographic factors, to avoid setting the values empirically.

In the case of a CMB detection method, the possible ways of improving the performance include utilising atlas-based anatomical information for FP reduction and varying patch-sizes. Also, the multiple scales in the screening step could be combined in a more efficient way as done in [59] for WMHs (where a single network was built by fusing different scales at the FC layer). Based on the distribution of FPs, the anatomical regions/structures/coordinates could be added as priors or could be used for weighing the loss functions in the initial candidate screening step. Also, the main challenge in the discrimination step is the imbalance in the CMB and non-CMB training samples. Possible developments in that direction could be to use synthetic sampling (SMOTE) and GANs to create more instances of CMB-like samples resulting in more balanced training data. More selective and rigorous data augmentation techniques and loss functions for class rectification [373] might also help overcome the class imbalance problem. Also, currently we refine the local maxima only once, using the highest scale of 64. However, this refinement step could be recurrently included in a training epoch, creating a more computationally efficient and anatomically relevant '*region proposal network*' similar to those used in [303, 304, 305] and [307]. Another possible direction to look into is to use adaptive thresholding techniques similar to LOCATE to estimate the thresholds for CMB probability map from the screening step at subject-level or region-level based on the dataset characteristics. Similar to the subject-level CMB detection pipeline, an immediate next step for the voxel-level

CMB detection could also include extensive validation on a large dataset such as the UKBB and data harmonisation by using various domain adaptation techniques.

Also, another more clinically-oriented future direction could be to study the relationship between the distribution of CMBs in the individual regions and various clinical/demographic factors, to further understand the risk factors and the impact of these SVD signs. The results could in turn be used for further improving the CMB detection.

6.8 Summary

1. Our subject-level CMB preselection pipeline predicts the CMB candidate subjects with subject-level accuracies of 74% and 78% on the TICH2 and OXVASC datasets respectively. On the UKBB dataset, the number of predicted CMBs/subject increases with increase in age and BP, which is in line with the literature [64, 285, 27].

2. Our DL-based CMB detection method consists of CMB candidate screening, CMB candidate discrimination and postprocessing for FP reduction steps. We obtained cluster-wise TPR and precision ≥ 0.80 on the OXVASC, TICH2 and HK datasets, with the lowest FPavg of 9.6 on T2*-weighted GRE images from the OXVASC dataset.

3. On comparing our method with existing ones, we observed that our method performed better on the OXVASC dataset when compared to most of the conventional ML methods. Also, our method provided precision comparable to the state-of-the-art method [301] on the OXVASC dataset.

4. We compared the diameter of manually segmented and predicted CMBs on the TICH2 dataset and achieved a mean absolute difference of 2.0 mm, with 35% of CMBs having a difference less than 1 mm.

5. Our automated rating of CMB using the structural atlases on the TICH2 dataset provided an overall accuracy of 95% during initial validation with manually segmented CMBs, with higher precision in the lobar regions. We observed similar a trend for the predicted true positive CMBs with an overall accuracy of 91%.

Chapter 7

Conclusions and Future Work

This thesis aims to automatically detect signs of cerebral small vessel disease (SVD) on MR images. SVD refers to pathological condition related to damage in small vessels. SVD has been associated with factors such as age, smoking, blood pressure (BP), and has been widely found to contribute to cognitive decline, neurodegenerative and cerebrovascular diseases. However, standard quantitative and automated measurements are not yet available, which would also help to further understand the relationship between the voxel-wise distribution of SVD signs and various clinical factors. In this work, we focused on two SVD signs: white matter hyperintensities (WMHs) and cerebral microbleeds (CMBs). Out of the two, automated segmentation of WMHs has been widely explored in the literature (including the FSL tool, BIANCA developed by FMRIB, Oxford), while automated detection of CMBs is still a novel area of research. Hence, we started our research work with WMH analysis and modelled the distribution of WMHs at the population-level with respect to age using Bayesian inference (chapter 3). We later explored various options for improving WMH segmentation using BIANCA, experimenting at various steps (chapter 4). We developed a deep learning method for improving the segmentation of WMHs, compared to BIANCA (chapter 5). We finally proceeded to develop a comprehensive automated end-to-end CMB analysis framework consisting of CMB candidate subject preselection, CMB detection and rating (chapter 6). The methods developed in this work were validated and applied to a wide range of datasets, from healthy ageing (e.g. healthy controls - HC and the UK Biobank study - UKBB), to vascular cohorts (e.g. the Oxford

vascular study - OXVASC, Tranexamic acid for IntraCerebral Haemorrhage 2 trial - TICH-2, and cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy cohort - CADASIL), to neurodegenerative populations (e.g neurodegenrative cohort - NDGEN) and publicly available datasets (e.g. the Hong Kong dataset - HK, MICCAI WMH segmentation challenge dataset - MWSC). For more details on the datasets used, refer to appendix A.

Our proposed method for obtaining a population-level lesion probability map with respect to a parametric factor (e.g. age, cognitive function) uses cubic spline approximation functions and Bayesian inference. It is computationally light and provided spatially continuous population-level probability maps. We obtained a mean squared error (MSE) value of 7.27×10^{-5} on the simulated data, which is comparable with existing methods (including [127] with MSE value of 5.22×10^{-5} with ≈ 15 times slower running time). On real datasets such as UKBB and OXVASC, the WMH probabilities increased with age (significant in the periventricular regions), as is well established in the literature [54, 87]. We also found increase in WMH probabilities with BP in the UKBB dataset. Regarding the less clear relationship between the spatial distribution of WMHs and increase in blood pressure, we found that in hypertensive subjects, probabilities of deep WMHs (DWMHs) are specifically higher, in line with [50]. Hence, the population-level probability maps could be used for understanding the progression of WMHs under various clinical/demographic factors. Moreover, population-level maps could also be helpful in improving WMH segmentation, which in turn could improve their characterisation. Therefore in the following work, where we explored the possibility of improving BIANCA, one of the strategies we applied involved the use of population-level lesion maps derived/estimated from this work.

Our experiments for improving BIANCA involve utilisation of the population-level WMH probabilities as additional feature, exploration of alternative classifiers (RF, SVM, NN and Adaboost) and adaptive estimation of the threshold values for binarising subject-level WMH probability maps. The use of population-level WMH

probabilities (estimated from the WMH modelling method), specific to the subject's age group, provided negligible improvement in BIANCA. However, in order to allow for the potential of using population-level maps with respect to other factors (apart from age) in a different dataset, this option will be included in a future release of BIANCA. On comparing the performance of various classifiers, the KNN classifier (currently used in BIANCA) still provided the best results on the NDGEN and MWSC datasets. However, RF and NN classifiers also provided comparable results, showing potential for improvement in different datasets. Hence, RF and NN will also be added as alternative options for classifiers in a future release of BIANCA.

We also proposed a locally adaptive threshold estimation (LOCATE) method, for obtaining more accurate subject-level binary WMH maps from BIANCA probability maps. LOCATE divides the image into smaller regions using Voronoi tessellation, and estimates the threshold for each region in an adaptive manner using supervised classifier (RF) based on the local characteristics (volume, intensity, distance from ventricles and probabilities). LOCATE provided improvement in WMH segmentation performance for 5 datasets (NDGEN, OXVASC and 3 cohorts of MWSC), without requiring extra training data, despite the changes in the WMH distribution from healthy to severe vascular pathology (e.g. CADASIL and OXVASC). LOCATE detected more true positive lesions. LOCATE is already available as beta version (`https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BIANCA/Userguide`) and will be integrated into BIANCA as an additional thresholding option in its future release. All the additional options will be added without increasing the computational complexity of BIANCA.

Over recent years, deep learning (DL) methods based on convolutional neural networks (CNNs) have proven to provide robust segmentation in medical imaging, as evident from the results of the MICCAI WMH segmentation challenge (MWSC) 2017 (`http://wmh.isi.uu.nl/`): 17 out of 20 methods from the initial submission (and the top ranking 10 methods) used CNN models. Ensemble models have shown that they can provide more accurate segmentation, with less noise. We proposed a

triplanar U-Net ensemble network (TrUE-Net) for accurate segmentation of WMHs. We chose a triplanar architecture since it has less dependence on the variation in resolution along different planes, and better segmentation of edges. TrUE-Net provided better segmentation compared to BIANCA with significantly higher F1-measure in 4 datasets (NDGEN and 3 cohorts of MWSC), especially in the low lesion load subjects (while most existing methods, including BIANCA+LOCATE, gave lower SI values in these subjects). Overall, TrUE-Net achieved higher SI values and lower Hausdorff distance values for all datasets compared to BIANCA, indicating a more precise WMH segmentation. Usually, one of the main drawbacks of DL methods is that they require large amount of data for training. We overcame this shortcoming, since in terms of data requirements, TrUE-Net does not require any additional data other than the same amount used for training BIANCA. This was possible due to the use of data augmentation techniques for data amplification and ensemble model for noise/FP reduction.

In order to contextualise the impact of the methods/improvements presented in this work, table 7.1 illustrates an indirect comparison with existing methods in the literature. Among the non-DL methods, the performance is high when anatomy and appearance based features are used in addition to intensity (including BIANCA and LOCATE). The SI values obtained by BIANCA+LOCATE and TrUE-Net are in line with the values reported in previous studies. Also, the performance of TrUE-Net is comparable to the top-performing DL methods in the challenge. However, it is not a direct comparison since the methods from the challenge were evaluated on unseen test data, excluding the labels corresponding to other pathologies.

The two methods used in this thesis (TrUE-Net, newly proposed in this work, and BIANCA, here improved from its first release) have their own advantages and limitations. TrUE-Net provides better segmentation compared to BIANCA, but it is computationally intensive (GPUs are required for training and might be required even for testing) and has high memory requirements, especially in a clinical setting. On the other hand, TrUE-Net does not require parameter tuning, as in the case of

Method	Type[1]	Population - study[2] (subjects)	Image modalities	Dice[3]	Sens[4]	Spec
Wang et al., [149]	I,U	Singapore ageing cohort (272)	T1, T2, FLAIR	0.77	0.81	0.97
Yoo et al., [146]	I,S	Longitudinal/ dementia - Korean study (32)	FLAIR	0.76	-	-
Jeon et al., [159]	IA,U	SVD - AMPETIS (45)	PD, T2, FLAIR	0.90	-	-
Yang et al., [162]	IA,U	Dementia - LEILA (30)	FLAIR	0.81	-	-
Shi et al., [163]	IAAp,U	Acute infarction (91)	DWI, T1, FLAIR	0.84	0.80	-
Samaille et al., [164]	IAAp,U	MCI, CADASIL (67)	T1, FLAIR	0.72	-	-
Kruggel et al., [167]	IAAp	HC, dementia - LEILA (116)		-	0.90	0.91
Khademi et al., [168]	IAAp,U	Subjects with lesions (24)	FLAIR	0.83	0.82	0.99
Admiraal-Behloul et al., [171]	IA,U	Vasc. disease - PROSPER (100)	PD, T2, FLAIR	0.75	-	-
Anbeek et al., [55]	S	Arterial vascular disease (20)	T1, IR, PD, T2, FLAIR	0.80	0.97	0.97
Li et al., [58]	DL	MWSC TS	T1, FLAIR	0.80	0.84	-
Andermatt et al., [238]	DL	MWSC TS	T1, FLAIR	0.78	0.83	-
Berseth et al., [227]	DL	MWSC TS	T1, FLAIR	0.77	0.73	-
Valverde et al., [235]	DL	MWSC TS	T1, FLAIR	0.77	0.75	-
Mehrtash et al., [216]	DL	MWSC TS	T1, FLAIR	0.77	0.59	-
Luna et al., [216]	DL	MWSC TS	T1, FLAIR	0.72	0.63	-
Xu et al., [239]	DL	MWSC TS	T1, FLAIR	0.73	0.63	0.67
Ghafoorian et al., [59]	DL	Elder SVD (50)	T1, FLAIR	0.78	-	-
BIANCA [53]	IAAp,S	NDGEN (21)	T1, FLAIR	0.77	0.80	-
		OXVASC (18)	T1, FLAIR, MD	0.74	0.71	-
		MWSC TrS - Sing (20), Ams (20), Utr (20)	T1, FLAIR	0.74^S, 0.70^A, 0.63^U	0.76^S, 0.80^A, 0.67^U	-
BIANCA + LOCATE (Chapter 4) [79]	IAAp,S	NDGEN (21)	T1, FLAIR	0.77	0.83	-
		OXVASC (18)	T1, FLAIR, MD	0.75	0.82	-
		MWSC TrS - Sing (20), Ams (20), Utr (20)	T1, FLAIR	0.73^S, 0.70^A, 0.64^U	0.78^S, 0.80^A, 0.70^U	-
TrUE-Net (Chapter 5)	DL	NDGEN (20)	T1, FLAIR	0.78	0.72	-
		OXVASC (18)	T1, FLAIR	0.78	0.70	-
		MWSC TrS - Sing (20), Ams (20), Utr (20)	T1, FLAIR	0.81^S, 0.78^A, 0.74^U	0.78^S, 0.71^A, 0.75^U	-

[1] Type: I - intensity, IA - intensity + anatomy, IAAp - intensity + anatomy + appearance, U - unsupervised, S - supervised, DL - deep learning

[2] Population-study: MWSC TS - MICCAI WMH segmentation Challenge test dataset, MWSC TrS - MICCAI WMH segmentation Challenge training dataset [216]

[3,4] Dice, sensitivity: ^{S} - Singapore, ^{A} - Amsterdam, ^{U} - Utrecht.

Table 7.1: Comparison of the existing methods with BIANCA and the methods/improvements proposed in this work.

BIANCA. Assuming the availability of pretrained weights/models and GPUs, testing times per subject for TrUE-Net ($<$ 1 min on GPU) and BIANCA (2-3 mins on CPU) are comparable, excluding the extra time taken for WM mask generation ($\approx$ 10 mins).

Regarding generalisability to different scanners and datasets, only a few methods (among those presented in the table [7,8,19] and BIANCA), have been evaluated on a variety of datasets (pathological population and/or healthy subjects). Most methods (e.g. all the remaining ones in the table) are tested on datasets from a specific population, and hence might require additional experimentation to validate/improve their generalisability (e.g. fine-tuning of parameters, change of cut-off/threshold values). Since different datasets have different image modalities, using multimodal methods (e.g. BIANCA) provides latitude to choose the modalities that yield better results and could handle datasets with missing modalities. On the other hand, the choice of input modalities becomes one additional parameter to tune when applied on an unknown dataset. For instance, for the OXVASC dataset, the best performance for BIANCA was achieved with T1+FLAIR+MD, while TrUE-Net provided better SI values by using only T1+FLAIR.

Data harmonisation (i.e. integration of data from multiple sources or *domains*) is another main consideration while evaluating a method. The use of domain adaptation techniques on TrUE-Net showed promising results without additional parameter tuning. We explored various domain adaptation techniques such as transfer learning, unsupervised and semi-supervised domain adversarial training of neural network (DANN) between *source* and *target* domains (differing in intensity and acquisition characteristics). Out of all the approaches tested, unsupervised DANN gave the best performance and was able to overcome the variation in lesion characteristics between the domains. DANN provided a good domain adaptation without requiring manual segmentations and large amount of data from the target domain. For instance, we used only 15 subjects from the target dataset for training DANN to achieve a good

segmentation performance. The results also showed that a suitable domain adaptation method could provide on par, if not better, results when compared to the target-trained models in the case of insufficient data.

It is worth noting that, although manual segmentation is used to train the models and as a gold standard for evaluation, there are often inter- and intra-observer variabilities amongst expert segmentations, due to lower contrast, fuzzy boundaries and lack of consensus regarding exclusion criteria (e.g. whether the thin lining around the ventricles should be considered as WMHs or not). These variabilities could affect the performance of automated segmentation methods.

For CMBs, the proposed CMB candidate subject preselection pipeline provided comparable results on the OXVASC and TICH2 datasets, irrespective of different modalities used for training (T2*-weighted GRE from OXVASC and SWI from TICH2). The pipeline predicted CMB candidate subjects with subject-level accuracies of 74% and 78% on TICH2 and OXVASC datasets respectively. On the UKBB dataset, the number of predicted CMBs/subject increased with age and BP, in line with the literature [64, 285, 27]. Our 3-step CMB detection method consists of initial voxel-wise candidate screening, object-level candidate discrimination and FP reduction. The proposed method can be trained on either SWI or T2*-weighted GRE images and provided good results despite the presence of other vascular pathologies and of CMB mimics. We obtained the best performance with a sensitivity of 0.85, precision of 0.96 and average FPs/subject of 9.6 on T2*-weighted GRE images from the OXVASC dataset (despite the lower sensitivity to CMBs of this sequence compared to SWI). On the TICH2 dataset, which included haemorrhagic lesions, we achieved a sensitivity of 0.80, a precision of 0.85 for an average of 25 FPs/subject. Our method provided better performance than most of the conventional ML methods [294, 296, 297], comparable to [297] and a better precision than the state-of-the-art DL method [301]. We also characterised the detected CMBs based on size and determined the automated rating based on the MARS scale. We compared the diameters

of manually segmented and detected true positive CMBs and achieved a mean absolute difference of 2mm. Our automated rating method provided 95% accuracy during initial evaluation with manually segmented CMBs. On the further validation with the results of CMB detection, we achieved an accuracy of 91% on true-positive CMBs, and observed that FPs mostly occurred in the lobar regions.

7.1 Future work

In the current work, we detected and analysed two signs of SVD, namely WMHs and CMBs. While providing advances in the field, the work developed and presented in this thesis also opens several potential avenues for future research.

WMH segmentation and analysis methods: As a future work, our proposed method based on Bayesian inference could be used to model the distribution of WMHs with respect to other individual or combination of factors that could be useful in a clinical context. For example, similar to our analysis between hypertensive and non-hypertensive groups, an analysis between cognitive groups (e.g. based on the MoCA scores) would be useful in analysing the difference in WMH progression in the normal and pathological ageing (dementia). Regarding LOCATE, the immediate next step could be to apply it on TrUE-Net probability maps to study the possible improvement in WMH segmentation. Also, further improvements, such as modification of features and training threshold values could be made to improve the performance of LOCATE, especially in low lesion load subjects. For TrUE-Net, a potential future direction would be to explore the possible ways to improve the performance by leveraging the features at different levels of abstraction and the loss functions used in the optimisation. Towards the first aim, we have explored the possible utility of attention gate at the coarsest level of U-Nets (TrUE-Net-Att). However, it would be useful to study the effect of attention gates in the finer level(s) on the segmentation performance. In the case of loss functions, the Dice loss function could be weighted with the values inversely proportional to the volume of individual lesions, so that the loss function becomes sensitive towards small lesions (similar to improving spatial sensitivity

towards deep regions with weighted CE loss). Similar to what we did for BIANCA, we could use population-level WMH probability maps as priors (or additional input channel) in TrUE-Net to explore the potential improvement in segmentation.

CMB segmentation and analysis methods: Similar to WMH segmentation, adding priors could also improve CMB detection at object-level. For example, applying the minimum-intensity projections (MIP) to a few consecutive slices along all three (*x-, y-* and *z-*) directions and averaging the MIP maps could provide a good prior map for CMB detection and would help to reject vessel-like structures. Another possibility for future work along this direction could be to use the atlas-based anatomical priors (e.g. region/structures/coordinates based on CMB distribution) for further reduction of FPs. A possible next step in CMB candidate screening could be to make the ROI identification more efficient than our current method but less intensive than region proposal network (RPN) by dynamic refinement of local maxima with each training epoch, thereby reducing the spurious peaks due to noise. Also, LOCATE could be applied on the CMB probability maps from the screening step to explore its generalisability with respect to the signs. If necessary, the algorithm could be modified to cope better with CMBs (e.g. adaptive adjustment of the sizes of Voronoi regions, changes in the training method). In the CMB candidate discrimination step, to tackle the imbalance in the detected CMB and non-CMB patches, possible next steps could be exploration of sampling techniques, synthesis of CMB-like patches, and selective augmentation techniques. Some of the adaptive sampling methods that could be explored include oversampling of the minority class (in our case, CMBs) with replacement and synthetic minority over-sampling technique (SMOTE, [374]) where synthetic examples of the minority class are generated during oversampling. Also, DL models such as GANs could be used for synthesising CMB-like patches to reduce the imbalance in the classes [375, 376]. Further improvements in data augmentation techniques could include selective and adaptive data augmentation [377], where the augmentation is integrated into the loss function for efficient classification. Also, the loss function itself could be modified for class rectification [373] to reduce the effect of

imbalance in CMB and non-CMB classes. In addition to the above possible options, efficient training techniques such as omni supervised learning (predicted outputs are used recurrently for training [360]) could be used in cases of both WMHs and CMBs where training data is scarce. For the CMB subjects preselection pipeline, we could replace the initial CMB candidate detection (based on hand-crafted features) with our DL-based CMB detection method and study the improvement in preselection performance.

Tool development: The immediate next step would be to integrate LOCATE within the main BIANCA release, along with other options (alternative classifiers and prior map). Another generic but useful next step in tool development would be to make the values of several parameters in the CMB pipeline and object-level detection adaptive to image characteristics (e.g. area/volume thresholds in CMB pipeline as the percentiles of values, intensity cut-offs based on histograms) rather than setting them manually.

Extended validation: We have currently evaluated TrUE-Net only on the publicly available training data from MWSC 2017. Since the challenge remains open, we will submit our method to the challenge to validate it on the unseen test data. This will help us in studying the model generalisability and would also provide a more direct comparison with other DL methods. Also, we plan to validate the preselection pipeline and object-level CMB detection on the UKBB dataset. Towards this aim, 60 subjects (50 predicted as CMB subjects, 5 predicted as non-CMB subjects and 5 borderline-cases by the proposed pipeline) will be manually labelled by expert neuroradiologists for the validation of our pipeline. Once the manual labelling is available, it will enable several evaluations and validations. First, we can compare the performance of the preselection pipeline with hand-crafted features and our DL-based CMB detection method. Hence, the better performing model (between the above two) could be used to preselect CMB candidate subjects from the UKBB dataset. Second, the manually labelled images can also be used to train the CMB detection method on the UKBB dataset, so that the complete end-to-end analysis could be applied on the entire

UKBB dataset. In the same way that WMH volume derived with BIANCA is one of the imaging derived phenotypes (IDPs) in UKBB, we could determine several CMB-related IDPs such as subject-level CMB presence/absence, CMB count and MARS ratings on the UKBB dataset.

Data harmonisation: A possible future work towards data harmonisation could be to obtain pre-trained models (CMB detection model and TrUE-Net, for WMH segmentation) on the UKBB data, so that these models could be adapted to smaller datasets with various domain adaptation methods. Another alternative approach would be to harmonise data from various centres (*targets*) to match the UKBB data characteristics, so that the pre-trained models (for CMB and WMH segmentation) could be applied to them directly. The latter approach would be highly convenient in a clinical setting since it would avoid additional training.

Clinical applications: The subjects preselected as CMB candidates could be flagged as possible CMB subjects for further investigation or for preliminary rating of a composite SVD score [372]. Regarding the modelling method, similar to WMH distribution, population-level analyses could be performed on CMBs, to understand their clinical impact. These analyses could be done either by modelling their spatial distribution (as in the case of WMHs) or their count per subject with respect to various clinical factors. Perhaps, these analyses could be done separately for microbleeds and macrobleeds for further understanding their role as biomarkers and their clinical impact. Also, the scope of our Bayesian modelling algorithm could be extended beyond SVD to study the distribution of other signs with different spatial pattern (e.g. MS lesions, atrophy) with respect to clinical factors.

Appendix A

Dataset details

This appendix provides the description of various datasets used in this work, their sample details and their acquisition parameters. In this work we used 8 main datasets, with different acquisition protocols, populations, WMH loads/CMB counts and characteristics, and occasionally their subsets for various analyses. Tables A.1 and A.2 provide demographic details of the datasets used in WMH and CMB analyses respectively.

A.1 The Oxford vascular study (OXVASC)

This dataset consists of eligible participants from the OXVASC study [323] who had recently experienced a minor non-disabling stroke or transient ischemic attack (TIA). The demographics of the subsets used for WMH and CMB analyses are provided in tables A.1 and A.2. Scanning was performed at the Oxford Acute Vascular Imaging Centre (AVIC) using MR acquisition parameters provided in table A.3. MRI acquisitions started in 2012 and the protocol was upgraded in January 2015 (indicated as protocol I and II in the table). The only sequence differing between protocols-I and II is the T2-weighted, FLAIR sequence as shown in the table.

We used three subsets of data from the OXVASC study:

1. The first subset, consisting of 474 subjects scanned using protocol-I, was used in chapter 3 for population-level modelling of the WMH distribution.

Datasets		Chapter	N	Age (yrs)		Sex (F:M)	Sample details
				Range	Mean±Std		
OXVASC		3	474	20 - 102	67.4 ± 14.3	240:234	Subjects with a recent minor non-disabling stroke or TIA
		4 & 5	18	50 - 91	73.3 ± 12.3	7:11	
UK Biobank		3	1000	45.4-78.4	64.1 ± 6.8	459:541	500 hypertensive and 500 non-hypertensive subjects
NDGEN		4	21	63 - 86	77.1 ± 5.8	10:11	9 with AD, 5 with MCI, 7 HC
		5	20	63 - 86	77.2 ± 5.7	10:10	8 with AD, 5 with MCI, 7 HC
CADASIL		4	15	33 - 70	53.7 ± 11.3	11:4	Subjects with CADASIL
HC		4	19	29 - 70	54.6 ± 11.2	6:13	Healthy controls, age matched with the CADASIL dataset
MWSC	**NUHS Singapore**	4 & 5	20	NP	70.1 ± 9.3	85:85	Provided [216] for 170 subjects from memory clinics (60 - training, 110 - testing, not used in this work)
	VU Amsterdam	4 & 5	20				
	UMC Utrecht	4 & 5	20				

Table A.1: Demographic and sample details of the datasets used for WMH segmentation and analyses, along with their corresponding chapters. N - total number of subjects, TIA - transient ischemic attack, AD - Alzheimer's disease, MCI - mild cognitive impairment, HC - healthy controls, CADASIL - cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy, NP - not provided.

2. The second subset of 18 subjects acquired using protocol-II was used in chapters 4 (T1-weighted, FLAIR and DWI) and 5 (T1-weighted and FLAIR) for the evaluation of WMH segmentation methods. Manual segmentations are available with WMH load range: 3530 - 83391 mm^3, median: 16906 mm^3.

3. The third subset consisting of T2*-weighted gradient recalled echo (GRE) images from 40 subjects was used in chapter 6 for CMB detection and charac-

Datasets		N	Age (yrs)		Sex (F:M)	Sample details
			Range	Mean±Std		
OXVASC		40	35.6-94.8	68.7 ± 15.5	15:25	Vascular cohort with a recent minor non-disabling stroke or TIA
TICH2		50	29 - 88	64.8 ± 15.5	25:26	SWI from the subjects with spontaneous intracerebral haemorrhage
HK		20	NP	69.3 ± 8.15	NP	Sample details specified in [301] for the dataset of 320 subjects (126 with stroke and 194 normal ageing subjects). Only 20 subjects (10 from each category) available publicly
UK Biobank	**UKBB-I**	180	45 - 75	58.5 ± 8.8	85:95	Selected randomly from UK Biobank data
	UKBB-II	287	74.1-80.3	75.6 ± 1.0	106:181	Selected out of ≈ 20,000 subjects, with SWI, and shortlisted based on demographic factors with no missing values

Table A.2: Demographic and sample details of the datasets used for CMB detection and analyses in chapter 6. N - total number of subjects, TIA - transient ischemic attack, SWI - susceptibility weighted images, NP - not provided.

terisation. Out of 40 subjects, 20 subjects have CMBs and the corresponding manual segmentations are available. Total number of CMBs are: 267, mean: 10.7 ± 1.13 CMBs/subject, median: 2 CMBs/subject.

A.2 The UK Biobank study (UKBB)

The largest publicly available MR imaging study ever, UK Biobank (UKBB) consists of MR scans from ≈ 48,000 subjects acquired currently (and will reach 100,000). Scanning was performed at three dedicated imaging centres to maximise data compatibility, using MR acquisition parameters specified in table A.4. The demographic

The OXVASC dataset - MR acquisition parameters

Parameters	FLAIR		T1	DWI	GRE
	Protocol-I	Protocol-II			
Scanner	3T Siemens Verio				
Sequence	T2 FLAIR	T2 FLAIR	3D MP-RAGE	DWI - EPI	2D single-echo T2*-GRE
TR/TE (ms)	9000/94	9000/88	2000/1.94	8000/86	504/15
TI (ms)	2500	2500	-	-	-
Flip angle (deg)	150	150	8	16	20
Voxel size (mm)	0.8 × 5 × 0.8	1 × 3 × 1	1 × 1 × 1	2 × 2 × 2	0.9 × 0.8 × 5
Matrix size (voxels)	464 × 28 × 512	174 × 52 × 192	208 × 256 × 256		640 × 640 × 25
Additional sequence specific parameters	-	-	-	b-value = 0, 500, 1000 s/mm2, 32 directions, GRAPPA factor = 2	GRAPPA factor = 2

Table A.3: MR acquisition parameters for FLAIR (protocol-I and II), T1, DWI and T2*-GRE in the OXVASC dataset. TR - repetition time, TE - echo time, TI - inversion time, DWI - diffusion weighted images.

details of the subsets of data used for our WMH and CMB experiments are specified in table A.1 and A.2 respectively.

We used three subsets of data from the 20,000 subject release of the UKBB dataset (in January 2018) as follows:

1. The first subset consists of 1000 subjects, with 500 hypertensive (HT) subjects and 500 non-hypertensive (non-HT) subjects randomly selected from $\approx$ 8,000 subjects. We used this dataset for modelling the distribution of WMHs with respect to age in HT and non-HT groups in chapter 3.

2. The second subset (indicated as UKBB-I in chapter 6) consists of SWI from 180 randomly sampled subjects for the initial evaluation of the CMB candidate

subject preselection pipeline. We obtained the relationship between subject-level CMB counts and factors such as age and BP on this dataset in chapter 6.

3. The third subset, UKBB-II in chapter 6, consists of 287 subjects (after rejecting subjects with missing data values) using SWI images. These subjects were shortlisted from ≈ 20,000 subjects by applying specific criteria on various demographic factors such as age, blood pressure, smoking, WMH volume and incidence of stroke. We applied our pipeline to this subset to preselect CMB candidate subjects in chapter 6 for the manual labelling of CMBs.

The UK Biobank dataset - MR acquisition parameters

Parameters	**FLAIR**	**T1**	**SWI**
Scanner	3T Siemens Magnetom Skyra syngo MR D13		
Sequence	T2 FLAIR	3D MP-RAGE	3D multi-echo GRE
TR/TE (ms)	5000/395	2000/2	27/9.4/20
TI (ms)	1800	880	-
Flip angle (deg)		8	15
Voxel size (mm)	$1.05 \times 1 \times 1$	$1 \times 1 \times 1$	1 $0.8 \times 0.8 \times 3$
Matrix size (voxels)	$192 \times 256 \times 256$	$208 \times 256 \times 256$	$256 \times 288 \times 48$

Table A.4: MR acquisition parameters for FLAIR, T1 and SWI in the UK Biobank dataset. TR - repetition time, TE - echo time, TI - inversion time.

A.3 The neurodegenerative cohort (NDGEN)

This dataset includes MRI data from 9 subjects with probable Alzheimers Disease (AD), 5 with amnestic mild cognitive impairment and 7 cognitively healthy control subjects recruited from the Oxford Project To Investigate Memory and Ageing (OPTIMA) study and from the Memory Assessment Clinic at the John Radcliffe Hospital in Oxford [334]. The demographics of the dataset are provided in table A.1 and the MR acquisition protocols and parameters are provided in table A.5. We used all

21 images in chapter 4 for improving BIANCA and to evaluate LOCATE. However, we excluded one subject with incompatible image dimensions from the WMH segmentation experiments using deep learning in chapter 5. Manual segmentations are available for this dataset, with the WMH load range: 1878 - 89259 mm^3, median: 20772 mm^3.

The NDGEN dataset - MR acquisition parameters

Parameters	**FLAIR**	**T1**
Scanner	3T Siemens Trio	
Sequence	T2 FLAIR	3D MP-RAGE
TE/TR (ms)	9000/89	2040/4.7
Flip angle (deg)	150	8
Voxel size (mm)	$1.1 \times 0.9 \times 3$	$1 \times 1 \times 1$
Matrix size (voxels)	$256 \times 256 \times 35$	$174 \times 192 \times 192$

Table A.5: MR acquisition parameters for FLAIR and T1 in the NDGEN dataset. TR - repetition time, TE - echo time.

A.4 The cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) cohort

This dataset consists of 15 subjects with CADASIL disease. The demographic details for this dataset are provided in table A.1. Scanning was performed using the same scanner and acquisition parameters as the OXVASC dataset with protocol-II (refer to table A.3). CADASIL is a genetic form of small vessel disease caused by mutations within the NOTCH-3 gene. It is the most common heritable cause of stroke and vascular dementia in adults and it is characterised by extensive damage to white matter brain regions [40]. These patients are younger than those with sporadic SVD, without confounding factors of co-existent neurodegenerative pathology [41]. They therefore provide a model of pure SVD in which to investigate WMHs [42]. Manual

segmentations are unavailable for this dataset. However, in order to obtain a binary WMH mask from the BIANCA probability maps, a lower threshold of 0.2 was empirically chosen (based on visual inspection of the lesion probability map) by an expert neurologist as the 'optimal' global threshold for this dataset. We used T1-weighted, FLAIR and DWI images from this dataset for visually validating the WMH segmentation with LOCATE in chapter 4. We also used T1-weighted and FLAIR images for training purposes in the domain adaptation experiments in chapter 5.

A.5 The healthy control (HC) cohort

This dataset consists of 19 healthy controls, age-matched with the CADASIL subjects. For the demographic details of this dataset refer to table A.1. Scanning was performed using the same scanner and acquisition parameters as the OXVASC protocol-II (table A.3). Manual segmentations are unavailable for this dataset. We used T1-weighted, FLAIR and DWI images from this dataset for visually validating LOCATE in chapter 4. We also used T1-weighted and FLAIR images for analysing data characteristics for domain adaptation in chapter 5.

A.6 The MICCAI WMH segmentation challenge Dataset (MWSC)

This dataset consists of 60 subjects, 20 each from three different sources: memory clinics of institutes from Utrecht, Amsterdam and memory clinics of the National University Hospital and St. Lukes Hospital in Singapore [216]. The dataset is publicly available as the training dataset of the MWSC challenge (`http://wmh.isi.uu.nl/`). The combined demographics of training and unseen test data has been provided in [216] (refer to table A.1), while the demographic details of the training data alone are currently unavailable. The MRI acquisition protocols and parameters for all the three datasets (UMC Utrecht, NUHS Singapore and VU Amsterdam) are provided in table A.6. Manual segmentations are available for all three datasets, with an additional exclusion label provided for other pathologies. The WMH volume ranges

(excluding other pathologies) are 845 - 74991 mm^3 (median: 26240 mm^3) for the UMC Utrecht cohort, 786 - 61332 mm^3 (median: 17795 mm^3) for the NUHS Singapore cohort and 1522 - 43528 mm^3 (median: 6015 mm^3) for the VU Amsterdam cohort.

We used T1-weighted and FLAIR images from this dataset for BIANCA experiments and evaluating LOCATE in chapter 4 and for evaluating the WMH segmentation performance of the deep learning method in chapter 5.

The MWSC dataset - MR acquisition parameters

Parameters	**NUHS Singapore**		**VU Amsterdam**		**UMC Utrecht**	
	FLAIR	**T1**	**FLAIR**	**T1**	**FLAIR**	**T1**
Scanner	3T Siemens TrioTim		3T GE Signa HDxt		3T Philips Achieva	
Sequence	2D FLAIR	3D T1	3D FLAIR	3D T1	3D FLAIR	3D T1
TR/TE (ms)	9000/82	2300/1.9	8000/126	7.8/3.0	11000/125	7.9/4.5
TI (ms)	2500	-	2340	-	2800	-
Voxel size (mm)	$1 \times 1 \times 3$	$1 \times 1 \times 1$	$0.95 \times 0.95 \times 3$	$1 \times 1 \times 1$	$0.98 \times 0.98 \times 1.2$	$0.94 \times 0.94 \times 1$

Table A.6: MR acquisition parameters specified for FLAIR and T1 in [216] for the NUHS Singapore, VU Amsterdam and UMC Utrecht cohorts in the MICCAI WMH segmentation challenge. TR - repetition time, TE - echo time, TI - inversion time.

A.7 The Tranexamic acid for IntraCerebral Haemorrhage 2 (TICH-2) trial dataset

This dataset, a subset of the data used in [80, 81], was obtained from our collaborators involved in the TICH2 study at the University of Nottingham. The dataset consists of SWI from 50 subjects with spontaneous intracerebral haemorrhage. The demographic details of the dataset are provided in table A.2. The standard manufacturer specific SWI protocol parameters (reported in table A.7) from the manufacturers protocol tree (ie, SWAN (General Electric); VEN_BOLD (Philips); SWI (Siemens)) were used, as specified in [80].

The TICH2 dataset - MR acquisition parameters

Scanner	**Siemens**		**Philips**		**GE**	
Field strength	1.5T	3T	1.5T	3T	1.5T	3T
TE (ms)	40	20	0 - 35		23 - 49.5	
TR (ms)	49	27	14.5 - 52		37 - 1950	
Flip angle (deg)	15		10 - 20		15 - 50	
Slice thickness (mm)	1 - 4		1 - 10		2.4 - 4	

Table A.7: MR acquisition parameters for the standard manufacturer specific SWI protocol specified in [80] for the TICH2 dataset. Ranges are provided for those parameters that vary subject-wise. TR - repetition time, TE - echo time.

Out of 50 subjects, 25 subjects have CMBs and the corresponding manual segmentations are available. The total number of CMBs are: 505, mean: 20 $\pm$ 32.6 CMBs/subject, median: 5 CMBs/subject. Additionally, for these 25 subjects with CMBs, MARS ratings (for more details, refer to the sample MARS rating form shown in appendix D) were provided for both '*definite*' and '*possible*' CMBs. For both training and testing purposes, we obtained the final manual segmentation maps by combining the manually segmented CMBs from both categories. We used SWI images from the dataset for training and evaluating the subject-level preselection pipeline, object-level CMB detection and automated rating of CMBs in chapter 6.

A.8 The Hong Kong dataset (HK)

This dataset consists of SWI from 20 subjects (10 normal ageing subjects and 10 subjects with stroke) that were used as training data in [301] and are available publicly (`http://www.cse.cuhk.edu.hk/~qdou/cmb-3dcnn/cmb-3dcnn.html`). Originally in [301], 320 images (194 normal ageing subjects and 126 subjects with stroke) were used and the demographic details for the full dataset (320 subjects) have been provided in [301] (table A.2). Currently, the demographic details for the subset of 20 subjects are unavailable. The SWI images were acquired using a 3T Philips Medical

System with 3D spoiled gradient echo sequence using venous blood oxygen level dependent series with the following acquisition parameters [301]: TR/TE = 17/24 ms, FOV of 230 mm, voxel-size of 0.45 × 0.45 × 2 mm, matrix size of 512 × 512 × 150 voxels, slice spacing of 1 mm.

Only CMB coordinates were available for the 20 subjects along with the dataset. Hence, another rater independently provided the manual segmentations, and we considered the combined information from both the coordinates and the manual masks as our final ground truth. The total number of CMBs are: 89, mean: 4.5 ± 4.4 CMBs/subject, median: 2 CMBs/subject. We used SWI from this dataset to evaluate object-level CMB detection in chapter 6.

Appendix B

Brief overview of classifiers

This appendix provides a brief overview of various classifiers used in chapter 4 for the comparative analyses within BIANCA.

B.1 K-nearest neighbour classifier (KNN)

K-nearest neighbour (KNN) [378] is a non-parametric supervised classification algorithm, where the classification of a voxel into the positive or negative (i.e. WMH or non-WMH) class is determined by the nature of its k neighbours in the feature space from the training set. Let us consider an input feature vector x_i corresponding to a voxel in training set T, then fitting a KNN model on x_i for getting $\hat{Y}$ is given by,

$$\hat{Y}(x) = \frac{1}{k} \sum_{n_i \in N_k(x)} y_i \tag{B.1}$$

where $N_k(x)$ is the neighbourhood of x_i consisting of the k closest feature points n_i. The chosen *closeness* or *similarity* measure plays a vital role in determining the final prediction. The most commonly used *closeness* measure is the *Euclidean distance* of the feature point to its neighbours. For determining the final prediction, in the case of regression, the average value of y_i of k nearest neighbours is determined, while in the case of classification, the majority voting of classes of the nearest neighbours is considered.

The choice of the value of k plays a major role in the final model generalisability. The effect of k on the classification boundaries is shown in figure B.1. The higher the

parameter k (e.g. $k > 100$), the higher the bias towards a specific class (especially with highly dense neighbours). On the other hand, the lower the parameter k (e.g. $k = 1$), the higher the variance of the model since the number of neighbours considered at a time might not be sufficient to arrive at a robust prediction. An optimal value of k (e.g. $10 \leq k \leq 50$) gives a good decision boundary or classification as shown in the figure.

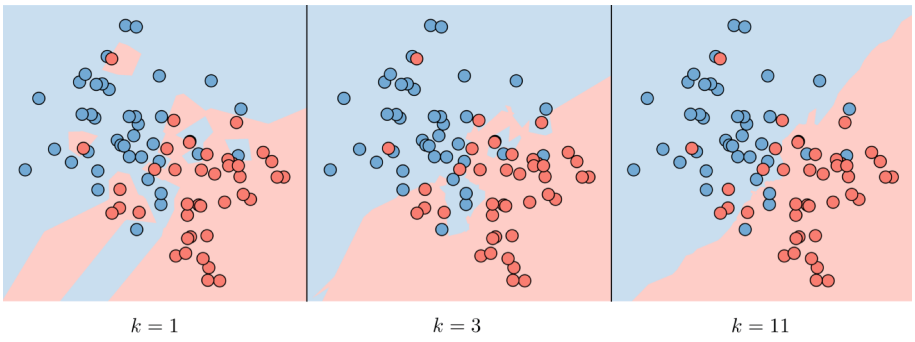


Figure B.1: The effect of k values on the classification. Image source: `https://stanford.edu/~shervine/teaching/cs-229/`.

KNN is currently used in BIANCA for WMH segmentation with $k = 40$ and the features specified in chapter 4.

Advantages: 1. the concept is intuitive and easy to understand, 2. easily extendable to multiple classes, 3. learns complex decision boundaries, handles missing data well and is robust to outliers.

Disadvantages: 1. performs poorly on high-dimensional feature space, 2. the prediction for each new instance is computationally expensive since the nearest neighbours need to be redetermined, 3. the distance metric is not often easy to choose and justify.

B.2 Support vector machine (SVM)

Support vector machine (SVM) [379] is a supervised kernel-based learning method. The SVM algorithm uses a set of mathematical functions, called kernels, that take data as input and transform them into the required form as shown in figure B.2. Out

of the different types of kernel functions available (e.g. linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid), the Gaussian RBF is most often used.

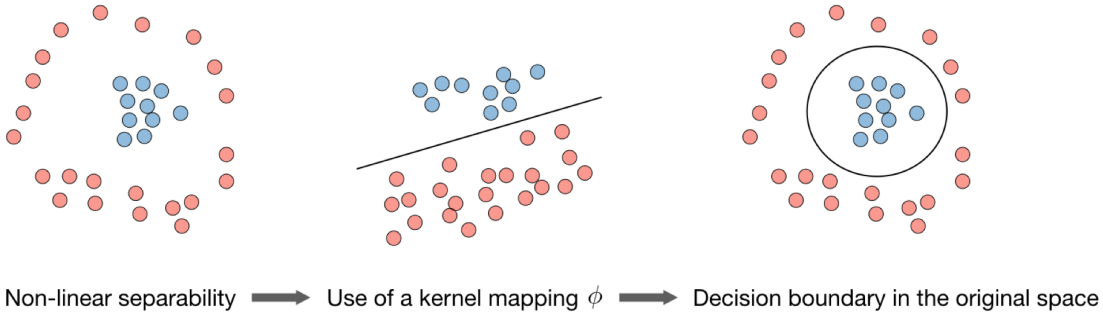


Figure B.2: Transformation of the input data into linear separable points using an appropriate kernel mapping function. Image source: `https://stanford.edu/~shervine/teaching/cs-229/`.

SVM forms a hyperplane between training vectors of different classes (i.e WMHs and non-WMHs) and maximises the margin, thus creating the largest possible distance of the vectors from the separating hyperplane (shown in figure B.3). Let the training data consist of N vectors in the feature space after kernel mapping, $x_i,\ i = 1...N$ where $x \in R^n$ and the corresponding targets be $y_i,\ i = 1...N$, where $y_i \in [-1, 1]$ ('-1' and '1' correspond to non-WMH and WMH classes respectively). The separating hyperplane is defined as,

$$x : f(x) = x^T b + b_0 = 0 \tag{B.2}$$

b is a unit vector $||b|| = 1$. The classifier function is given by,

$$C(x) = sign[x^T b + b_0] \tag{B.3}$$

Let M be the geometric margin between the vectors and the separating hyperplane given by,

$$M = \frac{F}{||b||} = \min_{i=1...N} \quad y_i\left(\frac{b}{||b||} \cdot x + \frac{b_0}{||b||}\right) \tag{B.4}$$

Hence, $2M$ is the maximum distance of separation between the two classes (figure B.3). Hence, finding the largest margin could be posed as an optimisation problem

given by,

$$\begin{aligned} &\max_{b,b_0,||b||} \quad M \\ &\text{Subject to } y_i(x^T b + b_0) \geq M, i = 1, ...N \end{aligned} \tag{B.5}$$

Based on the inverse relationship between M and $||b||$, the equation could be rephrased as the convex quadratic optimisation problem,

$$\begin{aligned} &\min_{b,b_0} \quad ||b||^2 \\ &\text{Subject to } y_i(x^T b + b_0) - 1 \geq 0, i = 1, ...N \end{aligned} \tag{B.6}$$

Hence, the SVM classifier operates by solving the above optimisation problem to maximise the separation between the classes, by minimising a loss/cost function. The most commonly used loss function for this optimisation is the Hinge loss function given by,

$$L_H(y) = max(0, 1 - yz) \tag{B.7}$$

where z is the classifier score. The band M is determined by the vectors (called '*support vectors*') that are closest to the hyperplane, forming the boundary of classes on either side of the hyperplane (figure B.3). The optimisation is done so that support vectors reduce the upper bound on the expected generalisation error [380].

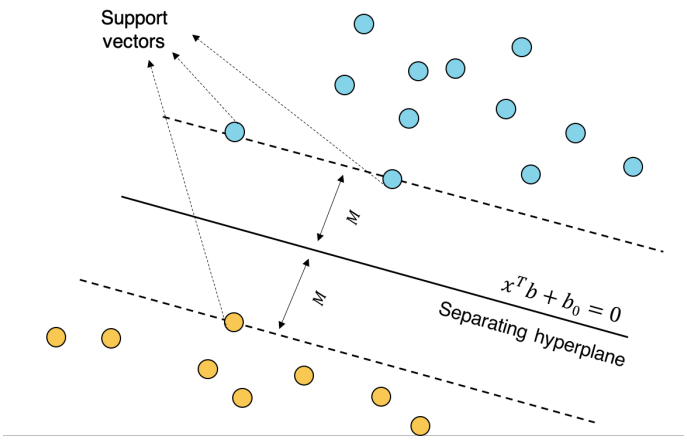


Figure B.3: Optimisation in SVM. Separating hyperplane between the vectors of different classes, and support vectors defining the margin of the vectors from the hyperplane.

The SVM classifier has two parameters: 1. C-value, a regularisation parameter, provides the trade-off between correct classification of samples and the margin M, 2.

the parameter γ is the inverse of radius of influence of support vectors in deciding the margin (for large values of γ, area of influence includes only the support vectors and hence results in overfitting).

We experimented with parameters of the SVM classifier for WMH segmentation in chapter 4 and found that the best results were achieved for a C-value and γ of 1.0.

Advantages: 1. quite robust to overfitting, especially in high-dimensional feature space, 2. maximisation of the margin is a convex optimisation problem, hence solvable using various available library packages and has a unique solution.
Disadvantages: 1. has many hyperparameters (e.g. kernel type, C-value, γ and optimisation parameters) to tune, 2. the algorithm is difficult to parallelise and hence is not easily scalable to large datasets.

B.3 Random forest (RF)

Random forest (RF) [341] is a supervised ensemble classification algorithm based on *bootstrap aggregating* or *bagging*. The idea of bagging is to average many noisy but approximately unbiased models, and hence reduce the variance. RF uses random sets of training data points and random subsets of features for forming a multitude of decision tree predictors as shown in figure B.4. With the independent sampling of random subsets of training data with identical distribution for all the individual trees, the bias of bagged trees is the same as that of the individual trees, and the final classification could be improved only by reducing the variance. Hence, bagging avoids overfitting and is robust to noise [341].

In the decision trees in RF, every node splits the data into two classes based on features (e.g. intensity), and the measure for determining the optimal split at each node is called its *impurity* measure. This measure is typically the *Gini impurity measure* (default in *scikit-learn* package) or *information gain/entropy* for classification and *variance* for regression. The Gini impurity measure (variation of entropy measure

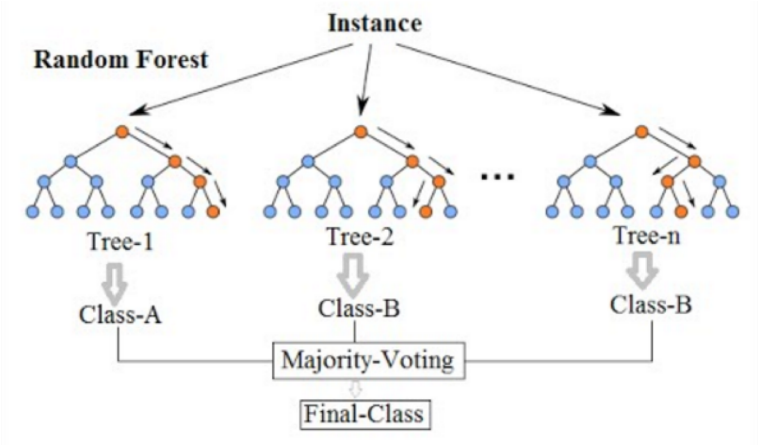


Figure B.4: A simplified version of an RF classifier. Image courtesy: Will Koehrsen, Medium; `https://medium.com/`.

of decision trees) for N classes could be given as,

$$g(p) = 1 - \sum_{i=1}^{N} p_i^2 \tag{B.8}$$

where p_i is the proportion of data labelled with class i. The information gain could be defined as the difference between the entropy in the parent node and the weighted sum of entropies in the child nodes. The split that provides the minimum impurity measure is taken as the best split and the features are ranked based on the overall average value of this measure. The final classification decision is by majority voting (in the case of classification) or average (in the case of regression) of the outputs of the individual trees. One of the popular ways of measuring the prediction error of RF is to determine out-of-bag (OOB) error. OOB error is defined as the mean of prediction errors on training samples x_i, using only the trees that did not have x_i in their bootstrap sample.

For WMH segmentation, we achieved the best results for the RF classifier with the specific range of parameter values for 50 trees of depth 25 for each tree, with 400 samples at each leaf node and minimum Gini impurity measure $g(p)$ at each split as 1×10^{-3}.

Advantages: 1. highly flexible and accurate, 2. can handle redundant features, since it estimates feature importance, 3. robust to overfitting and scales well to large

datasets (since random subsets of samples are used at a given time), 4. can handle class-imbalance by setting class weights, 5. can learn non-linear hypothesis functions. **Disadvantages:** 1. interpretation of results might be difficult, 2. feature importance/ranking may not be consistent to variations in the training data.

B.4 Adaptive boosting of classifiers (AdaBoost)

Boosting, one of the powerful learning methods among the committee-based approaches [381], is based on the idea that combining several weak learners or *classifiers* would lead to a stronger *committee* of learners. Adaptive boosting (AdaBoost) [382] sequentially applies a weak classifier (e.g. decision trees) to repeatedly modified versions of the data (with updated weights) to create a sequence of weak classifiers $C_k(x)$, where $k = 1...K$, as shown in the figure B.5. The predictions of all the weak classifiers are combined through a weighted majority vote to get the final prediction as given by,

$$C(x) = \text{sign}(\sum_{k=1}^{K} a_k C_k(x)) \tag{B.9}$$

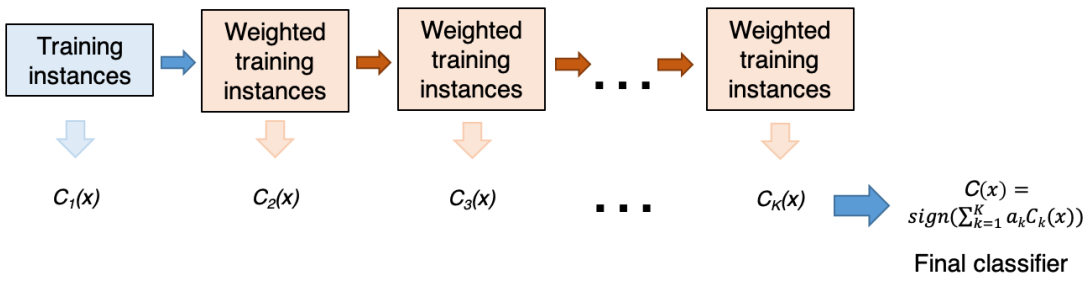


Figure B.5: A schematic diagram of an AdaBoost classifier.

The weights of the individual classifiers, a_k, $k = 1...K$ are determined by the boosting algorithm, in such a way to up-weight the more accurate classifiers. For each iteration, the data modification comprises of applying the weights a_k to each training instance (feature points of voxels), starting with equal weights in the first iteration. With each successive iteration, the weights are updated and the classifiers are applied again. At the m$^{\text{th}}$ iteration, the weights of the misclassified training

instances are increased (or up-weighted), while the weights of correctly classified samples are reduced. This forces the classifier to focus more on the instances that were missed in the previous iteration.

Boosting could be considered as an additive expansion of a set of *basis* functions (weak classifiers in this case). Hence, exponential loss functions are popular and more suitable loss functions for fitting a boosting model, due to computational convenience - it leads to simple re-weighting of the models. The exponential loss function is given by,

$$L(y, f(x)) = e^{-yf(x)} \tag{B.10}$$

Apart from the exponential loss function, the cross-entropy loss function is also commonly used. Boosting trees, where decision trees are used as weak classifiers in the AdaBoost algorithm, are often used for classification and regression tasks. Let R_i, $i = 1...N$ denote the disjoint specific region representing a terminal node of a tree and given the constant γ_i for each R_i, the decision tree could be mathematically formulated as,

$$T_i(x) = \sum_{i=1}^{N} \gamma_i I(x \in R_i) \tag{B.11}$$

And the boosted tree model is given by the sum of K tree models,

$$C_K(x) = \sum_{k=1}^{K} T_i(x) \tag{B.12}$$

As specified in chapter 4, after experimenting with the specific range of parameter values, we achieved the best results for the AdaBoost classifier by using 2 decision trees in each weak classifier, and with 30 such weak classifiers. The learning rate was 1.0 for fitting the classifier and the exponential loss function was used.

Advantages: 1. provides a wide variety of choices for weak classifiers and it is easy to implement, 2. does not need any prior knowledge of weak classifiers, 3. provides fairly good generalisation across tasks, 4. avoids overfitting, when a discriminative weak classifier is used.

Disadvantages: 1. does not handle noisy data very well, 2. not robust to outliers, 3. using weak classifiers that are too weak can lead to low margins between classes.

B.5 Neural networks (NNs)

Neural networks (NNs) (or *artificial neural networks*, [383]) are non-linear statistical models, with an architecture shown in the network diagram in figure B.6. As the name suggests, NNs were initially designed to emulate the functioning of the human brain. Hence, the units in each layer are called *neurons* and the connections act like synapses.

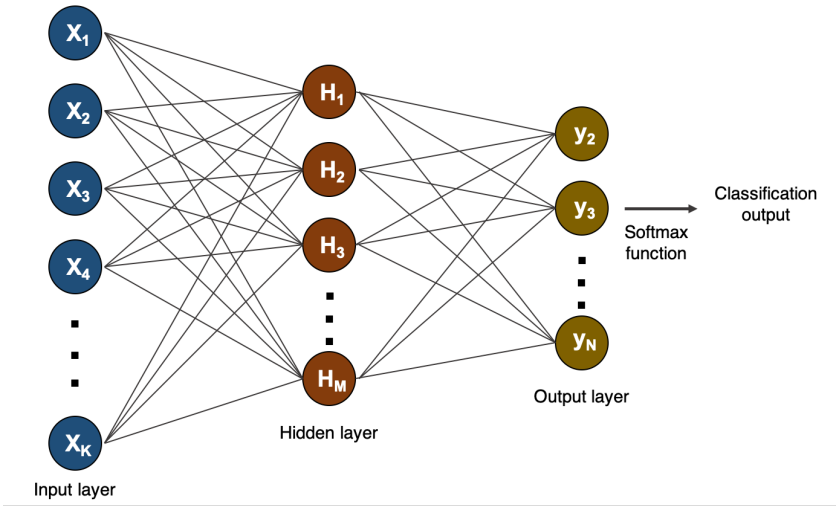


Figure B.6: Network diagram of a typical NN (shown for a sample model with two-stages), with K inputs, M hidden units and N output values.

For N classes, the output layer of a NN has N neurons, with the N^{th} unit indicating the probability of the instance belonging to class N. Let the target labels be $y_n \in [0,1]$, $n = 1...N$ (where '0' and '1' represent non-WMHs and WMHs respectively). The intermediate layer of the NN is called a hidden layer, and the features (intermediate output values, H_m, $m = 1...M$) at the hidden layer are derived as linear combinations of the inputs X, and the target outputs y_n at the final layer are modelled as functions of the linear combinations of H_m. Mathematically,

$$\begin{aligned} H_m &= \sigma(a_{0m} + a_m^T X),\ m = 1...M, \\ y_n &= b_{0n} + b_n^T H,\ n = 1...N \end{aligned} \tag{B.13}$$

where H = $(H_1, H_2, ...H_m)$, y = $(y_1, y_2, ...y_n)$ and $\sigma(\cdot)$ is the non-linear *activation function*. The additional *bias* units in the layers are indicated by the constant terms a_{0m} and b_{0n}. The outputs y_n are typically passed through the transformation $G(\cdot)$ to get the final output y. The function $G(\cdot)$ is usually the identity in the case of regression and a non-linear function in the case of classification, usually the *softmax* function given below:

$$G_n(y) = \frac{e^{y_n}}{\sum_{x=1}^{N} e^{y_x}} \tag{B.14}$$

The unknown parameters of the NN, often called weights, are learnt by fitting the model on the training data, using activation functions and an error/loss function $R(\theta)$ (cross-entropy function is often used). The most commonly used approach for minimising $R(\theta)$ is by gradient descent, called *back-propagation* in this context. In this approach, the gradients are derived using the chain rule and computed by a forward/backward passes, updating the weights with each step/epoch.

On experimenting with a range of parameter values, we achieved the best results for the NN classifier (chapter 4) by using 100 neurons in the hidden layer, with 1800 iterations/epochs and an initial learning rate of 1×10^{-3} for updating weights.

Advantages: 1. can easily identify complex non-linear relationship between data and dependent variables, 2. can handle noisy data.

Disadvantages: 1. difficult to interpret, 2. more prone to overfitting, 3. determines local maxima as opposed to a unique solution, 4. does not handle missing data well and is not robust to outliers, 5. if more neurons are used, it is not computationally scalable for large datasets.

Appendix C

Performance evaluation metrics

This appendix provides a brief description of various performance evaluation measures used in our work. We have categorised them into four categories: 1. voxel-wise overlap measures, 2. cluster-wise measures, 3. region-level measures and 4. subject-level measures.

C.1 Voxel-wise overlap measures

The voxel-wise overlap measures that are used mainly for evaluating WMH segmentation methods in chapters 4 and 5 are the following:

- **Dice Similarity Index (SI):**

$$SI = \frac{2 \times (\mathrm{TP})}{(\mathrm{T} + \mathrm{P})} = \frac{2 \times (\mathrm{TP})}{(2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN})} \tag{C.1}$$

  where TP, FP, FN, T and P refers to true positive, false positive, false negative, true and positive lesion voxels respectively. True lesion voxels refer to the lesion voxels in the manual segmentation and positive lesion voxels are the voxels labelled as lesions by the classifier.

- **Voxel-wise true positive rate (TPR):** the number of true positive lesion voxels divided by the number of true lesion voxels as given by,

$$TPR = \frac{TP}{(TP + FN)} \tag{C.2}$$

where TP and FN are the number of true positive and false negative (true lesion voxels but missed by the segmentation method) voxels respectively.

- **Voxel-wise false positive rate (FPR):** the number of false positive (FP) lesion voxels (voxels incorrectly labelled as lesion) divided by the number of non-lesion voxels, as given by,

$$FPR = \frac{FP}{(FP + TN)} \tag{C.3}$$

where TN is the number of true negative (non-lesion) voxels.

- **Absolute volume difference (AVD):** the absolute value of the difference between the volume of manually segmented lesion voxels and the volume of detected lesion voxels, as a percentage of the manually segmented lesion voxels. Mathematically,

$$AVD = \frac{\mid V_{mles} - V_{dles} \mid}{V_{mles}} \times 100\ (\%) \tag{C.4}$$

where V_{mles} and V_{dles} are the total volumes of manually segmented and detected lesions respectively

C.2 Cluster-wise measures

The cluster-wise metrics used in this work for the evaluation of WMH and CMB detection in chapters 4, 5 and 6 are as follows:

- **Cluster-level TPR:** the number of true positive clusters (i.e. WMH lesions or CMBs) divided by the total number of true clusters as given by,

$$\text{cluster-wise TPR} = \frac{TP_{clus}}{(TP_{clus} + FN_{clus})} \tag{C.5}$$

where TP_{clus} and FN_{clus} are true positive and false negative clusters respectively.

- **Cluster-wise precision:** the number of true positive clusters divided by the total number of detected clusters as given by,

$$\text{cluster-wise precision} = \frac{TP_{clus}}{(TP_{clus} + FP_{clus})} \tag{C.6}$$

where TP_{clus} and FP_{clus} are true positive and false positive clusters respectively.

- **Cluster-wise F1-measure:** Given the cluster-wise TPR (eqn. C.5) and cluster-wise precision (eqn. C.6), the value of cluster-wise F1- measure is calculated by,

$$\text{cluster-wise F1-measure} = 2 \times \frac{(\text{cluster-wise precision} \times \text{cluster-wise TPR})}{(\text{cluster-wise precision} + \text{cluster-wise TPR})} \tag{C.7}$$

$$= \frac{2 \times (TP_{clus})}{(2 \times TP_{clus} + FP_{clus} + FN_{clus})} = Dice_{clus} \tag{C.8}$$

where TP_{clus}, FP_{clus} and FN_{clus} are the number of true positive, false positive and false negative lesion clusters respectively. Hence, it can be observed that cluster-wise F1 measure is equivalent to SI (Dice) values at the cluster-level.

- **95$^{\text{th}}$ percentile of Hausdorff distance measure (H95):** Given two sets of points, the bidirectional Hausdorff distance (or *Pompeiu-Hausdorff* distance) is defined as the longest distance (in mm) from the points in one set to the closest point in the second set. Given the coordinates of edge voxels of manually segmented WMHs $W_m = (x_m, y_m, z_m)$ and the edge voxels of detected WMHs $W_m = (x_d, y_d, z_d)$, the one-sided Hausdorff distance from W_m to W_d is given by,

$$hd(W_m, W_d) = \underset{w_m \in W_m}{max} \; \underset{w_d \in W_d}{min} ||w_m - w_d||_2 \tag{C.9}$$

Similarly, the one-sided Hausdorff distance from W_d to W_m is given by,

$$hd(W_d, W_m) = \underset{w_d \in W_d}{max} \; \underset{w_m \in W_m}{min} ||w_d - w_m||_2 \tag{C.10}$$

Hence, the bidirectional Hausdorff distance between the two set of coordinates is given by,

$$HD(W_m, W_d) = max(hd(W_m, W_d), hd(W_d, W_m)) \tag{C.11}$$

We consider the 95^{th} (H95) percentile of the bidirectional Hausdorff distances $HD(W_m, W_d)$ as our evaluation metric.

- **Average number of FPs per subject (FPavg):** for a given dataset D, FPavg is defined as the ratio of the total number of detected FPs to the number of subjects (or images) in the dataset, as given by,

$$FPavg = \frac{\text{Total number of FPs}}{\text{Number of subjects in } D} \tag{C.12}$$

C.3 Region-level measures

We used region-level metrics for determining the performance of the automated CMB rating method in the individual structures or regions, based on the confusion matrices in chapter 6. The confusion matrix for a multi-class problem is shown in table C.1. In our case, the classes indicate the regions (e.g. infratentorial, deep and lobar). The region-level metrics used are as follows:

		Predicted labels			
	Class	1	2	···	N
Manual labels	1	$c_{(1,1)}$	$c_{(1,2)}$	···	$c_{(1,N)}$
	2	$c_{(2,1)}$	$c_{(2,2)}$	···	$c_{(2,N)}$
	···	···	···	···	···
	N	$c_{(N,1)}$	$c_{(N,2)}$	···	$c_{(N,N)}$

Table C.1: Confusion matrix for a multi-class problem. Manual labels are shown along the rows and the predicted labels along the columns for classes 1 to N. The values $c_{(i,j)}$ in each cell indicates the number of labels predicted as class j while they were manually labelled as class i.

- **Region-wise sensitivity:** For a given class, region-wise sensitivity is defined as the ratio of the number of predicted TP labels to the number of true labels

(TP + FN). Hence, for N classes, the sensitivity for class i is given by,

$$sens_i = \frac{c_{(i,i)}}{\sum_{j=1}^{j=N} c_{(i,j)}} \tag{C.13}$$

Hence, the overall region-wise sensitivity is the average of the sensitivities of all the classes as given by,

$$\text{region-wise sensitivity} = \frac{1}{N}\sum_{i=1}^{i=N} sens_i = \frac{1}{N}\sum_{i=1}^{i=N} \frac{c_{(i,i)}}{\sum_{j=1}^{j=N} c_{(i,j)}} \tag{C.14}$$

- **Region-wise precision:** For a given class, region-wise precision is defined as the ratio of the number of predicted TP labels to the number of predicted labels (TP + FP). Hence, for N classes, the precision for class i is given by,

$$prec_i = \frac{c_{(i,i)}}{\sum_{j=1}^{j=N} c_{(j,i)}} \tag{C.15}$$

Hence, the overall region-wise precision is the average of the precisions of all the classes as given by,

$$\text{region-wise precision} = \frac{1}{N}\sum_{i=1}^{i=N} prec_i = \frac{1}{N}\sum_{i=1}^{i=N} \frac{c_{(i,i)}}{\sum_{j=1}^{j=N} c_{(j,i)}} \tag{C.16}$$

- **Region-wise F1-measure:** Given the region-wise sensitivity and region-wise precision determined using eqns. C.14 and C.16 respectively, the region-wise F1-measure is given by,

$$\text{region-wise F1-measure} = 2 \times \frac{\text{region-wise sensitivity} \times \text{region-wise precision}}{\text{region-wise sensitivity} + \text{region-wise precision}} \tag{C.17}$$

Similarly, F1-measures for the individual regions i could be determined by,

$$\text{F1-measure}_i = 2 \times \frac{sens_i \times prec_i}{sens_i + prec_i} \tag{C.18}$$

- **Accuracy:** For N classes, the accuracy of prediction is defined as the ratio of

predicted TP labels to the total number of labels, as given by,

$$\text{Accuracy} = \frac{\sum_{i=1}^{i=N} c_{(i,i)}}{\sum_{j=1}^{j=N} \sum_{i=1}^{i=N} c_{(i,j)}} \times 100 \ (\%) \tag{C.19}$$

C.4 Subject-level measures

We used the following subject-level metrics for evaluating the performance of the CMB candidate subject preselection pipeline in chapter 6:

- **Subject-level true positive rate (TPR):** For a given dataset D, the subject-level TPR is the number of predicted TP CMB subjects (S_{TP}) divided by the number of true CMB subjects, as given by,

$$\text{subject-level TPR} = \frac{S_{TP}}{(S_{TP} + S_{FN})} \tag{C.20}$$

  where S_{FN} is the number of false negative subjects.

- **Subject-level specificity:** For a given dataset D, the subject-level specificity is the number of predicted true negative subjects (S_{TN}) divided by the number of non-CMB subjects, as given by,

$$\text{subject-level specificity} = \frac{S_{TN}}{(S_{TN} + S_{FP})} \tag{C.21}$$

  where S_{FP} is the number of false positive subjects.

- **Subject-level accuracy:** For a given dataset D consisting of S_N subjects, the subject-level accuracy is given by the number of correctly predicted CMB and non-CMB subjects (S_{TP} and S_{TN}) divided by the total number of subjects,

$$\text{subject-level accuracy} = \frac{S_{TP} + S_{TN}}{S_N} \times 100 \ (\%) \tag{C.22}$$

Appendix D

The Microbleed Anatomical Rating Scale (MARS): sample form

A sample form of the MARS rating scale for cerebral microbleeds (CMB) is provided in figure D.1.

We used the regions and structures as specified in this form for obtaining automated ratings of CMBs in chapter 6. We combined both the left (L) and right (R) divisions of the structures/regions to provide a single CMB count in each structure/region, thus avoiding any lateralisation. Also, we combined both possible and definite cases to get a single score for each structure/region.

Patient ID: Date of Birth _ _/_ _/_ _ _ _ Date of MRI _ _/_ _/_ _ _ _

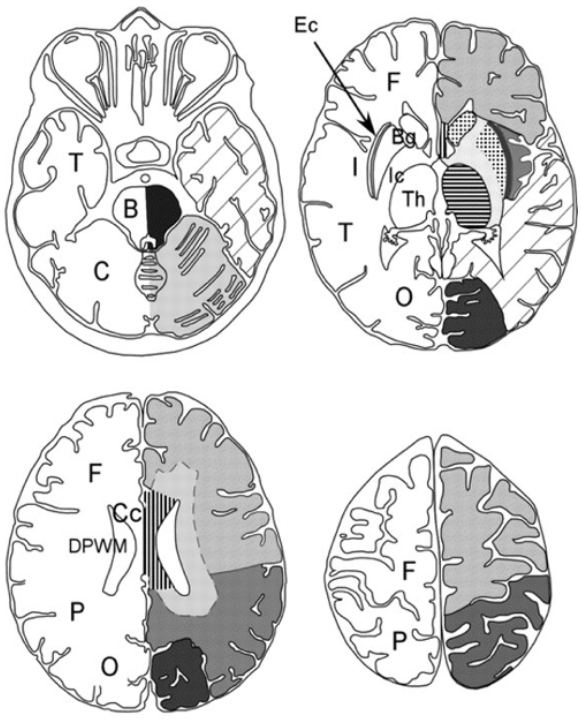
DEFINITE MICROBLEEDS: Small, round, well-defined, hypointense on GRE T2*; 2-10 mm; not well seen on T2

MICROBLEED MIMICS

- Vessels: linear / curvilinear lesions in subarachnoid space, usually cortical or juxta-cortical (visible on T2)
- Mineralization in globi pallidi or dentate nuclei: symmetrical hypointensities (may be bright flecks on CT)
- Haemorrhages within area of infarction (look at the T2, FLAIR or DWI sequences to identify infarction)
- Air-bone interfaces: frontal / temporal lobes (check adjacent GRE T2* slices to clarify)
- Partial volume artifact at the edges of the cerebellum (check adjacent GRE T2* to clarify)
- Small haemorrhages close to a large ICH (visible on GRE T2*) or to an infarct (visible on T2, FLAIR or DWI)

Right Left

		DEFINITE		POSSIBLE	
		R	L	R	L
Infratentorial TOTAL	Brainstem (B)				
	Cerebellum (C)				
Deep TOTAL	Basal Ganglia (Bg)*				
	Thalamus (Th)				
	Internal Capsule (Ic)				
	External Capsule (Ec)				
	Corpus Callosum (Cc)				
	Deep and periventricular WM (DPWM)				
Lobar TOTAL**	Frontal (F)				
	Parietal (P)				
	Temporal (T)				
	Occipital (O)				
	Insula (I)				
	TOTALS				



* (Caudate, Lentiform), **Lobar regions include cortex and subcortical white matter

Figure D.1: A sample Microbleed Anatomical Rating Scale (MARS) form for CMB rating.

Appendix E

Modelling the WMH distribution using Bayesian inference: additional experiments

E.1 Number of iterations in the optimisation step for convergence

For the optimisation of the Bayesian inference model, we specified the maximum number of iteration to 50 empirically, setting more number of iterations than needed. Figure E.1 shows the convergence curves of the norm values of the gradient (calculated from eqn. 3.20) for the OXVASC and UKBB datasets. As specified in section 3.2.4, we set the final convergence tolerance value for the decrease in the norm value of gradient as 1×10^{-4}. In case of both datasets, we model WMH distribution with respect to age. In case of the OXVASC dataset, the optimisation is faster since it consists of vascular cohort with relatively high WMH loads. Hence, R matrix is denser and its values increase for the higher age bins as expected, thus resulting in the quicker convergence.

For the UKBB dataset, the hypertensive (HT) group has higher WMH load in general (we showed in the results that the WMH probabilities increase significantly in the deep regions) when compared to the non-hypertensive (non-HT) group. Hence the non-HT group is a difficult case where the values of R matrix are quite lower and sparser, compared to the HT group, resulting in smaller changes in the gradient and

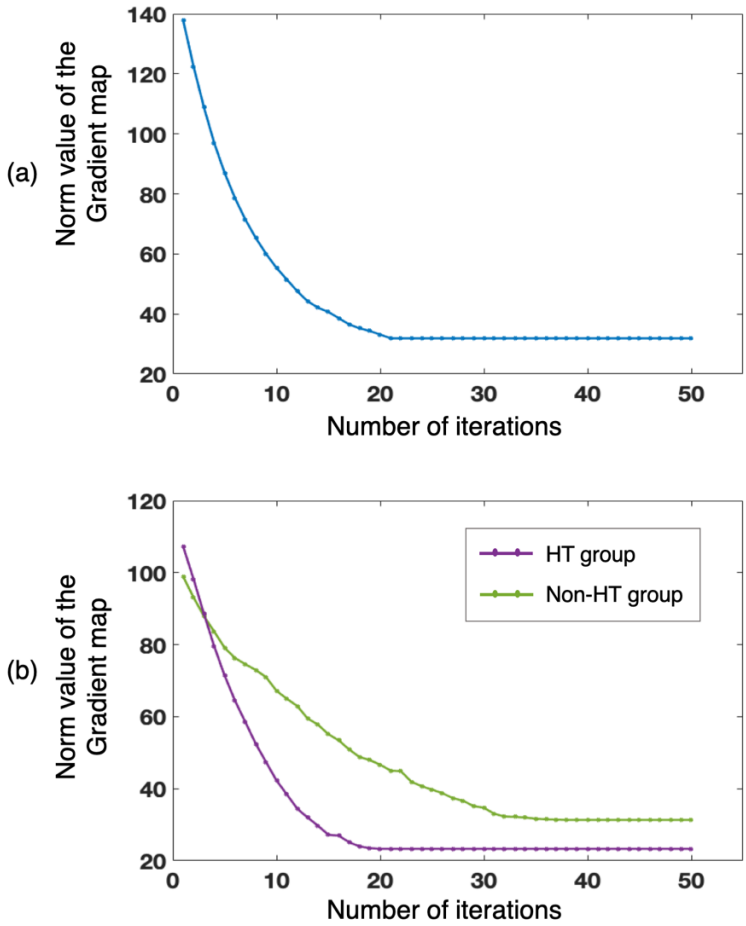


Figure E.1: Convergence curves for the optimisation of the Bayesian inference algorithm. The convergence of the norm value of the gradient maps shown for (a) the OXVASC dataset (blue) and (b) the UKBB dataset. The hypertensive (HT) and the non-hypertensive (non-HT) groups of the UKBB dataset are shown in purple and green colours respectively.

hence the slower convergence. Therefore, we can observe that while the optimisation algorithm converges at around 20 iterations for the straight-forward cases, it requires slightly more iterations for the difficult cases. Hence, the empirically chosen value of 50 would allow for the convergence of the algorithm for different cases.

E.2 Effect of the intermediate b-spline knot spacing values on simulated data

In section 3.5.1.1 of chapter 3, we have discussed the effect of b-spline knot spacing on the simulated data, with the knot spacing values of 2 and 6 voxels. The MSE values show that knot spacing of 2 voxels provided the best results on the simulated

data. In this appendix, we show the effect of intermediate knot spacing values (3, 4 and 5 voxels) and the standard deviation of the kernel K used in the ground truth estimation independently on the error values on the simulated data. Figure E.2 shows the effect of change in knot spacing values and the kernel K on the error values.

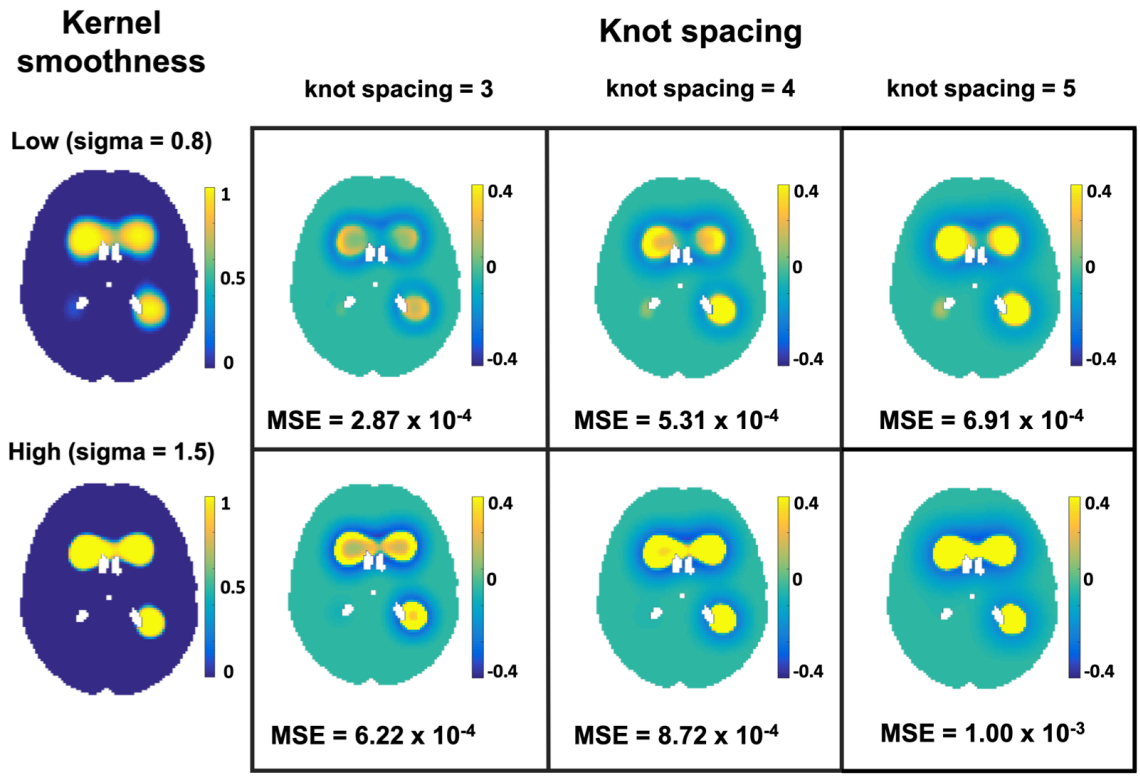


Figure E.2: Comparison of error maps for various values of knot spacing in our Bayesian spline method $\hat{\theta}$ and standard deviation (sigma) values for smoothing kernel in simulation ground truth θ_{true}. Within the table, the top row shows error maps and mean-squared error (MSE) values, when θ_{true}, generated using a smaller symmetric kernel (sigma = 0.8) is compared against $\hat{\theta}$ modelled with b-spline knot spacing of 3, 4 and 5 voxels respectively. Similarly, the bottom row shows error maps and MSE values for comparison of smoothing kernel K with larger symmetric kernel (sigma = 1.5) in θ_{true} against $\hat{\theta}$ with the three values of b-spline knot spacing.

As shown in the figure, all cases of knot spacing provide lower MSE values against the ground truth with smaller K with a sigma value of 0.8, compared to the larger kernel (sigma = 1.5). Also, the MSE values increases with increase in knot spacing values for both the cases of the ground truth kernels. From figure 3.5 in chapter 3, the MSE value corresponding to the knot spacing of 2 voxels is lower than that corresponding to the knot spacing of 3 voxels. Hence, as specified in the section 3.5.1.1,

we used the knot spacing of 2 for the simulated and real data experiments. The knot spacing values of 5 and 6 voxels provided high MSE values since they introduce spurious smoothness in the results, thus resulting in high values in the error maps.

Appendix F

Analysis of LOCATE threshold histograms

We analysed the characteristics of LOCATE thresholds by exploring (i) how representative the LOCATE threshold histogram is for a specific dataset and (ii) how LOCATE threshold histogram characteristics changes for various datasets.

F.1 Case 1: LOCATE threshold histogram characteristics within a specific dataset

We analysed the LOCATE threshold values for different subjects of a specific dataset. We chose 5 randomly sampled subjects from the vascular cohort of the OXVASC dataset with various WMH loads (27797, 16033, 6557, 4673 and 15774 mm^3) and determined their LOCATE threshold values. Figure F.1(a-e) shows the threshold maps and normalised histograms of estimated LOCATE threshold values for the above selected subjects and figure F.1f shows the superimposition of the histograms. The normalised histogram provides a fair comparison since the number of Voronoi subregions N_{Vor_reg} (mentioned in the histogram plots) varies between different subjects. From the figure, it can be observed that the distribution of LOCATE thresholds remains almost consistent for various subjects within a specific population. Irrespective of the WMH loads, the LOCATE threshold histograms were widely spread with a small peak at 0.5. The wider spread could be due to the presence of more low contrast deep WMHs in addition to the periventricular WMHs.

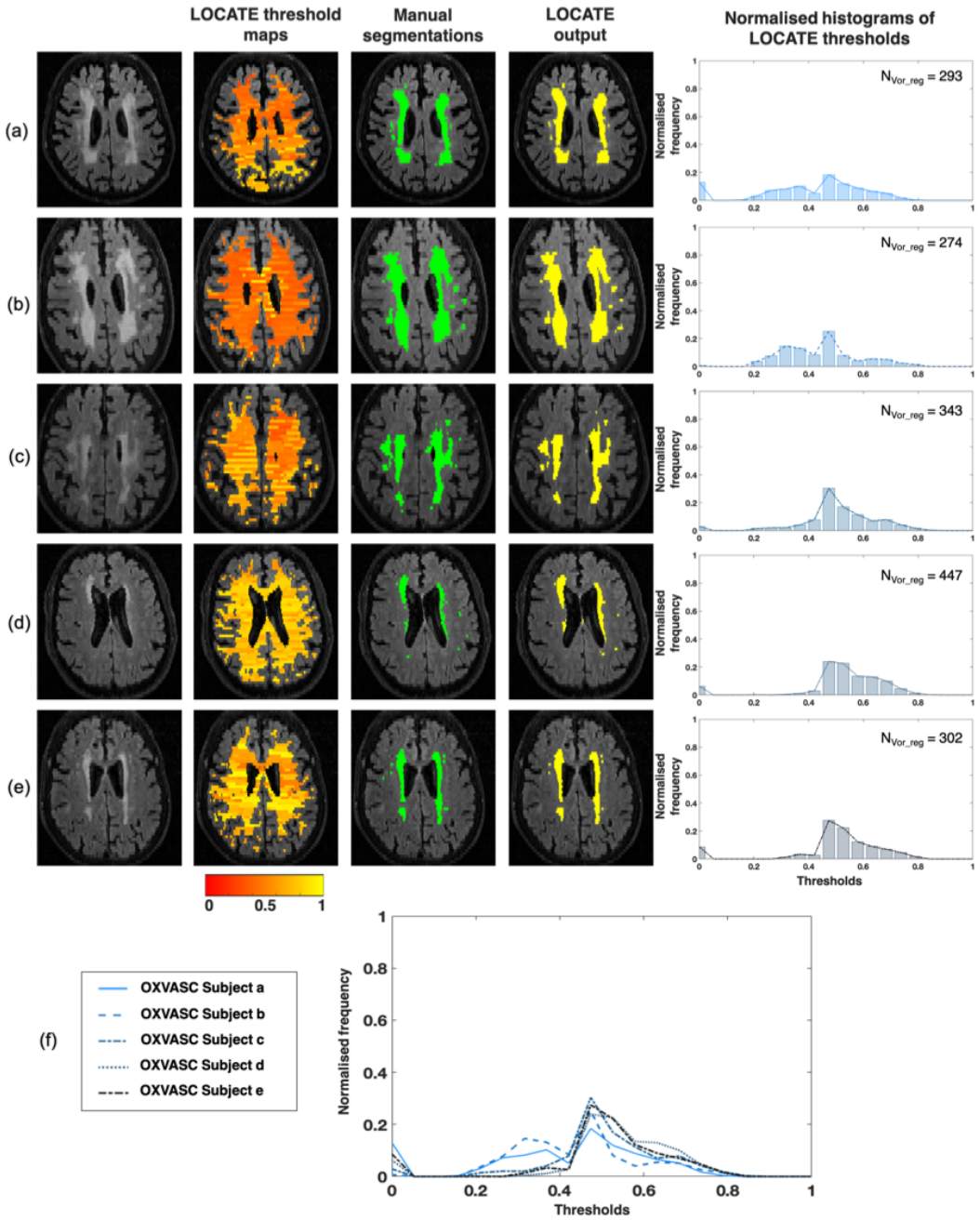


Figure F.1: Analysis of LOCATE thresholds within a dataset. Normalised histograms shown along with LOCATE threshold maps (the colour bar indicates the threshold values), manual segmentations and LOCATE outputs shown for 5 sample subjects from (a) to (e); the number of Voronoi sub-regions N_{Vor_reg} is specified in the histogram plot for each case. (g) The normalised histograms from cases (a - e) superimposed for the visual comparison.

F.2 Case 2: LOCATE threshold histogram characteristics across different datasets

We analysed the characteristics of threshold values estimated by LOCATE for various datasets. We considered 6 representative subjects with wide range of clinical and lesion load characteristics as follows: Alzheimers disease (NDGEN), recent minor TIA (OXVASC), CADASIL, healthy controls (HC), data from memory clinic (MWSC Amsterdam and Utrecht). Figure F.2(a-f) shows the threshold maps and normalised histograms (number of Voronoi sub-regions N_{Vor_reg} mentioned in the histogram plots) of estimated LOCATE threshold values for the above selected subjects and figure F.2g shows the superimposition of the histograms. The threshold maps and the histograms show that the range and distribution of threshold values varies for different datasets based on their lesion contrast and demographics of the population.

We found that healthy controls and subjects with more periventricular lesions have high probability values in general (e.g. HC, NDGEN, MWSC - Amsterdam). The histograms of OXVASC and MWSC - Utrecht have a wider spread since the selected sample subjects had more deep lesions with various contrast at different spatial locations compared to other cases. The histogram of the LOCATE thresholds predicted for CADASIL subject is quite interesting - it has two distinct peaks at 0.2 and 0.5. Despite of the fact that the LOCATE model for CADASIL is trained on the OXVASC dataset, the peak at 0.2 confirms that this value, which is also the recommended global threshold by the clinician based on the visual inspection, is a reasonable choice since it has the highest frequency. However, the other peak at 0.5 indicates that a considerable number of sub-regions had a threshold value of 0.5, resulting in more accurate and less noisy detection with LOCATE, when compared to a global threshold (which is also evident from other cases such as the one in row 3 in Figure 4.18 in Chapter 4).

Overall, we found that while the distribution of LOCATE thresholds is fairly consistent within a specific population, it varies across different populations depending

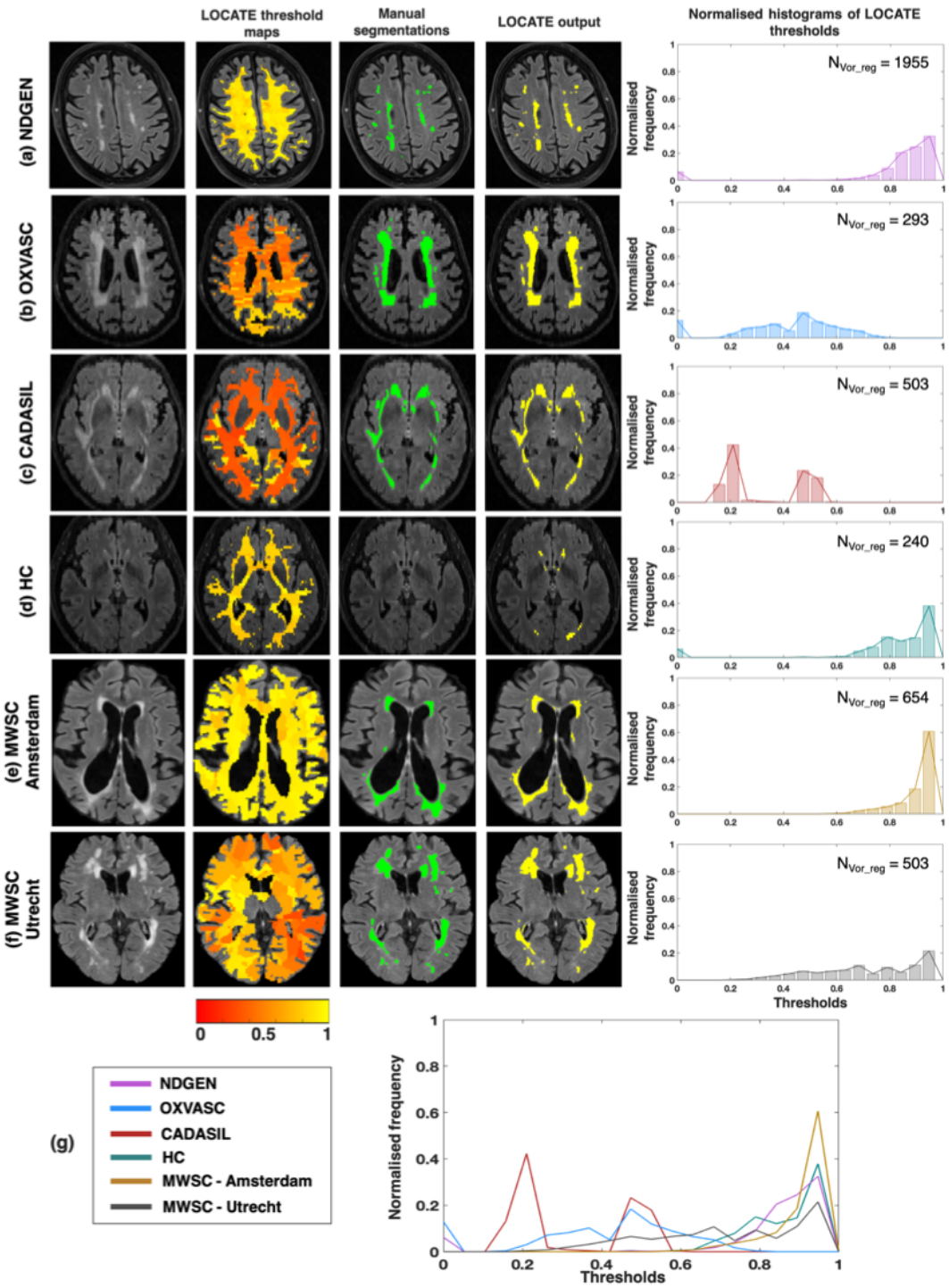


Figure F.2: Analysis of LOCATE thresholds for various datasets. Normalised histograms shown along with LOCATE threshold maps (the colour bar indicates the threshold values), manual segmentations and LOCATE outputs shown for 6 sample subjects, each from (a) NDGEN, (b) OXVASC, (c) CADASIL, (d) HC, (e) MWSC - Amsterdam and (f) MWSC - Utrecht datasets; the number of Voronoi sub-regions N_{Vor_reg} is specified in the histogram plot for each case. (g) The normalised histograms from cases (a - f) superimposed over one another for the visual comparison.

on the lesion characteristics (e.g. contrast and spatial distribution) and demographics of the population.

Bibliography

[1] Dixon M Moody, William R Brown, Venkata R Challa, and Randy L Anderson. Periventricular venous collagenosis: association with leukoaraiosis. *Radiology*, 194(2):469–476, 1995.

[2] Ewoud J van Dijk, Niels D Prins, Henri A Vrooman, Albert Hofman, Peter J Koudstaal, and Monique MB Breteler. Progression of cerebral small vessel disease in relation to risk factors and cognitive consequences: Rotterdam Scan study. *Stroke*, 39(10):2712–2719, 2008.

[3] Saima Hilal, Vincent Mok, Young Chul Youn, Adrian Wong, Mohammad Kamran Ikram, and Christopher Li-Hsian Chen. Prevalence, risk factors and consequences of cerebral small vessel diseases: data from three Asian countries. *J Neurol Neurosurg Psychiatry*, 88(8):669–674, 2017.

[4] Fred Rincon and Clinton Wright. Current pathophysiological concepts in cerebral small vessel disease. *Frontiers in aging neuroscience*, 6:24, 2014.

[5] Leonardo Pantoni. Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges. *The Lancet Neurology*, 9(7):689–701, 2010.

[6] Philip B Gorelick, Angelo Scuteri, Sandra E Black, Charles DeCarli, Steven M Greenberg, Costantino Iadecola, Lenore J Launer, Stephane Laurent, Oscar L Lopez, David Nyenhuis, et al. Vascular contributions to cognitive impairment and dementia: a statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*, 42(9):2672–2713, 2011.

[7] Joanna M Wardlaw, Eric E Smith, Geert J Biessels, Charlotte Cordonnier, Franz Fazekas, Richard Frayne, Richard I Lindley, John T O'Brien, Frederik Barkhof, Oscar R Benavente, et al. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *The Lancet Neurology*, 12(8):822–838, 2013.

[8] Dong-Wha Kang, Moon-Ku Han, Hye-Jin Kim, Sung-Cheol Yun, Sang-Beom Jeon, Hee-Joon Bae, Sun U Kwon, and Jong S Kim. New ischemic lesions coexisting with acute intracerebral hemorrhage. *Neurology*, 79(9):848–855, 2012.

[9] Gillian M Potter, Fergus N Doubal, Caroline A Jackson, Francesca M Chappell, Cathie L Sudlow, Martin S Dennis, and Joanna M Wardlaw. Counting cavitating lacunes underestimates the burden of lacunar infarction. *Stroke*, 41(2):267–272, 2010.

[10] Sarah E Vermeer, William T Longstreth Jr, and Peter J Koudstaal. Silent brain infarcts: a systematic review. *The Lancet Neurology*, 6(7):611–619, 2007.

[11] Philip Choi, Mandy Ren, Thanh G Phan, Michele Callisaya, John V Ly, Richard Beare, Winston Chong, and Velandai Srikanth. Silent infarcts and cerebral microbleeds modify the associations of white matter lesions with gait and postural stability: population-based study. *Stroke*, 43(6):1505–1510, 2012.

[12] Micaela Santos, Gabriel Gold, Enik Kvari, Francois R Herrmann, Vasilis P Bozikas, Constantin Bouras, and Panteleimon Giannakopoulos. Differential impact of lacunes and microvascular lesions on poststroke depression. *Stroke*, 40(11):3557–3562, 2009.

[13] Reinhold Schmidt, Franz Fazekas, Gertrude Kleinert, Hans Offenbacher, Kurt Gindl, Franz Payer, Wolfgang Freidl, Kurt Niederkorn, and Helmut Lechner. Magnetic resonance imaging signal hyperintensities in the deep and subcortical white matter: a comparative study between stroke patients and normal volunteers. *Archives of neurology*, 49(8):825–827, 1992.

[14] Karlijn F de Laat, Anil M Tuladhar, Anouk GW van Norden, David G Norris, Marcel P Zwiers, and Frank-Erik de Leeuw. Loss of white matter integrity is associated with gait disorders in cerebral small vessel disease. *Brain*, 134(1):73–83, 2010.

[15] William R Brown, Dixon M Moody, Venkata R Challa, Clara R Thore, and John A Anstrom. Venous collagenosis and arteriolar tortuosity in leukoaraiosis. *Journal of the neurological sciences*, 203:159–163, 2002.

[16] Alida A Gouw, Alexandra Seewann, Wiesje M Van Der Flier, Frederik Barkhof, Annemieke M Rozemuller, Philip Scheltens, and Jeroen JG Geurts. Heterogeneity of small vessel disease: a systematic review of MRI and histopathology correlations. *Journal of Neurology, Neurosurgery & Psychiatry*, 82(2):126–135, 2011.

[17] Malee S Fernando, Julie E Simpson, Fiona Matthews, Carol Brayne, Claire E Lewis, Robert Barber, Raj N Kalaria, Gill Forster, Filomena Esteves, Stephen B Wharton, et al. White matter lesions in an unselected cohort of the elderly: molecular pathology suggests origin from chronic hypoperfusion injury. *Stroke*, 37(6):1391–1398, 2006.

[18] Ashkan Shoamanesh, CS Kwok, and O Benavente. Cerebral microbleeds: histopathological correlation of neuroimaging. *Cerebrovascular diseases*, 32(6):528–534, 2011.

[19] David J Werring, Duncan W Frazer, Lucy J Coward, Nick A Losseff, Hilary Watt, Lisa Cipolotti, Martin M Brown, and H Rolf Jäger. Cognitive dysfunction in patients with cerebral microbleeds on T2*-weighted gradient-echo MRI. *Brain*, 127(10):2265–2275, 2004.

[20] Sang Won Seo, Byung Hwa Lee, Eun-Joo Kim, Juhee Chin, Yoon Sun Cho, Uicheul Yoon, and Duk L Na. Clinical significance of microbleeds in subcortical vascular dementia. *Stroke*, 38(6):1949–1951, 2007.

[21] H Pollock, M Hutchings, RO Weller, and E-T Zhang. Perivascular spaces in the basal ganglia of the human brain: their relationship to lacunes. *Journal of anatomy*, 191(3):337–346, 1997.

[22] Fergus N Doubal, Alasdair MJ MacLullich, Karen J Ferguson, Martin S Dennis, and Joanna M Wardlaw. Enlarged perivascular spaces on MRI are a feature of cerebral small vessel disease. *Stroke*, 41(3):450–454, 2010.

[23] AMJ MacLullich, JM Wardlaw, KJ Ferguson, JM Starr, JR Seckl, and IJ Deary. Enlarged perivascular spaces are associated with cognitive function in healthy elderly men. *Journal of Neurology, Neurosurgery & Psychiatry*, 75(11):1519–1523, 2004.

[24] Fei Han, Fei-Fei Zhai, Quan Wang, Li-Xin Zhou, Jun Ni, Ming Yao, Ming-Li Li, Shu-Yang Zhang, Li-Ying Cui, Zheng-Yu Jin, et al. Prevalence and risk factors of cerebral small vessel disease in a chinese population-based sample. *Journal of stroke*, 20(2):239, 2018.

[25] Andreas Charidimou and David J Werring. Cerebral microbleeds: detection, mechanisms and clinical challenges. *Future Neurology*, 6(5):587–611, 2011.

[26] Mani S Krishnan, John T O'Brien, Michael J Firbank, Leonardo Pantoni, Giovanna Carlucci, Timo Erkinjuntti, Anders Wallin, Lars-Olof Wahlund, Philip Scheltens, Elisabeth CW van Straaten, et al. Relationship between periventricular and deep white matter lesions and depressive symptoms in older people. The LADIS Study. *International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences*, 21(10):983–989, 2006.

[27] Steven M Greenberg, Meike W Vernooij, Charlotte Cordonnier, Anand Viswanathan, Rustam Al-Shahi Salman, Steven Warach, Lenore J Launer, Mark A Van Buchem, Monique MB Breteler, Microbleed Study Group, et al. Cerebral microbleeds: a guide to detection and interpretation. *The Lancet Neurology*, 8(2):165–174, 2009.

[28] Joanna M Wardlaw, Maria C Valdés Hernández, and Susana Muñoz-Maniega. What are white matter hyperintensities made of? relevance to vascular cognitive impairment. *Journal of the American Heart Association*, 4(6):e001140, 2015.

[29] Jan Cees De Groot, Frank-Erik De Leeuw, Matthijs Oudkerk, Jan Van Gijn, Albert Hofman, Jellemer Jolles, and Monique MB Breteler. Periventricular cerebral white matter lesions predict rate of cognitive decline. *Annals of neurology*, 52(3):335–341, 2002.

[30] Ki Woong Kim, James R MacFall, and Martha E Payne. Classification of white matter lesions on magnetic resonance imaging in elderly persons. *Biological psychiatry*, 64(4):273–280, 2008.

[31] F Fazekas, R Kleinert, H Offenbacher, Rea Schmidt, G Kleinert, F Payer, H Radner, and H Lechner. Pathologic correlates of incidental MRI white matter signal hyperintensities. *Neurology*, 43(9):1683–1683, 1993.

[32] Sarah E Vermeer, Monika Hollander, Ewoud J van Dijk, Albert Hofman, Peter J Koudstaal, and Monique MB Breteler. Silent brain infarcts and white matter lesions increase stroke risk in the general population: the Rotterdam Scan Study. *Stroke*, 34(5):1126–1129, 2003.

[33] Charles DeCarli, Evan Fletcher, Vincent Ramey, Danielle Harvey, and William J Jagust. Anatomical mapping of white matter hyperintensities (WMH) exploring the relationships between periventricular WMH, deep WMH, and total WMH burden. *Stroke*, 36(1):50–55, 2005.

[34] Wei Wen and Perminder Sachdev. The topography of white matter hyperintensities on brain MRI in healthy 60-to 64-year-old individuals. *Neuroimage*, 22(1):144–154, 2004.

[35] DMJ Van den Heuvel, VH Ten Dam, AJM De Craen, F Admiraal-Behloul, H Olofsen, ELEM Bollen, J Jolles, HM Murray, GJ Blauw, RGJ Westendorp, et al. Increase in periventricular white matter hyperintensities parallels decline in mental processing speed in a non-demented elderly population. *Journal of Neurology, Neurosurgery & Psychiatry*, 77(2):149–153, 2006.

[36] Ludovica Griffanti, Mark Jenkinson, Sana Suri, Enikő Zsoldos, Abda Mahmood, Nicola Filippini, Claire E Sexton, Anya Topiwala, Charlotte Allan, Mika Kivimäki, et al. Classification and characterization of periventricular and deep white matter hyperintensities on MRI: a study in older adults. *Neuroimage*, 170:174–181, 2018.

[37] Massimo Filippi and Maria A Rocca. Dirty-appearing white matter: a disregarded entity in multiple sclerosis, 2010.

[38] GRW Moore, C Laule, A Mackay, E Leung, DKB Li, G Zhao, AL Traboulsee, and DW Paty. Dirty-appearing white matter in multiple sclerosis. *Journal of neurology*, 255(11):1802, 2008.

[39] Campbell Le Heron, Sanjay Manohar, Olivia Plant, Kinan Muhammed, Ludovica Griffanti, Andrea Nemeth, Gwenaëlle Douaud, Hugh S Markus, and Masud Husain. Dysfunctional effort-based decision-making underlies apathy in genetic cerebral small vessel disease. *Brain*, 141(11):3193–3210, 2018.

[40] Hugues Chabriat, Anne Joutel, Martin Dichgans, Elizabeth Tournier-Lasserve, and Marie-Germaine Bousser. Cadasil. *The Lancet Neurology*, 8(7):643–653, 2009.

[41] Johannes Attems and Kurt A Jellinger. The overlap between vascular disease and Alzheimers disease-lessons from pathology. *BMC medicine*, 12(1):206, 2014.

[42] Rebecca A Charlton, RG Morris, A Nitkunan, and HS Markus. The cognitive profiles of CADASIL and sporadic small vessel disease. *Neurology*, 66(10):1523–1526, 2006.

[43] Philip Scheltens, Frederik Barkhof, Didier Leys, Jean Pierre Pruvo, JJP Nauta, Pierre Vermersch, Marc Steinling, and Jaap Valk. A semiquantative rating scale for the assessment of signal hyperintensities on magnetic resonance imaging. *Journal of the neurological sciences*, 114(1):7–12, 1993.

[44] LO Wahlund, F Barkhof, F Fazekas, L Bronge, M Augustin, M Sjögren, A Wallin, H Ader, Didier Leys, L Pantoni, et al. A new rating scale for age-related white matter changes applicable to MRI and CT. *Stroke*, 32(6):1318–1322, 2001.

[45] Riitta Mäntylä, Timo Erkinjuntti, Oili Salonen, Hannu J Aronen, Teemu Peltonen, Tarja Pohjasvaara, and Carl-Gustaf Standertskjld-Nordenstam. Variable agreement between visual rating scales for white matter hyperintensities on MRI: comparison of 13 rating scales in a poststroke cohort. *Stroke*, 28(8):1614–1623, 1997.

[46] Marco Duering, Benno Gesierich, Stephan Seiler, Lukas Pirpamer, Mariya Gonik, Edith Hofer, Eric Jouvent, Edouard Duchesnay, Hugues Chabriat, Stefan Ropele,

et al. Strategic white matter tracts for processing speed deficits in age-related small vessel disease. *Neurology*, 82(22):1946–1950, 2014.

[47] J Matthijs Biesbroek, Hugo J Kuijf, Yolanda van der Graaf, Koen L Vincken, Albert Postma, Willem PTM Mali, Geert J Biessels, Mirjam I Geerlings, SMART Study Group, et al. Association between subcortical vascular lesion location and cognition: a voxel-based and tract-based lesion-symptom mapping study. the SMART-MR study. *PloS One*, 8(4):e60541, 2013.

[48] Randall R Benson, CRG Guttmann, X Wei, SK Warfield, C Hall, JA Schmidt, RIWL Kikinis, and LI Wolfson. Older people with impaired mobility have specific loci of periventricular abnormality on MRI. *Neurology*, 58(1):48–55, 2002.

[49] Charles D Smith, David A Snowdon, Hao Wang, and William R Markesbery. White matter volumes and periventricular white matter hyperintensities in aging and dementia. *Neurology*, 54(4):838–842, 2000.

[50] Egill Rostrup, AA Gouw, Hugo Vrenken, Elisabeth CW van Straaten, Stefan Ropele, Leonardo Pantoni, Domenico Inzitari, Frederik Barkhof, Gunhild Waldemar, LADIS Study Group, et al. The spatial distribution of age-related white matter changes as a function of vascular risk factorsresults from the LADIS study. *Neuroimage*, 60(3):1597–1607, 2012.

[51] Maria Eugenia Caligiuri, Paolo Perrotta, Antonio Augimeri, Federico Rocca, Aldo Quattrone, and Andrea Cherubini. Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: A review. *Neuroinformatics*, 13(3):261–276, 2015.

[52] M Jenkinson, CF Beckmann, TE Behrens, and MW Woolrich. Smith SM. *Fsl Neuroimage*, 62:782–90, 2012.

[53] Ludovica Griffanti, Giovanna Zamboni, Aamira Khan, Linxin Li, Guendalina Bonifacio, Vaanathi Sundaresan, Ursula G Schulz, Wilhelm Kuker, Marco Battaglini, Peter M Rothwell, et al. BIANCA (Brain Intensity AbNormality Classification Algorithm): a new tool for automated segmentation of white matter hyperintensities. *NeuroImage*, 141:191–205, 2016.

[54] Michela Simoni, Linxin Li, Nicola LM Paul, Basil E Gruter, Ursula G Schulz, Wilhelm Küker, and Peter M Rothwell. Age-and sex-specific rates of leukoaraiosis in TIA and stroke patients: population-based study. *Neurology*, 79(12):1215–1222, 2012.

[55] Petronella Anbeek, Koen L Vincken, Matthias JP Van Osch, Robertus HC Bisschops, and Jeroen Van Der Grond. Probabilistic segmentation of white matter lesions in MR imaging. *NeuroImage*, 21(3):1037–1044, 2004.

[56] Renske De Boer, Henri A Vrooman, Fedde Van Der Lijn, Meike W Vernooij, M Arfan Ikram, Aad Van Der Lugt, Monique MB Breteler, and Wiro J Niessen. White matter lesion extension to automatic brain tissue segmentation on MRI. *NeuroImage*, 45(4):1151–1161, 2009.

[57] Tim B Dyrby, Egill Rostrup, William FC Baaré, Elisabeth CW van Straaten, Frederik Barkhof, Hugo Vrenken, Stefan Ropele, Reinhold Schmidt, Timo Erkinjuntti, Lars-Olof Wahlund, et al. Segmentation of age-related white matter changes in a clinical multi-center study. *NeuroImage*, 41(2):335–345, 2008.

[58] Hongwei Li, Gongfa Jiang, Jianguo Zhang, Ruixuan Wang, Zhaolei Wang, Wei-Shi Zheng, and Bjoern Menze. Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. *NeuroImage*, 183:650–665, 2018.

[59] Mohsen Ghafoorian, Nico Karssemeijer, Tom Heskes, Inge WM van Uden, Clara I Sanchez, Geert Litjens, Frank-Erik de Leeuw, Bram van Ginneken, Elena Marchiori, and Bram Platel. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Scientific Reports*, 7(1):5110, 2017.

[60] Sergi Martinez-Ramirez, Steven M Greenberg, and Anand Viswanathan. Cerebral microbleeds: overview and implications in cognitive impairment. *Alzheimer's research & therapy*, 6(3):33, 2014.

[61] MMA Conijn, MI Geerlings, G-J Biessels, T Takahara, TD Witkamp, JJM Zwanenburg, PR Luijten, and J Hendrikse. Cerebral microbleeds on MR imaging: comparison between 1.5 and 7T. *American journal of neuroradiology*, 32(6):1043–1049, 2011.

[62] David J Werring. *Cerebral microbleeds: pathophysiology to clinical practice*. Cambridge University Press, 2011.

[63] MW Vernooij, Aad van der Lugt, Mohammad Arfan Ikram, PA Wielopolski, WJ Niessen, Albert Hofman, GP Krestin, and MMB Breteler. Prevalence and risk factors of cerebral microbleeds: the Rotterdam Scan Study. *Neurology*, 70(14):1208–1214, 2008.

[64] Mariëlle MF Poels, Meike W Vernooij, M Arfan Ikram, Albert Hofman, Gabriel P Krestin, Aad van der Lugt, and Monique MB Breteler. Prevalence and risk factors of cerebral microbleeds: an update of the Rotterdam scan study. *Stroke*, 41(10_suppl_1):S103–S106, 2010.

[65] Franz Fazekas, Reinhold Kleinert, Gudrun Roob, Gertrude Kleinert, Peter Kapeller, Reinhold Schmidt, and Hans-Peter Hartung. Histopathologic analysis of foci of signal loss on gradient-echo T2*-weighted MR images in patients with spontaneous intracerebral hemorrhage: evidence of microangiopathy-related microbleeds. *American Journal of Neuroradiology*, 20(4):637–642, 1999.

[66] S-H Lee, J-M Park, S-J Kwon, H Kim, YH Kim, JK Roh, and BW Yoon. Left ventricular hypertrophy is associated with cerebral microbleeds in hypertensive patients. *Neurology*, 63(1):16–21, 2004.

[67] Matthew Schrag, Grant McAuley, Justine Pomakian, Arshad Jiffry, Spencer Tung, Claudius Mueller, Harry V Vinters, E Mark Haacke, Barbara Holshouser, Daniel Kido, et al. Correlation of hypointensities in susceptibility-weighted images to tissue histology in dementia patients with cerebral amyloid angiopathy: a postmortem MRI study. *Acta neuropathologica*, 119(3):291–302, 2010.

[68] S Sveinbjornsdottir, S Sigurdsson, T Aspelund, O Kjartansson, G Eiriksdottir, B Valtysdottir, OL Lopez, MA Van Buchem, PV Jonsson, V Gudnason, et al. Cerebral microbleeds in the population based AGES-Reykjavik study: prevalence and location. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(9):1002–1006, 2008.

[69] Heather C O'donnell, Jonathan Rosand, Katherine A Knudsen, Karen L Furie, Alan Z Segal, Rosaleen I Chiu, Deborah Ikeda, and Steven M Greenberg. Apolipoprotein E genotype and the risk of recurrent lobar intracerebral hemorrhage. *New England Journal of Medicine*, 342(4):240–245, 2000.

[70] Jonathan Rosand, Alona Muzikansky, Ashok Kumar, Jonathan J Wisco, Eric E Smith, Rebecca A Betensky, and Steven M Greenberg. Spatial clustering of hemorrhages in probable cerebral amyloid angiopathy. *Annals of neurology*, 58(3):459–462, 2005.

[71] Dymph J Mesker, Mariëlle MF Poels, M Arfan Ikram, Meike W Vernooij, Albert Hofman, Henri A Vrooman, Aad van der Lugt, and Monique MB Breteler. Lobar distribution of cerebral microbleeds: the Rotterdam Scan Study. *Archives of neurology*, 68(5):656–659, 2011.

[72] Jacqueline A Pettersen, Gayathri Sathiyamoorthy, Fu-Qiang Gao, Gregory Szilagyi, Neelesh K Nadkarni, Peter St George-Hyslop, Ekaterina Rogaeva, and Sandra E Black. Microbleed topography, leukoaraiosis, and cognition in probable alzheimer disease from the Sunnybrook dementia study. *Archives of neurology*, 65(6):790–795, 2008.

[73] SM Gregoire, UJ Chaudhary, MM Brown, TA Yousry, C Kallis, HR Jäger, and DJ Werring. The Microbleed Anatomical Rating Scale (MARS): reliability of a tool to map brain microbleeds. *Neurology*, 73(21):1759–1766, 2009.

[74] Charlotte Cordonnier, Gillian M Potter, Caroline A Jackson, Fergus Doubal, Sarah Keir, Cathie LM Sudlow, Joanna M Wardlaw, and Rustam Al-Shahi Salman. Improving interrater agreement about brain microbleeds: development of the Brain Observer MicroBleed Scale (BOMBS). *Stroke*, 40(1):94–99, 2009.

[75] R.N.K. Nandigam, A. Viswanathan, P. Delgado, M.E. Skehan, E.E. Smith, J. Rosand, S.M. Greenberg, and B.C. Dickerson. MR imaging detection of cerebral microbleeds: Effect of susceptibility-weighted imaging, section thickness, and field strength. *American Journal of Neuroradiology*, 30(2):338–343, 2009.

[76] R L Mittl, R I Grossman, J F Hiehle, R W Hurst, D R Kauder, T A Gennarelli, and G W Alburger. Prevalence of MR evidence of diffuse axonal injury in patients with mild head injury and normal head CT findings. *American Journal of Neuroradiology*, 15(8):1583–1589, 1994.

[77] Charlotte Cordonnier, Joanna Wardlaw, and Rustam Al-Shahi Salman. Spontaneous brain microbleeds: systematic review, subgroup analyses and standards for study design and reporting. *Brain*, 130(8):1988–2003, 02 2007.

[78] Vaanathi Sundaresan, Ludovica Griffanti, Petya Kindalova, Fidel Alfaro-Almagro, Giovanna Zamboni, Peter M Rothwell, Thomas E Nichols, and Mark Jenkinson. Modelling the distribution of white matter hyperintensities due to ageing on MRI images using bayesian inference. *NeuroImage*, 185:434–445, 2019.

[79] Vaanathi Sundaresan, Giovanna Zamboni, Campbell Le Heron, Peter M Rothwell, Masud Husain, Marco Battaglini, Nicola De Stefano, Mark Jenkinson, and Ludovica Griffanti. Automated lesion segmentation with BIANCA: impact of population-level features, classification algorithm and locally adaptive thresholding. *NeuroImage*, page 116056, 2019.

[80] Rob A Dineen, Stefan Pszczolkowski, Katie Flaherty, Zhe K Law, Paul S Morgan, Ian Roberts, David J Werring, Rustam Al-Shahi Salman, Tim England, Philip M Bath, et al. Does tranexamic acid lead to changes in MRI measures of brain tissue health in patients with spontaneous intracerebral haemorrhage? protocol for a MRI substudy nested within the double-blind randomised controlled TICH-2 trial. *BMJ open*, 8(2):e019930, 2018.

[81] Nikola Sprigg, Katie Flaherty, Jason P Appleton, Rustam Al-Shahi Salman, Daniel Bereczki, Maia Beridze, Hanne Christensen, Alfonso Ciccone, Ronan Collins, Anna Czlonkowska, et al. Tranexamic acid for hyperacute primary intracerebral haemorrhage (TICH-2): an international randomised, placebo-controlled, phase 3 superiority trial. *The Lancet*, 391(10135):2107–2115, 2018.

[82] Jan Cees De Groot, Frank-Erik De Leeuw, Matthijs Oudkerk, Jan Van Gijn, Albert Hofman, Jellemer Jolles, and Monique MB Breteler. Cerebral white matter lesions and cognitive function: the Rotterdam Scan Study. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 47(2):145–151, 2000.

[83] Leonardo Pantoni, Anna Maria Basile, Giovanni Pracucci, Kjell Asplund, Julien Bogousslavsky, Hugues Chabriat, Timo Erkinjuntti, Franz Fazekas, José M Ferro, Michael Hennerici, et al. Impact of age-related cerebral white matter changes on the transition to disability–the LADIS study: rationale, design and methodology. *Neuroepidemiology*, 24(1-2):51–62, 2005.

[84] Benjamin S Aribisala, Maria C Valdés Hernández, Natalie A Royle, Zoe Morris, Susana Muñoz Maniega, Mark E Bastin, Ian J Deary, and Joanna M Wardlaw. Brain atrophy associations with white matter lesions in the ageing brain: the Lothian Birth Cohort 1936. *European radiology*, 23(4):1084–1092, 2013.

[85] Anne M Grool, Mirjam I Geerlings, Sigurdur Sigurdsson, Gudny Eiriksdottir, Palmi V Jonsson, Melissa E Garcia, Kristin Siggeirsdottir, Tamara B Harris, Thordur Sigmundsson, Vilmundur Gudnason, et al. Structural MRI correlates of apathy symptoms in older persons without dementia: AGES-Reykjavik Study. *Neurology*, 82(18):1628–1635, 2014.

[86] Giovanna Zamboni, Ludovica Griffanti, Sara Mazzucco, Sarah T Pendlebury, and Peter M Rothwell. Age-dependent association of white matter abnormality with cognition after TIA or minor stroke. *Neurology*, pages 10–1212, 2019.

[87] Fidel Alfaro-Almagro, Mark Jenkinson, Neal K Bangerter, Jesper LR Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Stamatios N Sotiropoulos, Saad Jbabdi, Moises Hernandez-Fernandez, Emmanuel Vallee, et al. Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage*, 166:400–424, 2018.

[88] Anna Altermatt, Laura Gaetano, Stefano Magon, Lorena Bauer, Regina Feurer, Hans Gnahn, Julia Hartmann, Christian L Seifert, Holger Poppert, Jens Wuerfel, et al. Clinical associations of T2-weighted lesion load and lesion location in small vessel disease: Insights from a large prospective cohort study. *NeuroImage*, 189:727–733, 2019.

[89] Franz Fazekas, John B Chawluk, Abass Alavi, Howard I Hurtig, and Robert A Zimmerman. MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. *American Journal of Neuroradiology*, 8(3):421–426, 1987.

[90] MMB Breteler, JC Van Swieten, ML Bots, DE Grobbee, JJ Claus, JHW Van Den Hout, F Van Harskamp, HLJ Tanghe, PTVM De Jong, J Van Gijn, et al. Cerebral white matter lesions, vascular risk factors, and cognitive function in a population-based study: the Rotterdam Study. *Neurology*, 44(7):1246–1246, 1994.

[91] H Baezner, C Blahak, A Poggesi, L Pantoni, D Inzitari, H Chabriat, T Erkinjuntti, F Fazekas, JM Ferro, P Langhorne, et al. Association of gait and balance disorders with age-related white matter changes: the LADIS study. *Neurology*, 70(12):935–942, 2008.

[92] Giovanna Zamboni, Ludovica Griffanti, Mark Jenkinson, Sara Mazzucco, Linxin Li, Wilhelm Küker, Sarah T Pendlebury, and Peter M Rothwell. White matter imaging correlates of early cognitive impairment detected by the montreal cognitive assessment after transient ischemic attack and minor stroke. *Stroke*, 48(6):1539–1547, 2017.

[93] Donald M Lyall, Simon R Cox, Laura M Lyall, Carlos Celis-Morales, Breda Cullen, Daniel F Mackay, Joey Ward, Rona J Strawbridge, Andrew M McIntosh, Naveed Sattar, et al. Association between APOE e4 and white matter hyperintensity volume, but not total brain volume or white matter integrity. *Brain imaging and behavior*, pages 1–9, 2019.

[94] Jan Cees de Groot, F-E De Leeuw, M Oudkerk, A Hofman, J Jolles, and MMB Breteler. Cerebral white matter lesions and subjective cognitive dysfunction: the Rotterdam Scan Study. *Neurology*, 56(11):1539–1545, 2001.

[95] Stéphanie Debette and HS Markus. The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. *Bmj*, 341:c3666, 2010.

[96] Alan J Thomas, John T O'brien, Sue Davis, Clive Ballard, Robert Barber, Rajesh N Kalaria, and Robert H Perry. Ischemic basis for deep white matter hyperintensities in major depression: a neuropathological study. *Archives of general psychiatry*, 59(9):785–792, 2002.

[97] Paul Yates, Victor Villemagne, Kathryn Ellis, Patricia Desmond, Colin Masters, and Christopher Rowe. Cerebral microbleeds: A review of clinical, genetic, and neuroimaging associations. *Frontiers in Neurology*, 4:205, 2014.

[98] Gudrun Roob, Anita Lechner, Reinhold Schmidt, Erich Flooh, Hans-Peter Hartung, and Franz Fazekas. Frequency and location of microbleeds in patients with primary intracerebral hemorrhage. *Stroke*, 31(11):2665–2669, 2000.

[99] Y Tsushima, Y Tanizaki, J Aoki, and K Endo. MR detection of microhemorrhages in neurologically healthy adults. *Neuroradiology*, 44(1):31–36, 2002.

[100] Yoshito Tsushima, Jun Aoki, and Keigo Endo. Brain microhemorrhages detected on T2*-weighted gradient-echo MR images. *American journal of neuroradiology*, 24(1):88–96, 2003.

[101] Yoshifumi Horita, Toshio Imaizumi, Jun Niwa, J Yoshikawa, K Miyata, T Makabe, R Moriyama, K Kurokawa, M Mikami, and M Nakamura. Analysis of dot-like hemosiderin spots using brain dock system. *No shinkei geka. Neurological surgery*, 31(3):263–267, 2003.

[102] Tom Jeerakathil, Philip A Wolf, Alexa Beiser, John K Hald, Rhoda Au, Carlos S Kase, Joseph M Massaro, and Charles DeCarli. Cerebral microbleeds: prevalence and associations with cardiovascular risk factors in the Framingham Study. *Stroke*, 35(8):1831–1835, 2004.

[103] C Qiu, MF Cotch, S Sigurdsson, PV Jonsson, MK Jonsdottir, S Sveinbjrnsdottir, G Eiriksdottir, R Klein, TB Harris, MA Van Buchem, et al. Cerebral microbleeds, retinopathy, and dementia: the AGES-Reykjavik study. *Neurology*, 75(24):2221–2228, 2010.

[104] Tomofumi Nishikawa, Tetsuya Ueba, Motohiro Kajiwara, Ichro Fujisawa, Naomi Miyamatsu, and Kohsuke Yamashita. Cerebral microbleeds predict first-ever symptomatic cerebrovascular events. *Clinical neurology and neurosurgery*, 111(10):825–828, 2009.

[105] Yuki Takashima, Takahiro Mori, Manabu Hashimoto, Naoko Kinukawa, Akira Uchino, Takefumi Yuzuriha, and Hiroshi Yao. Clinical correlating factors and cognitive function in community-dwelling healthy subjects with cerebral microbleeds. *Journal of stroke and cerebrovascular diseases*, 20(2):105–110, 2011.

[106] Seung-Hoon Lee, Soon-Tae Lee, Beom Joon Kim, Hee-Kwon Park, Chi-Kyung Kim, Keun-Hwa Jung, and Jae-Kyu Roh. Dynamic temporal change of cerebral microbleeds: long-term follow-up MRI study. *PLoS One 6*, (10), 2011.

[107] Simone M Gregoire, Martin M Brown, Constantinos Kallis, H Rolf Jger, Tarek A Yousry, and David J Werring. MRI detection of new microbleeds in patients with ischemic stroke: five-year cohort follow-up study. *Stroke*, 41(1):184–186, 2010.

[108] Irmhild Altmann-Schneider, Stella Trompet, Anton JM de Craen, Adriaan CGM van Es, J Wouter Jukema, David J Stott, Naveed Sattar, Rudi GJ Westendorp, Mark A van Buchem, and Jeroen van der Grond. Cerebral microbleeds are predictive of mortality in the elderly. *Stroke*, 42(3):638–644, 2011.

[109] Wouter JP Henneman, Jasper D Sluimer, Charlotte Cordonnier, Merel ME Baak, Philip Scheltens, Frederik Barkhof, and Wiesje M van der Flier. MRI biomarkers of vascular damage and atrophy predicting mortality in a memory clinic population. *Stroke*, 40(2):492–498, 2009.

[110] Muhammad Ayaz, Alexander S Boikov, E Mark Haacke, Daniel K Kido, and Wolff M Kirsch. Imaging cerebral microbleeds using susceptibility weighted imaging: one step toward detecting vascular dementia. *Journal of Magnetic Resonance Imaging*, 31(1):142–148, 2010.

[111] C Cordonnier, WM Van der Flier, JD Sluimer, D Leys, F Barkhof, and PH Scheltens. Prevalence and severity of microbleeds in a memory clinic setting. *Neurology*, 66(9):1356–1360, 2006.

[112] Karlijn F de Laat, Heleen AC van den Berg, Anouk GW van Norden, Rob AR Gons, Marcel GM Olde Rikkert, and Frank-Erik de Leeuw. Microbleeds are independently related to gait disturbances in elderly individuals with cerebral small vessel disease. *Stroke*, 42(2):494–497, 2011.

[113] Anand Viswanathan, Jean-Pierre Guichard, Andreas Gschwendtner, Frederique Buffon, Rodica Cumurcuic, Carole Boutron, Eric Vicaut, Markus Holtmannspötter, Chahin Pachai, Marie-Germaine Bousser, et al. Blood pressure and haemoglobin a1c are associated with microhaemorrhage in cadasil: a two-centre cohort study. *Brain*, 129(9):2375–2383, 2006.

[114] Rivka van den Boom, Saskia A. J. Lesnik Oberstein, Michel D. Ferrari, Joost Haan, and Mark A. van Buchem. Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy: MR imaging findings at different ages3rd6th decades. *Radiology*, 229(3):683–690, 2003. PMID: 14593195.

[115] Martin Dichgans, Markus Holtmannsptter, Jrgen Herzog, Nils Peters, Michael Bergmann, and Tarek A. Yousry. Cerebral microbleeds in CADASIL. *Stroke*, 33(1):67–71, 2002.

[116] Steven M Greenberg, Jessica A Eng, MingMing Ning, Eric E Smith, and Jonathan Rosand. Hemorrhage burden predicts recurrent intracerebral hemorrhage after lobar hemorrhage. *Stroke*, 35(6):1415–1420, 2004.

[117] David J Werring, Simone M Gregoire, and Lisa Cipolotti. Cerebral microbleeds and vascular cognitive impairment. *Journal of the neurological sciences*, 299(1-2):131–135, 2010.

[118] Jeroen DC Goos, MI Kester, Frederik Barkhof, Martin Klein, Marinus A Blankenstein, Philip Scheltens, and Wiesje M van der Flier. Patients with Alzheimer disease with multiple microbleeds: relation with cerebrospinal fluid biomarkers and cognition. *Stroke*, 40(11):3455–3460, 2009.

[119] Yusuke Yakushiji, Masanori Nishiyama, Satomi Yakushiji, Tatsumi Hirotsu, Akira Uchino, Junko Nakajima, Makoto Eriguchi, Yusuke Nanri, Megumi Hara, Etsuo Horikawa, et al. Brain microbleeds and global cognitive function in adults without neurological disorder. *Stroke*, 39(12):3323–3328, 2008.

[120] Monzurul H Chowdhury, Atsushi Nagai, Hirokazu Bokura, Eisuke Nakamura, Shotai Kobayashi, and Shuhei Yamaguchi. Age-related changes in white matter lesions, hippocampal atrophy, and cerebral microbleeds in healthy subjects without major cerebrovascular risk factors. *Journal of Stroke and Cerebrovascular Diseases*, 20(4):302–309, 2011.

[121] Vincent Thijs, Robin Lemmens, Christophe Schoofs, Astrid Grner, Philip Van Damme, Maarten Schrooten, and Philippe Demaerel. Microbleeds and the risk of recurrent stroke. *Stroke*, 41(9):2005–2009, 2010.

[122] Elizabeth Bates, Stephen M Wilson, Ayse Pinar Saygin, Frederic Dick, Martin I Sereno, Robert T Knight, and Nina F Dronkers. Voxel-based lesion–symptom mapping. *Nature neuroscience*, 6(5):448–450, 2003.

[123] Arnaud Charil, Alex P Zijdenbos, Jonathan Taylor, Cyrus Boelman, Keith J Worsley, Alan C Evans, and Alain Dagher. Statistical mapping analysis of lesion location and neurological disability in multiple sclerosis: application to 452 patient data sets. *NeuroImage*, 19(3):532–544, 2003.

[124] Arnaud Charil, Alain Dagher, Jason P Lerch, Alex P Zijdenbos, Keith J Worsley, and Alan C Evans. Focal cortical atrophy in multiple sclerosis: relation to lesion load and disability. *NeuroImage*, 34(2):509–517, 2007.

[125] Alan E Gelfand, Hyon-Jung Kim, CF Sirmans, and Sudipto Banerjee. Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387–396, 2003.

[126] Dani Gamerman, Ajax RB Moreira, and Håvard Rue. Space-varying regression models: specifications and simulation. *Computational Statistics & Data Analysis*, 42(3):513–533, 2003.

[127] Tian Ge, Nicole Müller-Lenke, Kerstin Bendfeldt, Thomas E Nichols, and Timothy D Johnson. Analysis of multiple sclerosis lesions via spatially varying coefficients. *The annals of applied statistics*, 8(2):1095, 2014.

[128] Sudipto Banerjee, Alan E Gelfand, Andrew O Finley, and Huiyan Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848, 2008.

[129] Philip T Reiss and R Todd Ogden. Functional generalized linear models with images as predictors. *Biometrics*, 66(1):61–69, 2010.

[130] Chong-Zhi Di, Ciprian M Crainiceanu, Brian S Caffo, and Naresh M Punjabi. Multilevel functional principal component analysis. *The annals of applied statistics*, 3(1):458, 2009.

[131] Cari G Kaufman, Mark J Schervish, and Douglas W Nychka. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555, 2008.

[132] Ciprian M Crainiceanu, Ana-Maria Staicu, and Chong-Zhi Di. Generalized multilevel functional regression. *Journal of the American Statistical Association*, 104(488):1550–1561, 2009.

[133] Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295, 2016.

[134] Stefan Klöppel, Ahmed Abdulkadir, Stathis Hadjidemetriou, Sabine Issleib, Lars Frings, Thao Nguyen Thanh, Irina Mader, Stefan J Teipel, Michael Hüll, and Olaf Ronneberger. A comparison of different automated methods for the detection of white matter lesions in MRI data. *NeuroImage*, 57(2):416–422, 2011.

[135] Laurent Itti, Linda Chang, and Thomas Ernst. Segmentation of progressive multifocal leukoencephalopathy lesions in fluid-attenuated inversion recovery magnetic resonance imaging. *Journal of Neuroimaging*, 11(4):412–417, 2001.

[136] Martha E Payne, Denise L Fetzer, James R MacFall, James M Provenzale, Christopher E Byrum, and K Ranga R Krishnan. Development of a semi-automated method for quantification of MRI gray and white matter lesions in geriatric subjects. *Psychiatry Research: Neuroimaging*, 115(1-2):63–77, 2002.

[137] Ron Kikinis, Martha E Shenton, Guido Gerig, John Martin, Mark Anderson, David Metcalf, Charles RG Guttmann, Robert W McCarley, Ferenc A Jolesz, William Lorensen, et al. Routine quantitative analysis of brain and cerebrospinal fluid spaces with MR imaging. *Journal of Magnetic Resonance Imaging*, 2(6):619–629, 1992.

[138] Christopher E Byrum, James R MacFall, H Cecil Charles, Venkata R Chitilla, Orest B Boyko, Lucy Upchurch, Jean S Smith, Pradeep Rajagopalan, Theodore Passe, Dennis Kim, et al. Accuracy and reproducibility of brain and tissue volumes using a magnetic resonance segmentation method. *Psychiatry Research: Neuroimaging*, 67(3):215–234, 1996.

[139] Yasuo Kawata, Hidetaka Arimura, Yasuo Yamashita, Taiki Magome, Masafumi Ohki, Fukai Toyofuku, Yoshiharu Higashida, and Kazuhiro Tsuchiya. Computer-aided evaluation method of white matter hyperintensities related to subcortical vascular dementia based on magnetic resonance imaging. *Computerized Medical Imaging and Graphics*, 34(5):370–376, 2010.

[140] J Ramirez, Erin Gibson, Azhar Quddus, Nancy J Lobaugh, Anthony Feinstein, Brian Levine, CJM Scott, N Levy-Cooperman, FQ Gao, and Sandra E Black. Lesion explorer: a comprehensive segmentation and parcellation package to obtain regional volumetrics for subcortical hyperintensities and intracranial tissue. *NeuroImage*, 54(2):963–973, 2011.

[141] Emma J Burton, Rose Anne Kenny, John OBrien, Sally Stephens, Michael Bradbury, Elise Rowan, Raj Kalaria, Michael Firbank, Keith Wesnes, and Clive Ballard. White matter hyperintensities are associated with impairment of memory, attention, and global cognitive performance in older stroke patients. *Stroke*, 35(6):1270–1275, 2004.

[142] Charles DeCarli, DGM Murphy, Met al Tranh, CL Grady, JV Haxby, JA Gillette, JA Salerno, A Gonzales-Aviles, B Honvitz, SI Rapoport, et al. The effect of white matter hyperintensity volume on brain structure, cognitive performance, and cerebral metabolism of glucose in 51 healthy adults. *Neurology*, 45(11):2077–2084, 1995.

[143] Yvette I Sheline, Joseph L Price, S Neil Vaishnavi, Mark A Mintun, Deanna M Barch, Adrian A Epstein, Consuelo H Wilkins, Abraham Z Snyder, Lars Couture, Kenneth Schechtman, et al. Regional white matter hyperintensity burden in automated segmentation distinguishes late-life depressed subjects from comparison subjects matched for vascular risk factors. *American Journal of Psychiatry*, 165(4):524–532, 2008.

[144] RC McKinstry, YI Sheline, AZ Snyder, A Moinuddin, JJ Christensen, and MA Mintun. Bispectral analysis of white matter lesions (WML) using T2W and 3D T1W MRI. In *Abstr Soc Neurosci*, volume 33, pages 640–647, 2003.

[145] Clifford R Jack Jr, Peter C O'Brien, Daniel W Rettman, Maria M Shiung, Yuecheng Xu, Raja Muthupillai, Armando Manduca, Ramesh Avula, and Bradley J Erickson. FLAIR histogram segmentation for measurement of leukoaraiosis volume. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 14(6):668–676, 2001.

[146] Byung Il Yoo, Jung Jae Lee, Ji Won Han, Eun Young Lee, James R MacFall, Martha E Payne, Tae Hui Kim, Jae Hyoung Kim, Ki Woong Kim, et al. Application of variable threshold intensity to segmentation for white matter hyperintensities in fluid attenuated inversion recovery magnetic resonance images. *Neuroradiology*, 56(4):265–281, 2014.

[147] Rita Simões, Christoph Mönninghoff, Martha Dlugaj, Christian Weimar, Isabel Wanke, Anne-Marie van Cappellen van Walsum, and Cornelis Slump. Automatic segmentation of cerebral white matter hyperintensities using only 3D FLAIR images. *Magnetic resonance imaging*, 31(7):1182–1189, 2013.

[148] Xingchang Wei, Simon K Warfield, Kelly H Zou, Ying Wu, Xiaoming Li, Alexandre Guimond, John P Mugler III, Randall R Benson, Leslie Wolfson, Howard L Weiner, et al. Quantitative analysis of MRI signal abnormalities of brain white matter with high reproducibility and accuracy. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 15(2):203–209, 2002.

[149] Yanbo Wang, Joseree Ann Catindig, Saima Hilal, Hock Wei Soon, Eric Ting, Tien Yin Wong, Narayanaswamy Venketasubramanian, Christopher Chen, and Anqi Qiu. Multi-stage segmentation of white matter hyperintensity, cortical and lacunar infarcts. *Neuroimage*, 60(4):2379–2388, 2012.

[150] Kok Haur Ong, Dhanesh Ramachandram, Rajeswari Mandava, and Ibrahim Lutfi Shuaib. Automatic white matter lesion segmentation using an adaptive outlier detection method. *Magnetic resonance imaging*, 30(6):807–823, 2012.

[151] Eloy Roura, Arnau Oliver, Mariano Cabezas, Sergi Valverde, Deborah Pareto, Joan C Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, and Xavier Lladó. A toolbox for multiple sclerosis lesion segmentation. *Neuroradiology*, 57(10):1031–1043, 2015.

[152] E Herskovits, R Bryan, and F Yang. Automated bayesian segmentation of microvascular white-matter lesions in the ACCORD-MIND study. 2008.

[153] Pauline Maillard, Nicolas Delcroix, Fabrice Crivello, Carole Dufouil, Sebastien Gicquel, Marc Joliot, Nathalie Tzourio-Mazoyer, Annick Alpérovitch, Christophe Tzourio, and Bernard Mazoyer. An automated procedure for the assessment of white matter hyperintensities by multispectral (T1, T2, PD) MRI and an evaluation of its between-centre reproducibility based on two large community databases. *Neuroradiology*, 50(1):31–42, 2008.

[154] Carole H Sudre, M Jorge Cardoso, Willem H Bouvy, Geert Jan Biessels, Josephine Barnes, and Sebastien Ourselin. Bayesian model selection for pathological neuroimaging data applied to white matter lesion segmentation. *IEEE transactions on medical imaging*, 34(10):2079–2102, 2015.

[155] Carole H Sudre, M Jorge Cardoso, Sebastien Ourselin, Alzheimers Disease Neuroimaging Initiative, et al. Longitudinal segmentation of age-related white matter hyperintensities. *Medical image analysis*, 38:50–64, 2017.

[156] Saurabh Jain, Diana M Sima, Annemie Ribbens, Melissa Cambron, Anke Maertens, Wim Van Hecke, Johan De Mey, Frederik Barkhof, Martijn D Steenwijk, Marita Daams, et al. Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images. *NeuroImage: Clinical*, 8:367–375, 2015.

[157] Rasoul Khayati, Mansur Vafadust, Farzad Towhidkhah, and Massood Nabavi. Fully automatic segmentation of multiple sclerosis lesions in brain MR flair images using adaptive mixtures method and markov random field model. *Computers in biology and medicine*, 38(3):379–390, 2008.

[158] Christopher Schwarz, Evan Fletcher, Charles DeCarli, and Owen Carmichael. Fully-automated white matter hyperintensity detection with anatomical prior knowledge and without FLAIR. In *International Conference on Information Processing in Medical Imaging*, pages 239–251. Springer, 2009.

[159] Seun Jeon, Uicheul Yoon, Jun-Sung Park, Sang Won Seo, Jung-Hyun Kim, Sung Tae Kim, Sun I Kim, Duk L Na, and Jong-Min Lee. Fully automated pipeline for quantification and localization of white matter hyperintensity in brain magnetic resonance image. *International Journal of Imaging Systems and Technology*, 21(2):193–200, 2011.

[160] Dan Wu, Marilyn Albert, Anja Soldan, Corinne Pettigrew, Kenichi Oishi, Yusuke Tomogane, Chenfei Ye, Ting Ma, Michael I Miller, and Susumu Mori. Multi-atlas based detection and localization (MADL) for location-dependent quantification of white matter hyperintensities. *NeuroImage: Clinical*, 22:101772, 2019.

[161] Jang-Zern Tsai, Syu-Jyun Peng, Yu-Wei Chen, Kuo-Wei Wang, Chen-Hua Li, Jing-Yi Wang, Chi-Jen Chen, Huey-Juan Lin, Eric Edward Smith, Hsiao-Kuang Wu, et al. Automated segmentation and quantification of white matter hyperintensities in acute ischemic stroke patients with cerebral infarction. *PloS one*, 9(8):e104011, 2014.

[162] Faguo Yang, Zuyao Y Shan, and Frithjof Kruggel. White matter lesion segmentation based on feature joint occurrence probability and $\chi 2$ random field theory from magnetic resonance (MR) images. *Pattern Recognition Letters*, 31(9):781–790, 2010.

[163] Lin Shi, Defeng Wang, Shangping Liu, Yuehua Pu, Yilong Wang, Winnie CW Chu, Anil T Ahuja, and Yongjun Wang. Automated quantification of white matter lesion in magnetic resonance imaging of patients with acute infarction. *Journal of neuroscience methods*, 213(1):138–146, 2013.

[164] Thomas Samaille, Ludovic Fillon, Rémi Cuingnet, Eric Jouvent, Hugues Chabriat, Didier Dormont, Olivier Colliot, and Marie Chupin. Contrast-based fully automatic

segmentation of white matter hyperintensities: method and validation. *PloS one*, 7(11):e48953, 2012.

[165] Adam M Brickman, Joel R Sneed, Frank A Provenzano, Ernst Garcon, Lauren Johnert, Jordan Muraskin, Lok-Kin Yeung, Molly E Zimmerman, and Steven P Roose. Quantitative approaches for assessment of white matter hyperintensities in elderly populations. *Psychiatry Research: Neuroimaging*, 193(2):101–106, 2011.

[166] Balasrinivasa Rao Sajja, Sushmita Datta, Renjie He, Meghana Mehta, Rakesh K Gupta, Jerry S Wolinsky, and Ponnada A Narayana. Unified approach for multiple sclerosis lesion segmentation on brain MRI. *Annals of biomedical engineering*, 34(1):142–151, 2006.

[167] Frithjof Kruggel, Joseph Suresh Paul, and Hermann-Josef Gertz. Texture-based segmentation of diffuse lesions of the brains white matter. *NeuroImage*, 39(3):987–996, 2008.

[168] April Khademi, Anastasios Venetsanopoulos, and Alan R Moody. Robust white matter lesion segmentation in FLAIR MRI. *IEEE Transactions on biomedical engineering*, 59(3):860–871, 2011.

[169] Muhammad Febrian Rachmadi, Maria del C Valdés-Hernández, Hongwei Li, Ricardo Guerrero, Rozanna Meijboom, Stewart Wiseman, Adam Waldman, Jianguo Zhang, Daniel Rueckert, Joanna Wardlaw, et al. Limited one-time sampling irregularity map (LOTS-IM) for automatic unsupervised assessment of white matter hyperintensities and multiple sclerosis lesions in structural brain magnetic resonance images. *BioRxiv*, page 334292, 2019.

[170] Karen M Gosche, Robert P Velthuizen, F Reed Murtagh, John A Arrington, William W Gross, James A Mortimer, and Laurence P Clarke. Automated quantification of brain magnetic resonance image hyperintensities using hybrid clustering and knowledge-based methods. *International journal of imaging systems and technology*, 10(3):287–293, 1999.

[171] Faiza Admiraal-Behloul, DMJ Van Den Heuvel, Hans Olofsen, Matthias JP van Osch, Jeroen van der Grond, MA Van Buchem, and JHC Reiber. Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. *NeuroImage*, 28(3):607–617, 2005.

[172] Minjie Wu, Caterina Rosano, Meryl Butters, Ellen Whyte, Megan Nable, Ryan Crooks, Carolyn C Meltzer, Charles F Reynolds III, and Howard J Aizenstein. A fully automated method for quantifying and localizing white matter hyperintensities on MR images. *Psychiatry Research: Neuroimaging*, 148(2-3):133–142, 2006.

[173] Mohamed L Seghier, Anil Ramlackhansingh, Jenny Crinion, Alexander P Leff, and Cathy J Price. Lesion identification using unified segmentation-normalisation models and fuzzy clustering. *Neuroimage*, 41(4):1253–1266, 2008.

[174] Erin Gibson, Fuqiang Gao, Sandra E Black, and Nancy J Lobaugh. Automatic segmentation of white matter hyperintensities in the elderly using flair images at 3T. *Journal of Magnetic Resonance Imaging*, 31(6):1311–1322, 2010.

[175] M Anitha, P Tamije Selvy, and V Palanisamy. Wml detection of brain images using fuzzy and possibilistic approach in feature space. *WSEAS Transactions on computers, e-ISSN*, 22242872, 2012.

[176] Feroze B Mohamed, Simon Vinitski, Carlos F Gonzalez, Scott H Faro, Fred A Lublin, Robert Knobler, and Juan Esteban Gutierrez. Increased differentiation of intracranial white matter lesions by multispectral 3d-tissue segmentation: preliminary results. *Magnetic resonance imaging*, 19(2):207–218, 2001.

[177] Hugo J Kuijf, Pim Moeskops, Bob D de Vos, Willem H Bouvy, Jeroen de Bresser, Geert Jan Biessels, Max A Viergever, and Koen L Vincken. Supervised novelty detection in brain tissue classification with an application to white matter hyperintensities. In *Medical Imaging 2016: Image Processing*, volume 9784, page 978421. International Society for Optics and Photonics, 2016.

[178] Wei Wen, Perminder S Sachdev, Jason J Li, Xiaohua Chen, and Kaarin J Anstey. White matter hyperintensities in the forties: their prevalence and topography in an epidemiological sample aged 44–48. *Human brain mapping*, 30(4):1155–1167, 2009.

[179] Richard H Swartz, Sandra E Black, Anthony Feinstein, Conrad Rockel, Gal Sela, Fu Qiang Gao, Curtis B Caldwell, and Michael J Bronskill. Utility of simultaneous brain, csf and hyperintensity quantification in dementia. *Psychiatry Research: Neuroimaging*, 116(1-2):83–93, 2002.

[180] Martijn D Steenwijk, Petra JW Pouwels, Marita Daams, Jan Willem van Dalen, Matthan WA Caan, Edo Richard, Frederik Barkhof, and Hugo Vrenken. Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). *NeuroImage: Clinical*, 3:462–469, 2013.

[181] Alex P Zijdenbos, Reza Forghani, Alan C Evans, et al. Automatic" pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. *IEEE Trans. Med. Imaging*, 21(10):1280–1291, 2002.

[182] Richard Beare, Velandai Srikanth, Jian Chen, Thanh G Phan, Jennifer Stapleton, Rebecca Lipshut, and David Reutens. Development and validation of morphological segmentation of age-related cerebral white matter hyperintensities. *Neuroimage*, 47(1):199–203, 2009.

[183] Zhiqiang Lao, Dinggang Shen, Dengfeng Liu, Abbas F Jawad, Elias R Melhem, Lenore J Launer, R Nick Bryan, and Christos Davatzikos. Computer-assisted segmentation of white matter lesions in 3D MR images using support vector machine. *Academic radiology*, 15(3):300–313, 2008.

[184] Soheil Damangir, Amirhossein Manzouri, Ketil Oppedal, Stefan Carlsson, Michael J Firbank, Hogne Sonnesyn, Ole-Bjørn Tysnes, John T O'Brien, Mona K Beyer, Eric Westman, et al. Multispectral MRI segmentation of age related white matter changes using a cascade of support vector machines. *Journal of the neurological sciences*, 322(1-2):211–216, 2012.

[185] Bo-yong Park, Mi Ji Lee, Seung-hak Lee, Jihoon Cha, Chin-Sang Chung, Sung Tae Kim, and Hyunjin Park. Dews (DEep white matter hyperintensity segmentation

framework): a fully automated pipeline for detecting small deep white matter hyperintensities in migraineurs. *NeuroImage: Clinical*, 18:638–647, 2018.

[186] Azhar Quddus, Paul Fieguth, and Otman Basir. Adaboost and support vector machines for white matter lesion segmentation in MR images. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 463–466. IEEE, 2006.

[187] Vamsi Ithapu, Vikas Singh, Christopher Lindner, Benjamin P Austin, Chris Hinrichs, Cynthia M Carlsson, Barbara B Bendlin, and Sterling C Johnson. Extracting and summarizing white matter hyperintensities using supervised segmentation methods in Alzheimer's disease risk and aging studies. *Human brain mapping*, 35(8):4219–4235, 2014.

[188] Jiyang Jiang, Tao Liu, Wanlin Zhu, Rebecca Koncz, Hao Liu, Teresa Lee, Perminder S Sachdev, and Wei Wen. UBO detector–a cluster-based, fully automated pipeline for extracting white matter hyperintensities. *NeuroImage*, 174:539–549, 2018.

[189] Mariano Rincón, Estela Díaz-López, Per Selnes, K Vegge, M Altmann, Tormod Fladby, and Atle Bjørnerud. Improved automatic segmentation of white matter hyperintensities in MRI based on multilevel lesion features. *Neuroinformatics*, 15(3):231–245, 2017.

[190] Mohsen Ghafoorian, Nico Karssemeijer, Inge WM van Uden, Frank-Erik de Leeuw, Tom Heskes, Elena Marchiori, and Bram Platel. Automated detection of white matter hyperintensities of all sizes in cerebral small vessel disease. *Medical physics*, 43(12):6246–6258, 2016.

[191] Mohsen Ghafoorian, Nico Karssemeijer, Inge van Uden, Frank Erik de Leeuw, Tom Heskes, Elena Marchiori, and Bram Platel. Small white matter lesion detection in cerebral small vessel disease. In *Medical Imaging 2015: Computer-Aided Diagnosis*, volume 9414, page 941411. International Society for Optics and Photonics, 2015.

[192] Andrew Jesson and Tal Arbel. Hierarchical MRF and random forest segmentation of ms lesions and healthy tissues in brain MRI. *Proceedings of the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge*, pages 1–2, 2015.

[193] James R Stone, Elisabeth A Wilde, Brian A Taylor, David F Tate, Harvey Levin, Erin D Bigler, Randall S Scheibel, Mary R Newsome, Andrew R Mayer, Tracy Abildskov, et al. Supervised learning technique for the automated identification of white matter hyperintensities in traumatic brain injury. *Brain injury*, 30(12):1458–1468, 2016.

[194] Jose V Manjón, Pierrick Coupé, Parnesh Raniga, Ying Xia, Jurgen Fripp, and Olivier Salvado. HIST: hyperintensity segmentation tool. In *International Workshop on Patch-based Techniques in Medical Imaging*, pages 92–99. Springer, 2016.

[195] Paul Schmidt, Christian Gaser, Milan Arsic, Dorothea Buck, Annette Förschler, Achim Berthele, Muna Hoshi, Rüdiger Ilg, Volker J Schmid, Claus Zimmer, et al. An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. *Neuroimage*, 59(4):3774–3783, 2012.

[196] William D Penny, Karl J Friston, John T Ashburner, Stefan J Kiebel, and Thomas E Nichols. *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.

[197] Joseph A Maldjian, Christopher T Whitlow, BN Saha, G Kota, C Vandergriff, EM Davenport, J Divers, BI Freedman, and DW Bowden. Automated white matter total lesion volume segmentation in diabetes. *American Journal of Neuroradiology*, 34(12):2265–2270, 2013.

[198] Jesse Knight, Graham W Taylor, and April Khademi. Voxel-wise logistic regression and leave-one-source-out cross validation for white matter hyperintensity segmentation. *Magnetic resonance imaging*, 54:119–136, 2018.

[199] Elizabeth M Sweeney, Joshua T Vogelstein, Jennifer L Cuzzocreo, Peter A Calabresi, Daniel S Reich, Ciprian M Crainiceanu, and Russell T Shinohara. A comparison of supervised machine learning algorithms and feature vectors for MS lesion segmentation using multimodal structural MRI. *PloS one*, 9(4):e95753, 2014.

[200] Mahsa Dadar, Josefina Maranzano, Karen Misquitta, Cassandra J Anor, Vladimir S Fonov, M Carmela Tartaglia, Owen T Carmichael, Charles Decarli, D Louis Collins, Alzheimer's Disease Neuroimaging Initiative, et al. Performance comparison of 10 different classification techniques in segmenting white matter hyperintensities in aging. *NeuroImage*, 157:233–249, 2017.

[201] Muhammad Febrian Rachmadi, Maria del C Valdés-Hernández, Maria Leonora Fatimah Agan, Carol Di Perri, Taku Komura, Alzheimer's Disease Neuroimaging Initiative, et al. Segmentation of white matter hyperintensities using convolutional neural networks with global spatial information in routine clinical brain MRI with none or mild vascular pathology. *Computerized Medical Imaging and Graphics*, 66:28–43, 2018.

[202] R Guerrero, C Qin, O Oktay, C Bowles, L Chen, R Joules, R Wolz, M del C Valdés-Hernández, DA Dickie, J Wardlaw, et al. White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. *NeuroImage: Clinical*, 17:918–934, 2018.

[203] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[204] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[205] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[206] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[207] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[208] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[209] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[210] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[211] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

[212] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. Understanding the exploding gradient problem. *CoRR, abs/1211.5063*, 2, 2012.

[213] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[214] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

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

[216] Hugo J Kuijf, J Matthijs Biesbroek, Jeroen de Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adrià Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities; results of the WMH segmentation challenge. *IEEE transactions on medical imaging*, 2019.

[217] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[218] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[219] Christian Wachinger, Martin Reuter, and Tassilo Klein. Deepnat: Deep convolutional neural network for segmenting neuroanatomy. *NeuroImage*, 170:434–445, 2018.

[220] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, and Christian Wachinger. Quicknat: Segmenting MRI neuroanatomy in 20 seconds. *arXiv preprint arXiv:1801.04161*, 2018.

[221] Yu Wang, Quan Zhou, Jia Liu, Jian Xiong, Guangwei Gao, Xiaofu Wu, and Longin Jan Latecki. LEDNet: A lightweight encoder-decoder network for real-time semantic segmentation. *arXiv preprint arXiv:1905.02423*, 2019.

[222] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[223] YG Naresh, Suzanne Little, and Noel E O'Connor. A residual encoder-decoder network for semantic segmentation in autonomous driving scenarios. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1052–1056. IEEE, 2018.

[224] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[225] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[226] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.

[227] Matt Berseth. WMH Segmentation Challenge, MICCAI 2017.

[228] Yue Zhang, Wanli Chen, Yifan Chen, and Xiaoying Tang. A post-processing method to improve the white matter hyperintensity segmentation accuracy for randomly-initialized U-net. In *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*, pages 1–5. IEEE, 2018.

[229] Tom Brosch, Lisa YW Tang, Youngjin Yoo, David KB Li, Anthony Traboulsee, and Roger Tam. Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE transactions on medical imaging*, 35(5):1229–1239, 2016.

[230] Mohsen Ghafoorian, Nico Karssemeijer, Tom Heskes, IWM Van Uder, Frank-Erik de Leeuw, Elena Marchiori, Bram van Ginneken, and Bram Platel. Non-uniform patch sampling with deep convolutional neural networks for white matter hyperintensity segmentation. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 1414–1417. IEEE, 2016.

[231] Pim Moeskops, Jeroen de Bresser, Hugo J Kuijf, Adriënne M Mendrik, Geert Jan Biessels, Josien PW Pluim, and Ivana Išgum. Evaluation of a deep learning approach for the segmentation of brain tissues and white matter hyperintensities of presumed vascular origin in MRI. *NeuroImage: Clinical*, 17:251–262, 2018.

[232] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

[233] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

[234] Asier López-Zorrilla, Mikel de Velasco-Vázquez, Oscar Serradilla-Casado, Leire Roa-Barco, Manuel Graña, Darya Chyzhyk, and CC Price. Brain white matter lesion segmentation with 2D/3D CNN. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*, pages 394–403. Springer, 2017.

[235] Sergi Valverde, Mariano Cabezas, Eloy Roura, Sandra González-Villà, Deborah Pareto, Joan C Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Arnau Oliver, and Xavier Lladó. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *NeuroImage*, 155:159–168, 2017.

[236] Botian Xu, Yaqiong Chai, Cristina M Galarza, Chau Q Vu, Benita Tamrazi, Bilwaj Gaonkar, Luke Macyszyn, Thomas D Coates, Natasha Lepore, and John C Wood. Orchestral fully convolutional networks for small lesion segmentation in brain MRI. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 889–892. IEEE, 2018.

[237] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. In *Advances in Neural Information Processing Systems*, pages 9605–9616, 2018.

[238] Simon Andermatt, Simon Pezold, and Philippe Cattin. Multi-dimensional gated recurrent units for the segmentation of biomedical 3D-data. In *Deep Learning and Data Labeling for Medical Applications*, pages 142–151. Springer, 2016.

[239] Yongchao Xu, Thierry Géraud, Élodie Puybareau, Isabelle Bloch, and Joseph Chazalon. White matter hyperintensities segmentation in a few seconds using fully convolutional network and transfer learning. In *International MICCAI Brainlesion Workshop*, pages 501–514. Springer, 2017.

[240] Guodong Zeng and Guoyan Zheng. Deeply supervised multi-scale fully convolutional networks for segmentation of white matter hyperintensities, 2017.

[241] Richard McKinley, Alain Jungo, Roland Wiest, and Mauricio Reyes. Pooling-free fully convolutional networks with dense skip connections for semantic segmentation, with application to brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 169–177. Springer, 2017.

[242] Adriënne M Mendrik, Koen L Vincken, Hugo J Kuijf, Marcel Breeuwer, Willem H Bouvy, Jeroen De Bresser, Amir Alansary, Marleen De Bruijne, Aaron Carass, Ayman El-Baz, et al. MRBrainS challenge: online evaluation framework for brain image segmentation in 3T MRI scans. *Computational intelligence and neuroscience*, 2015:1, 2015.

[243] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2017.

[244] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[245] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018.

[246] Jason Kuen, Zhenhua Wang, and Gang Wang. Recurrent attentional networks for saliency detection. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3668–3677, 2016.

[247] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention U-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[248] Jo Schlemper, Ozan Oktay, Liang Chen, Jacqueline Matthew, Caroline Knight, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention-gated networks for improving ultrasound scan plane detection. *arXiv preprint arXiv:1804.05338*, 2018.

[249] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[250] Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283, 2017.

[251] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[252] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[253] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. GAN-based data augmentation for improved liver lesion classification. 2018.

[254] Xinyue Zhu, Yifan Liu, Jiahong Li, Tao Wan, and Zengchang Qin. Emotion classification with data augmentation using generative adversarial networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 349–360. Springer, 2018.

[255] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[256] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[257] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 2015.

[258] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In *Advances in neural information processing systems*, pages 1410–1418, 2009.

[259] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[260] Muhammad Ghifary, W Bastiaan Kleijn, and Mengjie Zhang. Domain adaptive neural networks for object recognition. In *Pacific Rim international conference on artificial intelligence*, pages 898–904. Springer, 2014.

[261] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010.

[262] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[263] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513–520, 2011.

[264] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016.

[265] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[266] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[267] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[268] Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *arXiv preprint arXiv:1812.02849*, 2019.

[269] Thomas Schlegl, Joachim Ofner, and Georg Langs. Unsupervised pre-training across image domains improves lung tissue classification. In *International MICCAI Workshop on Medical Computer Vision*, pages 82–93. Springer, 2014.

[270] Vanya V Valindria, Ioannis Lavdas, Wenjia Bai, Konstantinos Kamnitsas, Eric O Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker. Domain adaptation for MRI organ segmentation using reverse classification accuracy. *arXiv preprint arXiv:1806.00363*, 2018.

[271] Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles RG Guttmann, Frank-Erik de Leeuw, Clare M Tempany, Bram van Ginneken, et al. Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 516–524. Springer, 2017.

[272] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[273] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4068–4076, 2015.

[274] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *International conference on information processing in medical imaging*, pages 597–609. Springer, 2017.

[275] Sumit Chopra, Raia Hadsell, Yann LeCun, et al. Learning a similarity metric discriminatively, with application to face verification. In *CVPR (1)*, pages 539–546, 2005.

[276] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744, 1994.

[277] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5715–5725, 2017.

[278] Saifeng Liu, David Utriainen, Chao Chai, Yongsheng Chen, Lin Wang, Sean K Sethi, Shuang Xia, and E Mark Haacke. Cerebral microbleed detection using susceptibility weighted imaging and deep learning. *NeuroImage*, 198:271–282, 2019.

[279] Yunhee Jeong, Muhammad Febrian Rachmadi, Maria Del Carmen Valdés Hernández, and Taku Komura. Dilated saliency U-Net for white matter hyperintensities segmentation using irregularity age map. *Frontiers in Aging Neuroscience*, 11:150, 2019.

[280] IA Lee and KJ Preacher. Calculation for the test of the difference between two dependent correlations with one variable in common [computer software], 2013.

[281] James H Steiger. Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2):245, 1980.

[282] Maria del C Valdés Hernández, Peter J Gallacher, Mark E Bastin, Natalie A Royle, Susana Muñoz Maniega, Ian J Deary, and Joanna M Wardlaw. Automatic segmentation of brain white matter and white matter lesions in normal aging: comparison of five multispectral techniques. *Magnetic resonance imaging*, 30(2):222–229, 2012.

[283] Sigurdur Sigurdsson, Thor Aspelund, Lars Forsberg, Jesper Fredriksson, Olafur Kjartansson, Bryndis Oskarsdottir, Palmi V Jonsson, Gudny Eiriksdottir, Tamara B Harris, Alex Zijdenbos, et al. Brain tissue volumes in the general population of the elderly: the AGES-Reykjavik study. *Neuroimage*, 59(4):3862–3870, 2012.

[284] Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper LR Andersson, et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523, 2016.

[285] S Shams, J Martola, L Cavallin, T Granberg, M Shams, P Aspelin, LO Wahlund, and M Kristoffersen-Wiberg. SWI or T2*: which MRI sequence to use in the detection of cerebral microbleeds? the Karolinska Imaging Dementia Study. *American Journal of Neuroradiology*, 36(6):1089–1095, 2015.

[286] Meike W Vernooij, M Arfan Ikram, Piotr A Wielopolski, Gabriel P Krestin, Monique MB Breteler, and Aad van der Lugt. Cerebral microbleeds: accelerated 3D T2*-weighted GRE MR imaging versus conventional 2D T2*-weighted GRE MR imaging for detection. *Radiology*, 248(1):272–277, 2008.

[287] Mohamed L Seghier, Magdalena A Kolanko, Alexander P Leff, Hans R Jäger, Simone M Gregoire, and David J Werring. Microbleed detection using automated segmentation (MIDAS): a new method applicable to standard clinical MR images. *PloS one*, 6(3):e17547, 2011.

[288] Samuel RS Barnes, E Mark Haacke, Muhammad Ayaz, Alexander S Boikov, Wolff Kirsch, and Dan Kido. Semiautomated detection of cerebral microbleeds in magnetic resonance images. *Magnetic resonance imaging*, 29(6):844–852, 2011.

[289] TLA van den Heuvel, AW Van Der Eerden, R Manniesing, M Ghafoorian, T Tan, TMJC Andriessen, T Vande Vyvere, L Van den Hauwe, BM ter Haar Romeny, BM Goraj, et al. Automated detection of cerebral microbleeds in patients with traumatic brain injury. *NeuroImage: Clinical*, 12:241–251, 2016.

[290] Melanie A Morrison, Seyedmehdi Payabvash, Yicheng Chen, Sivakami Avadiappan, Mihir Shah, Xiaowei Zou, Christopher P Hess, and Janine M Lupo. A user-guided tool for semi-automated cerebral microbleed detection and volume segmentation: Evaluating vascular injury and data labelling for machine learning. *NeuroImage: Clinical*, 20:498–505, 2018.

[291] J De Bresser, M Brundel, MM Conijn, JJ Van Dillen, MI Geerlings, MA Viergever, PR Luijten, and GJ Biessels. Visual cerebral microbleed detection on 7T MR imaging: reliability and effects of image processing. *American Journal of Neuroradiology*, 34(6):E61–E64, 2013.

[292] Hugo J Kuijf, Jeroen de Bresser, Mirjam I Geerlings, Mandy MA Conijn, Max A Viergever, Geert Jan Biessels, and Koen L Vincken. Efficient detection of cerebral microbleeds on 7.0 T MR images using the radial symmetry transform. *NeuroImage*, 59(3):2266–2273, 2012.

[293] Hugo J Kuijf, Manon Brundel, Jeroen de Bresser, Susanne J van Veluw, Sophie M Heringa, Max A Viergever, Geert Jan Biessels, and Koen L Vincken. Semi-automated detection of cerebral microbleeds on 3.0 T MR images. *PLoS One*, 8(6):e66610, 2013.

[294] Wei Bian, Christopher P Hess, Susan M Chang, Sarah J Nelson, and Janine M Lupo. Computer-aided detection of radiation-induced cerebral microbleeds on susceptibility-weighted MR images. *NeuroImage: clinical*, 2:282–290, 2013.

[295] Amir Fazlollahi, Fabrice Meriaudeau, Victor L Villemagne, Christopher C Rowe, Paul Yates, Olivier Salvado, and Pierrick Bourgeat. Efficient machine learning framework for computer-aided detection of cerebral microbleeds using the radon transform. In *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*, pages 113–116. IEEE, 2014.

[296] Amir Fazlollahi, Fabrice Meriaudeau, Luca Giancardo, Victor L Villemagne, Christopher C Rowe, Paul Yates, Olivier Salvado, Pierrick Bourgeat, AIBL Research Group, et al. Computer-aided detection of cerebral microbleeds in susceptibility-weighted imaging. *Computerized Medical Imaging And Graphics*, 46:269–276, 2015.

[297] Qi Dou, Hao Chen, Lequan Yu, Lin Shi, Defeng Wang, Vincent CT Mok, and Pheng Ann Heng. Automatic cerebral microbleeds detection from MR images via independent subspace analysis based hierarchical features. In *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 7933–7936. IEEE, 2015.

[298] Babak Ghafaryasl, Fedde van der Lijn, Mariëlle Poels, Henri Vrooman, Mohammad Arfan Ikram, Wiro J Niessen, Aad van der Lugt, Meike Vernooij, and Marleen de Bruijne. A computer aided detection system for cerebral microbleeds in brain MRI. In *2012 9th IEEE international symposium on biomedical imaging (ISBI)*, pages 138–141. IEEE, 2012.

[299] Zhi-Wei Pan, Dao-Hong Xiang, Quan-Wu Xiao, and Ding-Xuan Zhou. Parzen windows for multi-class classification. *Journal of complexity*, 24(5-6):606–618, 2008.

[300] Hao Chen, Lequan Yu, Qi Dou, Lin Shi, Vincent CT Mok, and Pheng Ann Heng. Automatic detection of cerebral microbleeds via deep learning based 3D feature representation. In *2015 IEEE 12th international symposium on biomedical imaging (ISBI)*, pages 764–767. IEEE, 2015.

[301] Qi Dou, Hao Chen, Lequan Yu, Lei Zhao, Jing Qin, Defeng Wang, Vincent CT Mok, Lin Shi, and Pheng-Ann Heng. Automatic detection of cerebral microbleeds from MR images via 3d convolutional neural networks. *IEEE transactions on medical imaging*, 35(5):1182–1195, 2016.

[302] Yicheng Chen, Javier E Villanueva-Meyer, Melanie A Morrison, and Janine M Lupo. Toward automatic detection of radiation-induced cerebral microbleeds using a 3d deep residual network. *Journal of digital imaging*, pages 1–7, 2018.

[303] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[304] Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[305] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[306] Chenyi Chen, Ming-Yu Liu, Oncel Tuzel, and Jianxiong Xiao. R-CNN for small object detection. In *Asian conference on computer vision*, pages 214–230. Springer, 2016.

[307] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[308] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017.

[309] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[310] Xuming He, Richard S Zemel, and Miguel Á Carreira-Perpiñán. Multiscale conditional random fields for image labeling. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. IEEE, 2004.

[311] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016.

[312] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 447–456, 2015.

[313] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2012.

[314] Mohammadreza Mostajabi, Payman Yadollahpour, and Gregory Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3376–3385, 2015.

[315] Chunhui Gu, Joseph J Lim, Pablo Arbeláez, and Jitendra Malik. Recognition using regions. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1030–1037. IEEE, 2009.

[316] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[317] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[318] Linxin Li, Michela Simoni, Wilhelm Küker, Ursula G Schulz, Sharon Christie, Gordon K Wilcock, and Peter M Rothwell. Population-based case–control study of white matter changes on brain imaging in transient ischemic attack and ischemic stroke. *Stroke*, 44(11):3063–3070, 2013.

[319] Niels D Prins and Philip Scheltens. White matter hyperintensities, cognitive impairment and dementia: an update. *Nature Reviews Neurology*, 11(3):157, 2015.

[320] Domenico Inzitari, Giovanni Pracucci, Anna Poggesi, Giovanna Carlucci, Frederik Barkhof, Hugues Chabriat, Timo Erkinjuntti, Franz Fazekas, José M Ferro, Michael Hennerici, et al. Changes in white matter as determinant of global functional decline in older independent outpatients: three year follow-up of LADIS (leukoaraiosis and disability) study cohort. *Bmj*, 339:b2477, 2009.

[321] C Dufouil, A de Kersaint-Gilly, V Besancon, C Levy, E Auffray, L Brunnereau, A Alperovitch, and C Tzourio. Longitudinal study of blood pressure and white matter hyperintensities the EVA MRI cohort. *Neurology*, 56(7):921–926, 2001.

[322] Rebecca F Gottesman, Josef Coresh, Diane J Catellier, A Richey Sharrett, Kathryn M Rose, Laura H Coker, Dean K Shibata, David S Knopman, Clifford R Jack, and Thomas H Mosley. Blood pressure and white-matter disease progression in a biethnic cohort: Atherosclerosis risk in communities (ARIC) study. *Stroke*, 41(1):3–8, 2010.

[323] Change in stroke incidence, mortality, case-fatality, severity, and risk factors in Oxfordshire, UK from 1981 to 2004 (oxford vascular study.

[324] Philip Wolfe. Convergence conditions for ascent methods. *SIAM review*, 11(2):226–235, 1969.

[325] DM Bortz and CT Kelley. The simplex gradient and noisy optimization problems. In *Computational methods for optimal design and control*, pages 77–90. Springer, 1998.

[326] Anderson M Winkler, Gerard R Ridgway, Matthew A Webster, Stephen M Smith, and Thomas E Nichols. Permutation inference for the general linear model. *Neuroimage*, 92:381–397, 2014.

[327] Vaanathi Sundaresan, Giovanna Zamboni, Campbell Le Heron, Peter M. Rothwell, Masud Husain, Marco Battaglini, Nicola De Stefano, Mark Jenkinson, and Ludovica Griffanti. Automated lesion segmentation with BIANCA: impact of population-level features, classification algorithm and locally adaptive thresholding. *bioRxiv*, DOI: 10.1101/437608, 2018.

[328] Tracy D Vannorsdall, Shari R Waldstein, Michael Kraut, Godfrey D Pearlson, and David J Schretlen. White matter abnormalities and cognition in a community sample. *Archives of clinical neuropsychology*, 24(3):209–217, 2009.

[329] Maria Del C Valdés Hernández, Francesca M Chappell, Susana Muñoz Maniega, David Alexander Dickie, Natalie A Royle, Zoe Morris, Devasuda Anblagan, Eleni Sakka, Paul A Armitage, Mark E Bastin, et al. Metric to quantify white matter damage on brain magnetic resonance images. *Neuroradiology*, 59(10):951–962, 2017.

[330] Yifeng Ling, Eric Jouvent, Louis Cousyn, Hugues Chabriat, and François De Guio. Validation and optimization of BIANCA for the segmentation of extensive white matter hyperintensities. *Neuroinformatics*, pages 1–13, 2018.

[331] Ezequiel Geremia, Olivier Clatz, Bjoern H Menze, Ender Konukoglu, Antonio Criminisi, and Nicholas Ayache. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. *NeuroImage*, 57(2):378–390, 2011.

[332] Jhimli Mitra, Pierrick Bourgeat, Jurgen Fripp, Soumya Ghose, Stephen Rose, Olivier Salvado, Alan Connelly, Bruce Campbell, Susan Palmer, Gagan Sharma, et al. Lesion segmentation from multimodal MRI using random forest following ischemic stroke. *NeuroImage*, 98:324–335, 2014.

[333] Daisuke Yamamoto, Hidetaka Arimura, Shingo Kakeda, Taiki Magome, Yasuo Yamashita, Fukai Toyofuku, Masafumi Ohki, Yoshiharu Higashida, and Yukunori Korogi. Computer-aided detection of multiple sclerosis lesions in brain magnetic resonance images: False positive reduction scheme consisted of rule-based, level set method, and support vector machine. *Computerized Medical Imaging and Graphics*, 34(5):404–413, 2010.

[334] Giovanna Zamboni, Gordon K Wilcock, Gwenaelle Douaud, Erin Drazich, Ellen McCulloch, Nicola Filippini, Irene Tracey, Jonathan CW Brooks, Stephen M Smith, Mark Jenkinson, et al. Resting functional connectivity reveals residual functional activity in Alzheimers disease. *Biological psychiatry*, 74(5):375–383, 2013.

[335] Stephen M Smith. Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–155, 2002.

[336] Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57, 2001.

[337] Mark Jenkinson and Stephen Smith. A global optimisation method for robust affine registration of brain images. *Medical image analysis*, 5(2):143–156, 2001.

[338] Jesper LR Andersson, Mark Jenkinson, Stephen Smith, et al. Non-linear registration, aka spatial normalisation FMRIB technical report TR07JA2. *FMRIB Analysis Group of the University of Oxford*, 2:1–21, 2007.

[339] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

[340] Ludovica Griffanti, Mark Jenkinson, Sana Suri, Enikő Zsoldos, Abda Mahmood, Nicola Filippini, Claire E Sexton, Anya Topiwala, Charlotte Allan, Mika Kivimäki, et al. Classification and characterization of periventricular and deep white matter hyperintensities on MRI: A study in older adults. *NeuroImage*, 2017.

[341] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[342] DMJ Van Den Heuvel, F Admiraal-Behloul, VH Ten Dam, H Olofsen, ELEM Bollen, HM Murray, GJ Blauw, RGJ Westendorp, AJM De Craen, MA Van Buchem, et al. Different progression rates for deep white matter hyperintensities in elderly men and women. *Neurology*, 63(9):1699–1701, 2004.

[343] Behnam Sabayan, Jeroen van der Grond, Rudi G Westendorp, Mark A van Buchem, and Anton JM de Craen. Accelerated progression of white matter hyperintensities and subsequent risk of mortality: a 12-year follow-up study. *Neurobiology of aging*, 36(6):2130–2135, 2015.

[344] Hongwei Li, Jianguo Zhang, Mark Muehlau, Jan Kirschke, and Bjoern Menze. Multi-scale convolutional-stack aggregation for robust white matter hyperintensities segmentation. *arXiv preprint arXiv:1807.05153*, 2018.

[345] 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.

[346] Hao Dong, Guang Yang, Fangde Liu, Yuanhan Mo, and Yike Guo. Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks. In *annual conference on medical image understanding and analysis*, pages 506–517. Springer, 2017.

[347] Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 683–687. IEEE, 2019.

[348] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[349] S Winzeck, SJT Mocking, R Bezerra, MJRJ Bouts, EC McIntosh, I Diwan, P Garg, A Chutinet, WT Kimberly, WA Copen, et al. Ensemble of convolutional neural networks improves automated segmentation of acute ischemic lesions using multiparametric diffusion-weighted MRI. *American Journal of Neuroradiology*, 40(6):938–945, 2019.

[350] David Opitz and Richard Maclin. Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, 11:169–198, 1999.

[351] Adhish Prasoon, Kersten Petersen, Christian Igel, François Lauze, Erik Dam, and Mads Nielsen. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In *International conference on medical image computing and computer-assisted intervention*, pages 246–253. Springer, 2013.

[352] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE international conference on computer vision*, pages 2200–2207, 2013.

[353] Xuezhi Wang and Jeff Schneider. Flexible transfer learning under support and model shift. In *Advances in Neural Information Processing Systems*, pages 1898–1906, 2014.

[354] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[355] Laurens Van Der Maaten. Barnes-hut-sne. *arXiv preprint arXiv:1301.3342*, 2013.

[356] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

[357] D Randall Wilson and Tony R Martinez. The general inefficiency of batch training for gradient descent learning. *Neural networks*, 16(10):1429–1451, 2003.

[358] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[359] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[360] Ruobing Huang, J Alison Noble, and Ana IL Namburete. Omni-supervised learning: Scaling up to large unlabelled medical datasets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 572–580. Springer, 2018.

[361] Alejandro F Frangi, Wiro J Niessen, Koen L Vincken, and Max A Viergever. Multiscale vessel enhancement filtering. In *International conference on medical image computing and computer-assisted intervention*, pages 130–137. Springer, 1998.

[362] Wolfgang Förstner. A framework for low level feature extraction. In *European Conference on Computer Vision*, pages 383–394. Springer, 1994.

[363] Gareth Loy and Alexander Zelinsky. A fast radial symmetry transform for detecting points of interest. In *European Conference on Computer Vision*, pages 358–368. Springer, 2002.

[364] Steven M Greenberg, RN Kaveer Nandigam, Pilar Delgado, Rebecca A Betensky, Jonathan Rosand, Anand Viswanathan, Matthew P Frosch, and Eric E Smith. Microbleeds versus macrobleeds: evidence for distinct entities. *Stroke; a journal of cerebral circulation*, 40(7):2382, 2009.

[365] Jean A Frazier, Sufen Chiu, Janis L Breeze, Nikos Makris, Nicholas Lange, David N Kennedy, Martha R Herbert, Eileen K Bent, Vamsi K Koneru, Megan E Dieterich, et al. Structural brain magnetic resonance imaging of limbic and thalamic volumes in pediatric bipolar disorder. *American Journal of Psychiatry*, 162(7):1256–1265, 2005.

[366] Jill M Goldstein, Larry J Seidman, Nikos Makris, Todd Ahern, Liam M OBrien, Verne S Caviness Jr, David N Kennedy, Stephen V Faraone, and Ming T Tsuang. Hypothalamic abnormalities in schizophrenia: sex effects and genetic vulnerability. *Biological psychiatry*, 61(8):935–945, 2007.

[367] D Louis Collins, Colin J Holmes, Terrence M Peters, and Alan C Evans. Automatic 3-D model-based neuroanatomical segmentation. *Human brain mapping*, 3(3):190–208, 1995.

[368] John Mazziotta, Arthur Toga, Alan Evans, Peter Fox, Jack Lancaster, Karl Zilles, Roger Woods, Tomas Paus, Gregory Simpson, Bruce Pike, et al. A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (ICBM). *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1412):1293–1322, 2001.

[369] Susumu Mori, Setsu Wakana, Peter CM Van Zijl, and LM Nagae-Poetscher. *MRI atlas of human white matter*. Elsevier, 2005.

[370] Setsu Wakana, Arvind Caprihan, Martina M Panzenboeck, James H Fallon, Michele Perry, Randy L Gollub, Kegang Hua, Jiangyang Zhang, Hangyi Jiang, Prachi Dubey, et al. Reproducibility of quantitative tractography methods applied to cerebral white matter. *Neuroimage*, 36(3):630–644, 2007.

[371] Kegang Hua, Jiangyang Zhang, Setsu Wakana, Hangyi Jiang, Xin Li, Daniel S Reich, Peter A Calabresi, James J Pekar, Peter CM van Zijl, and Susumu Mori. Tract probability maps in stereotaxic spaces: analyses of white matter anatomy and tract-specific quantification. *Neuroimage*, 39(1):336–347, 2008.

[372] Julie Staals, Stephen DJ Makin, Fergus N Doubal, Martin S Dennis, and Joanna M Wardlaw. Stroke subtype, vascular risk factors, and total MRI brain small-vessel disease burden. *Neurology*, 83(14):1228–1234, 2014.

[373] Qi Dong, Shaogang Gong, and Xiatian Zhu. Imbalanced deep learning by minority class incremental rectification. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1367–1381, 2018.

[374] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[375] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using GAN for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 289–293. IEEE, 2018.

[376] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2642–2651. JMLR. org, 2017.

[377] Alhussein Fawzi, Horst Samulowitz, Deepak Turaga, and Pascal Frossard. Adaptive data augmentation for image classification. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3688–3692. Ieee, 2016.

[378] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[379] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[380] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24, 2007.

[381] Ray Liere and Prasad Tadepalli. Active learning with committees for text categorization. In *AAAI/IAAI*, pages 591–596, 1997.

[382] Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[383] Tom M Mitchell. Artificial neural networks. *Machine learning*, 45:81–127, 1997.